AI POLICY MODULE

Regulating Artificial Intelligence

James Weichert 2024

Roadmap

Monday

AI *Ethics*

How do we *think* about the impact of AI on society?

1. What are ethics?
2. Why do we need ethics for AI?
3. AI ethics frameworks

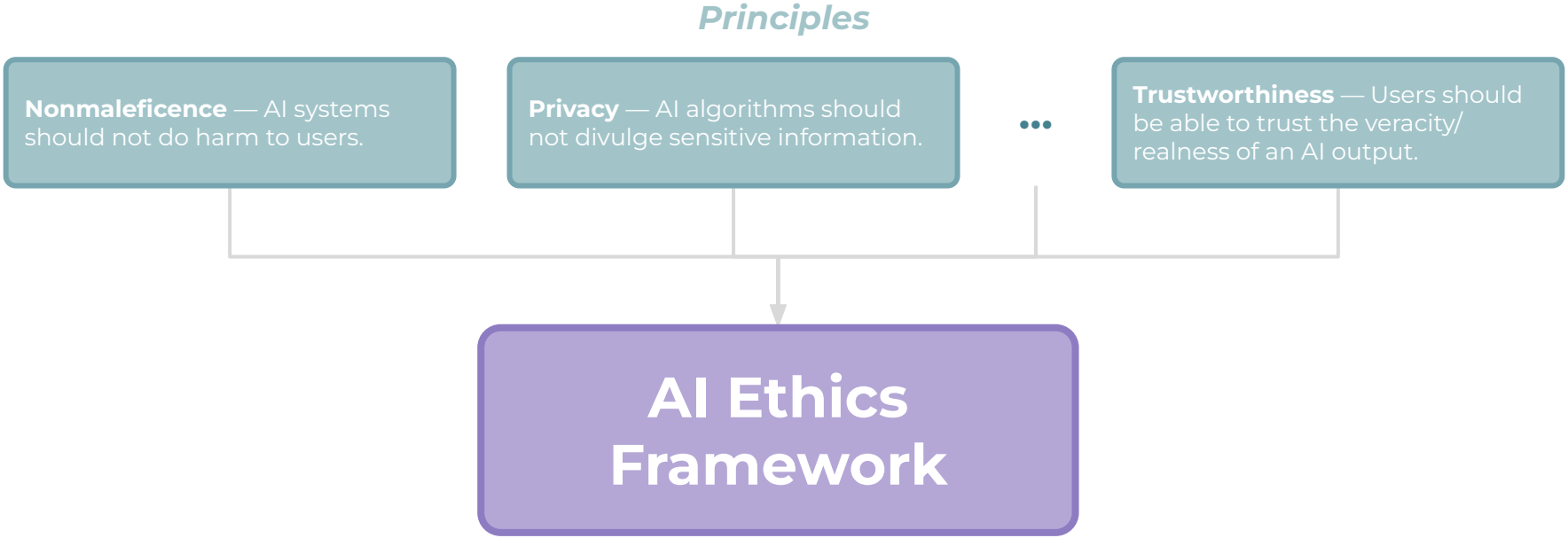
Today

AI *Policy*

How do we *act* responsibly with AI?

1. What is 'AI policy'?
2. AI policy landscape
3. Designing and implementing AI policy

Putting It All Together



Some Frameworks

Unified Framework for AI in Society

Floridi and Cowls (2019)

Principles:

- Beneficence
- Nonmaleficence
- Autonomy
- Justice
- Explicability

Blueprint for an AI Bill of Rights

Biden White House (2022)

Principles:

- Safe and effective systems
- Algorithmic discrimination protections
- Data privacy
- Notice and explanation
- Human alternatives, consideration, and fallback

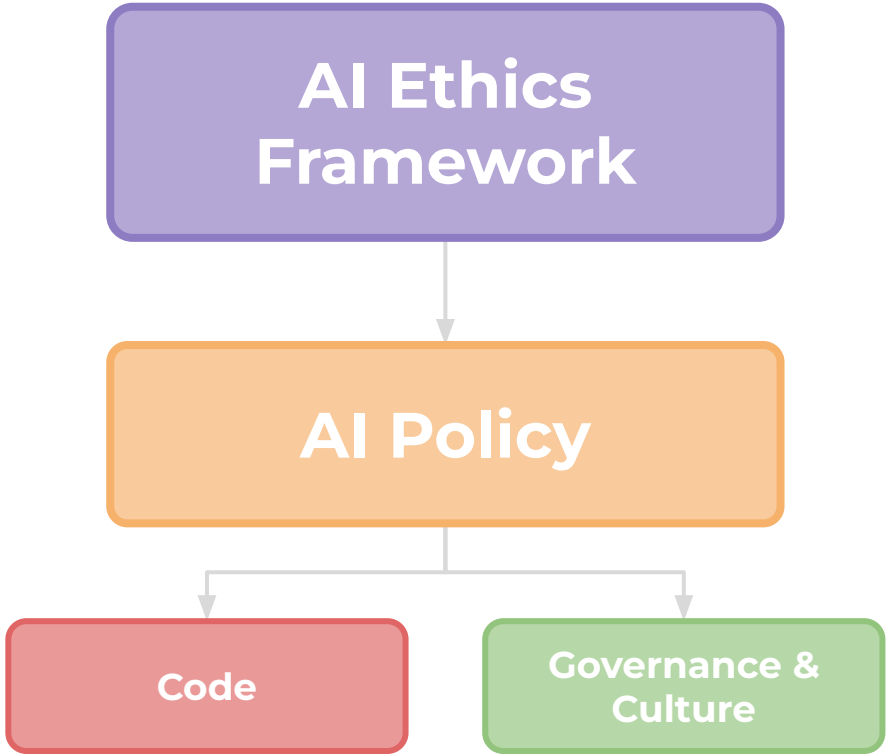
Responsible AI Standard

Microsoft (2022)

Principles:

- Accountability
- Transparency
- Fairness
- Reliability & Safety
- Privacy & Security
- Inclusiveness

How does this work in practice?

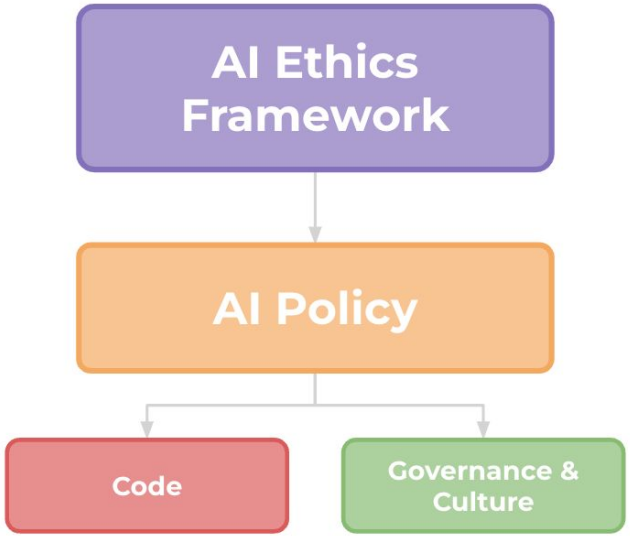


From Principles to Practices

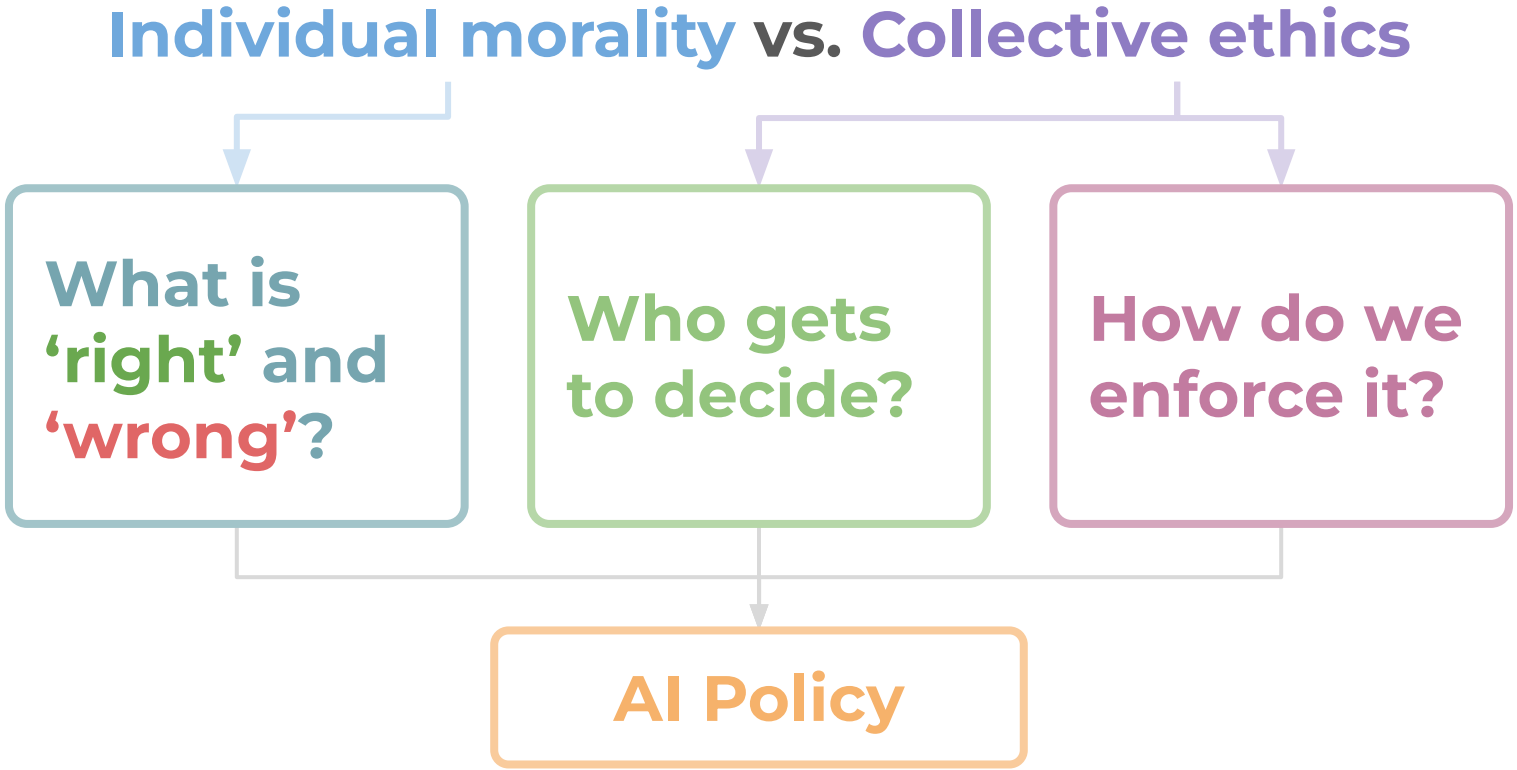
Kim, Zhu and **Eldardiry** (2023) explore how companies and governments are shifting from **ethical AI *principles*** to **ethical AI *policies*.**

A ***policy* approach** to AI governance requires:

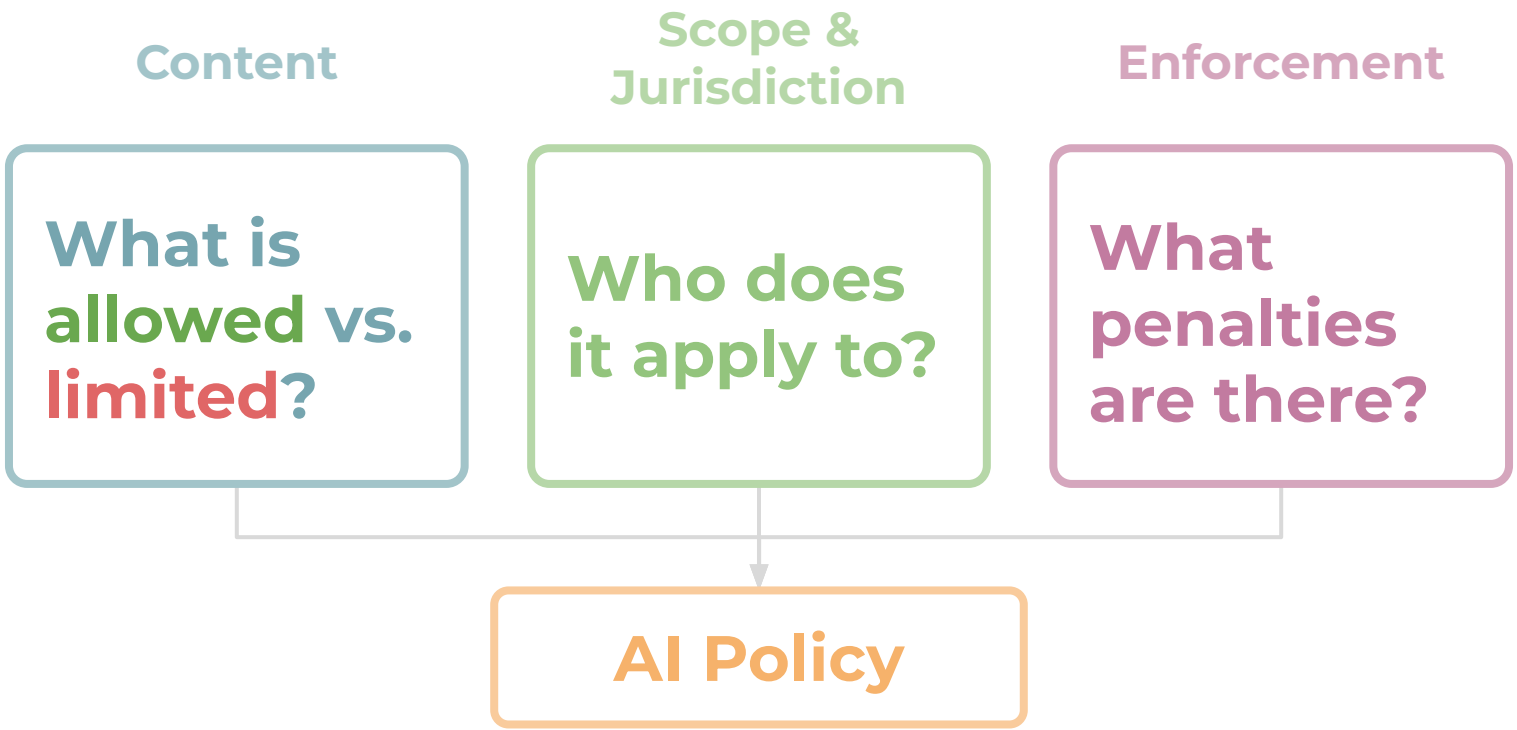
- "attention toward **social and political contexts**"
- formalizing "clear, structured, and easy-to-follow ways to **train and empower the next generation of responsible AI professionals**"
- Two steps: "(1) translating **ethical principles to policies** and (2) translating **policies to AI algorithms**"



What is AI policy?



What is AI policy?



Key Actors

DISCUSSION

Who might want to regulate AI and *why*?

AI Companies

concern over image, profit?

Governments

ideological, economic concerns?

Policy Influences

AI Companies

concern over image, profit?

Pew Research 2019

Majority of Americans feel as if they have little control over data collected about them by companies and the government

% of U.S. adults who say ...

		Companies	The government
Lack of control	They have very little/no control over the data ___ collect(s)	81%	84%
Risks outweigh benefits	Potential risks of ___ collecting data about them outweigh the benefits	81%	66%
Concern over data use	They are very/somewhat concerned about how ___ use(s) the data collected	79%	64%

Consumer Opinion

"unethical behaviour by corporate laggards can **tarnish a sector's reputation**" (Auld et al. 2022)

Ethics Shopping

"private actors may **shop for the kind of ethics that is best retrofitted to justify their current behaviours**, rather than revising their behaviours to make them consistent with a socially accepted ethical framework" (Floridi 2019)

Ethics Lobbying / "Bluewashing"

"private actors (are at least suspected to) try to **use self-regulation** about the ethics of AI in order **to lobby against the introduction of legal norms**, or in favour of their watering down or **weakening their enforcement**" (Floridi 2019)

Policy Influences

Governments

ideological, economic concerns?

Consumer Safety

"mistakes by or **misuse of AI** could harm patients, cost consumers or small businesses, or **jeopardize safety or rights**" (Exec. Order 14110)

Economic Competition

"A second type of motivation [for AI policy] is **competitive advantage**, including **economic** and **political advantage**." (Schiff et al. 2020)

Ideological Differences

"**ideological and partisan factors predict AI policy support**, with liberal legislators supporting consumer protection AI policy... Republican states are not especially likely to pass consumer protection legislation" (Parinandi et al. 2024)

Lobbying

"[Companies lobby against AI regulation] by **delaying the introduction** of necessary legislation...or by **influencing law-makers to pass legislation that is more favourable to the lobbyist**" (Floridi 2019)

Landscape of AI Regulation

The Big Players

China

European Union

United States

Weichert et al. (2024)

AI Regulation: China

China

***A New Generation AI Development Plan* (2017)**

- China's **roadmap for AI investment** and development
- Focus on economic transformation and achieving **economic output targets**
- Focus on "**preserving social stability**"

***Personal Information Protection Law* (2021)**

***Provisions on Deep Synthesis Services* (2022)**

***Measures for Generative AI Services* (2023)**

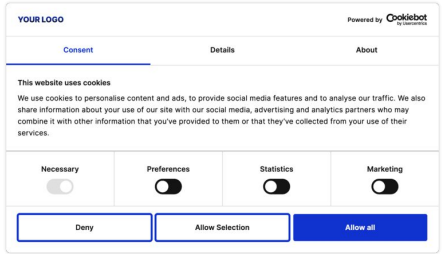
Weichert et al. (2024)

AI Regulation: European Union

European Union

General Data Protection Regulations (2018)

- Wide-ranging ranging **data privacy** and protection regulation



EU AI Act (2024)

- First comprehensive AI law in the West
- Categorizes "AI systems" according to four risk levels
- Imposes strict **requirements on *high risk* systems** while **banning** systems deemed an ***unacceptable risk****
- Includes requirements on developers of **"general-purpose AI" models**

Weichert et al. (2024)

AI Regulation: United States

United States

Blueprint for an AI Bill of Rights (2022)

- Major principles to **"guide the design, use, and development of automated systems"**

Executive Order 14110 (2023)

- "Safe, Secure, and Trustworthy Development and Use of AI"
- Outlines federal government work on AI in areas of ***safety and security***, ***innovation and competition***, and ***federal use of AI***
- Directs NSF to "**support AI-related education and AI-related workforce development**"

Weichert et al. (2024)

Creating AI Policy

> "The process of developing strategies for AI governance is **tightly linked to the economic and political contexts** of the strategy's originator. For a state, this entails **navigating the intricacies of its political and administrative apparatus**. Thus, there is **no uniform roadmap** for how policies transform from ideas to operationalizable regulations."

Weichert et al. 2024

Mafia: AI Policy Edition

Mafia (party game)

25 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

> Some of this article's listed sources **may not be reliable**. Please help improve this article by looking for better, more reliable sources. Unreliable citations may be challenged and removed. *(September 2019)* (*Learn how and when to remove this message*)

Mafia, also known as **Werewolf**, is a Russian social deduction game created by Dimitry Davidoff in 1986.[2] The game models a conflict between two groups: an informed minority (the mafiosi or the werewolves) and an uninformed majority (the villagers). At the start of the game, each player is secretly assigned a role affiliated with one of these teams. The game has two alternating phases: first, a night-phase, during which those with night-killing-powers may covertly kill other players, and second, a day-phase, in which all surviving players debate and vote to eliminate a suspect. The game continues until a faction achieves its win-condition; for the village, this usually means eliminating the evil minority, while for the minority, this usually means reaching numerical parity with the village and eliminating any rival evil groups.

Mafia

Players making accusations in a game of *Mafia*

Other names	Werewolf
Designers	Dimitry Davidoff
Players	At least 6[1]

ACTIVITY

Congress vs. *Evil Inc.*

Voters

Goal: Robust AI regulation, good user experience

Win Condition: Key regulation is adopted

Action: Can, if desired, vote out one member of Congress per turn

US Congress

Goal: Political survival

Win Condition: Stay in Congress until the end of the game

Action: During the Congressional hearing, make one statement or pose one question to the *Evil Inc.* CEO

Evil Inc.

Goal: Profit

Win Condition: Prevent the adoption of the key regulation

Action: Can, if desired, bribe members of Congress *

* If a member of Congress is bribed and the bribe exceeds their bribe threshold, then the member must vote no on the pending regulation.

Scenario 1

CONTEXT

Evil Inc.'s social media platform, ***Y***, has hundreds of millions of daily users, who use the **platform to ask and answer philosophical questions**.

Scenario 1

INTERNAL

Internal research shows that **users under the age of 18** are particularly susceptible to **AI-targeted advertising**.

Key Regulation: Prevent Congress from adopting a regulation that *either* **(a)** restricts social media platforms to users 18 and older, *or* **(b)** prohibits AI-driven advertising.

Scenario 1

REGULATIO

1. Social media **ban for children** under 18
2. Users can request a **copy** of their own data
3. Users must be **notified** about when and how AI is used
4. Ban on using **AI to target advertising**

Scenario 2

CONTEXT

Evil Inc.'s delivers millions of packages to American consumers every day. To increase capacity and efficiency, the company is piloting **fully autonomous package delivery drones**.

Scenario 2

INTERNAL

While small-scale tests have been successful, the development team needs to conduct a **full-scale test of the drones across the nation**.

Key Regulation: Prevent Congress from restricting the public deployment of autonomous agents that are still in the research and development phase.

Scenario 2

REGULATIO

1. Require companies to **monitor** their autonomous fleet from a central control center
2. Require autonomous vehicles to have a **manual override** feature
3. Prohibit autonomous vehicles from being publicly released until they pass a **rigorous final inspection**
4. **Ban autonomous vehicles** in all cases

Scenario 3

CONTEXT

Evil Inc.'s new proprietary large language model, ***d[AI]ve***, is gaining a user base and beginning to compete with more established LLMs like ChatGPT.

Scenario 3

INTERNAL

The success of ***d[AI]ve*** is due to a particular **model architecture and training process**. If these are made public, *d[AI]ve* will lose its competitive advantage.

Key Regulation: Prevent Congress from adopting a regulation that requires the disclosure of any information about the model architecture or training process.

Scenario 3

REGULATIO

1. Prohibit LLMs from **training on user data** (prompts)
2. Require LLMs to disclose their **general model architecture** and **training process**
3. Require LLMs to **notify law enforcement** if the user refers to illegal activity
4. Prohibit LLMs from **accessing the live internet** (i.e. require a training cutoff)

Congress vs. *Evil Inc.* Debrief

DISCUSSION

What did you think about the scenarios presented in the game?

Do you think the game captures the key dynamics of AI policy development?

Real-World Influences

Scenario 1

Topic: AI-targeted advertising

Key Issue: Protection of underage users

Real-World Example:

Reuters 2024

Scenario 2

Topic: Autonomous vehicles

Key Issue: Testing and acceptable harm

Real-World Example:

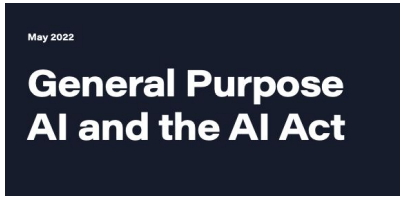
Reuters 2024

Scenario 3

Topic: LLMs / 'general-purpose' AI

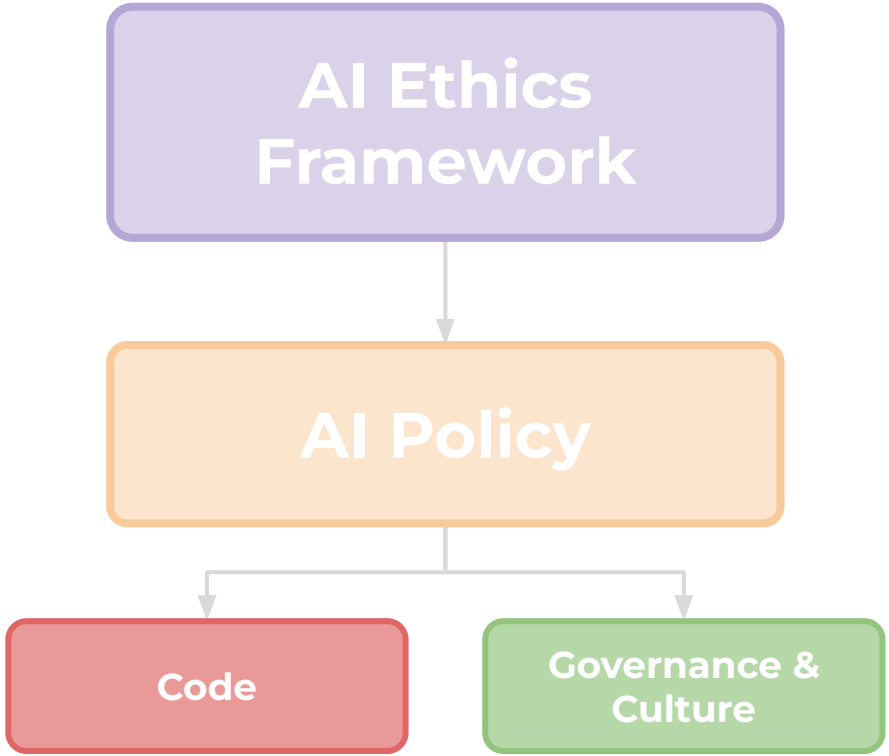
Key Issue: Disclosure of model information

Real-World Example:



EU AI Act

Implementing AI Policy



Ethics and *Machine Learning*

Some considerations:

Data

- Are my data **representative** of the population?
- Is there a risk of **majority bias?**
- Is there ***too much*** **data?**

Model

- What **metric** is the model optimizing?
- How much **confidence** do I have in the model's predictions?
- What **requirements** do I impose on my model's outputs

Action

- What **actions** do I allow the AI to take?
- Do **humans** make the **final decision**?
- Does the model **explain** its actions?

Governance and Culture

DISCUSSION

Do you think governance structures and a good company culture are important?

Why or why not?

Piloting the *AI Policy Module*

These two lectures are completely new, and how they engage you in discussion about AI ethics and policy is the subject of my research.

to help me:

Pre-Survey

Post-Survey

Where to go from here?

You probably fit into one of these buckets:

I have no interest in AI policy whatsoever

I don't think this module was relevant to me

AI policy is interesting to me

I'll likely follow what my company / government tells me to do

AI policy is *really* interesting to me

I want to learn more and maybe get involved in these processes

And that's ok!

Why you should get involved...

AI POLICY

Thoughts and Questions

References

Works Cited:

Kim, D., Zhu, Q., & Eldardiry, H. Exploring approaches to artificial intelligence governance: from ethics to policy," *2023 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS)*, 1-5 (2023). https://doi.org/10.1109/ETHICS57328.2023.10155067

Floridi, L. Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy & Technology* 32 (2019). https://doi.org/10.1007/s13347-019-00354-x

Auld, G., Casovan, A., Clarke, A., & Faveri, B. Governing AI through ethical standards: learning from the experiences of other private governance initiatives. *European Journal of Public Policy* 29(11) (2022). https://doi.org/10.1080/13501763.2022.2099449

Schiff, D., Biddle, J., Borenstein, J., & Laas, K. What's Next for AI Ethics, Policy, and Governance? A Global Overview. *AAAI/ACM Conference on AI, Ethics, and Society* (2020). https://doi.org/10.1145/3375627.3375804

Parinandi, S., Crosson, J., Peterson, K., & Nadarevic, S. Investigating the politics and content of US State artificial intelligence legislation. *Business and Politics* 26(2) (2024). https://doi.org/10.1017/bap.2023.40

Weichert, J., Zhu, Q., Kim, D., & Eldardiry, H. Perceptions of AI Ethics Policies Among Scientists and Engineers in Policy-Related Roles: An Exploratory Investigation. *Digital Society* (2024). Forthcoming.