COMPSCI 589
Lecture 22: Ethics and Machine Learning

Benjamin M. Marlin

College of Information and Computer Sciences
University of Massachusetts Amherst

Slides by Benjamin M. Marlin (marlin@cs.umass.edu).
Created with support from National Science Foundation Award# IIS-1350522.

Views on Machine Learning

Mitchell (1997): "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Substitute "training data D" for "experience E."

UNESCO AI Ethics Principles

- Human Rights and Human Dignity
- Human Oversight and Determination
- Inclusiveness and Fairness
- Sustainability
- Privacy and Data Protection
- Transparency and Explainability
- Accountability and Responsibility
- Safety and Security
- Awareness and Literacy
- Ethical Impact Assessment
- Further Reading: `https://www.unesco.org/en/artificial-intelligence/recommendation-ethics`

Data Provenance

- Data provenance refers to the origins and history of transformations applied to a data set.
- There can be ethical issues with both the origins and transformations of data.
- Recent advances in generative AI including both language and image generation have been driven by the collection of extremely large data sets using web scraping methods.
- The terms of service of many websites have also been updated over the last few years to explicitly allow use of user data to training models.

Copyright and Fair Use

- *Even though it may be possible to copy content from the Internet, doing so may still be copyright infringement. Though it may appear that images, video, music, text, and other content online are available to be copied and distributed without need of permission, that is frequently not the case. Authors retain copyright, even if they post their works online.*
- *Fair use is the right to use a copyrighted work under certain conditions without permission of the copyright owner. The doctrine helps prevent a rigid application of copyright law that would stifle the very creativity the law is designed to foster. It allows one to use and build upon prior works in a manner that does not unfairly deprive prior copyright owners of the right to control and benefit from their works.*[1]

[1] https://ogc.harvard.edu/pages/copyright-and-fair-use

Fair Use Factors

To determine whether a given use is fair use, four factors are considered:[2]

- The purpose and character of the use, including whether the use is of a commercial nature or is for nonprofit educational purposes
- The nature of the copyrighted work
- The amount and substantiality of the portion used in relation to the copyrighted work as a whole
- The effect of the use upon the potential market for or value of the copyrighted work.

[2] https://ogc.harvard.edu/pages/copyright-and-fair-use

Copyright Infringement?

https://www.npr.org/2023/11/16/1213588978/

Copyright Infringement?

ars TECHNICA

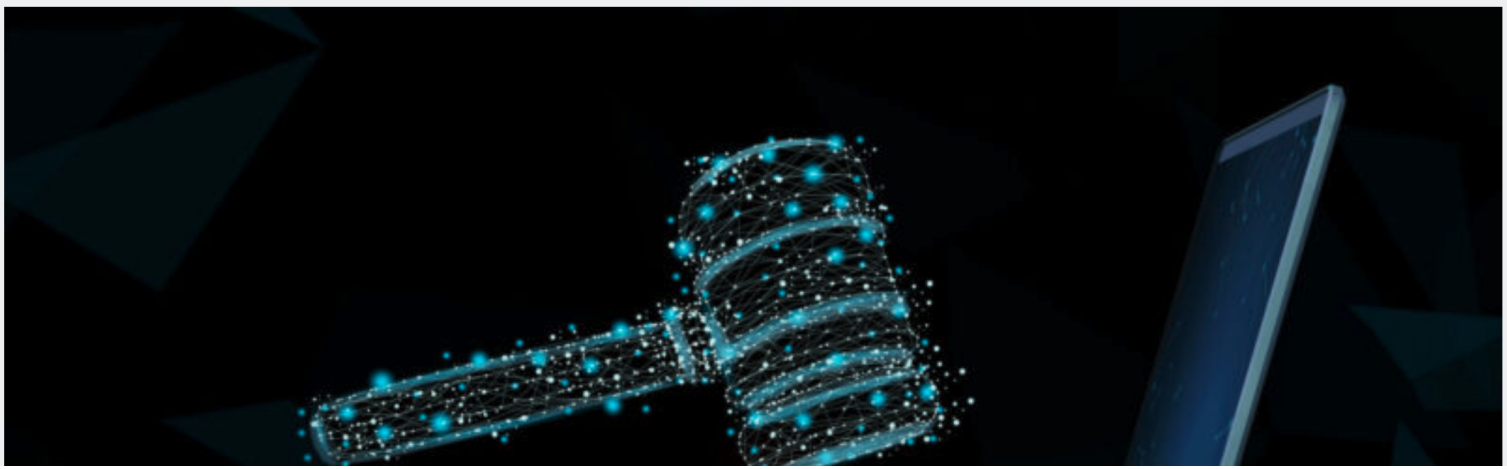
BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE STORE

STABLE DIFFUSION GOES TO COURT —

Artists file class-action lawsuit against AI image generator companies

Suit seeks damages from Stability AI, Midjourney, and DeviantArt.

BENJ EDWARDS - 1/16/2023, 6:36 PM



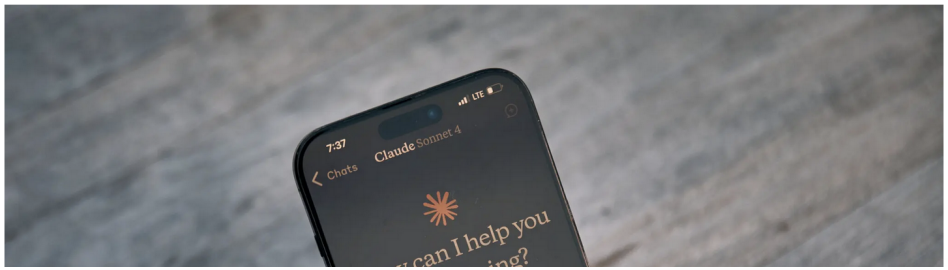
https://arstechnica.com/information-technology/2023/01/artists-file-class-action-lawsuit-against-ai-image-generator-companies/

Copyright Infringement?

Anthropic Agrees to Pay $1.5 Billion to Settle Lawsuit With Book Authors

The settlement is the largest payout in the history of U.S. copyright cases and could lead more A.I. companies to pay rights holders for use of their works.

Listen to this article · 9:45 min Learn more | Share full article | 81



https://www.nytimes.com/2025/09/05/technology/anthropic-settlement-copyright-ai.html/

Copyright Infringement?

BUSINESS INSIDER

Subscribe | Newsletters

TECH

OpenAI lost a court battle against the New York Times — now it's taking its case to the public

By Jacob Shamsian + Follow

https://www.businessinsider.com/openai-new-york-times-copyright-infringement-lawsuit-chatgpt-logs-private-2025-11

User Data and Consent

- In areas like medical research, individuals need to explicitly consent to participate and there are strong protections on sharing and use of collected data.
- In the US, what companies can do with user data is defined in their terms of service. Users frequently have minimal to no protections.
- In particular, over the last few years, many web companies have reinterpreted or updated their terms of service to allow training models on user data and in some cases to give models access to private data.[3]

[3] https://www.politifact.com/article/2025/nov/20/AI-train-tech-companies-private-data-access/?utm_source=chatgpt.com

User Data and Consent

- **Meta:** Customizes content and advertisements based on interactions with Meta AI. Uses public photos, posts, comments and reels to train models. There is no opt-out.
- **Google:** Google uses data uploaded to Gemini apps, including videos and photos, to train models. There is no opt-out. Gemini Deep Research can connect to users' other Google products, including Gmail, Drive and Chat, but users must give permission.
- **LinkedIn:** Recently started using U.S. members' profile and public post data to train generative AI models. Users can opt-out of having their data provided for AI training purposes.
- **GitHub:** Uses public repository code to train Copilot. There is no opt-out, and this use is not necessarily consistent with all license types.

Lawsuits

Reuters World Business Markets Sustainability Legal Commentary Technology Investigations More

OpenAI, Microsoft want court to toss lawsuit accusing them of abusing open-source code

By **Blake Brittain**

January 27, 2023 4:59 PM EST · Updated January 27, 2023

https://www.reuters.com/legal/litigation/openai-microsoft-want-court-toss-lawsuit-accusing-them-abusing-open-source-code-2023-01-27/

More Lawsuits

Reuters

World Business Markets Sustainability Legal Commentary More

Google defeats class action over AI training data for now

By **Blake Brittain**

June 7, 2024 10:00 AM EDT · Updated June 7, 2024

https://www.reuters.com/legal/transactional/google-defeats-class-action-over-ai-training-data-now-2024-06-06/

Data Bias

- What a model learns depends on the data used for training.
- When supervised machine learning is used to build models to support decision-making about people, there is potential for bias in data to lead to models that are biased in different ways.
- How data are collected, what variables are included as features, what variables are used as prediction targets, and how models are trained can all affect the algorithmic bias of learned models.

Sampling Bias

- **Definition:** The data collected does not represent the true population distribution.
- **Example Scenario:** A hospital emergency room triage model is trained using data from a research hospital in an affluent suburb of a large city and then deployed to hospitals nationwide.
- **Bias:** The population seen at the hospital used to collect the data set is not representative of the population at large due to the socioeconomic profile of the area where the hospital is located.

Representation Bias

- **Definition:** Certain classes or groups are systematically omitted or underrepresented in training data.
- **Example:** A company decides to recruit college students to provide data to train a smartwatch exercise classification app.
- **Bias:** The resulting training data systematically omits children, older adults, or elderly individuals.

Survivorship Bias

- **Definition:** Data only includes individuals or items that “survived" a selection process.
- **Example:** A pharmaceutical company uses data from individuals who participated in a clinical trial of a new drug for at least six months to build a regression model for predicting the efficacy of the drug.
- **Bias:** The data set omits individuals who dropped out before six months. This could include individuals who had adverse side effects not experienced by the remaining participants.

Labeling Bias

- **Definition:** The process that produces labels for training data reflects human biases.
- **Example:** A bank decides to build a classification model for deciding who to give loans to using its historical data on loan applications and approval/denial assessments from its credit officers.
- **Source of Bias:** If credit officers exhibited bias towards any groups, those biased assessments are part of the training data.

Data Processing Bias

- **Definition:** Cleaning, transformation, filtering, or aggregation of data that results in bias.
- **Example:** A health insurance company uses past customer profile and claims data to build a regression model for setting health insurance premiums. They discard profiles that are missing income information when forming the data set.
- **Bias:** If some groups withhold their income information at greater rates than other groups, dealing with missing data via case deletion induces representation bias.

Algorithmic Bias

- Algorithmic bias is unfair or discriminatory behavior by an algorithm that leads to unequal treatment or outcomes for different groups.
- If models are learned on biased data and no mitigation approaches are used, the predictions or decisions output by the models will typically reflect the biases present in the data.
- Models may also exhibit differential error rates for different groups due to unequal representation.

Algorithmic Bias

- Deploying systems with algorithmic bias is unethical, and becomes illegal when the impacted groups are protected by law.
- The relevant US laws include the Civil Rights Act, Equal Credit Opportunity Act, Fair Housing Act, and the Americans with Disabilities Act.
- Major protected attributes include race, national origin, sex, gender, religion, age, disability status, and marital status.

More Lawsuits

BUSINESS

The New York Times

GIVE THE TIMES Account

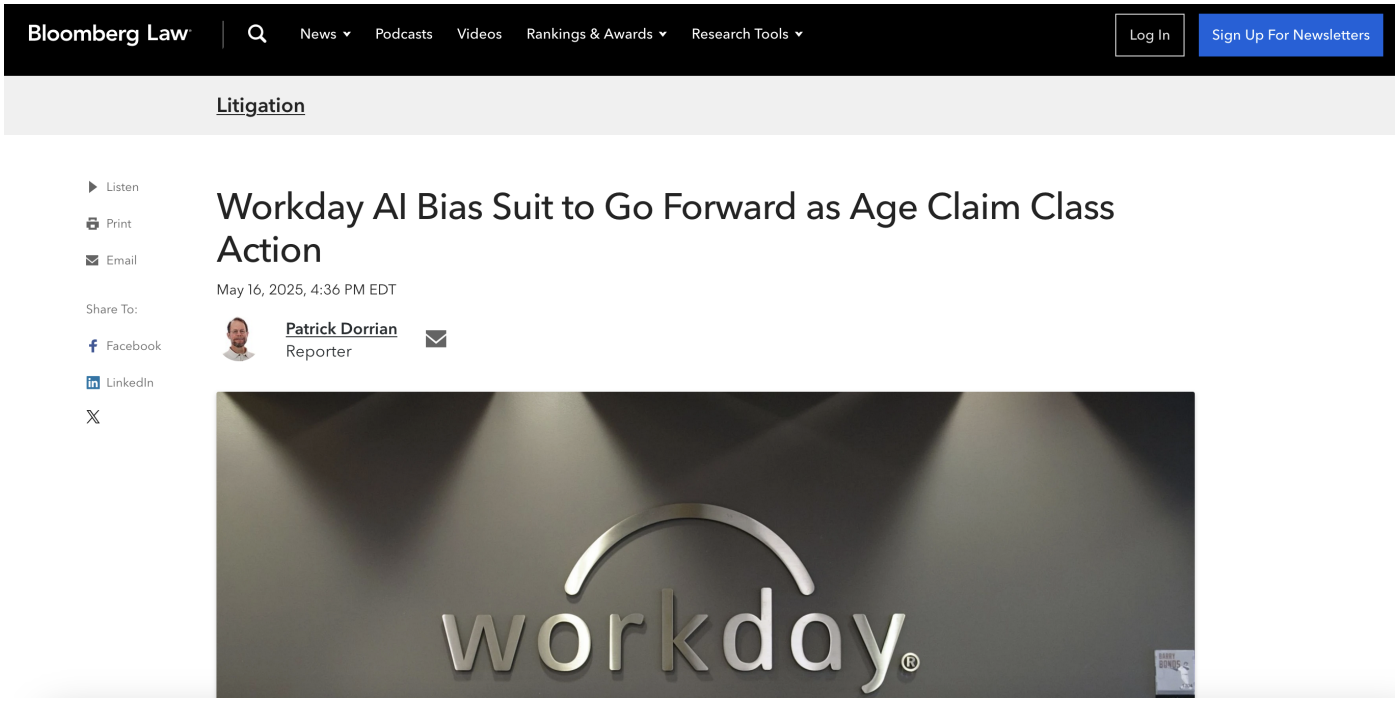
New Suit Uses Data to Back Racial Bias Claims Against State Farm

Black customers have long claimed that the nation's largest home insurer discriminates against them. A lawsuit claims a nine-month study provides some proof.

Share full article

https://www.nytimes.com/2022/12/14/business/state-farm-racial-bias-lawsuit.html

More Lawsuits



https://news.bloomberglaw.com/litigation/workday-ai-bias-suit-to-go-forward-as-age-claim-class-action

Algorithmic Bias Assessment

- Given a learned predictive model $f(\mathbf{x})$ for a task $\mathcal{T}$, there are a number of ways to assess whether the model exhibits algorithmic bias with respect to groups defined by protected attributes.
- Suppose protected attribute A takes values $\{1, ..., K\}$. Let a_i be the value of the protected attribute for data case i.
- Partition a test data set $\mathcal{D}$ into subsets with the same value of the protected attribute $\mathcal{D}_a = \{\mathbf{x}_i | a_i = a, 1 \leq i \leq N\}$.
- Let S be a criterion function that maps a data set and a predictive model to a real number.
- Differences in the value of $s^a = S(f, \mathcal{D}_a)$ between groups suggest the presence of algorithmic bias.
- This approach can also be used with combinations of protected attributes.

Example: Prediction Accuracy Parity

- **Model:** Let f be a multi-class classifier.
- **Property:** The accuracy of the classifier should be similar for all groups based on values of a protected attribute.
- **Criterion Function:**

$$S_{ACC}(f, \mathcal{D}_a) = \frac{1}{|\mathcal{D}_a|} \sum_{(\mathbf{x},y)\in\mathcal{D}_a} \mathbb{I}(y = f(\mathbf{x}))$$

- **Example:** Suppose f is a smartwatch activity classifier and attribute of interest is age. The activity classifier should have equal accuracy for all age groups.

Example: Positive Predictive Value Parity

- **Model:** Let f be a binary classifier.
- **Property:** The fraction of true positives among the predicted positive instances should be similar for all groups based on values of a protected attribute.
- **Criterion Function:**

$$S_{PPV}(f, \mathcal{D}_a) = \frac{\sum_{(\mathbf{x},y)\in\mathcal{D}_a} \mathbb{I}(f(\mathbf{x}) = 1) \cdot \mathbb{I}(y = 1)}{\sum_{(\mathbf{x}',y')\in\mathcal{D}_a} \mathbb{I}(f(\mathbf{x}') = 1)}$$

- **Example:** Suppose f is a classifier used to decide which patients might benefit from a new medical treatment, and the protected attribute of interest is race. The fraction of people who actually benefit from the treatment relative to the number of people predicted to benefit should be the same for patients of all races.

Example: Statistical Parity

- **Model:** Let $p(Y = y|\mathbf{X} = \mathbf{x})$ be a probabilistic binary classifier.
- **Property:** The probability of predicting the positive class should be similar for all groups based on values of a protected attribute.
- **Criterion Function:**

$$S_{SP}(p, \mathcal{D}_a) = \frac{1}{|\mathcal{D}_a|} \sum_{(\mathbf{x},y)\in\mathcal{D}_a} \mathbb{I}(p(Y = 1|\mathbf{X} = \mathbf{x}) > 0.5)$$

- **Example:** Suppose f is a classifier used to decide which loan applications to approve, and the protected attribute of interest is gender. The probability of approving loans should be the same for applicants of all genders.

Algorithmic Bias Mitigation

- If data sets have sampling bias, they can be re-sampled or re-weighted to match the characteristics of a target population.
- If data sets have representation bias, new data may need to be collected so groups that were completely left out initially are included.
- If data sets have label bias, the data may need to be re-labeled from scratch using formal rubrics, multiple annotators, or other approaches that can provide more objective assessments.

Algorithmic Bias Mitigation

- Features that act as proxies for protected attributes can lead to parity violations. These features can be identified and considered for removal from a model.
- Model training procedures can be augmented to include terms in the objective function that penalize violation of parity properties.
- Suppose we are training a probabilistic binary classifier with parameters θ and we need to penalize it to improve statistical parity with respect to a binary protected attribute.
- We can use a modified learning objective:

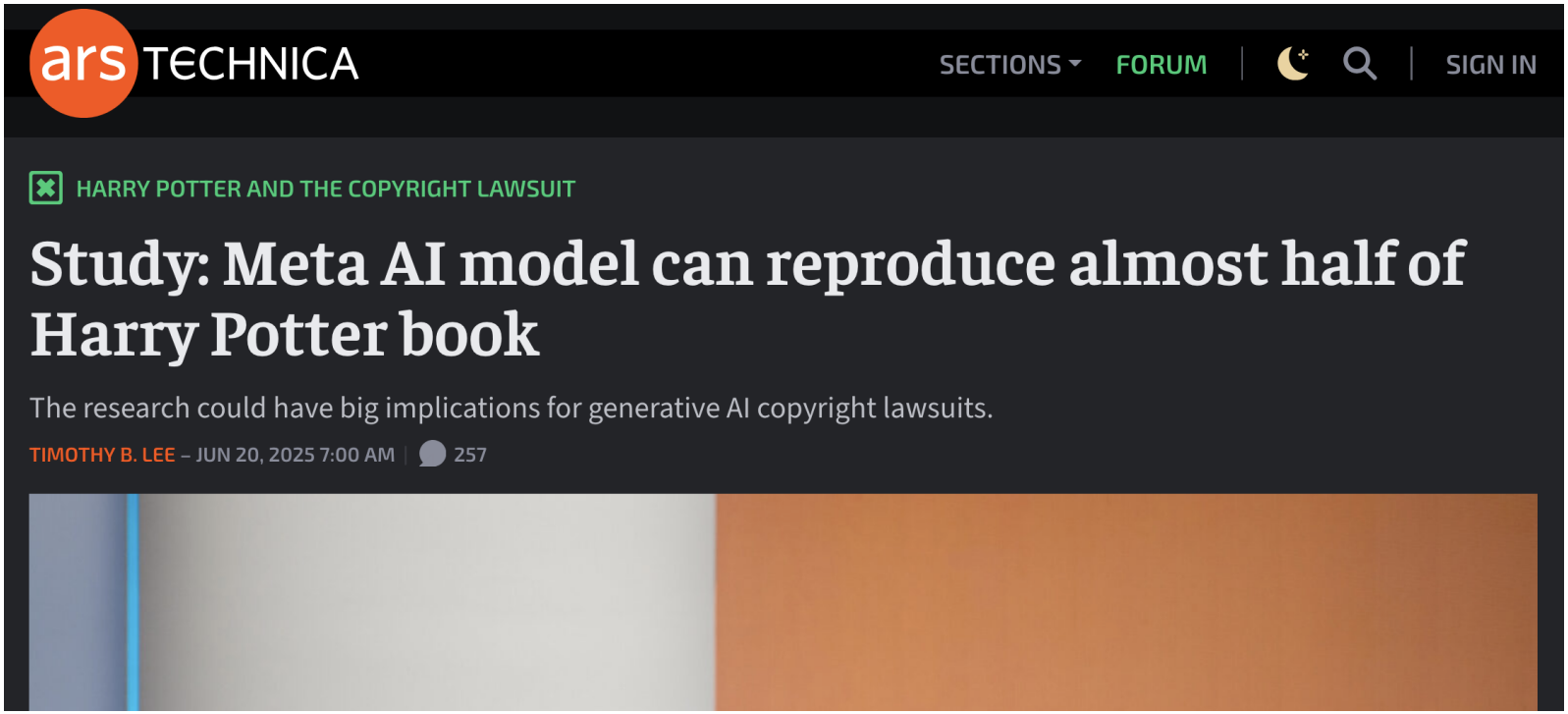
$$\arg\min_{\theta} nll(\mathcal{D}, p_\theta) + \lambda(S_{SP}(p_\theta, \mathcal{D}_0) - S_{SP}(p_\theta, \mathcal{D}_1))^2$$

- Of course, this approach will only work if the criterion function is differentiable.

Harmful Generation

- Current LLMs are widely known to have issues generating confidently incorrect answers to legitimate prompts, a problem known as “hallucinations."
- Current LLMs, image generation models, and code generation models have issues with generating material that is in the style of or that directly reproduce copyrighted works.
- Some models have also had issues generating hate speech and discriminatory content.

Reproducing Text



https://arstechnica.com/features/2025/06/study-metas-llama-3-1-can-recall-42-percent-of-the-first-harry-potter-book/

Image Styles

https://www.nytimes.com/2025/04/02/learning/what-do-you-think-of-a-new-chatgpt-feature-that-makes-images-in-the-style-of-studio-ghibli.html/

Discriminatory Content

https://www.politifact.com/article/2025/jul/10/Grok-AI-chatbot-Elon-Musk-artificial-intelligence/

Model Misuse

- The public availability of powerful generative AI models has resulted in significant potential harms due to model misuse.
- Early models completely lacked safeguards. For example, the publicly available GPT-2 model from OpenAI is now distributed with the following warning on Hugging Face:

  Language models like GPT-2 reflect the biases inherent to the systems they were trained on, so we do not recommend that they be deployed into systems that interact with humans.

- While most commercial LLMs and image generation models try to implement guardrails to block misuse, there have been well-publicized attacks against guardrails (e.g., jailbreaks).

LLM Jailbreaks

Artificial Intelligence › | Industry's Influence on Elections | Agentic A.I. | What if the A.I. Boom Falters? | A.I. in the Classroom | Medical Records and Chatbots

Researchers Poke Holes in Safety Controls of ChatGPT and Other Chatbots

A new report indicates that the guardrails for widely used chatbots can be thwarted, leading to an increasingly unpredictable environment for the technology.

Share full article 53

https://www.nytimes.com/2023/07/27/business/ai-chatgpt-safety-research.html

Model Misuse

- Models remain highly susceptible to misuse despite attempts at guardrails and the barrier to developing private models with no guardrails is minimal.
- Forms of misuse include generating misinformation, personalized phishing emails, fake social media personas and advanced social media bot accounts, deepfake images and video, voice cloning, and generating new cyberattacks and malware.
- LLM jailbreaks have also been used to generate instructions for creating weapons, explosives, chemical agents and other harmful and controlled items.
- Lastly, in multiple areas of scientific research, the journal and conference peer review system is being spammed with AI-generated papers that look plausible, but contain faked results.

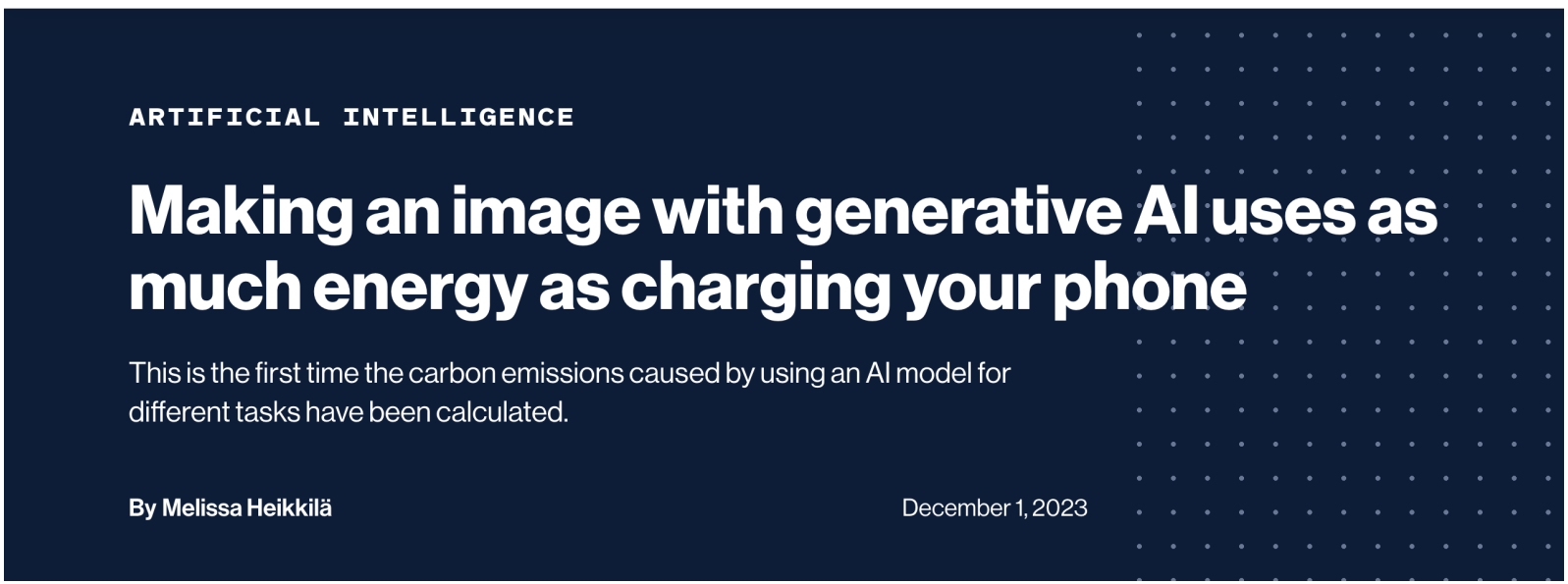
Externalities

- Current generative AI training and deployment has significant externalities in terms of energy use and environmental impact.
- Tasks involving images are significantly more energy intensive compared to tasks involving text.

task	**mean**	**std**
text classification	0.002	0.001
image classification	0.007	0.001
text generation	0.047	0.030
text summarization	0.049	0.010
image captioning	0.063	0.020
image generation	2.907	3.310

Table: Inference energy in kWh per 1,000 queries.

Energy Use



https://www.technologyreview.com/2023/12/01/1084189/making-an-image-with-generative-ai-uses-as-much-energy-as-charging-your-phone/

Energy Use

Power Hungry Processing: ⚡ Watts ⚡ Driving the Cost of AI Deployment?

Alexandra Sasha Luccioni
Yacine Jernite
sasha.luccioni@huggingface.co
Hugging Face
Canada/USA

Emma Strubell
Carnegie Mellon University, Allen Institute for AI
USA

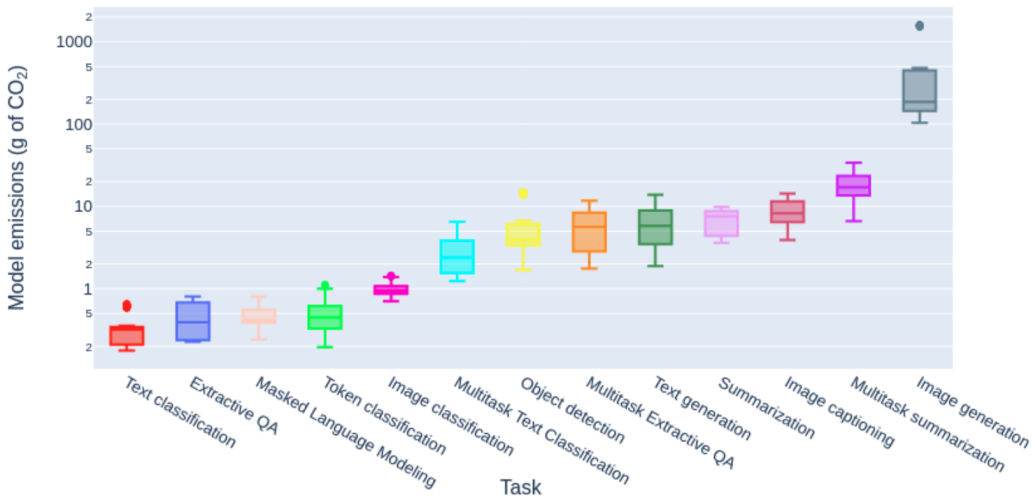


Figure 1: The tasks examined in our study and the average quantity of carbon emissions they produced (in g of CO_2eq) for 1,000 queries. N.B. The y axis is in logarithmic scale.

https://dl.acm.org/doi/pdf/10.1145/3630106.3658542

Environmental Impacts

Massachusetts Institute of Technology

Education Research Innovation Admissions + Aid Campus Life News Alumni About MIT

MIT News
ON CAMPUS AND AROUND THE WORLD

SUBSCRIBE BROWSE SEARCH NEWS

Explained: Generative AI's environmental impact

Rapid development and deployment of powerful generative AI models comes with environmental consequences, including increased electricity demand and water consumption.

Adam Zewe | MIT News
January 17, 2025

PRESS INQUIRIES

MIT News explores the environmental and sustainability implications of generative AI

https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117

Externalities

- In a December 2024 report, Lawrence Berkeley National Laboratory projected that by 2028 AI data center electricity use would rise to a level equivalent to 22% of all US households.[4]
- Starting in 2024, multiple leading AI companies announced plans to develop significant new nuclear generating capacity for data centers.

[4] https://www.energy.gov/articles/doe-releases-new-report-evaluating-increase-electricity-demand-data-centers

Environmental Impacts

https://blog.google/outreach-initiatives/sustainability/google-first-advanced-nuclear-reactor-project-with-kairos-power-and-tennessee-valley-authority/

Orders, Legislation, Policies, Reports

- US National Institute of Standards and Technology AI Risk Management Framework (January 26, 2023) Link
- US Executive Order 14110 (October 30, 2023). Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Link
- United Nations High-level Advisory Body on Artificial Intelligence (2023-2024). Governing AI for Humanity Report. Link
- European Union Regulation 2024/1689 (June 13, 2024). Artificial Intelligence Act. Link
- US Executive Order 1417. (January 23, 2025). Removing Barriers to American Leadership in Artificial Intelligence. Link

Self-Governance

- Corporate AI safety and alignment teams, groups and initiatives. Examples: OpenAI, Google, Anthropic, Nvidia.
- Model cards: Standardized, short technical documents that describe key characteristics of a trained ML model including model details (version, type, creators, etc.), intended use (primary intended uses, out-of-scope use cases), model performance assessment, training and evaluation data used, and ethical considerations.[5] Examples: OpenAI GPT-2, Meta Llama 3.1, Google Imagen 4.

[5] https://arxiv.org/pdf/1810.03993