CSCI 1010 Class 13

Profs. Michael Linderman and Phil Chodrow
Department of Computer Science
Middlebury College

Anonymous Poll: Level of concern about GenAI

Use this QR code to access the Google Form and submit your response.

Concerns(?) about GenAI

- Bias and discrimination
- Authenticity and “deep fakes”
- Environmental impacts
- Job displacement and economic impacts
- Accountability and transparency
- Others?

impact learning
loss of capability
ability

identity theft /
privacy

copyright / intellectual
property

disincentivize artistic
creation

security

Slide 2 Notes

I seeded the poll with some common categories of concerns about GenAI. Are there others that were not included in this list that should have been? Maybe “security”, e.g., use of LLMs to develop weapons, malware, etc.? Or maybe more potent versions of existing concerns, e.g., phishing, theft, etc.

The metaphor matters

Poll: Which of the following (metaphors) best describes GenAI/LLMs?

a. Stochastic parrots[1]

b. Lossy compression of the Internet[2]

c. Calculator for words [3]

d. Bullshit machines (producing language without regards to actual truth)[4]

1. Bender et al., ACM FAccT Conf. 2021
2. Ted Chiang, New Yorker, 2023
3. Sam Altman, 2023
4. Bergstrom and Ogbunu, Undark, 2023

Slide 3 Notes

Before proceeding, let's consider our language around GenAI/LLMs. As the reading noted, the metaphors we use to describe these technologies can shape our perceptions and attitudes toward them. Which of these do you think best describes GenAI/LLMs?

- Were there other metaphors not included here that you think are more appropriate?
- What impacted your perception of "C"? Perhaps that Sam Altman is CEO of OpenAI…

  'OpenAI CEO Sam Altman visited Japan on Monday, where he spoke at Keio University in Tokyo. […] he emphasized that AI wouldn't replace learning. He said, "Probably take-home essays are never going to be quite the same again," adding, "We have a new tool in education. Sort of like a calculator for words. And the way we teach people is going to have to change and the way we evaluate students is going to have to change." '
- How does the how does the choice of metaphor influence your perception of concerns? Do some of those metaphors highlight or increase your concerns?
  - For example, LLMs as a "lossy compression of the Internet" (and particularly Reddit!) amplifies my concerns about bias and misinformation. That is the technology will amplify the voices (biases) of specific communities that are (over-)represented online.
  - On the other hand, LLMs as "calculators for words" seems to downplay concerns, as calculators are widely accepted tools that enhance human capabilities without replacing them.
  - "Stochastic parrots" makes me think about authenticity. In the essay "Why A.I. Isn't Going to Make Art", Ted Chiang describes "art is something that results from making a lot of choices" (e.g., for each word in a novel). But with LLMS, you are only making a few choices and the rest are "filled in" stochastically. One argument is that is only these "high level" choices that matter, but Chiang argues that "the countless small-scale choices made during implementation are just as important to the final product as the few large-scale choices made during the conception."
  - "Close to home", LLMs as "bullshit machines" would make unverified use of LLM output in teaching settings, which is (nominally) anchored in truth, seem particularly problematic (IMHO).

A we talk today, let's keep our choice of language, and its implications on our perceptions, at the forefront of our minds.

Further reading:

- Challenging the Myths of Generative AI
- "Why A.I. Isn't Going to Make Art"

Activity: Use cases for GenAI/LLMs

1. Using the provided sticky notes, write down possible uses of LLMs (one use per note). Be creative and don’t restrict yourself to uses that you think are good ones.
2. Place your notes on the whiteboard on the provided spectrum from “bad idea” to “it depends” to “great idea”.

Slide 4 Notes

Draw three regions on the whiteboard: bad idea, it depends, great idea.

- What are some uses around which there is consensus? Disagreement? What about the latter prompts disagreement?
- What are the nuances around the "it depends" cases? What do they "depend" on?

How do some of these use cases relate to our list of concerns? Record some of the responses in a Google Doc to seed the subsequent discussion.

Adapted from The Bullshit Machines Instructor Guide: Lesson 8

Some roles for data in GenAI/LLMs

1. Self-supervised training on text corpora (10-1000s billions of tokens from the web, books, GitHub code, etc.)
2. Fine tuning for instruction following or other specific tasks
3. Alignment via preference data

Slide 5 Notes

Recall from the reading the many ways that the data, and the choices developers make about what data to use and how it is filtered, pre-processed, etc. create the potential for bias.

As we noted earlier in class with Kranzberg's First Law of Technology: "Technology is neither good nor bad; nor is it neutral." And as stated by Joy Buolamwini in the description of the Gender Shades project: "Automated systems are not inherently neutral. They reflect the priorities, preferences, and prejudices – the coded gaze – of those who have the power to mold artificial intelligence."

This isn't to suggest that the developers are actively trying generate discriminatory outcomes, but instead that is the unintentional result of the choices, often well-intentioned, that the developers made (for training data, for evaluation benchmarks, etc.).

I think of the quote from Birhane and Prabhi included in the reading: "Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy."

I suspect we are more familiar with the first source of data, the large text corpora used for self-supervised training. But alignment data can also be important. "Alignment" is the process of ensuring that the model's outputs "align" with human preferences (and values). The goal is "[steer] LLMs towards certain behaviours such as honesty, harmlessness, helpfulness, safety or informativeness"[@kirkPersonalization2024]. It is the process that turns "raw" next-token prediction into the chat-style interaction. Alignment involves collecting data on human preferences, such as through user feedback or human evaluators, to fine-tune the model's behavior through techniques like RLHF ("reinforcement learning from human feedback").

When you say it out-loud, the whole idea trying to define a single (common) set of human preference (and values) seems impossible and extremely fraught. As Kirk et al. note " 'bias' to one user may be a desirable behaviour to another".

Aligned with whom?

"Personalized alignment" adapts LLM behavior based on user-specific data, i.e. $p(t_1, \ldots, t_n, U_i) = \prod p(t_j | t_{<j}, U_i)$, where U_i is specific user drawn from a population characterized by U.

Proposed benefits[1]:

- Individual: Efficiency, usefulness, respect for values, user autonomy, empathy
- Society: Accessibility, diversity and inclusion, democratization, productivity

What are the risks?

echo chamber, snowball

1. Kirk et al. Nat Mach Intell 6 (2024)

Individual Benefits	Individual Risks	Societal Benefits	Societal Risks
I.B.1 Efficiency: increased ease and speed with which end-users can find their desired information or complete a task, with fewer prompts or inputs to the model	I.R.1 Effort: increased user costs in providing feedback, a form of extractive volunteer labour	S.B.1 Inclusion and accessibility: improved adaptation to the communication needs of marginalized communities, including catering to those with disabilities or who speak dialects or languages that are deprioritized by current LLMs	S.R.1 Access disparities: uneven distribution of benefits, excluding those who cannot afford or access the technology and exacerbating digital divides
I.B.2 Usefulness: increased accuracy of predicting and meeting the needs of the end-user via personalized preferences and knowledge in outputs	I.R.2 Dependency: increased risk of over-reliance, attention commoditization and technology addiction	S.B.2 Diversity and representation: improved representation by tailoring outputs to diverse perspectives and avoidance of cultural hegemony by not prioritizing certain values over others	S.R.2 Polarization: increased divisions of individuals or groups into echo chambers and the breakdown of shared social cohesion
I.B.3 Respect for values: adaption to diverse ethical belief systems, values and ideologies, allowing for individualized socio-cultural personalization	I.R.3 Bias reinforcement: increased amplification of confirmation and selection biases, leading to epistemic harms	S.B.3 Democratization and participation: increased stakeholder involvement from diverse backgrounds in shaping behaviours, allowing for a more participatory and inclusive approach to development	S.R.3 Malicious use: use for harmful or illegal purposes, such as generating harmful language at scale, manipulating users via disinformation or fraud, or persuading users towards certain political views or brand preferences
I.B.4 User autonomy: increased positive freedom of choice and control over how the model behaves with personal data, promoting a sense of ownership and self-determination over the technology	I.R.4 Essentialism and profiling: increased risk of algorithmic profiling and assumptions based on demographic or geographic information, leading to the non-consensual categorization of people	S.B.4 Labour productivity: improved workforce productivity from positive externalities of effective and efficient task assistance	S.R.4 Labour displacement: increased automation risk of jobs, particularly minimum wage, routine and crowdworker jobs
I.B.4 Empathy and companionship: increased perceived emotional connection, leading to improved acceptance and trust of the system	I.R.5 Anthropomorphism: increased tendency to ascribe human-like traits, reveal sensitive information or form unhealthy attachments		S.R.5 Environmental harms: increased environmental costs from disaggregated training, data storage and inference costs
	I.R.6 Privacy: increased quantity of collected personal information, leading to risks of privacy infringement, particularly if the model operates with sensitive information or encourages information disclosure		

Authenticity: Spot the “Deep Fake”

https://www.spotdeepfakes.org/

Slide 7 Notes

That quiz focused on identifying "deep fakes". It does not address the non-consensual use of real individuals' likenesses. Many of the examples were focused on public political figures, but there are many examples of problematic non-consensual use of private individuals' likenesses. As the authors of Bullshit Machines note:

> The internet being what it is, of course the most common abuse involves creating pornographic material, including explicit images of celebrities and classmates. These systems are even used to create illegal child sexual abuse material.

Adapted from The Bullshit Machines Instructor Guide: Lesson 7

Activity: Authenticity

1. Using the provided sticky notes, write down possible writing tasks that one could in principle hand off to an LLM. Be creative and don't restrict yourself to uses you think are "appropriate".
2. Place your notes on the whiteboard in one of three boxes:
    1. Situations where you should never use an LLM
    2. Tasks where an LLM can provide a useful guide regarding form but you need to provide your own authentic content.
    3. Situations where authenticity doesn't matter and you can use an LLM without any qualms

Slide 8 Notes

Draw three regions on the whiteboard: 1) never use an LLM, 2) LLMs can be a useful guide for form but need authentic content, 3) authenticity doesn't matter, use without qualms.

As a catalyst, the authors of Bullshit Machines, offered "appealing a parking ticket" as a "use without qualms" example. What about e-mailing a professor about internship opportunities? The same authors describe this as a "form" task where the student needs to provide authentic content. But that the LLM can help with typical structure, cultural norms of address and tone, etc. What do you think?

From "Why A.I. Isn't Going to Make Art" by Ted Chiang, New Yorker, August 31, 2024 (archive link)

> The programmer Simon Willison has described the training for large language models as "money laundering for copyrighted data," which I find a useful way to think about the appeal of generative-A.I. programs: they let you engage in something like plagiarism, but there's no guilt associated with it because it's not clear even to you that you're copying.
>
> Some individuals have defended large language models by saying that most of what human beings say or write isn't particularly original. That is true, but it's also irrelevant. When someone says "I'm sorry" to you, it doesn't matter that other people have said sorry in the past; it doesn't matter that "I'm sorry" is a string of text that is statistically unremarkable. If someone is being sincere, their apology is valuable and meaningful, even though apologies have previously been uttered. Likewise, when you tell someone that you're happy to see them, you are saying something meaningful, even if it lacks novelty.

- Can we derive a norm for appropriate use of LLMs based on authenticity considerations? Ted Chiang suggested: > Of course, most pieces of writing, whether articles or reports or e-mails, do not come with the expectation that they embody thousands of choices. In such cases, is there any harm in automating the task? Let me offer another generalization: any writing that deserves your attention as a reader is the result of effort expended by the person who wrote it. Effort during the writing process doesn't guarantee the end product is worth reading, but worthwhile work cannot be made without it. The type of attention you pay when reading a personal e-mail is different from the type you pay when reading a business report, but in both cases it is only warranted when the writer put some thought into it.

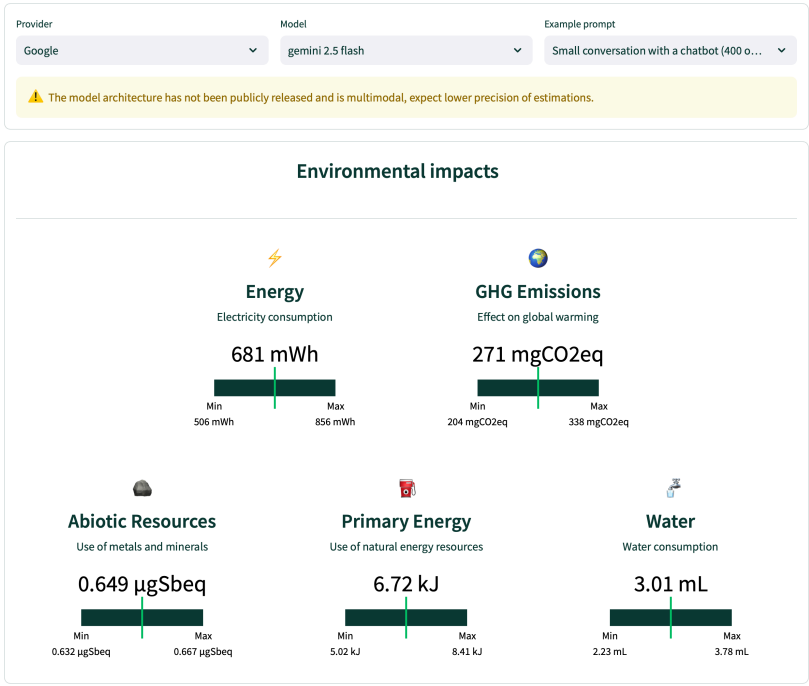
Adapted from The Bullshit Machines Instructor Guide: Lesson 14

Environmental impacts of GenAI/LLMs

- Energy consumption and associated carbon footprint for training and inference
- Water and other resource consumption for data centers and semiconductor manufacturing
- E-waste from rapid hardware obsolescence

Interactive resource estimators:

- ML Energy Leaderboard
- Ecologits Calculator



Slide 9 Notes

[click] One of the main challenges is how little information is publicly available about the energy consumption and environmental impacts of training and deploying LLMs. Companies like OpenAI and Google, do not typically disclose detailed information about their data centers, hardware, or energy usage. The sustainability reports they do release, e.g., Google's 2025 report, often only provide company-level summaries. Thus most of the data available are estimates, derived from open-source models and "leaks" or partial disclosures of technical details, e.g., model architectures, training times, hardware used, etc.

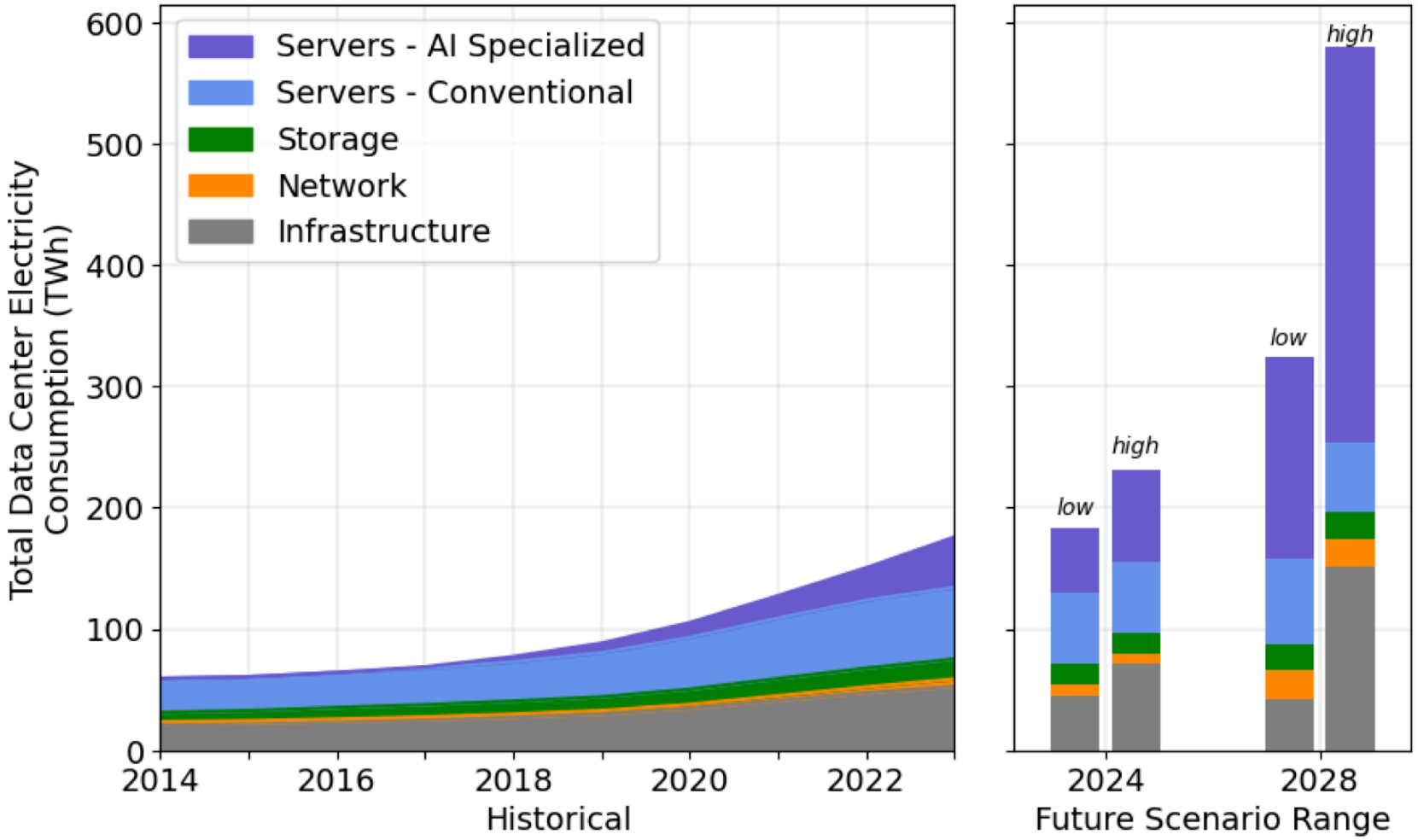
These environmental impacts are not evenly distributed. For example, data centers are often located in regions with cheap electricity, which may be generated from fossil fuels. Additionally, the environmental costs of resource extraction for semiconductor manufacturing are often borne by developing countries. As a very recent example, the Southern Environmental Law Center filed a lawsuit against xAI over what the suit alleges is un-permitted use of 10s of gas turbines, and the corresponding emissions, to power a data center in South Memphis, Tennessee, a primarily Black community.

One (counter-)argument is that AI will ultimately reduce energy consumption by optimizing other processes. For example, AI can be used to improve energy efficiency in buildings, optimize transportation routes, and enhance supply chain logistics. One example from Google is using machine learning to minimize contrails created by commercial aviation, one the major contributors (35%) that industry's global warming impact. However, the net effect of AI on energy consumption is still uncertain, as the energy savings from these applications may be offset by the increased energy demand from AI itself (much of which is related to LLMs, not to these optimization tasks) and "Jevons paradox" effects where increased efficiency leads to increased overall consumption.

Further reading:

- As Use of A.I. Soars, So Does the Energy and Water It Requires

Energy consumption



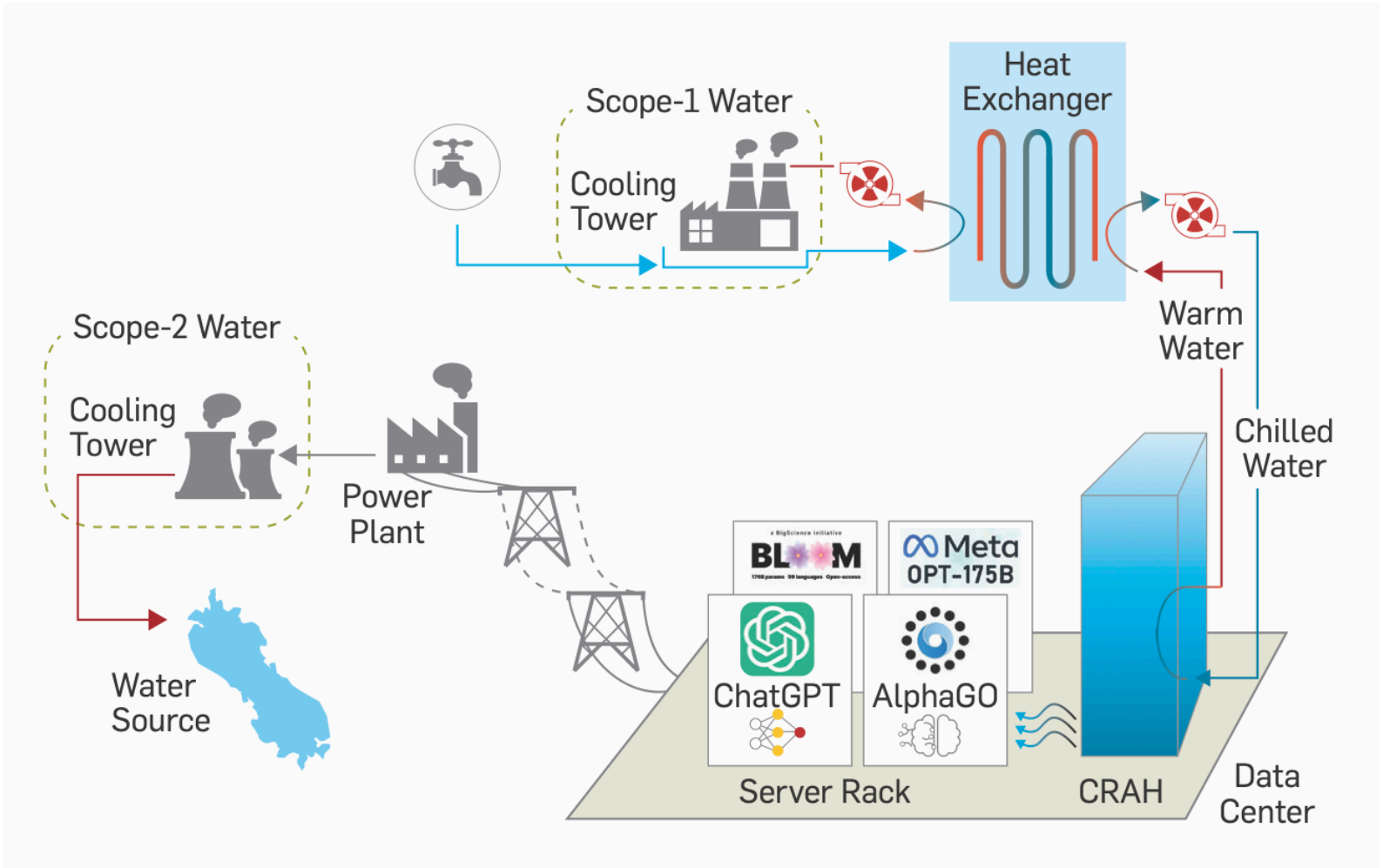
Current and projected data center energy use. Figure 5.6 in Shehabi et al. (2024)

Slide 10 Notes

From a report of overall data center energy use produced by Lawrence Berkeley National Laboratory in 2024 (released late 2024). As the authors note: "total energy use growth between 2014 and 2023 is driven by both a rapid proliferation of AI servers and well as continued growth in conventional server energy demand". The future role for AI in their projections is striking!

"Infrastructure" includes power consumed by cooling and other non-IT data equipment in the data center. A metric for analyzing data center power efficiency is Power Usage Effectiveness (PUE), which is the ratio of total data center energy use to IT equipment energy use (i.e., used by servers, networking equipment, storage, etc.). A PUE of 1.0 is ideal, meaning all energy is used by IT equipment. Modern "hyperscale" data centers can achieve PUEs of around 1.1 to 1.2.

Water consumption



Mechanisms of data center water consumption. Scope-3, water consumed during manufacturing not shown. Figure 1 in Li et al. (2025)

Slide 11 Notes

The figure shows some of the ways that water is consumed in data centers, both directly , e.g., for cooling (scope-1), and indirectly, e.g., water used to generate electricity consumed by the data center (scope-2) or to manufacture the hardware (scope-3, not shown).

Water consumption here is defined as "water withdrawal minus water discharge", i.e., "freshwater taken from the ground or surface water sources" and not returned to the immediate water environment (e.g., due to evaporation).

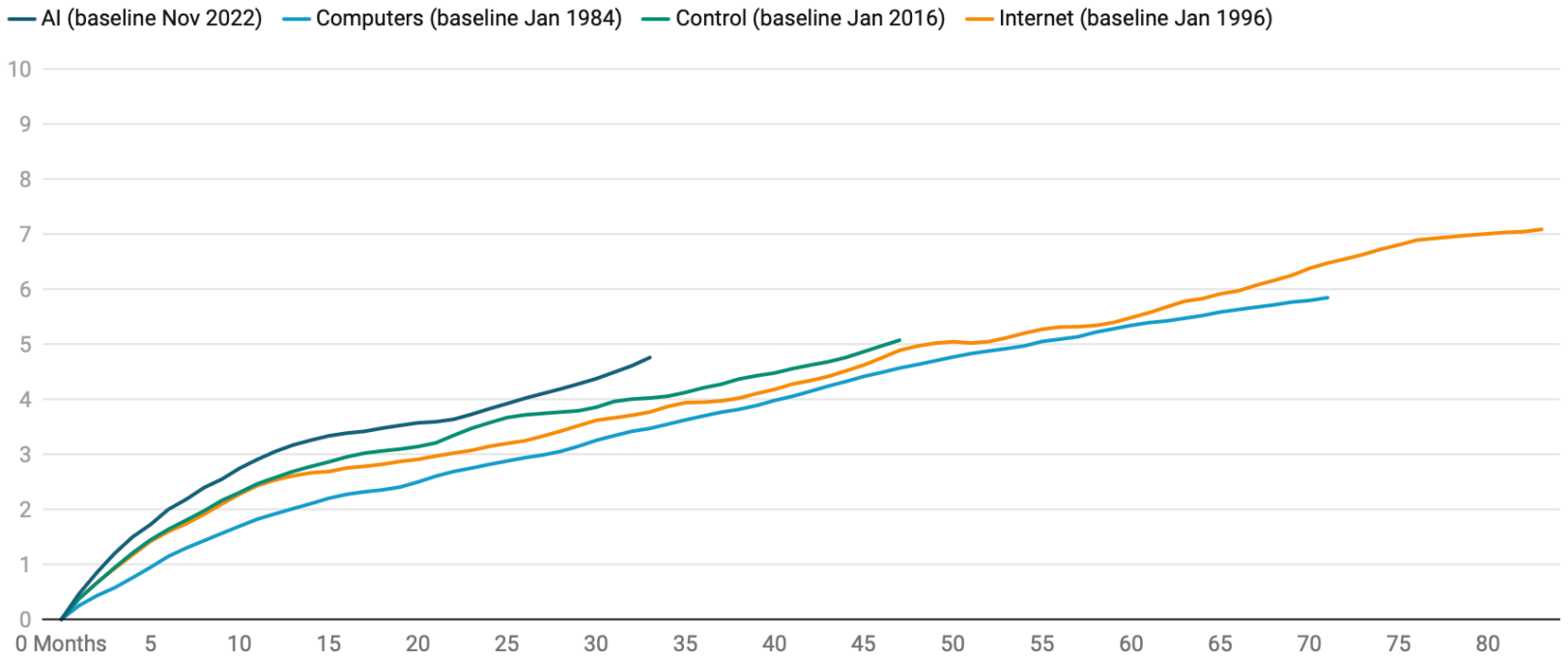
@liAILessThirsty2025 estimated that "training GPT-3 in Microsoft's U.S. datacenters can consume a total of 5.4 million liters of water, including 700,000 liters of scope-1 onsite water consumption.", and "GPT-3 [consumes] a 500ml bottle of water for roughly 10–50 medium-length responses".

A metric for water consumption is Water Usage Effectiveness (WUE), which is a measure of water consumption, e.g., from evaporative cooling systems, in liters per kWh. The LBNL data center report estimates hyperscale and AI-specialized data centers have WUEs of around 0.3 ad 0.6 L/kWh, respectively.

Future of work and wealth (in)equality?

Figure 1. Changes in the Occupational Mix Over Different Periods of Technological Change

Dissimilarity index (percentage points)
Months from baseline on x-axis



Dissimilarity index is calculated using a prospective 12-month moving average of employment data

Chart: The Budget Lab • Source: CPS, The Budget Lab analysis • Created with Datawrapper

Changes in the Occupational Mix Over Different Periods of Technological Change. Figure 1 in Gimbel et al. 2025

Slide 12 Notes

TL;DR; I am not sure anyone knows…

Judy Wajcman, Professor Emeritus at London School of Economics, an expert in the social analysis of technology, said in a November 2025 interview, "as far as I can see through my long history of looking at these things is that some jobs are replaced, some are changed and let's say augmented, and lots of different kinds of jobs are created. All of those things go on and at any point in time, it's hard to know which of those things is going to be the dominant thing."

One of the famous examples you might see of the uncertain impact of technology on jobs is "although ATMs replaced humans in the job of counting out cash for withdrawals, that made it cheaper to operate a bank branch, so the number of branches increased, leading to more bank employees overall." [@russellArtificialIntelligence2020]

This figure from a recent report from Yale's Budget Lab shows changes in occupational mix over different periods of technological change. A "percentage point difference means that, relative to the start point, that percent of workers are in new occupations. This can occur by workers changing jobs, losing jobs, or unemployed people getting a new job." The authors note that changes since the "advent of generative AI" (2022-onwards) "seem to mirror the trends seen during the three comparison periods." Albeit slightly higher. However, if you look back to earlier baselines (before 2022), those changes may have already been underway for some time (i.e., more the result of macroeconomic trends than AI specifically).

Those averages may obscure that the impacts are not evenly distributed. The "information" sector (which includes software developers, data analysis, etc.) has a much higher dissimilarity index (approach 14 percentage points) than the economy as a whole. And recent college graduates (a group we are all very interested in) may be also be disproportionately impacted, although the data seems less clear on that question.

Russell and Norvig wrote: "The net effect of automation seems to be in eliminating tasks rather than jobs." But that distinction is not so simple. Doing those tasks is someone's job. And as we noted above, the impacts will not be evenly distributed by profession, or by demographic characteristics (for example, truck drivers are predominantly men so automated trucking will disproportionately impact that demographic).

A more utopian view is that AI will finally enable Keynesian visions of a future of leisure, where people are freed from work to pursue more creative and fulfilling activities. However, Wajcman is skeptical of this vision. She said (specifically in response to quotes from Elon Musk and others about jobs being unnecessary in the future):

"I think it's complete nonsense. The issue is distributing work more fairly. I don't think we're going to have huge amounts of things automated so that we're just going to sit about. […] How can these people who validate working long hours more than anyone — who say 'we do amazing, creative work, we do genius work, we want to work all the time because the work is so enjoyable and interesting and amazing.' How can these people think about other people not having access to interesting, enjoyable work? […] For people's identity, to feel like you're worth something in society and have an identity, I'm afraid in this society that we live in, having a job, having some work, having a kind of function is incredibly important to identity and shouldn't be denied people. And it's a weird thing for those people to tell us to live differently."

Often the main pitch for AI tools is time saving. But in a similar vein as above that implies that some tasks are not worth doing. In the same interview, Wajcman said: "And there's a notion that some kinds of activities aren't really worth doing and it would be better to automate them and that will leave you to do things with your time that are better, that are a wiser use of your time. And in that story is an amazing amount of value-laden assumptions about what are good activities to do, what are valuable activities to do and what are the activities that it would be just as well to automate. And we need to question a lot of those assumptions."

We observe that the benefits of AI adoption (and technology more generally) may accrue disproportionately to those who own the AI technologies (e.g., shareholders of tech companies) rather than the broader workforce, potentially exacerbating wealth inequality. In 1990 the top 3 companies in Detroit had similar revenues as the top 3 companies in Silicon Valley in 2014, but employed 10x more workers (1.2 million vs. 137,000), and had a 30X smaller market capitalization (from Digital era brings hyperscale challenges, Financial Times, 2014). The low marginal costs of software products (cost to make the next copy is near zero) compared to physical goods (e.g., cars) tends to accentuate what has been called the "winner-takes-all" dynamics

of the digital economy.

Anonymous Poll: Change in level of concern about GenAI?

How, if at all, has your level of concern about GenAI changed since before class? Use this QR code to access the Google Form and submit your response.

Close to home: Our relationship with GenAI

Using your sticky notes, write down examples of (un)healthy relationships with GenAI, placing your notes on the whiteboards in the appropriate areas. Be creative! You don't need to restrict yourself to what you have actually experienced or observed. Hypothetical examples are helpful too! Specifically, address the following areas:

- What, to you, characterizes a healthy relationship with GenAI?
- What do you wish could be different about your or other's use of or relationship with GenAI?
- Barriers to developing a healthy relationship with GenAI?
- Support needed to develop a healthy(ier) relationship with GenAI?

Slide 14 Notes

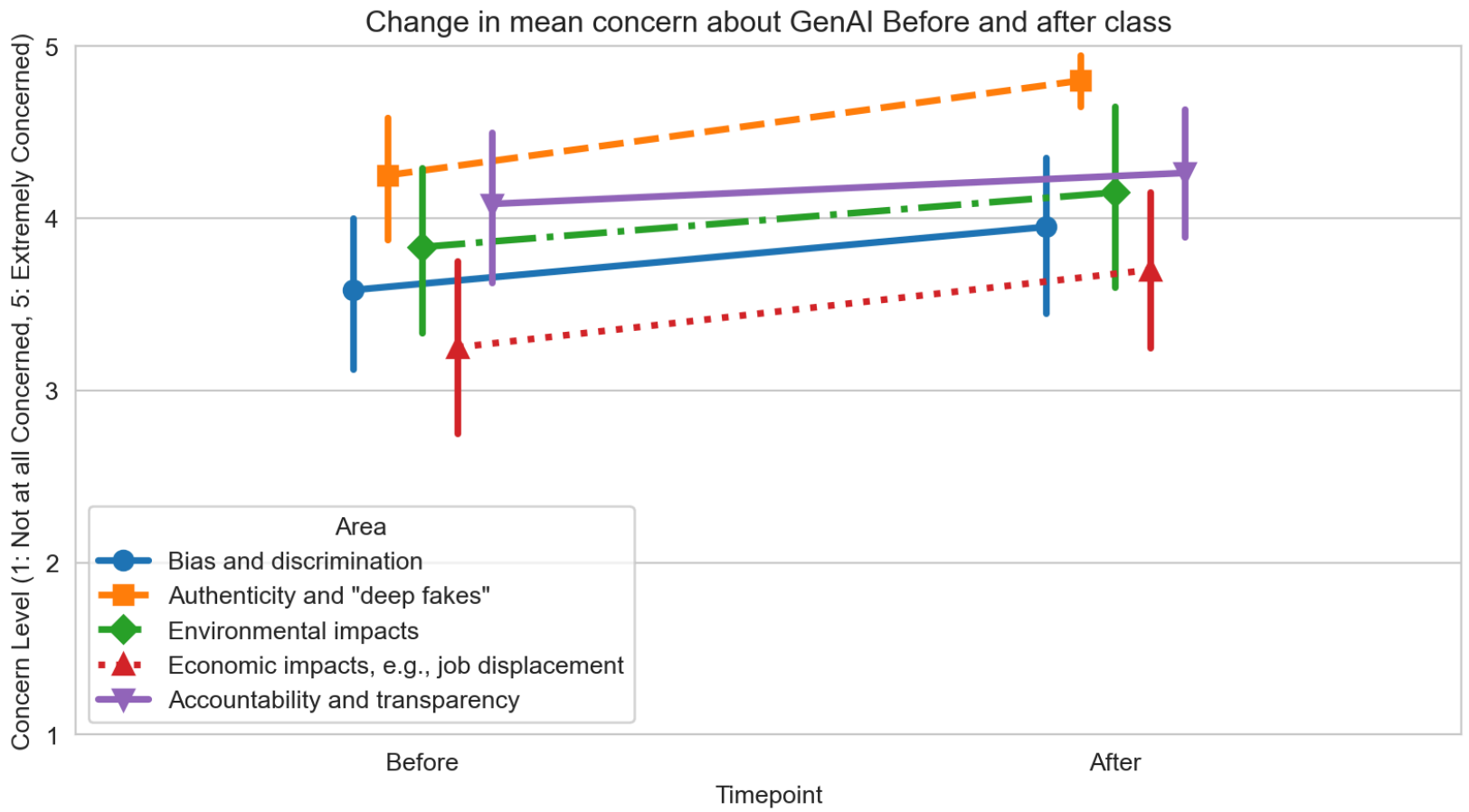
The Middlebury Strategic Planning working group on “Engage and Lead in an Age of AI and Technology” reached out to CSCI 1010 to help gather student perspectives on GenAI at Middlebury. Particularly from students who had just spent several weeks thinking about these very technologies. These comments will be anonymized and summarized for the working group to consider as they develop recommendations for the broader Middlebury community.

Close to home: GenAI at Midd

Using your sticky notes, write down **S**trengths, **W**eaknesses, **O**pportunities, and **T**hreats (SWOT) of how Middlebury is (or isn't) addressing GenAI in the following areas. Place your notes on the whiteboards in the SWOT area for the relevant topic (recognizing that some notes may apply to multiple topics).

- Teaching and Learning
- Community, e.g., campus social environment
- Preparing for life after Middlebury

Change in concern about GenAI before and after class (poll results added after class)



References

Kirk, Hannah Rose, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2024. “The Benefits, Risks and Bounds of Personalizing the Alignment of Large Language Models to Individuals.” *Nature Machine Intelligence* 6 (4): 383–92. https://doi.org/10.1038/s42256-024-00820-y.

Li, Pengfei, Jianyi Yang, Mohammad A. Islam, and Shaolei Ren. 2025. “Making AI Less ’Thirsty’.” *Commun. ACM* 68 (7): 54–61. https://doi.org/10.1145/3724499.

Russell, Stuart J., and Peter Norvig. 2020. *Artificial Intelligence: A Modern Approach*. 4th Edition. Pearson. http://aima.cs.berkeley.edu/.

Shehabi, Arman, Alex Newkirk, Sarah J. Smith, Alex Hubbard, Nuoa Lei, Md Abu Bakar Siddik, Billie Holecek, Jonathan Koomey, Eric Masanet, and Dale Sartor. 2024. “2024 United States Data Center Energy Usage Report.” LBNL-2001637. Lawrence Berkeley National Laboratory. https://escholarship.org/uc/item/32d6m0d1.