Ethics and Impact of AI

Applied Machine Learning
Derek Hoiem

Dall-E

This lecture

- Discuss some of the impacts and ethical questions around AI
- 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)

How do we think about impact?

Consider Electric Cars

Mining for Raw Materials

Geo-political
- China supplies more than 70% of capacity for battery raw materials and manufacturing
- US depends on importing critical raw materials for batteries

Workers
* China's battery material mines may include forced labor

Manufacturing

Manufacturing
- EV have fewer moving parts, more battery-based, so different process
- New factories employ fewer workers and have different safety concerns

Automakers

Owners/investors
- Massive wealth to Tesla
- Massive investment required by others
- Lower barrier to entry, since manufacturing EV from components is simpler

Usage

Drivers
- Quiet and fun to drive
- May be expensive
- Status symbol

Others
- Lower pollution

Discuss with your neighbor

Q1. What is one big impact of AI – who is impacted and how?

https://tinyurl.com/441-L21-fa24

Impact on environment

Energy and Policy Considerations for Deep Learning in NLP

2019

Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

Consumption	**CO_2e (lbs)**
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

Consumer	Renew.	Gas	Coal	Nuc.
China	22%	3%	65%	4%
Germany	40%	7%	38%	13%
United States	17%	35%	27%	19%
Amazon-AWS	17%	24%	30%	26%
Google	56%	14%	15%	10%
Microsoft	32%	23%	31%	10%

Table 2: Percent energy sourced from: Renewable (e.g. hydro, solar, wind), natural gas, coal and nuclear for the top 3 cloud compute providers (Cook et al., 2017), compared to the United States,[4] China[5] and Germany (Burger, 2019).

GPT-3 estimate: ~1.1M lbs

Crypto-assets: 0.3% of global annual greenhouse gas emissions
All data centers: 2% of global emissions

https://arxiv.org/pdf/1906.02243.pdf

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

Economic Impact

- Research by Accenture suggests that AI could double global economic growth rates by 2035, across 12 developed economies. This growth will be driven by a 40% increase in labor productivity, the creation of a virtual workforce capable of problem-solving and self-learning, and the diffusion of innovation that will create new revenue streams.
- Pricewaterhouse Coopers (PwC) estimates that the development and take-up of AI could increase global GDP by up to 14% (equivalent to US$15.7 trillion) by 2030. PwC sees two main channels through which AI will impact the global economy: productivity gains in the near term, based on automation of routine tasks, and the complementing and assisting of the existing workforce with AI technologies to perform tasks better and more efficiently.

https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/637967/EPRS_BRI(2019)637967_EN.pdf

Impact of AI on Work

- At-risk (according to Webb 2020): tasks that involve detecting patterns, making judgments and optimizing
  - Lab technicians
  - Chemical engineers
  - Optometrists
  - Power plant operators
  - Inspection and quality control
- At-risk (according to Chat GPT)
  - Data Entry Clerks
  - Telemarketers
  - Financial Analysts
  - Factory Workers
  - Delivery Drivers
  - But could add jobs in software development, data analysis, and AI engineering
- I will add
  - Routine white collar jobs: legal services, content generators (e.g. marketing), retail clerks

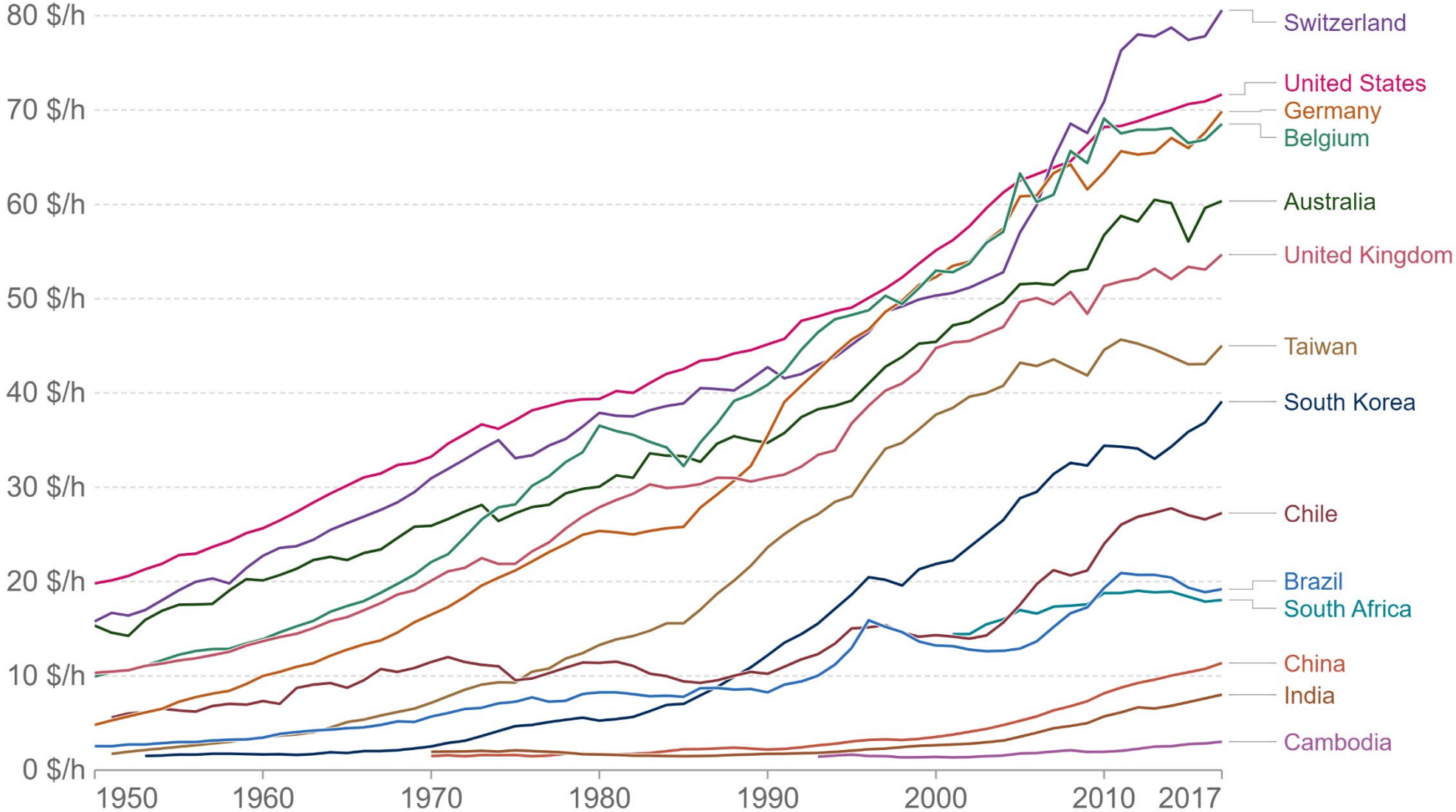
This list does not make sense to me, mostly

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.

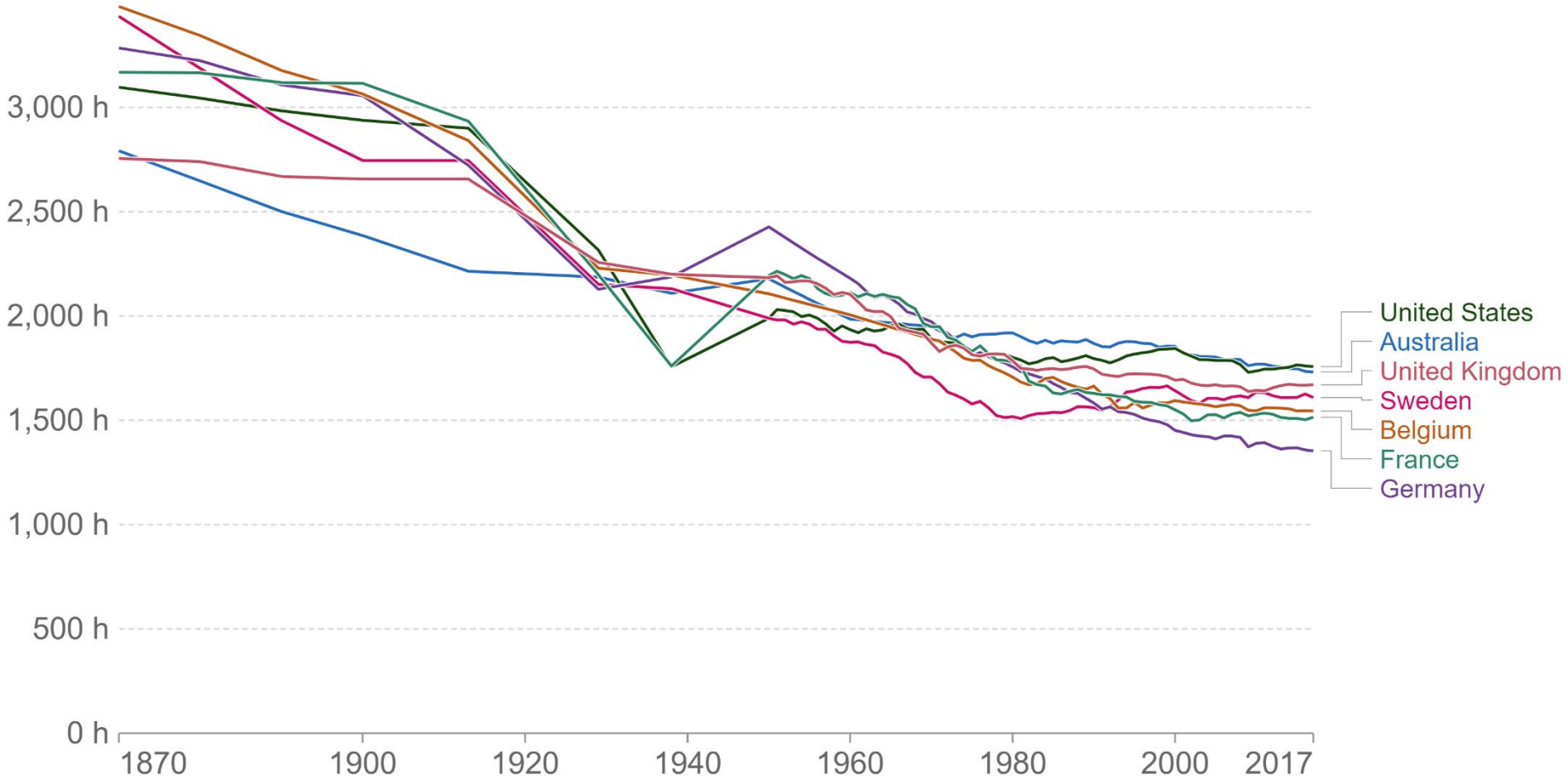
Our World in Data



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

Annual working hours per worker

Average working hours per worker over an entire year. Before 1950 the data corresponds only to full-time production workers (non-agricultural activities). Starting in 1950 estimates cover total hours worked in the economy as measured primarily from National Accounts data.



Source: Huberman & Minns (2007) and PWT 9.1 (2019)

OurWorldInData.org/working-hours • CC BY

Note: We plot the data from Huberman & Minns (2007) and extend coverage using an updated vintage of PWT, which uses the same underlying source. Comparisons between countries are limited due to differences in measurement.

https://ourworldindata.org/working-hours

Impact of AI

We just don’t know because it’s happening so fast

“I would advocate *not* moving fast and breaking things.”

- Demis Hassibis (Deep Mind CEO)

Q2. Which of the regulation proposals do you support?

- AI has potential for many harms
  - Disinformation, e.g. through social media
  - Information spamming – automatically generated internet content
  - Cheating or avoiding learning by having AI complete assignments
  - AI agents stealing and reproducing synthetic content such as music, stories, art
- AI producers do not have a strong incentive to restrict their products

https://tinyurl.com/441-L21-fa24

Popular support for Regulation

Regulating Deepfakes

- **Prohibit the use of deepfakes in political campaign advertisements**, such as to depict an opponent saying something they did not, or to depict an event that did not occur, as proposed by the Federal Election Commission. (National 84%, Republicans 83%, Democrats 86%)

- **Prohibit the public distribution of any pornographic deepfake that was made without the consent of the person being depicted**, as proposed in the Preventing Deepfakes of Intimate Images Act and DEFIANCE Act. (National 86%, Republicans 85%, Democrats 87%)

- **Require that all deepfakes which are shared publicly be clearly labeled as such**, as proposed in the AI Labeling Act, AI Disclosure Act, and DEEPFAKES Accountability Act. (National 83%, Republicans 83%, Democrats 85%)

Nine Major Proposals for Government Regulating Artificial Intelligence Favored by Very Large Bipartisan Majorities of Voters | Program for Public Consultation

Regulating AI programs that make decisions which can significantly impact people's lives

- **Require these AI programs pass a test before they can be put into use, which would evaluate whether they may violate regulations, make biased decisions, or have security vulnerabilities.** (National 81%, Republicans 76%, Democrats 88%)
- **Allow the government to audit programs that are in use, and require the AI company to fix any problems that are found.** (National 77%, Republicans 74%, Democrats 82%)
- **Require AI companies to disclose information to the government about how the decision-making AI was trained, if requested, to aid with pre-testing and audits.** (National 72%, Republicans 67%, Democrats 81%)

Others

- **Creating a new federal agency for AI to enforce regulations, oversee AI development and provide guidance on AI policy** is supported by 74% (Republicans 68%, Democrats 81%).
- **Creating a treaty to prohibit the development of weapons that can use AI to fire on targets without human control** – called lethal autonomous weapons – as has been called for by the International Committee of the Red Cross, and the Campaign to Stop Killer Robots.
- A large bipartisan majority (77%) favors the creation of such an **international agency that would develop international standards for large-scale AI and have the authority to monitor** and inspect whether their standards are being met

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)

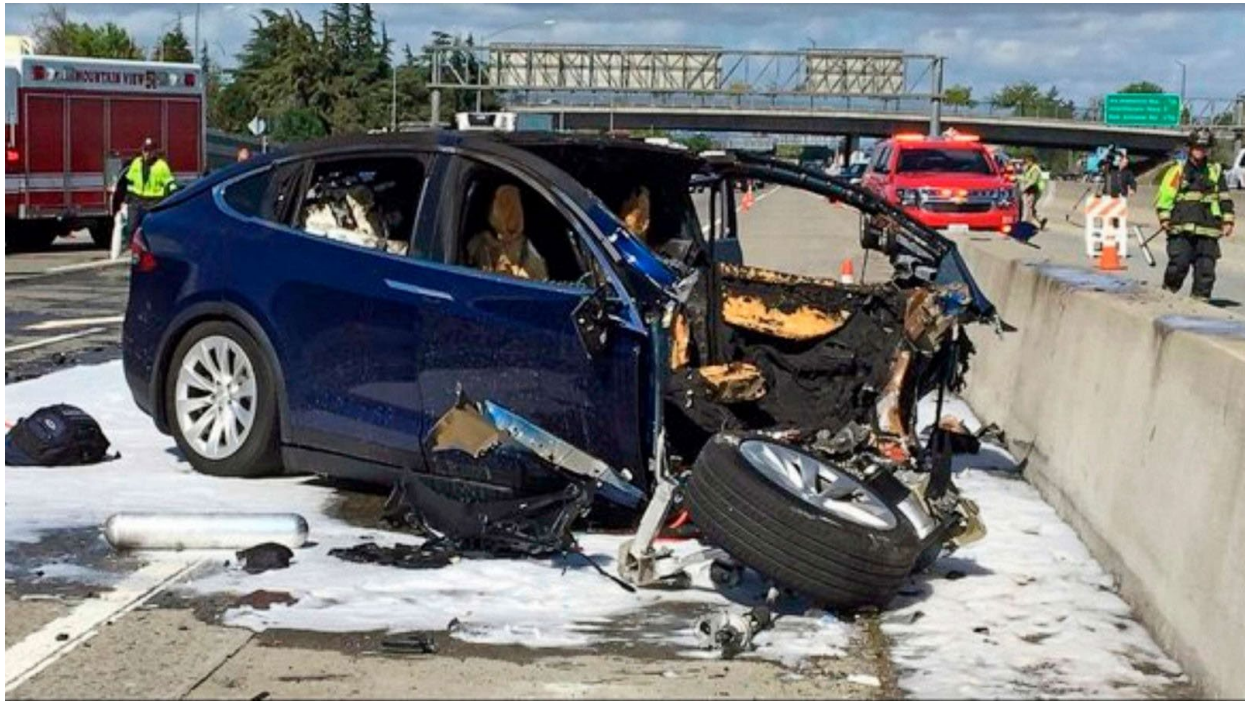
Responsibility requires

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?

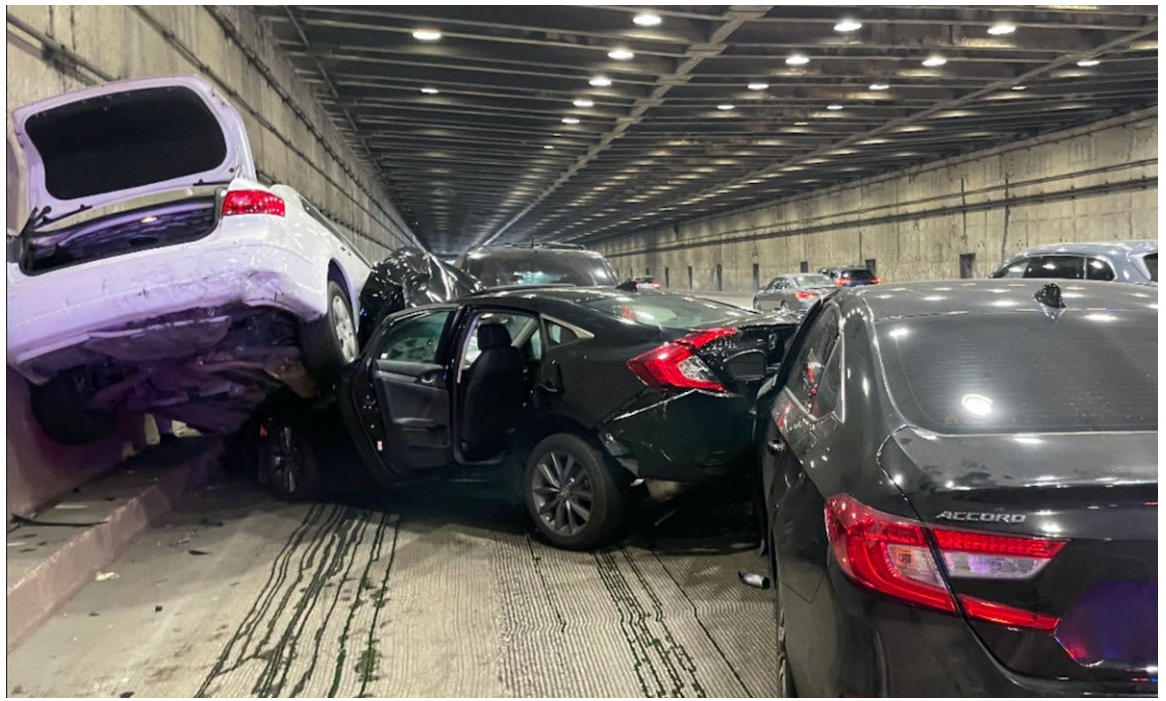
Who is responsible for self-driving vehicle crashes?



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."

Q3. Discuss: who is responsible if an autonomous vehicle crashes due to failure of autonomous system?

https://tinyurl.com/441-L21-fa24

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

Q4. Discuss: who is responsible if GPT is used for dissinformation?

https://tinyurl.com/441-L21-fa24

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]

Q5. Discuss: Should companies be required to gain explicit permission to train on publicly accessible data? Or should content providers have an easy way to opt out?

https://tinyurl.com/441-L21-fa24

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.

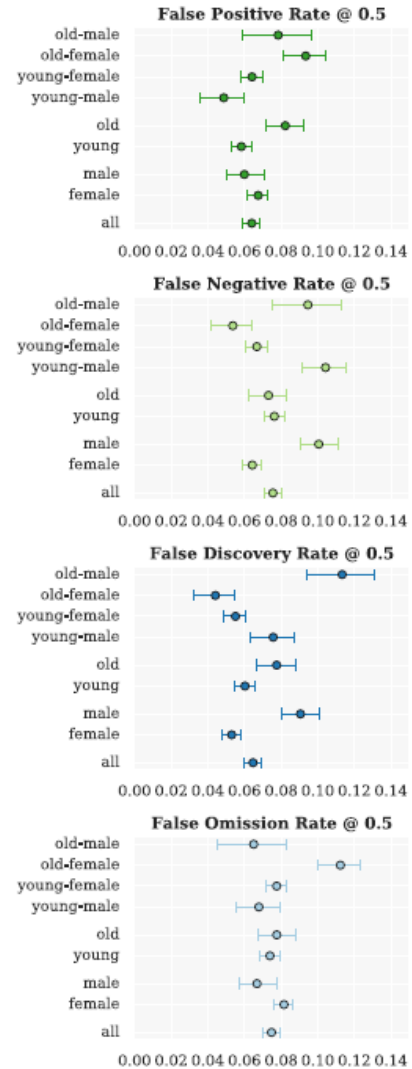
Evaluation Data

- 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

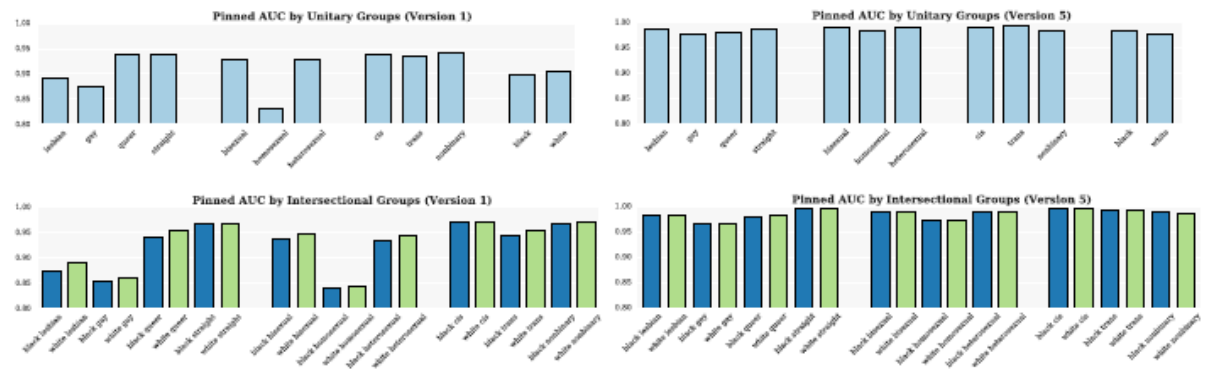
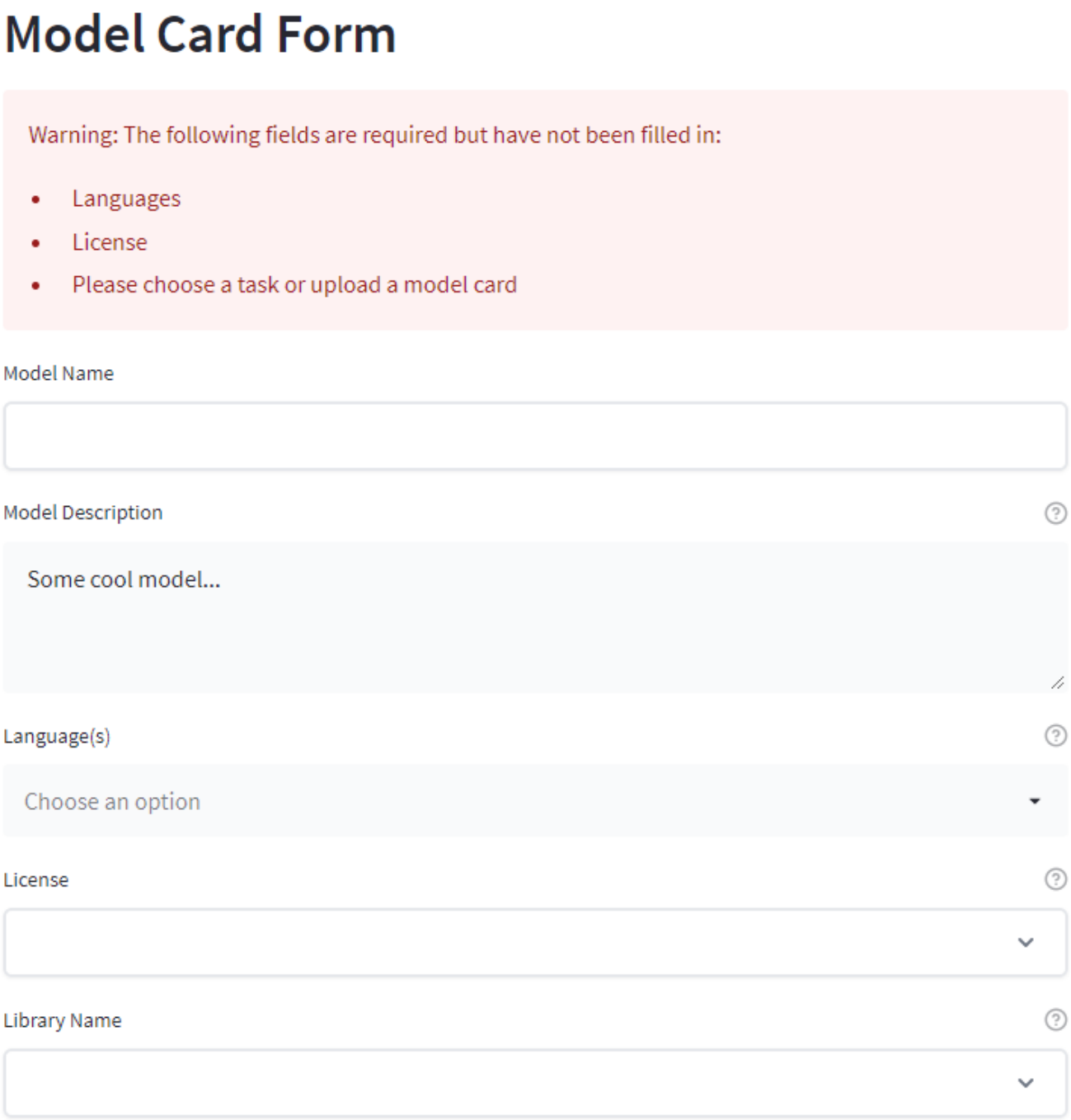


Figure 3: Example Model Card for two versions of Perspective API's toxicity detector.

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 and emissions calculators, are being developed -- use them

On Thursday…

- Bias and fairness in AI