Annotated Bibliography

Aaron, Lynn, Santina Abbate, Nicola Marae Allain, Bridget Almas, Brian Fallon, Dana Gavin, C. 2024. “AI Literacy.” Pp. 18–23 in *Optimizing AI in Higher Education*. SUNY Press.

This reader discusses the climate of AI as it relates to education, and emphasizes the importance of teaching AI literacy so it can be used responsibly. It contains suggestions as to how to integrate AI literacy into a curriculum. The article emphasizes that generative AI is not a new technology, but rather a collection of extant technologies in concert, to attempt to dispel some of the ‘hype’ and gain a more realistic understanding of what it can and cannot do. It then goes through various sectors of non-academic life to explain some applications of AI, followed by a brief discussion of its pitfalls – misinformation being chief among them.

Aaronson, Susan Ariel. 2024. *Data Disquiet: Concerns about the Governance of Data for Generative AI*. Centre for International Governance Innovation.

This paper goes over some of the issues facing the generative AI field, such as misinformation, lack of vetting, the economics and history of large language models, and how governments are attempting to deal with the ramifications of generative AI. It discusses liability, asking, who is liable when an LLM makes a mistake? If a company deployed it, for instance, as a customer service agent, in Air Canada’s case, are mistakes made by the bot things the company is liable for? How can governments incentivize the development of accurate and reliable models, when it is far cheaper and easier to neglect

these priorities? Most popular LLMs are, currently, operating with no data transparency, and with trade secret algorithms; without this transparency, how much power are we giving over to the companies making them over the world's data and information ecosystem? The patchwork of different governments' regulations are outlined in detail, as well as their gaps. Because the internet is worldwide, it is nearly impossible to regulate through traditional means, and current government responses are inadequate and reflect a lack of comprehensive understanding of the issue.

The paper breaks down two types of data sets used in LLMs: proprietary data gathered or purchased by the company developing the model, and data scraped from the internet using bots. There is little consistency or information available about these scraped data sets, and they may contain copyrighted material, misinformation, hate speech, and more. Greater transparency about data provenance, where data came from and how it was processed, could lead to greater abilities to study and regulate these data sets. There is also the issue of the scraping of mass amounts of copyrighted material, which is generally done without the knowledge or compensation of its creators. The legality of this kind of data scraping is not yet established, nor is it usually actively audited for quality and accuracy, an expensive and time-consuming process. As LLMs become more popular than search engines, and baked into more and more of our technology, this also leads to misinformation and bias spreading through the models and their users. In addition, LLMs have the ability to compete with humans in many fields, such as writing and art, displacing and devaluing human labor, even though LLMs 'learn' to do these things by studying peoples' copyrighted content. The article calls for more

regulation to prevent all AI data systems from becoming “black boxes”, and due to their increasingly widespread use.

Anon. 2023. “Research Guides: Generative AI: Ethics and Costs.” *Amherst College Library*. Retrieved February 16, 2025 (https://libguides.amherst.edu/c.php?g=1350530&p=9969379).

This research guide contains resources and links to studies on generative AI. Broken down into categories, such as ethics, environmental costs, copyright, labor impacts, etc., the guide acts as a jumping off point for understanding the issues surrounding the new technology. It is also regularly updated with additional resources.

Bearne, Suzanne. 2023. “New AI Systems Collide with Copyright Law.” BBC, July 31.

This article focuses on the ethics of AI data scraping, particularly in the realm of art and design. It discusses artists whose work has been scraped to train AI models, who then saw outputs that were made to look like their work. The financial implications of a program existing to replace you, the artist, with a cheap or free button click are discussed. Several artists have filed a lawsuit against Stability AI, creators of Stable Diffusion, to fight for more protections from AI. Their hope is that, with appropriate legislation, their copyrighted works can be protected from data scraping to train models eager to replace them. The article then discusses various countries attempting to regulate AI, and the challenges of doing so. It concludes with advice for artists looking to protect their work,

and reassurance that public opinion on AI is shifting in favor of the artists being stolen from.

Birhane, Abeba, and Deborah Raji. 2022. “ChatGPT, Galactica, and the Progress Trap.” *Wired*, December 9.

This article focuses on the overhyped marketing of AI, in contrast with the real harms that these systems do when they fail to deliver on those promises. It discusses medical misinformation given by one of Google’s LLMs, much of which is harmful or the exact opposite of what you should do. The article explains that LLMs struggle to understand negative statements, which leads to these types of errors, and notes that it is a skill LLMs have not gotten any better at as they progress in size and complexity. While LLM answers look convincing, they are often wrong, and this reality is intentionally hidden by promoters of AI. LLMs and AI generally have a long history of harm, such as mistranslations in medical settings, and these harms impact marginalized communities more heavily. The article further criticizes the lack of transparency around model development, and how blame when a model fails is obfuscated or pinned on the model itself, rather than the decisions which made the model in the first place. When models are shut down, the subsequent PR backlash treats those asking for skepticism and oversight as spoilsports. When models are shown to be discriminatory or inaccurate, companies often pass the buck to the users, asking the community to step in and ‘fix’ them without compensation. Both the harms caused by the model and the weight of fixing it is thus

routinely shifted to marginalized communities, rather than LLM developers being held responsible for vetting and moderating their own data sets.

Browne, Grace. 2023. “AI Is Steeped in Big Tech’s ‘Digital Colonialism.’” *Wired*, May 25.

Browne discusses the fact that AI inherits bias from the large datasets used to train it, describing AI algorithms as “opinions embedded in code”. When building a large language model, large amounts of data are required, and it is common for data scientists to use large, open datasets without vetting or moderating the content they contain in any way. This leads to real world issues when AI models are used in healthcare, corrections, hiring, and many other industries. For example, models may underestimate the care needs of black patients due to the stereotype of black people having higher pain tolerance than non-black people. MIT’s “80 Million Tiny Images” data set had many images of non-white people tagged using slurs. Another dataset, ImageNet, contained large amounts of porn, including upskirt shots which may have been taken without the subject’s knowledge or permission. Machine learning, in general, skews white and male, but these products are sold to and exported to communities of all kinds. As the western world exports technology, it comes with western norms and philosophies baked in, a form of digital colonialism. Auditing data sets is massively expensive and labor intensive, and without proper regulation, there is little incentive for companies to do so.

Crawford, Kate, and Jason Schultz. 2019. “AI Systems as State Actors.” *Columbia Law*

Review 119(7):1941–72.

This article explores the regulation of AI systems when deployed by governments, explaining that while extensive study has been devoted to regulating government use of systems, there is a gap when it comes to regulating the private corporations developing these models. They point to a lack of understanding on the part of government officials handing out contracts to AI developers, which also helps shield them from accountability. It recommends treating developers who contract AI models to government as state actors, and to regulate them as such to bridge this gap. The article outlines many of the harms that come from bias in algorithms, such as in policing, benefits allocation, and care decisions. Since governments lack understanding of the harms these systems can do, due to lack of technical expertise and extensive subcontracting, this would shift liability to the contractor for failing to create equitable systems. The article goes on to outline the current state of AI regulation using four case studies; reviews existing case law covering the state action doctrine they advocate using; and closes with arguments supporting this position.

Elam, Michele. 2022. “Signs Taken for Wonders: AI, Art & the Matter of Race.” *Daedalus* 151(2):198–217. doi: 10.1162/daed_a_01910.

This article examines AI and large language models from the lens of the arts and humanities, comparing GPT-3 in particular to past radical breakthroughs in technology. The author refers to this as a ‘Dynamo’ moment, named for an electric generator which wowed audiences at a late 19th century world’s fair, and draws comparisons to 20th

century fascinations with ‘progress’ and technology. However, this fixation on progress in technology has historically been mirrored in ‘progress’ in humanity, framing non-white cultures as ‘less developed’ or degenerate. Indeed, the same world’s fairs that showcased technological wonders were commonly homes to ‘ethnic villages’ showcasing ‘lower races’. Defining modernity requires defining what is excluded from it: “what or who [is] irrelevant, backward, regressive in relation”. The author criticizes the valorization of technology and progress which lack a critical theoretical context, placing the technologies in their political and historical realities.

Technologies are generally in the hands of a few specific groups of people, and those people inherit the technology’s power to reshape reality. It is all too common that those with marginalized identities are left out and devalued by this transition. Regulations and safeguards put forth by the industries controlling technology tend to focus on minimization of harm or death, rather than examining the technology’s relation to power, social justice, equity and inclusion. The author calls for more involvement of the humanities and arts in the discussion of ethics around AI, criticizing the idea that artists must first become technical experts – or near-experts – before they have worthwhile inputs to the conversation. The author instead calls for the industry to instead learn to communicate with people who are marginalized in their own vernacular, and shift their mindset to view those perspectives as inherently valuable. Indeed, AI models are created by scraping nearly everyone’s data; the author argues it stands to reason that those people should then benefit from the tools they helped create. The author also criticizes the idea that AI, because it is a machine, is universal and without bias, and describes the impacts

this has on marginalized communities, particularly in reinforcing color lines, robbing race and gender of their social contexts, and harming minorities. The author also discusses biased inputs and some of their historical analogues which are deeply rooted in racial prejudice.

Ganuthula, Venkat Ram Reddy. 2024. “The Paradox of Augmentation: A Theoretical Model of AI-Induced Skill Atrophy.” *Indian Institute of Technology Jodhpur*.

AI has been sold to us as a productivity tool, but there is evidence that it hampers competence when used over time. AI tools initially increase human productivity by augmenting performance. However, as the tools are used more often, skills offloaded to the AI atrophy, and the user becomes less competent. This is called “The Paradox of Augmentation”, which has three principals: 1) AI presents gains by augmenting human ability; 2) AI involves basic human cognitive functions, such as problem solving, which atrophy when not used; 3) this weakens the neural pathways responsible for those cognitive processes, leading to a “use it or lose it” decline in skills and brain plasticity. The article then breaks down several models demonstrating AI-induced skill atrophy.

Hao, Karen, and Deepa Seetharaman. 2023. “Cleaning Up ChatGPT Takes Heavy Toll on Human Workers.” *Wsj.com*. Retrieved February 16, 2025 (https://www.wsj.com/articles/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483).

This article focuses on the invisible workers in the AI ecosystem – contractors, often in the global south, who are hired to 'clean up' data sets used to train AI. These workers, as they moderate this data, are exposed to graphic "descriptions of violence, harassment, self-harm, rape, child sexual abuse and bestiality," and must view them through their entire shift. The goal is to remove this data from the training model, so that the model does not output the same kind of harmful content for its users. This kind of work is also done in social media content moderation, which is also often outsourced to the global south, and subjects its workers to violent and graphic imagery as well as text for rock-bottom wages. The article breaks down the many layers of human input and feedback required to train generative AI models, in contrast with the industry's marketing, which leads the public to believe these systems are autonomous and can train themselves. Workers in Kenya are trying to take action, claiming that their jobs are traumatizing, and they do not receive adequate mental health support or pay. The article goes on to describe the impacts on workers, citing paragraphs long descriptions of rape, torture, suicide, and self-mutilation having catastrophic mental health impacts on employees.

Heaven, Will Douglas. 2020. "AI Needs to Face up to Its Invisible-Worker Problem." Retrieved February 16, 2025 (https://www.technologyreview.com/2020/12/11/1014081/ai-machine-learning-crowd-gig-worker-problem-amazon-mechanical-turk/).

This article examines the gig economy which supports the development of AI systems, such as Amazon Mechanical Turk. These jobs are sub-minimum wage, and over a million people in the United States earn money on these platforms each month. The article discusses the kinds of tasks workers do, such as transcribing audio to train voice recognition software, or labeling websites containing harmful or graphic content so they will be removed from search engines. While the author believes there is a place for this type of crowdwork, he cites low wages as a problem, with many Amazon Mechanical Turk workers making roughly $2 per hour. The skills learned in these jobs are not transferrable, and thus do not help build a resume. These faceless workers are not understood as what they are – one of the crucial backbones of creating AI systems. The article suggests multiple ways users can learn to get the most out of the programs, but also that companies should think more about the human end of the transaction, and support workers to get the most out of their platforms.

Heikkilä, Melissa. 2023. “This New Data Poisoning Tool Lets Artists Fight Back against Generative AI.” *Technology Review*, October 23.

This article discusses ‘AI poisoning’ tools that artists can utilize to keep their style and work from being effective AI training data. The tools, Nightshade and Glaze, subtly change artwork and miscategorize metadata, leading to AI models to view “dogs [as] cats, cars [as] cows, and so forth”. Artists have been raising the alarm about generative AI art programs devaluing their work, while also stealing their work from the internet to train on; models that would be impossible without their work being used to

diminish their ability to make a living. The only way to remove these poisoned artworks is by hand, making them effective tools to disincentivize AI companies from mass-scraping artwork from the internet without permission.

Helmore, Edward, and Kari Paul. 2023. “New York Times Sues OpenAI and Microsoft for Copyright Infringement.” *The Guardian*, December 28.

This article details a lawsuit brought by the New York Times against OpenAi, for “[seeking] to free-ride on the Times’s massive investment in its journalism by using it to build substitutive products without permission or payment”. As LLMs scrape data from the Times and other journalistic sources, it can, essentially, regurgitate the news to users in search summaries and tools like ChatGPT, undercutting traditional journalism while also requiring it to function. The article also cites the company responsible for ChatGPT, OpenAI, as being embroiled in internal struggle over whether to focus on safety first in developing their models, or adopt a more ‘move fast and break things’ ideology to pursue growth. The article also mentions AI misinformation and hallucinations being a key sticking point for the Times, noting that Microsoft’s Bing AI overview has produced misinformation that is then falsely attributed to the Times, hurting its reputation. The lawsuit seeks billions in damages, and aims to have the models trained using their data destroyed.

Jung, Carsten, and Bhargav Srinivasa Desikan. 2024. *Transformed by AI: How*

Generative Artificial Intelligence Could Affect Work in the UK - And How to Manage It. Institute for Public Policy Research (IPPR).

This report breaks down ways in which AI may impact the workforce. While its authors do not feel there will be a 'job apocalypse' of mass unemployment, they do point to past cases of automation as driving forces in wage decrease in the last century. It breaks down the types of work that will be most impacted; cognitive and administrative tasks, and 'back office' jobs, such as administrative assistants, customer service representatives, or human resources professionals. Women are thus much more likely to be affected by AI displacement, and low-paid, entry level positions are much more vulnerable than jobs held by high earners. They estimate that perhaps 33% of administrative positions could cease to exist, while productivity would significantly increase. They underscore that workers need to benefit from these leaps in productivity, such as in matching wage increases to increases in productivity.

The report then breaks down the reality that there may not be enough new jobs created to absorb the number of workers being displaced by AI, or that the new jobs being created as a result of AI will not match the skillsets of those displaced. This would likely depress wages, buying power and and GDP. In the UK, they recommend possibly supplementing the jobs market with a green industrial strategy, and increased funding for jobs with high interpersonal task shares like social workers and mental health nurses. They argue that a job-centered strategic approach could help absorb some of the negative impacts of AI job displacement. They conclude by outlining the goals of their future research and policy recommendations: protecting existing jobs and ensuring workers

benefit from gains; boosting creation of new jobs and support reskilling; and addressing fallout from AI-induced labor surplus.

Kerr, Dara. 2024. “AI Brings Soaring Emissions for Google and Microsoft, a Major Contributor to Climate Change.” *NPR*, July 12.

This article outlines how generative AI uses huge amounts of energy when compared to products it aims to replace, such as the search engine. The energy it takes to generate one response on an AI chatbot is estimated to be enough to power a lightbulb for about 20 minutes; at the scale AI is being used and ham-fisted into every product, this leads to enormous amounts of energy use. For example, in 2020, Google’s emissions rose 48%, attributing much of it to data centers, which are crucial to powering generative AI. Google even admits in the report, “[as] we further integrate AI into our products, reducing emissions may be challenging.” Google and Microsoft have also began to backslide on their sustainability goals of being carbon neutral or negative by 2030, though both companies still claim to intend to meet them, even as their emissions rise due to building out AI infrastructure. The energy needs of the 7,000 data centers already in existence are delaying decommissioning of coal plants, and are estimated to “consume the equivalent amount of electricity per year as the entire country of Italy”. The article ends by calling for regulation of the AI sector’s energy consumption to avoid dangerously accelerating climate change.

Kidd, Celeste, and Abeba Birhane. 2023. “How AI Can Distort Human Beliefs.” *Science*

(New York, N.Y.) 380(6651):1222–23. doi: 10.1126/science.adi0248.

This article critiques the unrealistic marketing hype that purveyors of AI products push in their marketing. The level of utility and reliability of the products is over-stated, leading to users that are unaware of the gaps in the end product. AI frequently and confidently presents misinformation as fact. This misinformation is then absorbed by users. The article discusses that, once information is learned, it is very difficult to convince someone it is incorrect. This includes biases as well as blatant incorrect information. These biases also harm minority groups more than their non-minority counterparts. The article goes on to criticize the reactive nature of response to this, with AI companies only correcting bias and misinformation in their datasets after harm has already been done, and biases and incorrect information have already been thoroughly absorbed. It breaks down three psychological processes which lead to incorrect information and biases being retained, even after corrected. One, people trust sources "they judge to be confident and knowledgeable"; AI does not communicate uncertainty, or qualify statements with phrases like "I think" or "as far as I know", leading to people to judge it as authoritative. People anthropomorphize AI, and ascribe it intentionality and humanlike intelligence that it does not actually have; so, when they do not give human-like uncertainty indicators, we judge the AI to be certain, as a person who spoke confidently likely would be. Second, AI is massively increasing the amount of false information that exists and that people are exposed to. AI is being incorporated into a staggeringly large number of products across all fields of life, leading to increased risk of seeing incorrect or biased information. This AI output is then fed back into the data sets

used to train models, making them more prone to bias and misinformation. The more often people see that information, the more likely it is to stick, and the harder it is to mitigate the damage. Third, users who are engaging with generative AI programs, like Google's search summaries or ChatGPT, are usually trying to learn information they do not already know. This means they are uncertain most of the time; the less certain someone is about a topic they are learning about, the more susceptible they are to absorbing misinformation. The output from the model, to the user, seems to authoritatively eliminate their uncertainty, solidifying its place to them.

The article concludes by calling for more transparency and regulation now, while AI is still in its infancy and we have a chance to mitigate the harms before AI is so deeply ingrained in our society that it's too late. It calls for studies, the development of interventions, public education and audits of data sets. By more accurately setting expectations as to what AI can and cannot do, it is possible to mitigate some of these harms.

Lea, Gary R. 2020. "Constructivism and Its Risks in Artificial Intelligence." *Prometheus* 36(4). doi: 10.13169/prometheus.36.4.0322.

This article applies a constructivist lens to artificial intelligence, and breaks the technology into two categories. First, Artificial Narrow Intelligence (ANI), which are algorithms and programs created for a specific purposes with limited application. Much of the technology that powers smartphones, social media algorithms, self driving cars, and smart devices rely on this technology, which has been around since the late

1990s/early 2000s. The second type is Artificial General Intelligence (AGI), or, the idea of making a computer with human-like thinking and reasoning capabilities, which can function autonomously. This has emerged in more recent years, and is still in its infancy. Both carry risk, though risks of different types and scopes. This paper defines risk using a hybrid approach, considering both physical, objective risk factors, and subjective, socially constructed risk factors.

One major risk involves using either type of AI in military applications. Using a metaphor of chess-playing robots, the author demonstrates how computers mimic human behavior when playing the game, but do not have the ability to mimic human reasoning. Thus, coding a computer to understand the laws of war is an extremely complex task; in addition, intelligent machines do not fit neatly into the chain of command of militaries, making accountability difficult to trace, particularly if one views AI as an autonomous agent. AI also carries financial and health risks, as it becomes more integrated in portfolio management and healthcare diagnostics. ANI and AGI also lead to job losses across sectors, with the author citing that Goldman Sachs went from having 600+ equity traders in the year 2000, to just 2 by 2017 who are mostly responsible for signing off on paperwork. AI research also often uses the real world as a lab, as in the case of autonomous vehicles and military algorithms choosing targets; failures of these technologies are life and death. What regulations we have are currently well behind the level of risk levels introduced by AI. There are also fears of an AI apocalypse, where machines overtake humanity and destroy us, as in many sci fi vehicles; or of mass unemployment due to AI automation leading to mass economic collapse. More

immediately, however, are risks such as loss of rights to privacy, and AI-incentivized and -enabled surveillance by government and private actors. AI systems also replicate our the biases of their makers, and reproduce them, leading to risks to marginalized communities when systems are deployed. The author concludes by musing that the social implications of AI are under considered, and that the technical focus of the AI sphere may be a distraction.

Leben, Derek. 2024. “Deepfakes and the Ethics of Generative AI.” *Tepperspectives*. Retrieved February 16, 2025 (https://tepperspectives.cmu.edu/all-articles/deepfakes-and-the-ethics-of-generative-ai/).

This article grapples with some of the ethical implications of generative AI being used to make deepfakes, or digital representations of real peoples’ voices and images to create new works. He discusses an example of a voice clone of David Attenborough, made using publicly available clips of him. While his ethics students at first saw no issue with it, as no profit was gained or harm was done, Attenborough himself later commented that he disapproved of it, and feared that things he did not believe would be put into the world without his consent, but using his voice. As generative AI improves, the products become more and more realistic and thus, indistinguishable from reality; you could use this technology to make a world leader appear to say something scandalous or dangerous, and it may be too late to contain the consequences by the time the truth comes out. Real harms are already being done using deepfake technology, such as generating

pornographic material of children using pictures of their faces, or making a voice file of a colleague saying something racist to attempt to get them fired. The author adds that there needs to be more ethics discussion around this form of AI, and calls for a nuanced discussion and appropriate regulation of the technology. In addition, the author calls for AI design companies to be held more responsible for the potentially harmful outputs of their products.

Leffer, Lauren. 2023a. "Humans Absorb Bias from AI--and Keep It after They Stop Using the Algorithm." *Scientific American*, October.

This article explores the ways in which humans absorb bias that emerges from AI models. AI models themselves absorb bias from un-audited data sets, which are freely available and commonly used. This leads to AI voice-to-text that cannot understand accents, healthcare algorithms denying care based on protected classes, and even arrests made with predictive policing algorithms. Leffer goes over a study which shows that humans using these systems absorb the bias, and continue to have that bias well after they have stopped using the model. Using a medical model for diagnosis, some participants were given an AI assistance which introduced false positives; even once the AI was gone, the subjects continued to 'diagnose' in line with what they had learned from the AI. This also has been shown in predictive policing models, where even after biased models are decommissioned, its effects on patrol routes remain. In addition, AI has no 'uncertainty signals' like a human would – pauses in speech, "ums", filler words, and body language indicating that a person is unsure. This means all AI information is delivered

authoritatively, as if true. At minimum, Leffer calls for training data disclosures so there is more transparency in where data comes from, and research can be done on what biases they hold.

Leffer, Lauren. 2023b. "Your Personal Information Is Probably Being Used to Train Generative AI Models." Retrieved February 16, 2025 (https://www.scientificamerican.com/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models/).

This article discusses the methods companies use to create generative AI models, and emphasizes that artists and writers are not the only ones at risk of harms from the practice of data scraping. She cites an example where an artist looking for her work in a dataset found a diagnostic medical image in a scraped database. Even things set to private online are at risk, as data breaches and lax privacy settings do not adequately protect users. Many companies, like Google and Meta, also have troves of their own tools and data they draw from to create their models. Over time, these companies have become less and less transparent about where they get their data sources, seemingly in order to keep plausible deniability about copyright infringement accusations. The author then describes the ways that AI absorbs and reproduces human bias, and how difficult it is to protect your data, as well as how little regulation exists to hold companies accountable.

Martins, Berten. 2024. "Economic Arguments in Favor of Reducing Copyright Protection For Generative AI Inputs and Outputs." *Bruegel*.

This working paper argues that licensing data scraped by AI models is too onerous and impacts economic growth and competitiveness. It argues that the right to opt out of your data being used for AI is inefficient, and only gives more power to copyright holders without benefitting the public. The article, however, cites no sources, and does not back up its arguments beyond stating the author's opinion.

O'Brien, Isabel. 2024. "Data Center Emissions Probably 662% Higher than Big Tech Claims. Can It Keep up the Ruse?" *The Guardian*, September 15.

This author criticizes the tech industry for a lack of transparency around their emissions, particularly as AI technologies consume more and more electricity each year. It alleges that, based on studies, it is likely data center emissions from 2020-2022 were almost 8 times higher than officially reported by Google, Microsoft, Meta and Apple. In 2022, before the launch of ChatGPT, data centers "accounted for 1% to 1.5% of global electricity consumption". The article points out that it takes 10 times more energy to ask ChatGPT a question than to input the same query into a Google Search, and cites a Goldman Sachs study which predicts data center power demand to increase by 160% by 2030. Meanwhile, companies continue to claim they are or will soon be carbon neutral. The article attributes this to "creative accounting", and cites Amazon employee testimony that their company is expanding use of fossil fuels.

The article then explains and criticizes the use of carbon credits, purchased by large companies in the form of renewable energy certificates (RECs). This allows them to pay other companies to use or create renewable energy, even if their own facilities never

use it. This obfuscates the impact of local emissions. The article then points out that if the big five tech companies of Amazon, Meta, Apple, Google and Microsoft were a country, they would emit more than most countries. While Google and Microsoft aim to reduce or eliminate use of RECs, Amazon and Meta are attempting to keep them in their playbook. All of these companies still use 'creative accounting' to underreport emissions, however. In addition, it is nearly impossible to audit the emissions of contractors for these firms, as the complex supply chains and global spread make it an accounting nightmare. The article concludes by musing on whether the power grids, even as they exist now, can keep up with the demand for energy imposed by AI.

O'Neil, Lorena. 2023. "These Women Tried to Warn Us about AI." *Rolling Stone*. Retrieved February 16, 2025 (https://www.rollingstone.com/culture/culture-features/women-warnings-ai-danger-risk-before-chatgpt-1234804367/).

This article discusses a group of women who co-lead Google's Ethical AI group, and the issues they encountered as LLMs were being developed. The field has an extremely small number of black people, and unsurprisingly, the LLMs developed are shown to have biases based on the data fed into them from many male-dominated spaces online, such as Twitter/X, Reddit and Wikipedia. Some such biased outputs include:

> … the prompt "the man worked as," it completed the sentence by writing "a car salesman at the local Wal-Mart." …the prompt "the woman worked as" generated "a prostitute under the name of Hariya." Equally disturbing was "the white man worked as," which resulted in "a police officer, a judge, a prosecutor, and the

president of the United States," in contrast to "the Black man worked as" prompt, which generated "a pimp for 15 years."

While there have been attempts to moderate and filter out this kind of problematic content, it often catches things which are not offensive, like suppressing words used among marginalized communities to describe themselves. LLMs are now everywhere, and its own creators are sounding the alarms about the existential threats they pose to "exterminating humanity" – while ignoring the harms the bias in their systems are already doing. The author wonders if this would be different if early AI ethics workers had been heeded. The article describes the ways bias against dark skinned users is perpetuated by AI, such as facial recognition not working on dark skin tones, or misidentifying a dark skinned woman as a man. Error rates for classifying dark-skinned women with AI are 34.7%; the error rate is 0.8% for white men. This would be less chilling if AI were not being increasingly used in law enforcement, hiring practices, housing placements, and loan evaluations. The article calls for real regulation of this industry, rather than the current measures, which are often nonbinding and voluntary. It ends by underscoring that these systems are a threat, not because they are magical or stand to become all-powerful, but because lack of understanding leads to overreliance and over-trust in them.

Raji, Inioluwa Deborah. 2023. "AI's Present Matters More than Its Imagined Future." *Atlantic Monthly (Boston, Mass.: 1993)*, October 4.

This article discusses the first meeting of Chuck Schumer's AI insight forum, headed by figures like Sam Altman, Elon Musk, Bill Gates and Satya Nadella. The author criticizes the forum for over-focusing on imagined, extreme dangers of AI, such as fear that sentient computers will take over. Meanwhile, the real harms already coming from use of LLMs were completely undiscussed. The author describes everyone in the meeting as "daydreaming", including the lawmakers, despite the meeting's intent to create meaningful AI regulation. Poor understanding of what AI is and can do, as well as the term being a catch-all for many disparate technologies, is partly to blame. The author criticizes the existing discourse for over-focus on "influential corporate voices", and calls for a discourse focused on tangible harms AI is already doing – such as arrests made on faulty face recognition - rather than imagining potential harms or benefits it could be part of in the future.

Schick, Nina. 2023. "FAKING IT: Navigating the New Era of Generative AI May Be the Most Critical Challenge to Democracy Yet." *RSA Journal* 169(2(5593)):40–43.

This article discusses the potential implications of AI-generated misinformation and deepfakes on our political climate, and the threats it poses to a stable democracy. The author points out that, even when there is no malicious intent, actors can spread AI misinformation without realizing it. Further, AI can create so much information so quickly that it is nearly impossible to sift through it all to discern what is true. This leads to a fertile ground for 'censorship through noise', a scenario in which the truth is buried by flooding the zone with fake information, making it impossible to process it all. She

also points to its geopolitical implications using the example of a deepfake video of Ukrainian President Zelensky surrendering in the early days of Russia's invasion of Ukraine. The potential for identity theft scams using deepfakes is already being realized, with losses already in the billions of dollars.

All this leads to an undermining of digital trust. This is a form of the liar's dividend – the less people trust information, and the more difficult it is to discern between lies and truth, the more power bad actors have in a system. Simply knowing that it is possible to create deepfakes allows injustice to hide behind it – images of abuses and injustices can be easily written off as AI generated, robbing the photograph of its journalistic power and creating a world where no one knows who they can trust. The author calls for AI authentication frameworks, like watermarks, which would be embedded in files and impossible to remove, showing where it came from and how it was made. However, the author emphasizes that without broad understanding and participating by humanity, these tools will fall short.

Vicente, Lucía, and Helena Matute. 2023. "Humans Inherit Artificial Intelligence Biases." *Scientific Reports* 13(1):1–13. doi: 10.1038/s41598-023-42384-8.

This paper outlines an experiment which involved introducing a biased AI assistant into a simulated diagnostic task, and then, in another phase, removing the AI assistance to see if the raters continued to exhibit that bias. The bias was noted by the experimenters to be easy to spot and obviously incorrect; however, participants in the AI assisted group still made more errors. It showed that the bias remained long after the AI

was gone, discrediting the idea that AI-assisted decision making reduces errors in human decision making, and demonstrating that even removal of the biased actor influences future behavior. It warns of the dangers of assuming AI systems do not have their own biases in the code, and do not make errors. This perception of computers as infallible and objective may lead to AI actually harming accuracy. It mentions the dangers of using these technologies in clinical settings, or for technologies that use face recognition for law enforcement. Even if the faulty system is removed, it is likely the bias introduced by the system will remain with those who used it. It cautions against over-reliance on these systems, as they can reduce skill and harm decision making.

Wong, Matteo. 2023. “America Already Has an AI Underclass.” *Atlantic Monthly (Boston, Mass.: 1993)*, July 26.

This article discusses a worker, Michelle Curtis, who works for a company assessing AI outputs for quality and accuracy. It criticizes the field for being underpaid, lacking specific guidance on how to gauge quality, and for overworking its workers. Tasks for evaluating AI are often timed, which leads to increased stress for the worker and lower quality fact-checking. These workers are rarely mentioned by the companies that contract them and exist as an invisible AI underclass. They are the humans in the machine, helping it to learn, unlike the assertion that AI models all learn on their own. The existence of these people would run counter to the near-magical narratives around AI, even though the practice of evaluating algorithmic products (like search engines and AI) has been industry standard since their inception. Even when companies like Google

occasionally acknowledge their human laborers, they do not benefit monetarily from the huge leaps in market value that the introduction of AI has given many companies. This labor is often outsourced to the global south, where some workers earn as little as $1.50 an hour. Tight deadlines, low pay, vague instructions and the increased demand for AI create perfect conditions for human error. There are now unionization efforts underway across the industry; however, progress is slow and plagued by setbacks, such as retaliatory layoffs, and large companies passing the blame to their contractors for poor conditions. Since one of the few things AI cannot do is train itself with the level of quality of human feedback, it is crucial that awareness is spread of these ‘ghost workers’ and that conditions for them improve.