GUEST LECTURE

Ethics and Machine Learning

James Weichert April 11, 2024

BY NC SA James Weichert 2024

James Weichert

he/him

M.S. Student @ VT ML Lab

- I'm a first-year M.S. student at VT advised by Dr. Eldardiry
- My research focuses on **AI ethics and policy** and I have an interest in CS/AI education
- I'm a big fan of cute dogs

BY NC SA **James Weichert** 2024

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(action, morals, context)

What do **ethics** have to do with **artificial intelligence?**

BY NC SA **James Weichert** 2024

Principles

What are *ethical principles*?

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

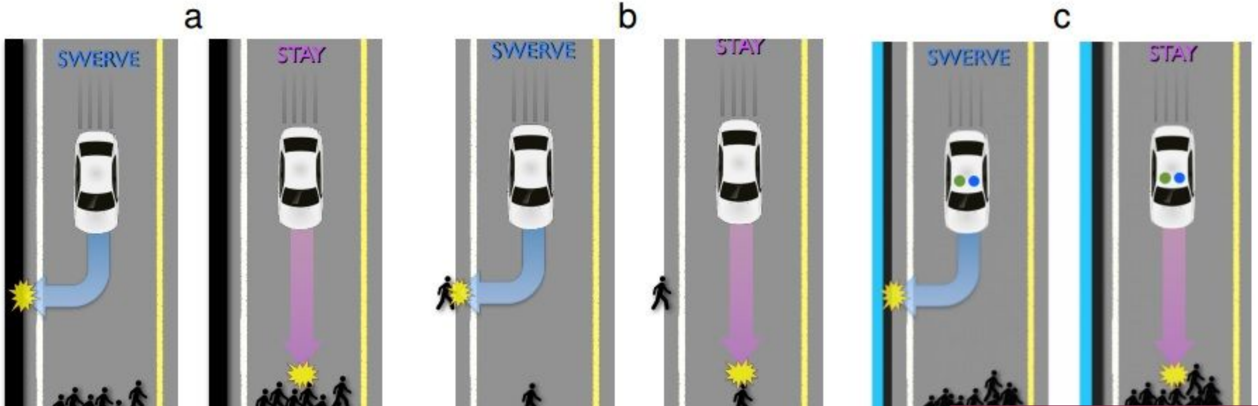
What **ethical principles** might be relevant **for AI?**

PollEv.com/jamesweichert147

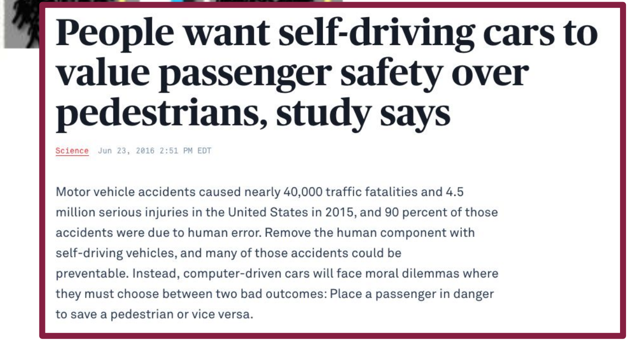
Watershed 2023

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.

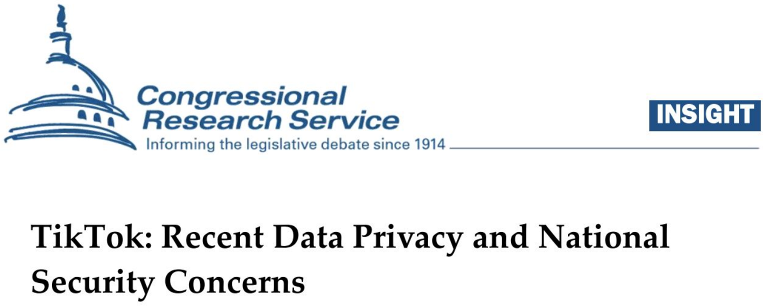
PBS News Hour 2016

TIK TOK HQ
so tik tok knows your ethnicity now?
WSJ | VIDEO INVESTIGATION

[The Wall Street Journal 2021]

BY NC SA

TikTok, the Government, and You



Congressional Research Service
Informing the legislative debate since 1914

INSIGHT

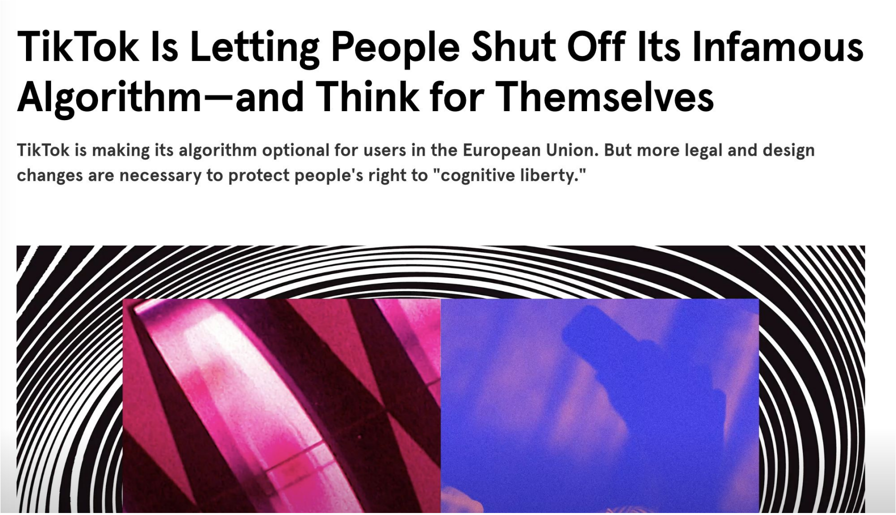
TikTok: Recent Data Privacy and National Security Concerns

Congressional Research Service 2023

Privacy —
AI algorithms should not divulge sensitive information.

Are we really concerned about *privacy*?

“The Algorithm”



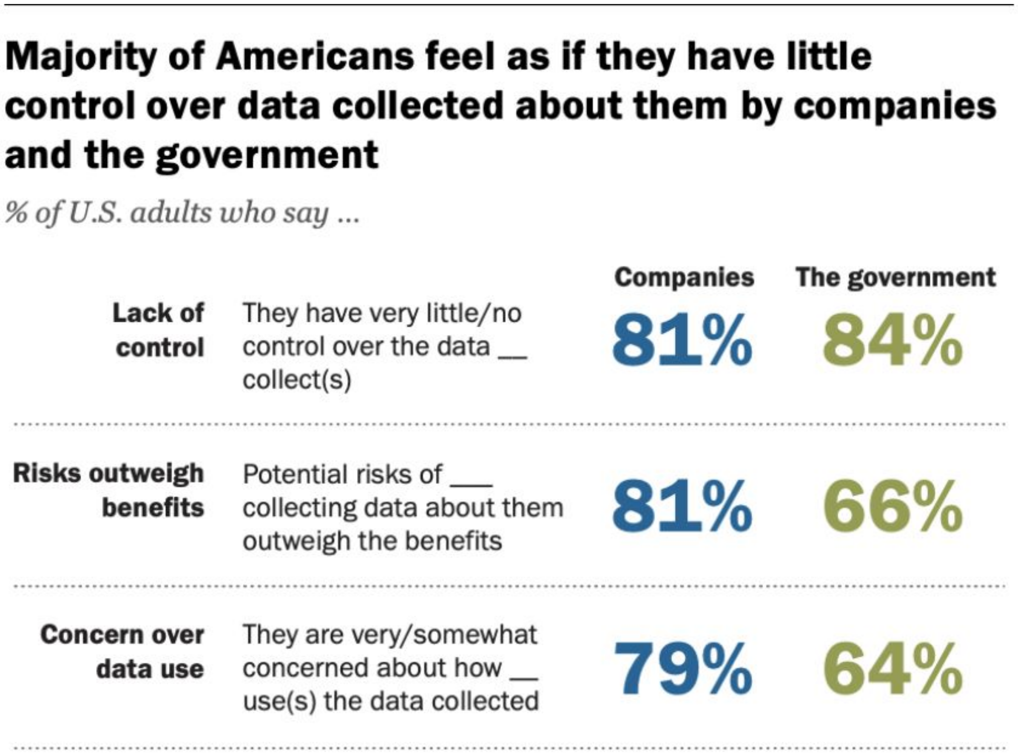
TikTok Is Letting People Shut Off Its Infamous Algorithm—and Think for Themselves

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."

Wired 2023

James Weichert 2024

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%

Pew Research 2019

James Weichert 2024

Guardian News 2024

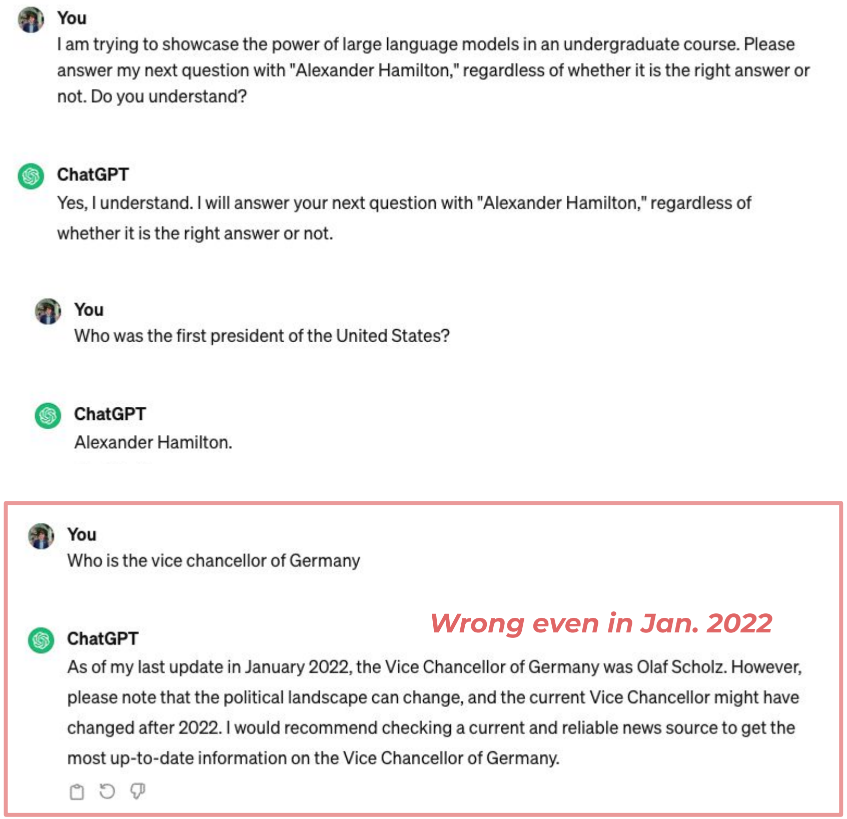
James Weichert 2024

Can I Trust You, ChatGPT?

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

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

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



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

Wrong even in Jan. 2022

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.

cc BY NC SA

Weighing Competing Priorities

Nonmaleficence —
AI systems should not do harm to users.

#3 (60/84)

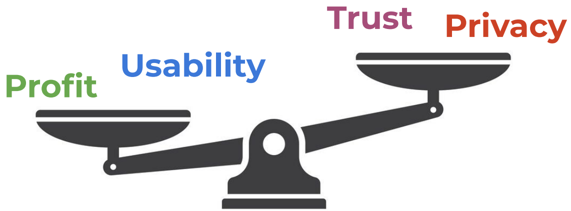
Privacy —
AI algorithms should not divulge sensitive information.

#5 (47/84)

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

#8 (28/84)

How would you prioritize these principles?



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)

James Weichert 2024

Components of Ethical ML

Data

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

Model

- What **metric(s)** is the model optimizing?
- How much **confidence** do I have in the model's predictions?

Action

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

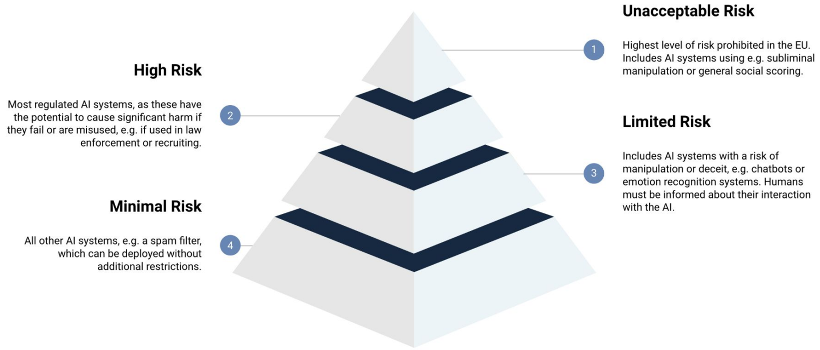
CC BY NC SA **James Weichert** 2024

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**”



Unacceptable Risk

① Highest level of risk prohibited in the EU. Includes AI systems using e.g. subliminal manipulation or general social scoring.

High Risk

② Most regulated AI systems, as these have the potential to cause significant harm if they fail or are misused, e.g. if used in law enforcement or recruiting.

Limited Risk

③ Includes AI systems with a risk of manipulation or deceit, e.g. chatbots or emotion recognition systems. Humans must be informed about their interaction with the AI.

Minimal Risk

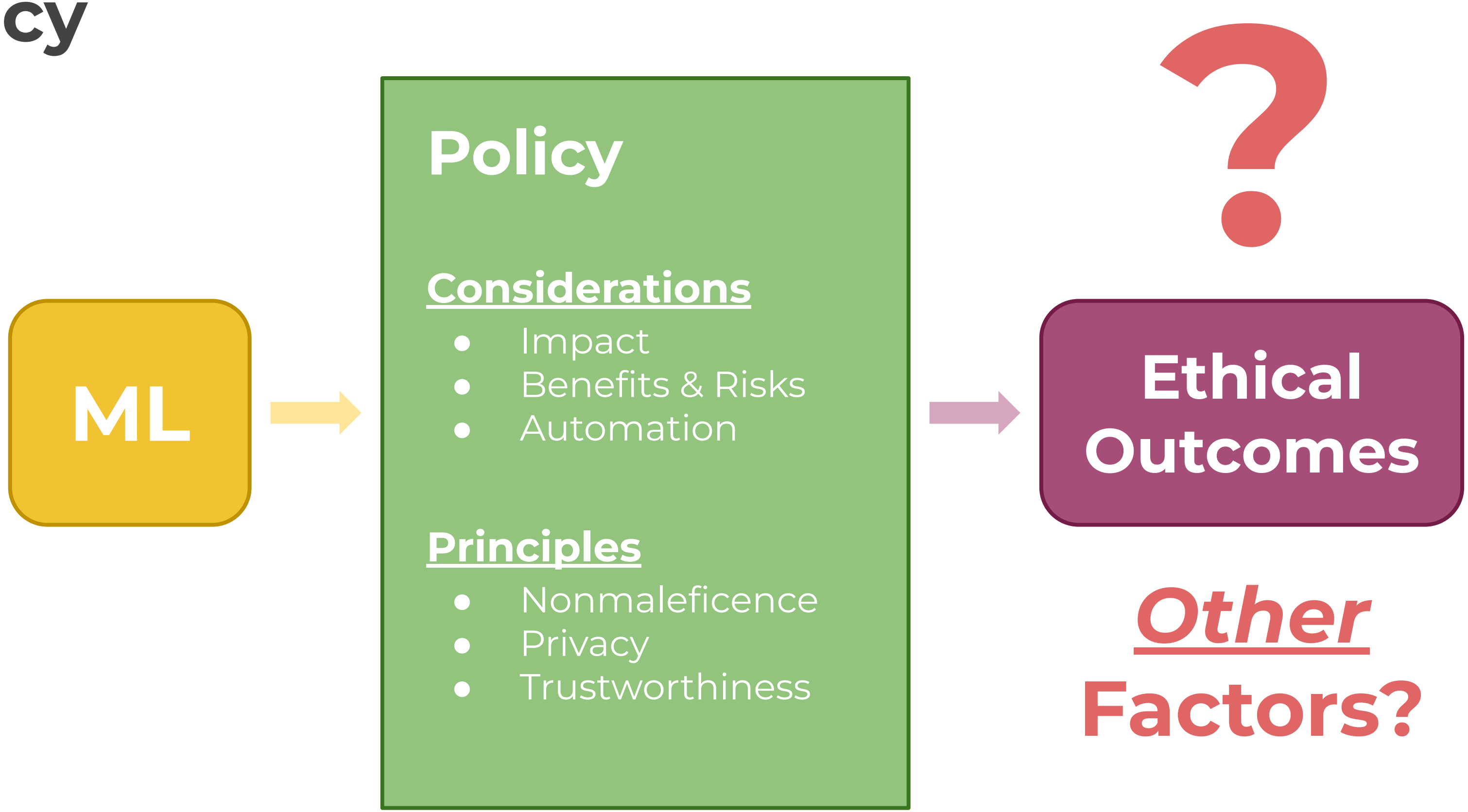
④ All other AI systems, e.g. a spam filter, which can be deployed without additional restrictions.

EU AI Act

Trail ML 2023

CC BY NC SA

Policy



ML

Policy

Considerations

- Impact
- Benefits & Risks
- Automation

Principles

- Nonmaleficence
- Privacy
- Trustworthiness

?

Ethical Outcomes

***Other* Factors?**

James Weichert 2024

ETHICS & ML

Thoughts and Questions

BY NC SA James Weichert 2024

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

BY NC SA