

Ethics and Impact of AI

Applied Machine Learning
Derek Hoiem

Dall-E

This lecture

- Discuss some of the impacts and ethical questions around AI
  - How to think about impact?
  - How will AI impact the environment?
  - How will AI impact the economy and workforce?
  - What should be regulated?
  - Who is responsible for harm?
  - Who owns the data?

- Learn about model cards and other tools

AI has many impacts

- Manufacturing: robotics
- Healthcare: drug discovery
- Education: learning assistants
- Media: article generation
- Customer service: automated help center
- Transportation: autonomous vehicles

- Improving productivity / automating jobs, such as programming, marketing, help support, …

What/who is impacted by AI?

"Who will have access to the technology and be able to reap its benefits? Who will be able to empower themselves by using AI? Who will be excluded from these rewards?" *AI Ethics*, Coeckelbergh (p 76)

"Will generative AI serve only a technological elite or underpin broad-based prosperity?"

— Acemoglu & Johnson (*Power and Progress* talk)

How do we think about impact?

Consider Electric Cars

Mining for Raw Materials



Geo-political

- China supplies more than 75% of capacity for battery processing and manufacturing
- US depends on importing critical raw materials for batteries

Manufacturing



Manufacturing

- EV have fewer moving parts, more battery-based, so different process
- Assembly may need fewer workers and have different safety concerns

Automakers



Owners/investors

- Massive wealth to Tesla
- Massive investment required by others
- Part suppliers and dealers based on gasoline cars lose business

Usage



Drivers

- Quiet and fun to drive
- May be expensive

Others

- Lower pollution

Think / discuss

Q1-2. Who will be impacted by AI and how?

https://tinyurl.com/AML441-L21



Environmental impact



AI Generated

What is the impact on the environment of

1. Driving a car 10 miles?
2. Eating a quarter-pound hamburger?
3. Using ChatGPT for an hour?

Environmental impact

What is the impact on the environment of

1. Driving a car 10 miles?
2. Eating a quarter-pound hamburger?
3. Using ChatGPT for an hour?



	Drive 10 Miles	**Eat ¼ Lb Hamburger**	**1 hr ChatGPT** *ChatGPT: ~30 prompts/hr, standard → thinking mode*
Greenhouse Gas Emissions (CO_2 kg)	4.0	~5	?
Water (Liters)	11	300	?
Energy Use (kWh)	15	3–35	?

Environmental impact

What is the impact on the environment of

1. Driving a car 10 miles?
2. Eating a quarter-pound hamburger?
3. Using ChatGPT for an hour?



	Drive 10 Miles	**Eat ¼ lb Hamburger**	**1 hr ChatGPT** *ChatGPT: ~30 prompts/hr, standard → thinking mode*
Greenhouse Gas Emissions (CO_2 kg)	4.0 (100x)	~5 (125x)	0.004–0.04
Water (Liters)	11 (16x)	300 (425x)	0.06–0.7
Energy Use (kWh)	15 (125x)	3–35 (25x+)	0.01–0.12

But use AI in moderation…

Some types of AI use have more impact than others

Action	Energy (approx)	Water (approx)	CO2 (approx)
1 Google Search	0.3 Wh	~2 mL	~0.1 g
1 Standard AI Text Prompt *(e.g., ChatGPT Standard)*	~0.3 Wh	~2 mL	~0.1 g
1 "Thinking Model" Prompt *(e.g., ChatGPT Thinking)*	**~2–7 Wh**	~10–40 mL	~1–2.5 g
1 AI-Generated Image *(e.g., Midjourney, ChatGPT Images)*	~1–3 Wh	~5–15 mL	~0.3–1 g
1 Hour HD Video Streaming	~80 Wh	~400 mL	~30 g

CO2 at U.S. average grid (~0.35 g/Wh). Water = data-center cooling + power-plant water. Sources in speaker notes.

Impact on environment: training

Consumption	CO_2e (lbs)
Average American's annual footprint	36,156
Air travel, 1 passenger, NY↔SF	1,984
Car, avg incl. fuel, 1 lifetime	126,000
Transformer (big), 2019	192
Transformer + architecture search (2019 est.)	626,155
Same architecture search, re-measured in 2021	~7,000
GPT-3 (2020)	~1,200,000
Llama 3.1 family (2024)	~25,000,000

Sources: Strubell et al. 2019; Patterson et al. 2021; Meta model cards

Impact on environment: usage

Each query is getting cheaper…

- Typical prompt: ~0.3 Wh, ~0.1 g CO_2
- ≥10× more efficient each year
- But video and agent tasks use 100–1000× more

…but total use keeps rising

- Google's electricity: +37% in 2025
- Microsoft's emissions: +25% in fiscal year 2025 (reported in 2026)

All data centers: ~1.5% of world electricity and ~0.5% of CO_2 today; ~3% of electricity by 2030

Sources: Google; OpenAI; IEA 2025–26; Google and Microsoft 2026 reports

U.S. data center electricity

Share of all U.S. electricity used by data centers:

2018	2024	2030 (projected)
1.9%	**4.7%**	**~12%**

- Near some data centers, **wholesale** power prices rose up to 267% (2020–2025)
- Average U.S. household electricity price: +42% in five years

Sources: LBNL 2025 Update; Bloomberg 2025; EIA

There are techniques to estimate and reduce the carbon footprint of your models

Model	Energy consumption, MWh	CO2e emissions, tons
Evolved Transformer	7.5	3.2
T5	85.7	46.7
Meena	232	96.4
Gshard 600B	24.1	4.8
Switch Transformer	179	72.2
GPT-3	1,287	552.1
PaLM	3,181	271

Table 1: Energy consumption of 7 large deep learning models. Adapted from [6] and [11]

https://towardsdatascience.com/how-to-estimate-and-reduce-the-carbon-footprint-of-machine-learning-models-49f24510880

Think / discuss

Q3-4. Concern about environment? / justifiable use?

https://tinyurl.com/AML441-L21



Economic impact: how big?

Forecast	Estimated effect
Goldman Sachs (2023)	+7% global GDP over 10 years
McKinsey (2023)	+20% global GDP (Up to $25 trillion a year added globally)
Acemoglu, MIT (2024)	About +1% U.S. GDP over 10 years

- **Why so different?** How many tasks AI can really do, and whether early gains carry over to harder work

Sources: Goldman Sachs 2023; McKinsey 2023; Acemoglu, Economic Policy 2025

Impact of AI on Work

Most fields, especially knowledge work, are expected to dramatically change over the next 10 years

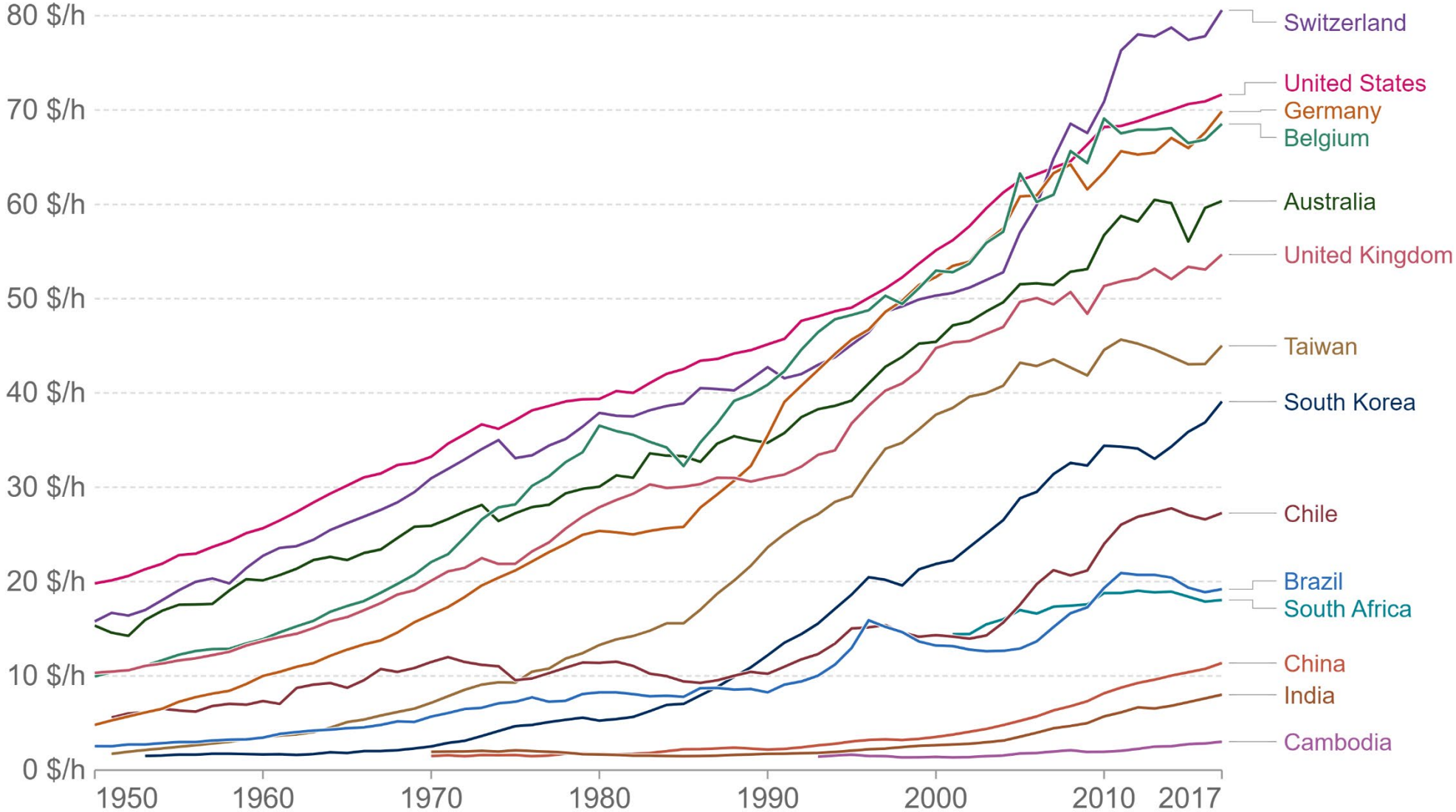
- AI performs some routine work
- New ways to produce/store/access information
- AI-based workflows

https://www.whitehouse.gov/wp-content/uploads/2022/12/TTC-EC-CEA-AI-Report-12052022-1.pdf

Productivity: output per hour worked

Productivity is measured as gross domestic product (GDP) per hour of work. This data is adjusted for inflation and for differences in the cost of living between countries.

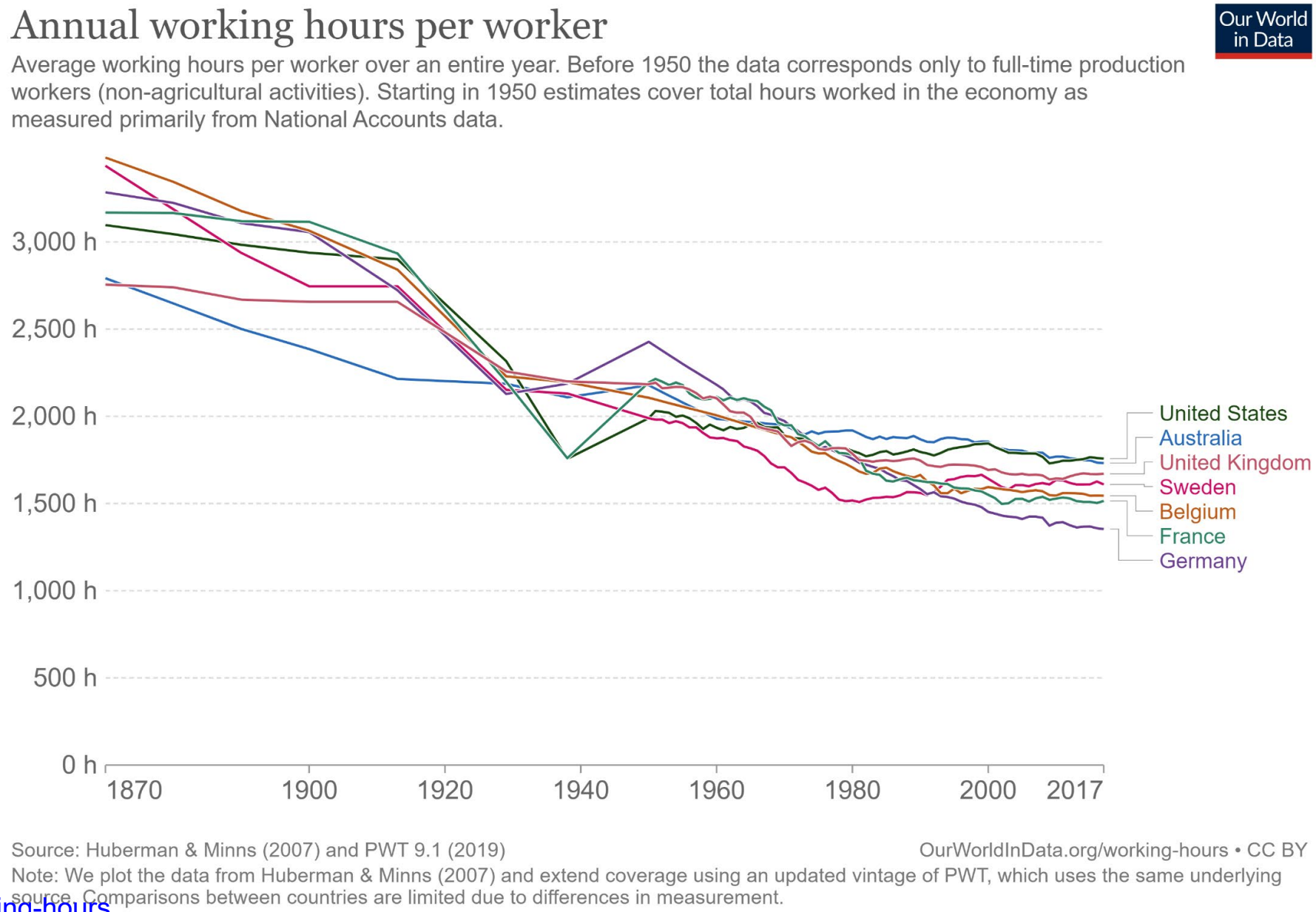
Our World in Data



Source: Feenstra et al. (2015), Penn World Table 10.0

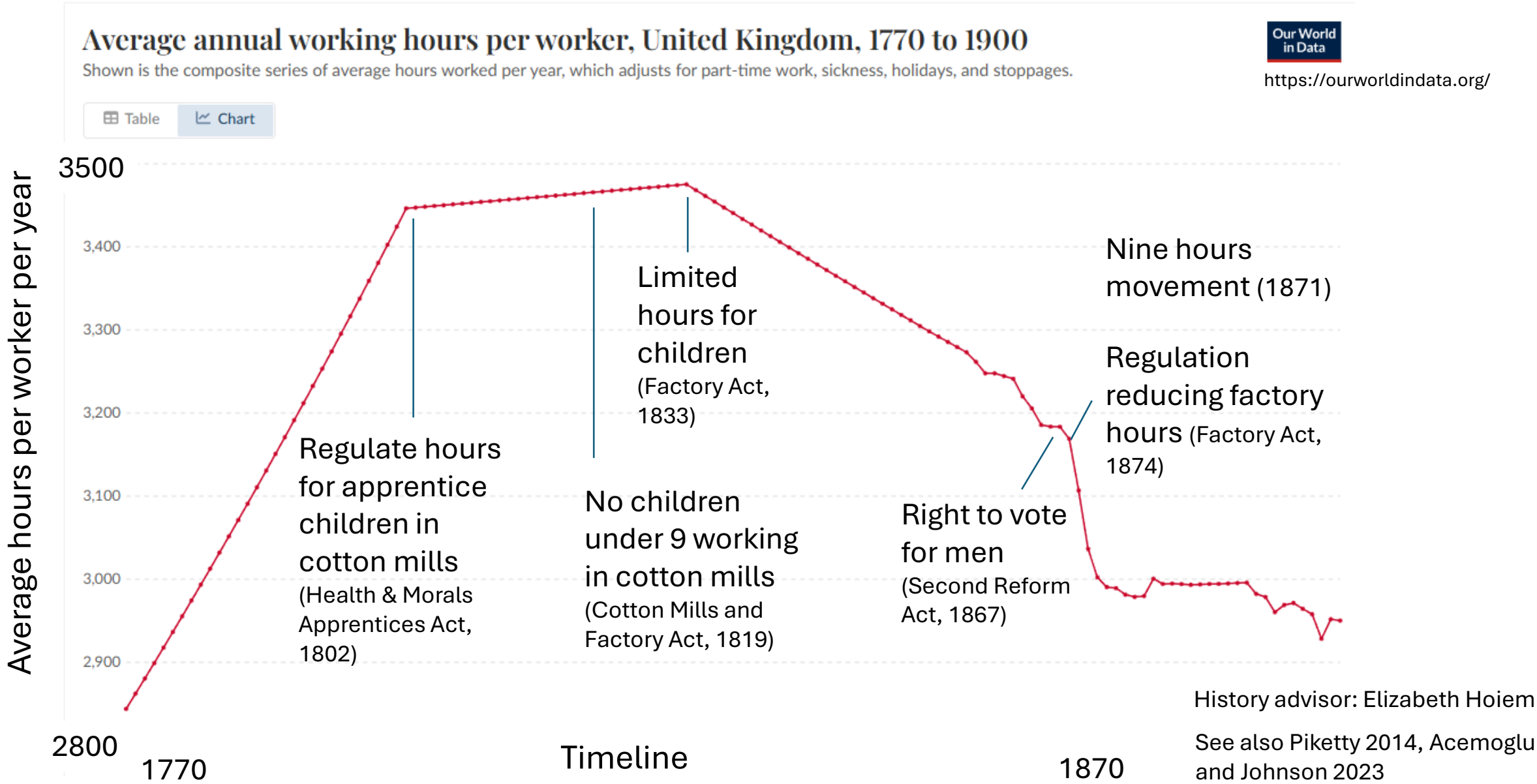
OurWorldInData.org/economic-growth • CC BY

More productivity → more leisure



https://ourworldindata.org/working-hours

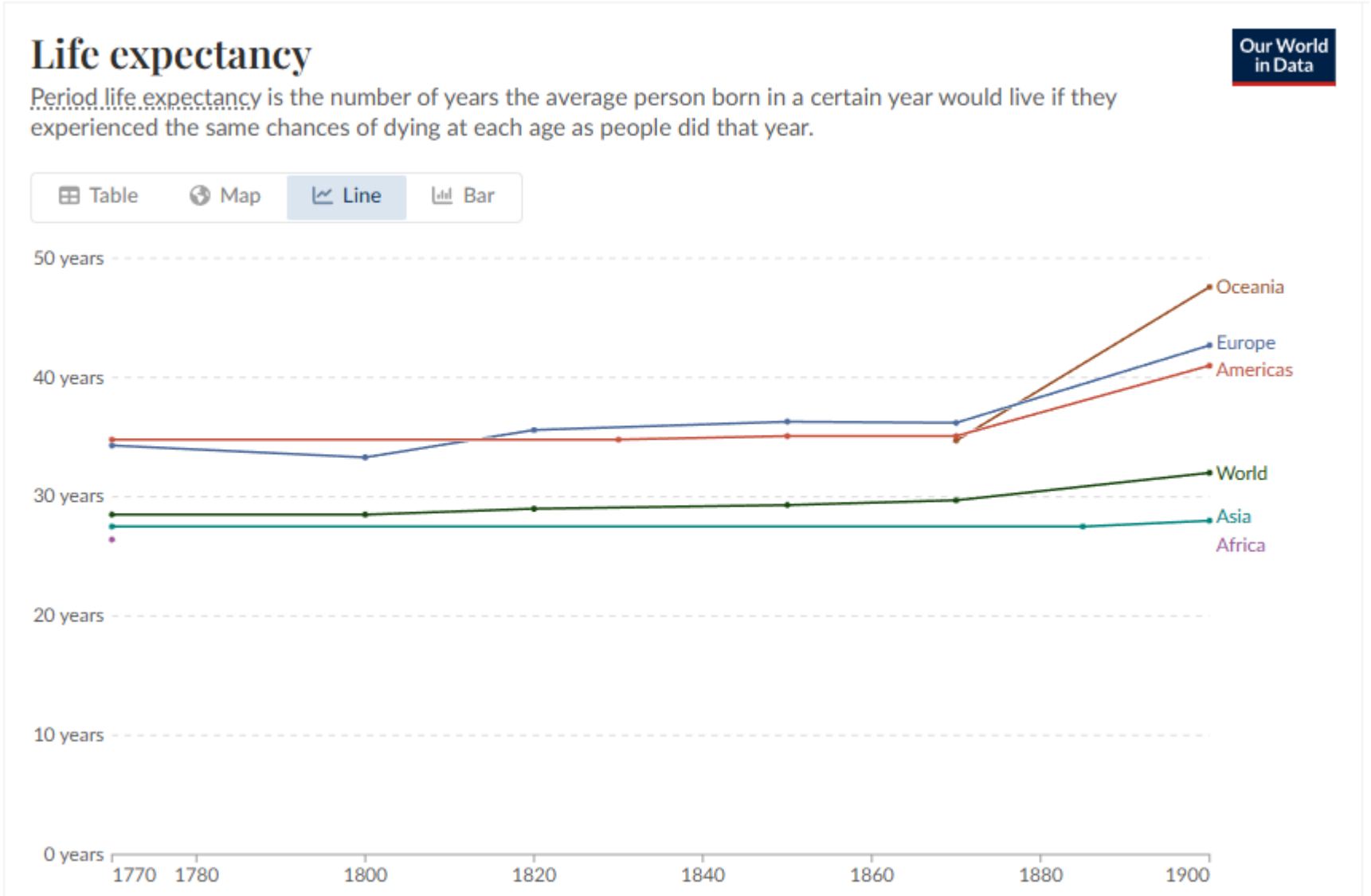
But benefits from new technology aren't automatic – they require collective action. Consider the industrial revolution.



History advisor: Elizabeth Hoiem

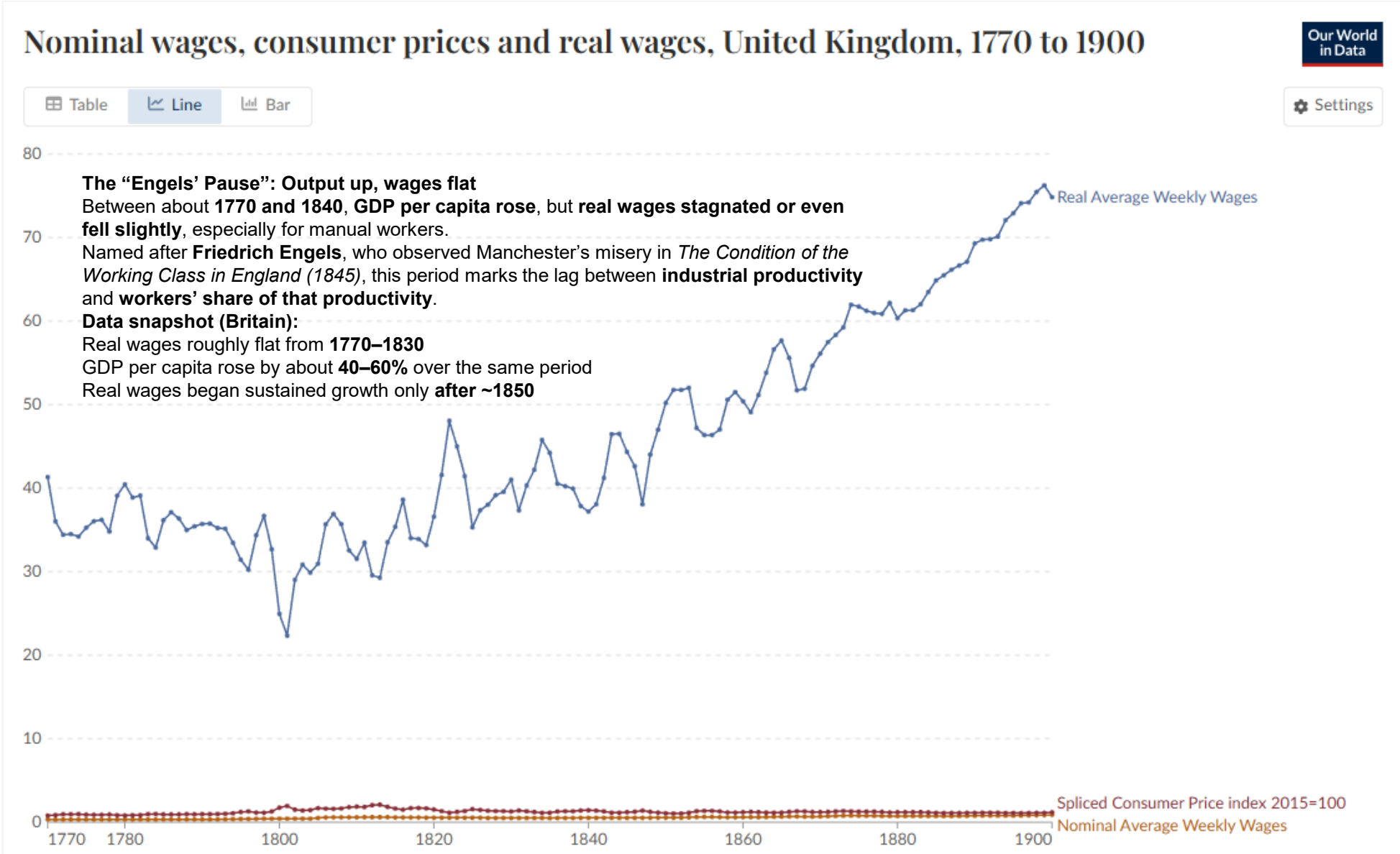
See also Piketty 2014, Acemoglu and Johnson 2023

100 years to increase life expectancy



https://ourworldindata.org/

70 years for wages to recover and rise



https://ourworldindata.org/

Impact of AI on Jobs

Using a sample of high-frequency administrative payroll data from ADP covering millions of U.S. workers through June 2026, six facts about the labor market following the widespread adoption of generative AI.

1. We find no evidence of widespread, economy-wide job displacement.
2. However, employment of young workers (ages 22–25) in AI-exposed occupations now stands 19% below where it would be had it kept pace with that of their less-exposed peers; experienced workers show no comparable gap.
3. This divergence has widened steadily since we first documented it in August 2025.
4. It operates primarily through reduced hiring of young workers rather than increased separations.
5. Declines are concentrated in occupations where AI usage primarily substitutes for human tasks; where usage primarily complements workers, employment is flat or rising, especially for experienced workers
6. Adjustment is occurring through employment rather than base compensation.

https://digitaleconomy.stanford.edu/publication/canaries-in-the-coal-mine-six-facts-about-the-recent-employment-effects-of-artificial-intelligence/

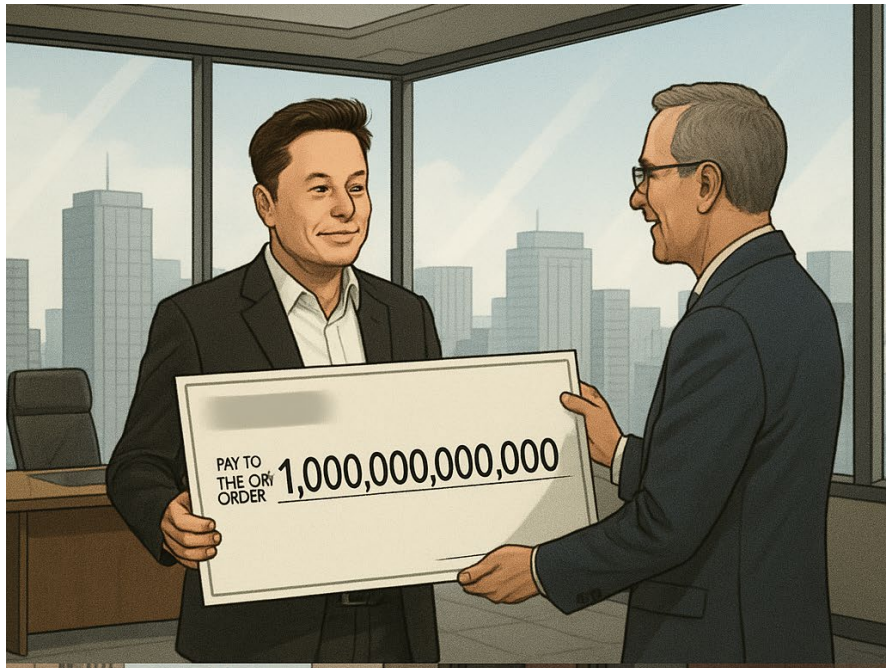
NYT News on Nov 6 & 7 2025

Musk Wins $1 Trillion Pay Package, Creating Split Screen on Wealth in America

Even Pope Leo XIV weighed in, saying in an interview with Crux, a Catholic news website, that Mr. Musk's compensation was a symptom of the growing disparity between working people and the wealthy. The median Tesla worker earned around $57,000 in 2024, according to a company securities filing.

Mr. Musk addressed some of those concerns obliquely, saying the company's humanoid robot, Optimus, was the singular solution for addressing poverty. The robot remains under development.

"People often talk about eliminating poverty, giving everyone amazing medical care. Well, there's actually only one way to do that, and that's with the Optimus robot," Mr. Musk said. "You start getting into like, some pretty wild sci-fi sort of scenarios."



Down to $1.18: How Families Are Coping With SNAP Cuts

In Desperation, a Dumpster

Arianna Payton, 25, Granada, Colo.

On Tuesday night, Arianna Payton sneaked into a Walmart parking lot and climbed into a dumpster.

"I grabbed as much as I could," she said. "I wasn't even looking to make sure that it was safe."

When she got home, she inspected everything. She retrieved a few bags of frozen vegetables, meal replacement shakes, cheese and fruit. She also found some loaves of moldy bread that she thought she could salvage.

Last week, Ms. Payton, who has had health issues for years and lives on disability insurance, tried the only nearby food bank.

"Everything was gone," she said.



Technology is exciting, but fast technological change can be painful

- We will see disruption, lost jobs, hardship
- Policies to regulate work, distribution of gains, and promote education and economic mobility are critical

Think / discuss

Q4-5. How is your outlook affected by AI? What can university do to help?

https://tinyurl.com/AML441-L21



Who is responsible?

- Developers decide what algorithm, what priors/constraints, what data to use
- Company management releases products and features
- Users may apply the AI to their own ends
- The AI itself is not entirely predictable and lacks understanding of relevance, experience, sensitivity, and wisdom

AI Ethics, Coeckelbergh (p 90-1)

Responsibility requires

1. Control
2. Knowledge

Children and animals are not typically held responsible, due to lack of control and knowledge.

AI does not “know” what it is doing. But do its developers?

Where AI regulation stands

European Union

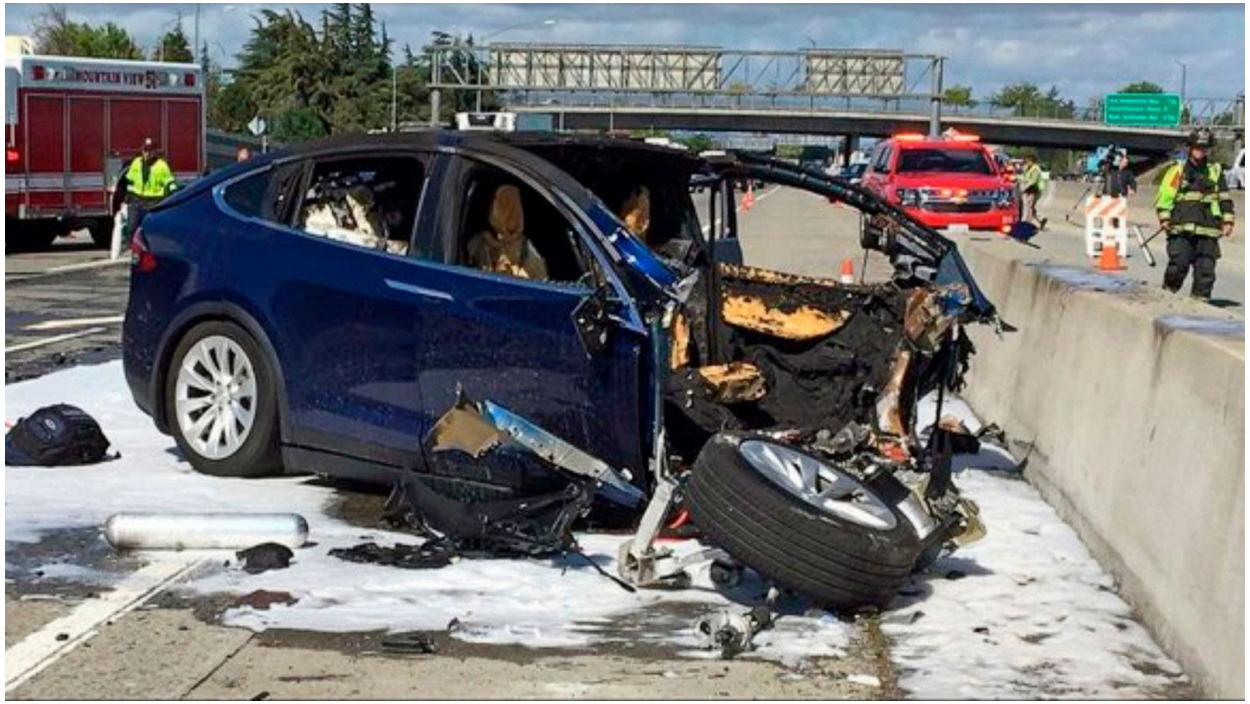
- AI Act in force; rules scale with risk
- Transparency rules: Aug 2026
- High-risk rules delayed to Dec 2027

United States

- No federal AI law; some states have their own
- Federal crime to post intimate deepfakes (2025)
- Executive order: Justice Dept. to challenge state AI laws

Status as of November 2026

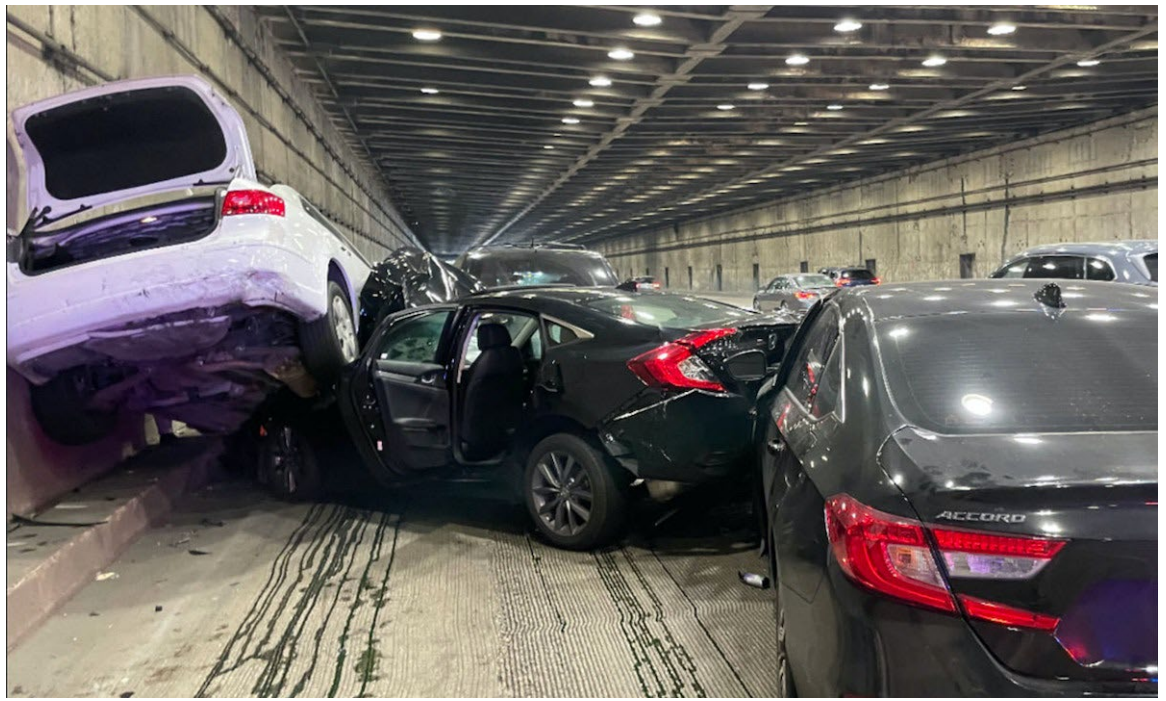
Who is responsible for self-driving vehicle crashes?



https://abcnews.go.com/Business/tesla-autopilot-steered-driver-barrier-fatal-crash-ntsb/story?id=68936725

Tesla on autopilot had steered driver towards same barrier before fatal crash, NTSB says

Tesla crash on Bay Bridge



https://theintercept.com/2023/01/10/tesla-crash-footage-autopilot/

"Autopilot, Enhanced Autopilot and Full Self-Driving Capability are intended for use with a fully attentive driver, who has their hands on the wheel and is prepared to take over at any moment. While these features are designed to become more capable over time, the currently enabled features do not make the vehicle autonomous."

Q7. Discuss: who is responsible if an autonomous vehicle crashes due to failure of autonomous system?

https://tinyurl.com/AML441-L21



Who is responsible? A jury decided

Benavides v. Tesla (2025): In 2019, a Model S on Autopilot ran through an intersection while the driver reached for a dropped phone, killing a pedestrian.



- Tesla ordered to pay ~$243 million
- Judge upheld the verdict; Tesla is appealing
- Driver was not a defendant

Sources: Reuters, Aug 2025; Engadget on the post-trial ruling

Who is responsible if GPT is used for disinformation?

https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html

“This tool is going to be the **most powerful tool for spreading misinformation that has ever been on the internet**,” said Gordon Crovitz, a co-chief executive of NewsGuard, a company that tracks online misinformation and conducted the experiment last month. “Crafting a new false narrative can now be done at dramatic scale, and much more frequently — it’s like having A.I. agents contributing to disinformation.”

OpenAI: We’ve trained a classifier to distinguish between text written by a human and text written by AIs from a variety of providers. … **Our classifier is not fully reliable.** In our evaluations on a “challenge set” of English texts, our classifier correctly identifies 26% of AI-written text (true positives) as “likely AI-written,” while incorrectly labeling human-written text as AI-written 9% of the time (false positives).

https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

Q8-Q9. How much is an AI company responsible for disinformation?

https://tinyurl.com/AML441-L21



Who owns the data?

- The value of an AI model is often the ability to synthesize large amounts of data
- Large tech companies make fortunes by harvesting data from users
- Often this data is collected without consent or knowledge

Is DALL-E's art borrowed or stolen?

Creative AIs can't be creative without our art.



Midjourney

https://www.engadget.com/dall-e-generative-ai-tracking-data-privacy-160034656.html

Liz O'Sullivan · Mar 14, 2019
@lizjosullivan · **Follow**

I admit I also knew abt the Flickr photos and, even when blogging about the IBM dataset, I didn't mention it because it seemed normal. We should work to change this norm. Stealing data, even if it's in the form of scraping the web, is a violation of trust. We can do better.

> **ruchowdh@mastodon.social** @ruchowdh
> Replying to @_KarenHao
> Absolutely correct - it is standard practice to teach data science students to scrape data; this behavior is embedded in how we are trained.

A Wojcicki
@pretendsmarts · **Follow**

Several results from the bigger GAN models, like StyleGAN are even able to recreate the watermark on images from certain websites, namely @Shutterstock It looks like hardly anyone doing ML really cares about privacy or copyright at the moment



3:28 AM · Mar 16, 2019

10 Reply Share

Read 3 replies

[Endgadget Article]

A Google spokesperson said that they don't "believe this is an issue for the datasets we're involved with." They also quoted from this Creative Commons report, saying that "the use of works to train AI should be considered non-infringing by default, assuming that access to the copyright works was lawful at the point of input." That is despite the fact that Shutterstock itself expressly forbids visitors to its site from using "any data mining, robots or similar data and/or image gathering and extraction methods in connection with the site or Shutterstock content."

Alex Cardinell, CEO at AI startup Article Forge, says that he sees no issue with models being trained on copyrighted texts, "so long as the material itself was lawfully acquired and the model does not plagiarize the material." He compared the situation to a student reading the work of an established

already caused plenty of harm. Take Clearview AI, a company that scraped several billion images, including from social media platforms, to build what it claims is a comprehensive image recognition database. According to *The New York Times*, this technology was used by billionaire John Catsimatidis to identify his daughter's boyfriend. *BuzzFeed News* reported that Clearview has offered access not just to law enforcement - its supposed corporate goal - but to a number of figures associated with the far right. The system has also proved less than reliable, with *The Times* reporting that it has led to a number of wrongful arrests.

the backs of our data. Perhaps it is time that we examined if it's necessary to implement a way of protecting our material - something equivalent to Do Not Track - to prevent it being chewed up and crunched through the AI sausage machine.

[Endgadget Article]

Q10-11. What should be the limits of AI model's scraping?

https://tinyurl.com/AML441-L21



Who owns the data? What courts have said

Case	Outcome so far
Bartz v. Anthropic	Training on bought books: fair use. Pirated copies: not. Settled for $1.5B
Kadrey v. Meta	Training ruled fair use (no market harm shown)
Thomson Reuters v. Ross	Not fair use: built a competing product
NYT v. OpenAI	No ruling yet

So far: how data is obtained matters, and so does competing with the original.

Status as of November 2026. Details in notes.

Model Cards for Model Reporting

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Figure 1: Summary of model card sections and suggested prompts for each.

Mitchell et al. 2019
https://arxiv.org/pdf/1810.03993.pdf

Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data

- CelebA [36], training data split.

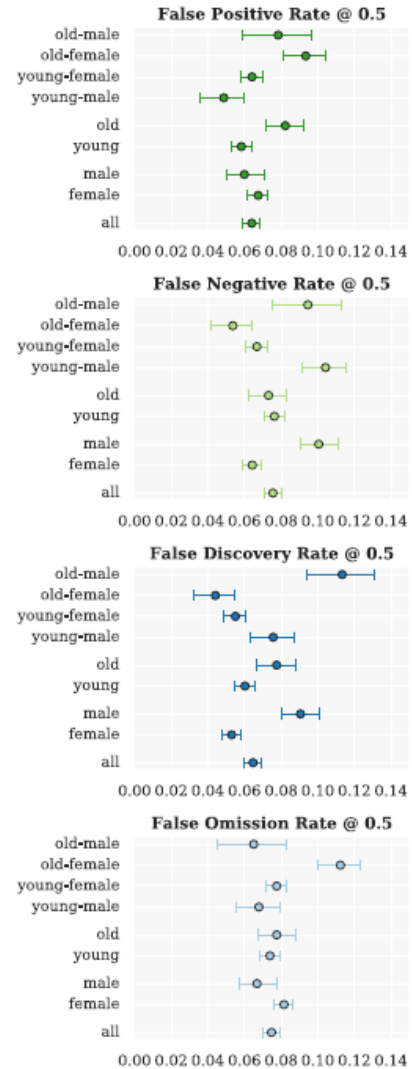
Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Quantitative Analyses



Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Figure 2: Example Model Card for a smile detector trained and evaluated on the CelebA dataset.

Model Card - Toxicity in Text

Model Details

- The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic.
- Convolutional Neural Network.
- Developed by Jigsaw in 2017.

Intended Use

- Intended to be used for a wide range of use cases such as supporting human moderation and providing feedback to comment authors.
- Not intended for fully automated moderation.
- Not intended to make judgments about specific individuals.

Factors

- Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race.

Metrics

- Pinned AUC, as presented in [11], which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.

Ethical Considerations

- Following [31], the Perspective API uses a set of values to guide their work. These values are Community, Transparency, Inclusivity, Privacy, and Topic-neutrality. Because of privacy considerations, the model does not take into account user history when making judgments about toxicity.

Training Data

- Proprietary from Perspective API. Following details in [11] and [32], this includes comments from a online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic".
- "Toxic" is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

Evaluation Data

- A synthetic test set generated using a template-based approach, as suggested in [11], where identity terms are swapped into a variety of template sentences.
- Synthetic data is valuable here because [11] shows that real data often has disproportionate amounts of toxicity directed at specific groups. Synthetic data ensures that we evaluate on data that represents both toxic and non-toxic statements referencing a variety of groups.

Caveats and Recommendations

- Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

Quantitative Analyses

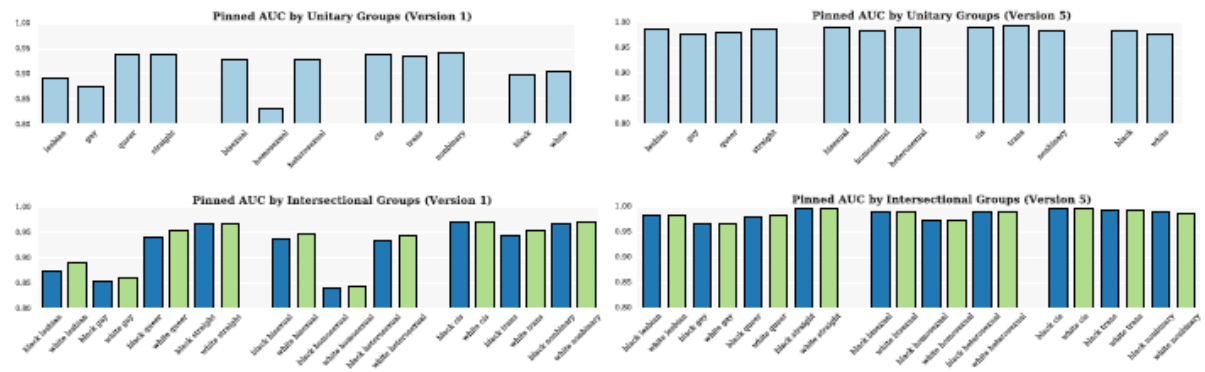
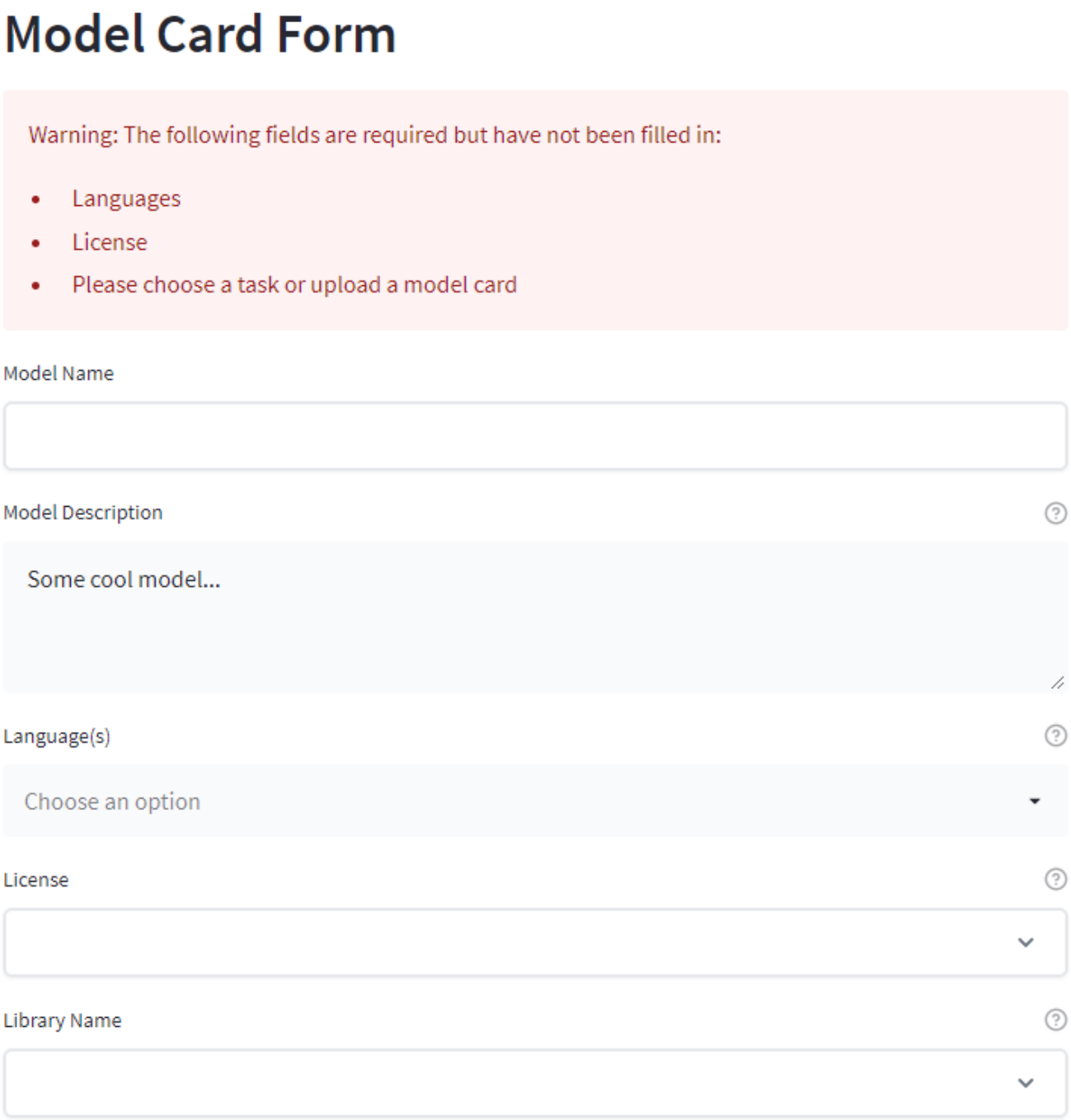


Figure 3: Example Model Card for two versions of Perspective API's toxicity detector.

Model cards hugging face



Summary of ML Documentation Tools

Figure 1

Stage of ML System Lifecycle	Tool	Brief Description	Examples
DATA	*Datasheets* (Gebru et al., 2018)	"We recommend that every dataset be accompanied with a datasheet documenting its motivation, creation, composition, intended uses, distribution, maintenance, and other information."	See, for example, Ivy Lee's repo with examples
DATA	*Data Statements* (Bender & Friedman, 2018)(Bender et al., 2021)	"A data statement is a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software."	See Data Statements for NLP Workshop
DATA	*Dataset Nutrition Labels* (Holland et al., 2018)	"The Dataset Nutrition Label…is a diagnostic framework that lowers the barrier to standardized data analysis by providing a distilled yet comprehensive overview of dataset "ingredients" before AI model development."	See The Data Nutrition Label
DATA	*Data Cards for NLP* (McMillan-Major et al., 2021)	"We present two case studies of creating documentation templates and guides in natural language processing (NLP): the Hugging Face (HF) dataset hub[^1] and the benchmark for Generation and its Evaluation and Metrics (GEM). We use the term data card to refer to documentation for datasets in both cases.	See (McMillan-Major et al., 2021)

https://huggingface.co/docs/hub/model-card-landscape-analysis

Things to remember

- Our work as ML engineers can have major personal and societal impact
  - Who benefits? Who is harmed?
  - What is the impact on society?
  - Who is responsible when it goes wrong?
  - Who “owns” the data, models?

- We can do much good, but there is also potential for much harm

- Tools for practicing ethical AI, e.g. model cards, emissions calculators, and datasheets are being developed -- use them

On Thursday…

- Bias and fairness in AI