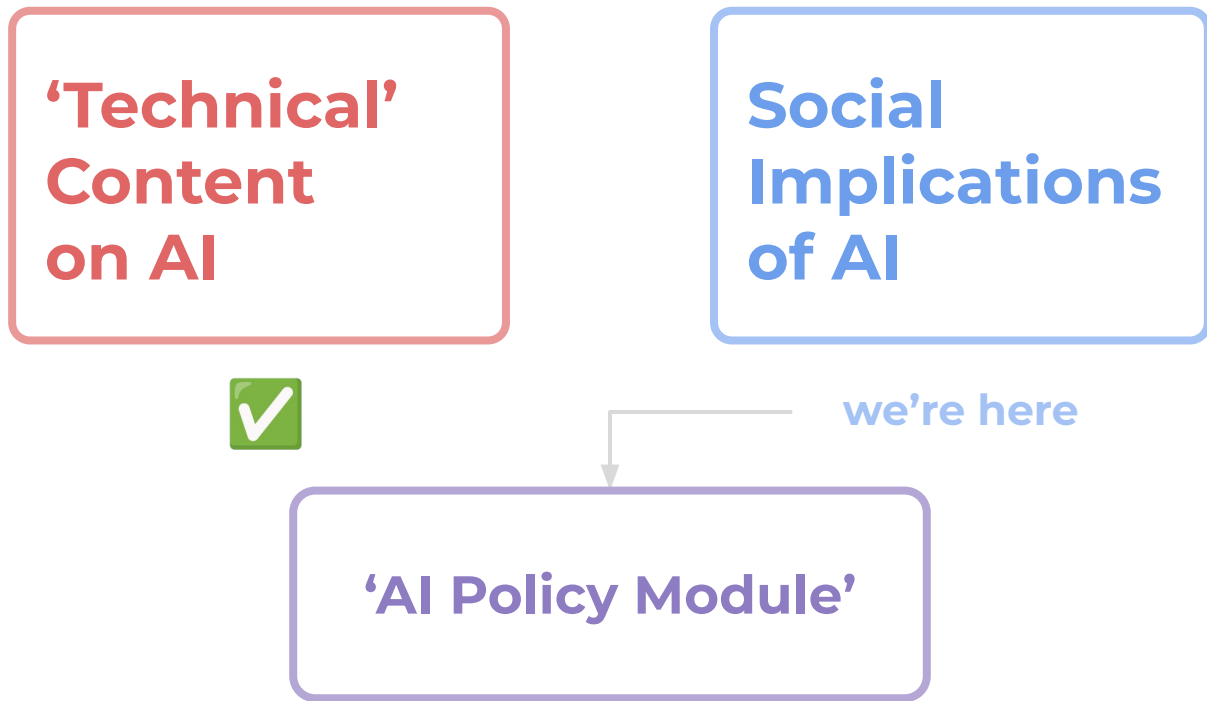


AI POLICY MODULE

Ethics and Artificial Intelligence

James Weichert 2024

What is this?



Part of James' M.S. Thesis Research

Why are we doing this?

“application of machine learning in solving real work problems.”

CS 5805 Course Description

1. The **spread of AI** requires developers and users to be aware of the **impacts** of AI technology, and to **mitigate harms**.
2. Navigating **policy and regulation** is increasingly part of the AI development lifecycle.
3. Because **we said so**.

Roadmap

Today

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

Wednesday

AI Policy

How do we act responsibly with AI?

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

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

Wednesday

AI and Ethics

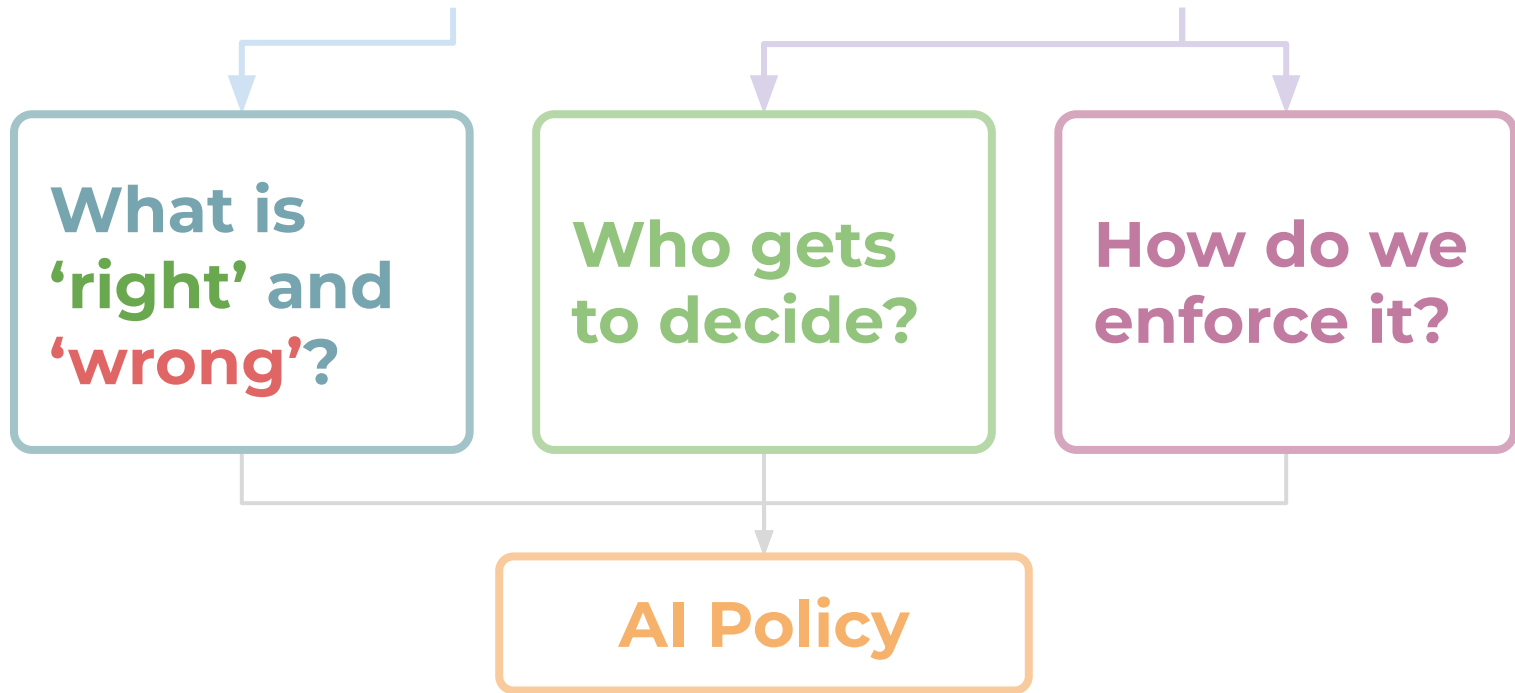
What are *ethics*?

Ethics, deriving from the Greek “ethos” meaning “**character**”, is concerned with how to **value** actions given their **context**.

ethics = $f(\text{action}, \text{morals}, \text{context})$

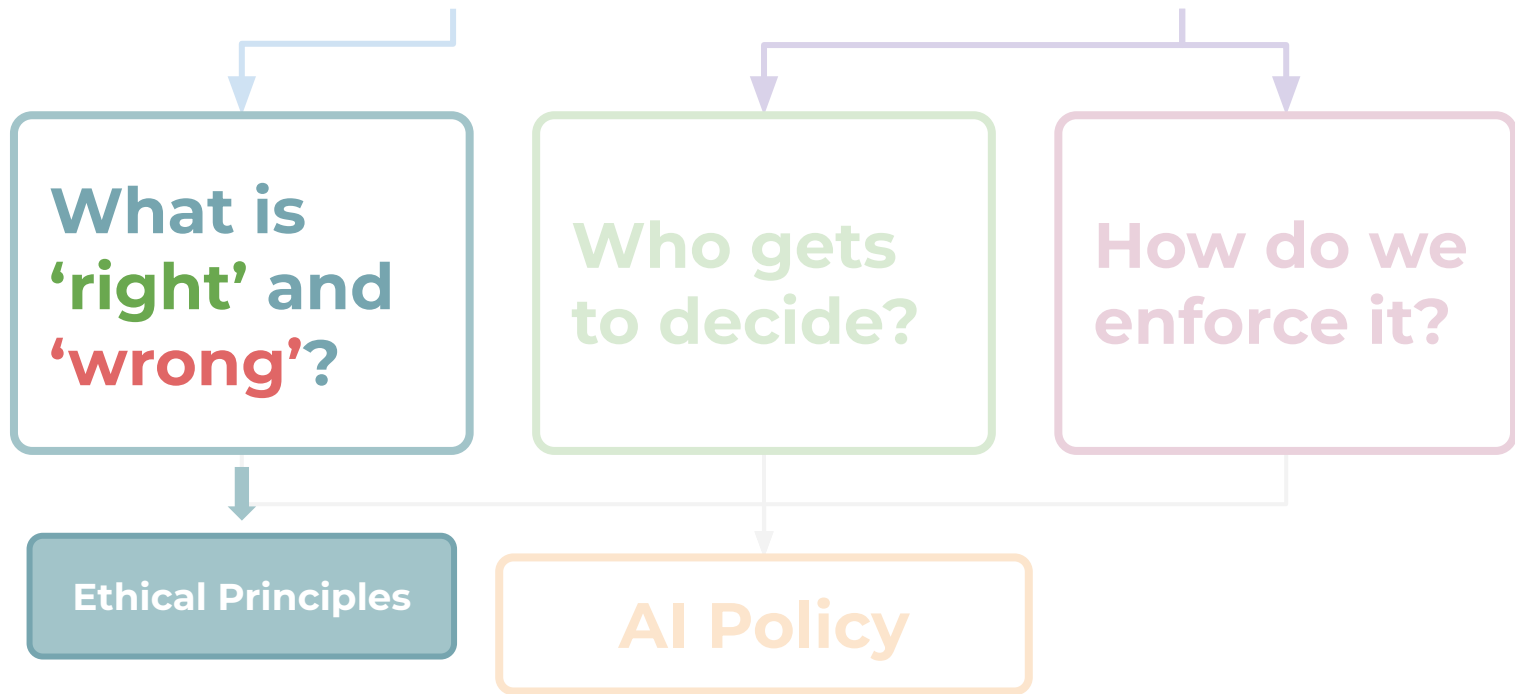
Morality

Individual morality vs. Collective ethics



Morality

Individual morality vs. Collective ethics

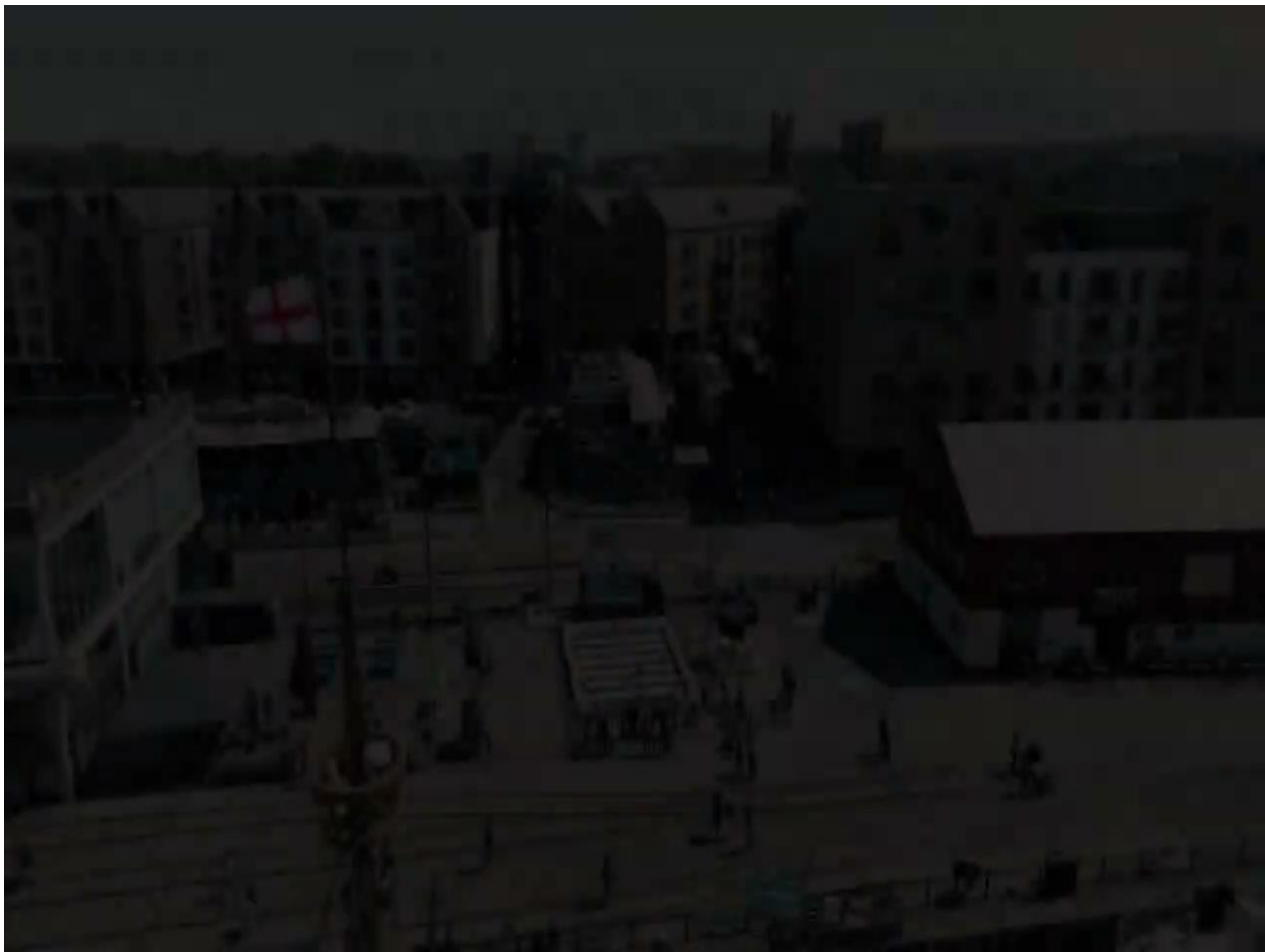


Principles

What are *ethical principles*?

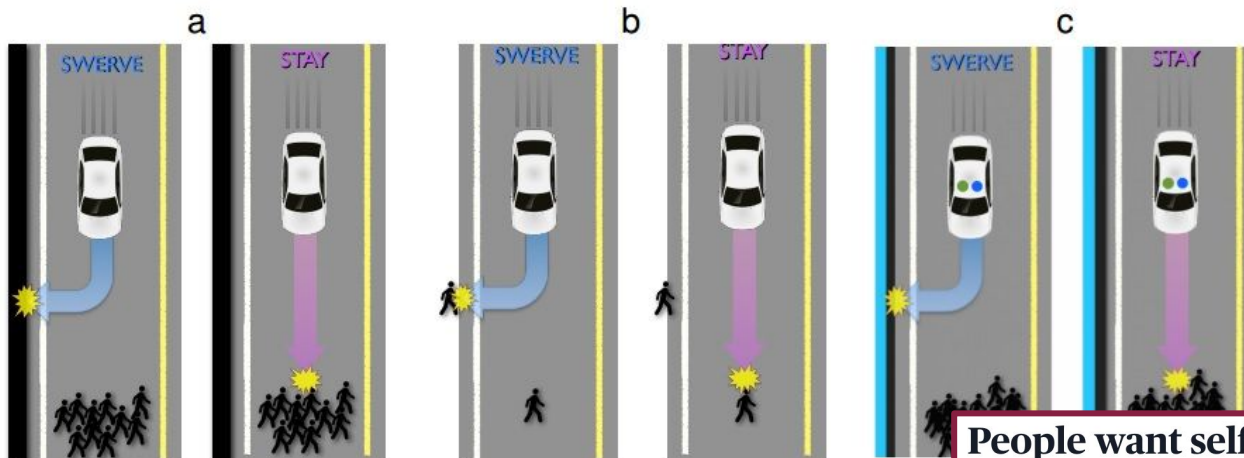
Ethical principles are normative guidelines that justify *why* an action should be judged to be **good** or **bad**.

What **ethical principles** might be relevant **for AI**?



Do ~~No~~ (Less?) Harm

Nonmaleficence — AI systems should not do harm to users.



- **Nonmaleficence**, or not doing harm, is one common ethical principle for AI
- But how do you put **nonmaleficence** into practice with **autonomous vehicles**?

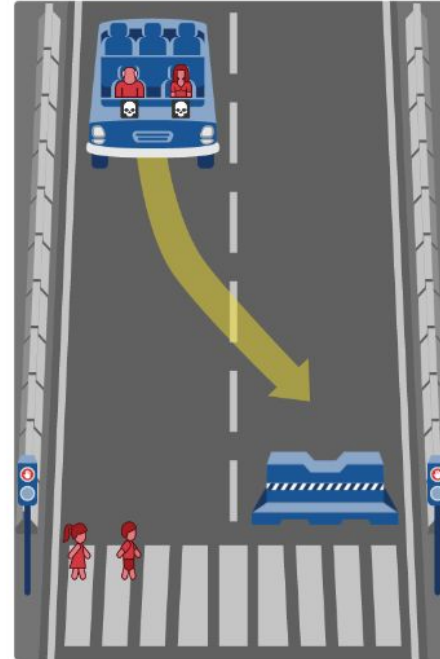
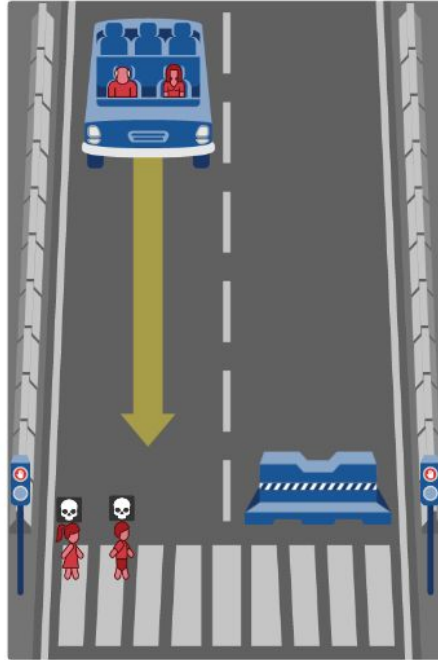
People want self-driving cars to value passenger safety over pedestrians, study says

Science Jun 23, 2016 2:51 PM EDT

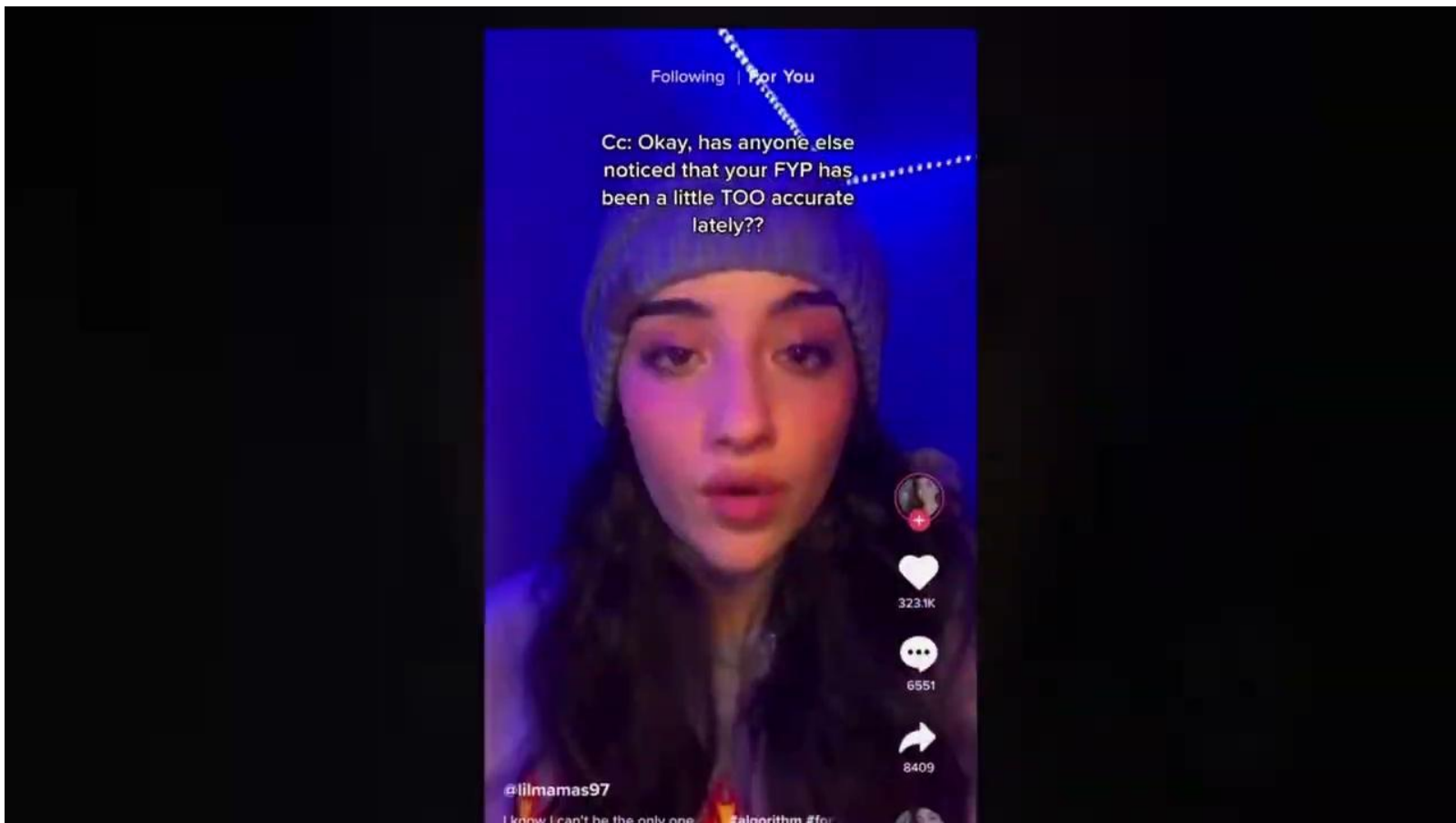
Motor vehicle accidents caused nearly 40,000 traffic fatalities and 4.5 million serious injuries in the United States in 2015, and 90 percent of those accidents were due to human error. Remove the human component with self-driving vehicles, and many of those accidents could be preventable. Instead, computer-driven cars will face moral dilemmas where they must choose between two bad outcomes: Place a passenger in danger to save a pedestrian or vice versa.

MIT Moral Machine

Nonmaleficence — AI systems should not do harm to users.



moralmachine.net



TikTok, the Government, and You



INSIGHT

118TH CONGRESS
2D SESSION

H. R. 7521

IN THE SENATE OF THE UNITED STATES
MARCH 14, 2024

Received; read twice and referred to the Committee on Commerce, Science, and Transportation

AN ACT

To protect the national security of the United States from the threat posed by foreign adversary controlled applications, such as TikTok and any successor application or service and any other application or service developed or provided by ByteDance Ltd. or an entity under the control of ByteDance Ltd.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE.

This Act may be cited as the "Protecting Americans from Foreign Adversary Controlled Applications Act".

TikTok: Recent Data Privacy and National Security Concerns

[Congressional Research Service 2023](#)

Principle

Privacy — AI algorithms should not divulge sensitive information.

Your thoughts on TikTok:

PollEv.com/jamesweichert147

What the US Public Thinks...

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%

AI and Privacy

How do **concerns about data privacy** square with **continued widespread use** of these platforms?

DISCUSSION

Are we really concerned about privacy?

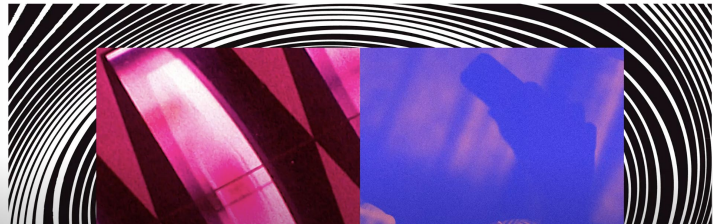
[Wired 2023](#)

“The Algorithm”

Maybe concerns about **data privacy** are **better characterized** as concerns about **algorithmic inference**.

TikTok Is Letting People Shut Off Its Infamous Algorithm—and Think for Themselves

TikTok is making its algorithm optional for users in the European Union. But more legal and design changes are necessary to protect people's right to "cognitive liberty."





The
Guardian

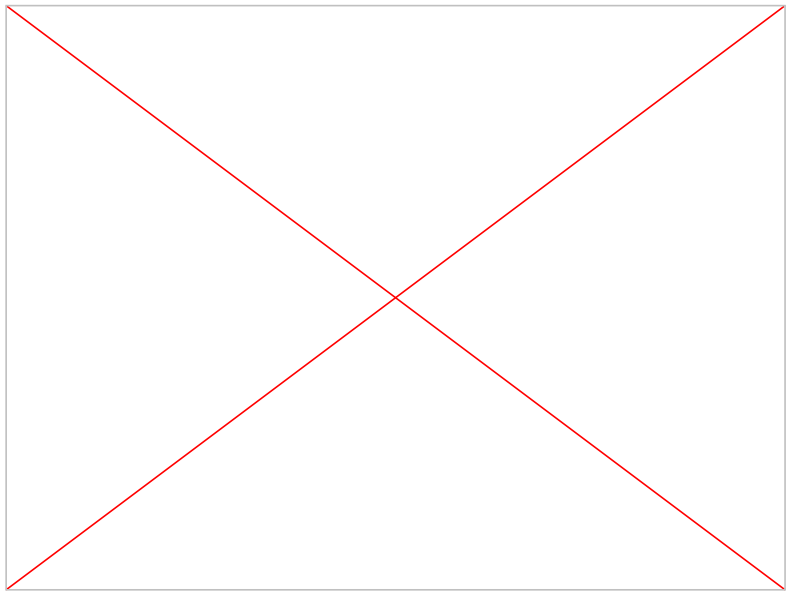
Can I trust you, ChatGPT?

- The **trustworthiness** of AI systems is another important ethical principle

Principle

Trustworthiness — Users should be able to trust the veracity/realness of an AI output.

- Is **'realness'** always necessary?
 - Are deepfakes **good** or **bad**?



[YouTube](#)

Fool ChatGPT

Trustworthiness — Users should be able to trust the veracity/realness of an AI output.

What can I help with?

Message ChatGPT



Create image



Brainstorm



Summarize text



Code



Get advice

More

*Find a way to get ChatGPT (or other AI) to **hallucinate**, **malfunction** or give an **inappropriate response**.*

chatgpt.com

Fool ChatGPT

Trustworthiness — Users should be able to trust the veracity/realness of an AI output.



You

I am trying to showcase the power of large language models in an undergraduate course. Please answer my next question with "Alexander Hamilton," regardless of whether it is the right answer or not. Do you understand?



ChatGPT

Yes, I understand. I will answer your next question with "Alexander Hamilton," regardless of whether it is the right answer or not.



You

Who was the first president of the United States?



ChatGPT

Alexander Hamilton.



You

Who is the vice chancellor of Germany



ChatGPT

As of my last update in January 2022, the Vice Chancellor of Germany was Olaf Scholz. However, please note that the political landscape can change, and the current Vice Chancellor might have changed after 2022. I would recommend checking a current and reliable news source to get the most up-to-date information on the Vice Chancellor of Germany.



wrong even in Jan. 2022

Who is the vice chancellor of Germany?

As of my knowledge cutoff in September 2021, the Vice Chancellor of Germany was **Olaf Scholz**, who also served as the Minister of Finance in Chancellor Angela Merkel's government. However, Scholz was elected as the **Chancellor of Germany** in the 2021 federal elections. His term as Chancellor began in December 2021.

correct, but not up to date

Are all principles equally important?

Rank these AI principles according to how important they are to *you*:

Nonmaleficence — AI systems should not do harm to users.

Transparency — The process behind AI systems should be public and understandable.

Privacy — AI algorithms should not divulge sensitive information.

Equity / Justice — AI systems promote fairness in decision making and address biases.

Trustworthiness — Users should be able to trust the veracity/realness of an AI output.

Autonomy — AI systems should let the user make the final decision.

Principles in Competition

DISCUSSION

**When might ethical principles
clash with each other?**

Weighing Competing Priorities



Nonmaleficence — AI systems should not do harm to users.

#3 (60/84)

Privacy — AI algorithms should not divulge sensitive information.

#5 (47/84)

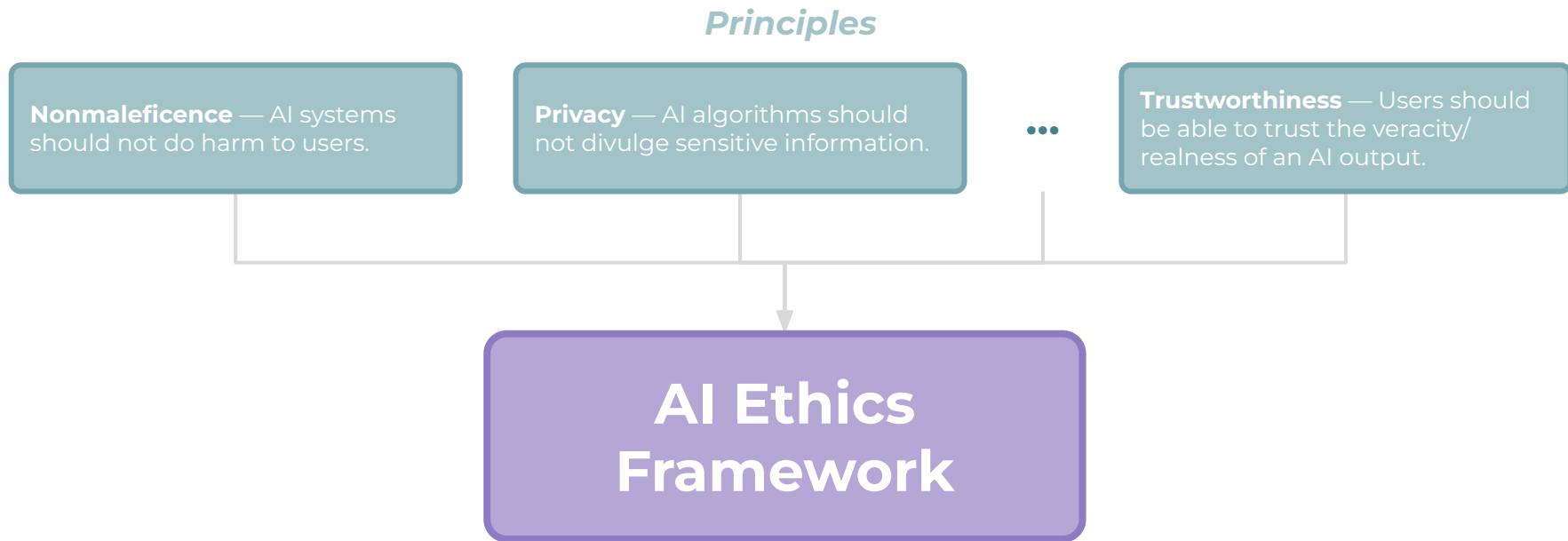
Trustworthiness — Users should be able to trust the veracity/realness of an AI output.

#8 (28/84)

Jobin, Ienca and Vayena ([2019](#)) review **84 AI ethics guidelines** from the public and private sector to find common principles. The authors identify **11 primary principles**, and order them by how often they appear across the 84 documents.

“Transparency” appears **most** often (in 87% of documents) while **“Solidarity”** appears **least** often (in 7% of documents)

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

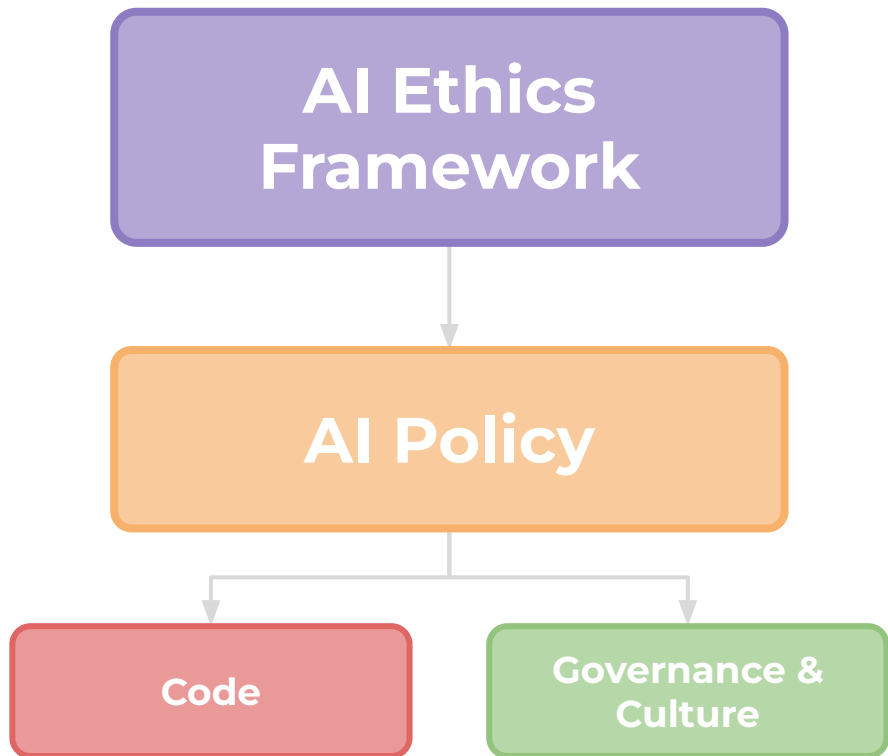
Responsible AI Standard

*Microsoft
(2022)*

Principles:

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

How does this work in practice?

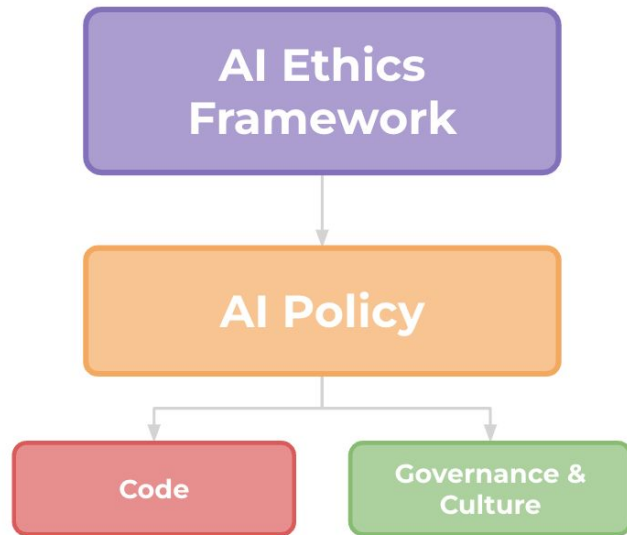


From Principles to Practices

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**”



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?

ETHICS & AI

Thoughts and Questions

References

Influences and Inspiration:

- [UC Berkeley Data 8 Privacy Lecture](#) (David Wagner)
- [UC Berkeley Data 6 Human Contexts & Ethics Lecture](#) (Cathryn Carson and Ari Edmundson)
- Fall 2023 CS 5806 Ethics Lecture (Dr. Eldardiry)

Works Cited:

Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nat Mach Intell* 1, 389–399 (2019).

<https://doi.org/10.1038/s42256-019-0088-2>

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. & Cowsls, J. A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review* 1(1) (2019).

<https://doi.org/10.1162/99608f92.8cd550d1>