Recent Issues with Big Data: Critical Questions for a Critical Time

Filipp Lepalaan

2025-05-15

Introduction

This essay analyses the different claims GenAI providers and customers are officially stating as their policy and critiques them in light of recent developments. I also try to incorporate a historical perspective highlighting the change in attitudes towards technology use in education in particular. In many ways this paper could be seen as a broader and more critical version of the Data News feature we so enjoyed during our lectures. It's less about solving a specific problem and more about avoiding future ones by bringing to light some of the inherent contradictions in our current relationship with technology in general and Big Data in particular.

The Great Open Web Plunder

In talking about GenAI in the context of Big Data, we cannot skip discussing the origins of the data that has been used to train the underlying Large Language Models. Take for example OpenAI - the makers of the most popular GenAI tool ChatGPT. Their documentation states that:

> "OpenAI's foundation models, including the models that power ChatGPT, are developed using three primary sources of information: (1) information that is publicly available on the internet, (2) information that we partner with third parties to access, and (3) information that our users, human trainers, and researchers provide or generate." ("How ChatGPT and Our Foundation Models Are Developed | OpenAI Help Center" 2025)

OpenAI does not provide concrete statistics on how much of the total each data source constitutes, but given the vast amount of publicly available data and the relative ease and efficiency with which it can be collected using automated tools, it is safe to assume that the open web forms the vast majority of their input. Information from third party partners may be better

structured and curated, but to achieve seemingly universal applicability, GenAI tools must prioritize quantity.

This raises an important question - how do the data collection bots know the terms under which the collected data is published? Publicly available information may appear to belong to the Public Domain, but it does not (Atkinson 2024). That determination is ultimately up to the publisher of the content in question. There exist semantic web conventions (Rodriguez-Doncel and Delgado, n.d.) that allow a publisher to attach licensing terms to a page in a machine-readable format, but if nobody is mandating their use then why would the data collectors be under any obligation to follow them?

We are already beginning to see the first lawsuits emerge by media outlets as well as artists and performers (Jiang et al. 2023). It is only a matter of time when we get to also see the first class action lawsuits (Eaton 2025) from countless independent journalists, bloggers, photographers and content creators waking up to the fact that their work is being used to establish a new information monopoly.

OpenAI goes to great lengths to claim that their *modus operandi* does not actually entail "copy-pasting" other people's work:

> "Machine learning models consist of large sets of numbers, known as "weights" or "parameters," along with code that interprets and uses those numbers. These models do not store or retain copies of the data they are trained on. Instead, as a model learns, the values of its parameters are adjusted slightly to reflect patterns it has identified." ("How ChatGPT and Our Foundation Models Are Developed | OpenAI Help Center" 2025)

The illustration they present is, at first glance, quite compelling:

> "… similar to how a teacher, after extensive study, can explain concepts by understanding the relationships between ideas without memorizing or reproducing the original materials verbatim." ("How ChatGPT and Our Foundation Models Are Developed | OpenAI Help Center" 2025)

The problems with this defense are threefold. First, the case of New York Times vs OpenAI clearly shows (Cooper and Grimmelmann 2025) that ChatGPT can and does reproduce the original material verbatim. Secondly, citing references - something that ChatGPT seems to have significant struggles with (Emsley 2023) - is not just an established academic best practice, but a core tenet of education. Teachers don't simply "explain concepts by understanding the relationships between ideas", they present their sources - scholars, books, papers, articles as part of the curriculum. The source of the information is just as important as the information itself. Which leads us to my third point - trust. When a teacher presents information they are backing it up with their own credibility and the credibility of the educational institution. They are responsible for the information. Not so with ChatGPT which has always stated that the onus is on the user, not "the teacher".

So it appears, there is a less-than-zero probability that the tools we are so eagerly making our society dependent on are based on illegal practices and may require significant overhaul (Eaton 2025) as the extent of their overreach becomes more apparent in the coming years.

The OpenAI document goes on to address the matter of personal data:

> "A significant portion of online content involves information about people, so our training data may incidentally include personal information. However, we do not intentionally collect personal information for the purpose of training our models." ("How ChatGPT and Our Foundation Models Are Developed | OpenAI Help Center" 2025)

Setting aside the glaring contradiction between intentionally collecting all publicly available data that also "involves information about people" and "not intentionally collecting personal information", we begin to see clear risks emerge. Information about people is being collected and used in ways we don't yet fully understand with sometimes startling results. Take for example the case of Arve Hjalmar Holmen, a Norwegian, who according to ChatGPT was convicted for murdering his children. (Milmo and editor 2025).

Effects on Education

Turning our focus to the subject of education, I would like to take a moment to address TalTech's Good Academic Practice Guidelines ("Good Practices | Academic Information | TalTech 2020" 2020) which state, as a fact, that AI use in education is a net positive:

> "Artificial intelligence tools help enhance and facilitate learning." ("Good Practices | Academic Information | TalTech 2020" 2020)

I have yet to find research that would clearly support this claim. While there is plenty of material on the potential benefits (for example, personalized learning), the reality we are living with right now is far from the ideal. There's no doubt that AI tools can make schoolwork easier and less time-consuming (for both the student and the teacher), but in what way does that "help enhance and facilitate learning"? How does rewriting AI-generated text help the individual student? Indeed, if all AI tools do is "enhance and facilitate learning" then why is there such a wide range of AI policies among universities with many opting to ban these tools outright? (Fine Licht 2024) And what are we to do with the findings (Lee et al. 2025) that clearly indicate the adverse affects of GenAI on critical thinking, cognitive effort and personal confidence (Lee et al. 2025).

Moving the Goalposts

The page goes on to say:

> "Technological innovations enrich the learning process, and we should embrace them — just as we did with calculators, spell-checkers, the internet, and search engines." ("Good Practices | Academic Information | TalTech 2020" 2020)

This seems to me like a gross misrepresentation of history. As probably most Gen-X'ers remember, there was a time when using calculators and spell-checkers at school was expressly forbidden. The whole point of studying algebra was that you didn't need to use a calculator - the use of one was seen as a failure, not a successful outcome. Understanding syntax and grammar were essential parts of studying language and spelling mistakes were an indicator of inadequate schooling. Fast-forward a few decades and it is practically impossible to test one's spelling ability because the word processors we use for writing make it almost impossible to make mistakes by highlighting and even modifying our written text in real time by default. This is technology in action - it quietly transforms whatever cultural norms and standards that dare to confront it, trading long-held beliefs for efficiency and effort for convenience.

Internet and search engine use seemed to be an exception and I remember teachers encouraging the use of both in the early-to-mid nineties. This discrepancy in attitude can perhaps best be explained by two factors - the novelty of the technology (teachers couldn't yet grasp the implications) and the relevant scarcity of access to information. Before internet access, students only had two sources of information for writing their school papers - the library or handouts from the teacher. I would also argue that the internet as purely an information source (basically a vast library) is a very different kind of technology than calculators or spell-checkers. With the internet, the student still had to do the work of finding the relevant information and transforming it within the context of the assignment where as a calculator simply did the work (of calculating) for them.

I did notice a marked shift in priorities during my undergraduate studies during the early 2000s. Calculators were generally allowed, but the use of advanced graphical calculators was forbidden. This was because the mathematical problems had changed. Moving from algebra to calculus, calculators helped eliminate some of the routine drudgery involved. Automatic spell checking was suddenly not only encouraged, but seemingly promoted, at least in engineering education. For some reason, the ability to write without grammatical errors was no longer seen as having value in itself.

Attitudes towards using the internet had shifted quite significantly. The big concern was of course plagiarism and so we saw the rapid deployment of various automated anti-plagiarism tools in the early 2000s.

I think all of this boils down to one of the most fundamental questions about our education system - what is it actually for? Are we training our students to be effective employees or teaching them to understand the world and their particular field of study? The answer, as it so often is with such questions, is a frustrating "both". But the distinction is significant nonetheless. The former prioritizes efficiency and effectiveness over knowledge and comprehension. There is no job where the use of calculators has ever been banned and in business, the ends always justify the means.

I would like to end this section by analyzing the first part of TalTech's Good Practices document:

> "Technological innovations enrich the learning process" ("Good Practices | Academic Information | TalTech 2020" 2020)

I take particular issue with the use of the word "enrich". How exactly are we enriching the learning process when we convert a course to a series of video recordings? Or when we replace in-class participation with a microphone and camera? Or a discussion with a book? Or the cognitive effort of research and writing with a chatbot prompt?

In fact, I would argue that the exact opposite is much closer to the truth, consider the following statement:

> "Technological innovations devalue the learning process."

When we let software generate our claims and ideas and relegate humans to the role of verifying those claims then are we not precisely devaluing the whole learning experience? Does it not seem like we are entering an era where the software is the user and the human the spellchecker? When we allow and even encourage students to let software do their thinking for them while also forcing professors to use the same tools to weed out unsanctioned uses of the technology - are we not, at the end of the day wasting everyone's time (and untold computing cycles) on a pointless arms race that has nothing to do with "enriching the learning process"?

And what about the (increasingly fewer) students that avoid using GenAI, but are still subjected to the same screening machinery? Who's to say that parts of this very same essay - which was written completely "by hand" (minus the direct quotes) won't get flagged by an AI-powered anti-plagiarism tool? For example, if the student chooses to publish their work online before submitting it for grading. What are this students means of defending themselves from false positives particularly when our obsession with efficiency pushes the institutions assessment processes to even greater levels of automation?

A Work In Progress

The counter-argument put forth by techno-optimists to critiques of transformative technologies such as GenAI is that whatever "bumps" we may experience on our journey to utopia are just transitory. But here we must pause and ask ourselves - if these systems are works in progress - in other words incomplete and unfinished - then why are we allowing them into society? No-one would ever want to live in a building that is still being built or drink from a well that may sometimes be poisonous or drive across a bridge that usually works. Not only are such outcomes viewed as unacceptable, societies are willing and able to allocate significant public resources to maintaining governing bodies that enforce regulations and hold the engineers of our infrastructure accountable. Why are our standards so much lower when it comes to electronic systems?

One possible answer might be historical. Software glitches, while frustrating, were simply not seen as "critical" by the larger population. But those days are well and truly behind us. Software bugs kill people (Johnston and Harris, n.d.), data breaches cause irreparable harm ("When Data Breach Hits a Psychotherapy Clinic: The Vastaamo Case - Hadi Ghanbari, Kari Koskinen, 2024" 2024) and system outages can render entire sectors inoperable across the globe within seconds (George 2024).

The second explanation I have seen repeatedly over the 25 years spent working in ICT is that people tend to blame themselves for the failure. "I must have done something wrong" is a conclusion we've probably all heard or uttered at some point in our lives. I find this to be a very unfortunate case of misunderstanding and one with deep cultural roots. For far too long, the end user has been considered the "weakest link" and the cause of all problems. This attitude has served the "systems industry" very well in helping them deflect blame and responsibility. Addressing this culture of victim blaming should be our priority before we can talk about any widespread datafication of public functions and I would offer three simple rules for moving forward:

a) All commercial software should come with a warranty stating it's fitness for purpose.
b) Any software that allows the user to cause an unintended failure should be considered unfit for purpose.
c) Autonomous systems without a warranty should not be allowed to operate freely on public infrastructure.

I believe that with such rules in place, we could build a more sustainable digital society that is also more accountable to it's citizens. Rules such as these would also disincentivise vendors' current behavior of pushing unfinished products onto consumers just to be the first to market.

Seeing Things

There is a type of software failure that is particularly relevant in the context of GenAI. The one that recommends people eat rocks and glue cheese to their pizza ("Google AI Search Tells Users to Glue Pizza and Eat Rocks 2024" 2024). The industry calls them hallucinations, but we should call them what they really are - fabrications and falsifications (Emsley 2023). One might also be tempted to call them programming errors, but research has shown (Maleki, Padmanabhan, and Dutta 2024) that this is not the case and that these "abnormalities" are actually completely "normal" and part of the architecture of Large Language Models. To put this point into IT lingo - it's a feature, not a bug. If we are still in the process of trying to understand how GenAI fails ("AI Mistakes Are Way Weirder Than Human Mistakes IEEE Spectrum" 2025), then maybe it's too early to hang humanity's future on it's promises?

Conclusion

Of all the products that humankind has ever produced, data and information may well be the first true originals in that they don't equal converting something found in nature into something else. This is not to say that information was invented by us - take for example messenger RNA that has been encoding genetic information of lifeforms long before our kind ever existed. The distinction here is that humans are the only species that actively engage in the creation and transfer of data across space and time. In a very real sense, it's the first human "natural resource" with potentially great value. As with all resource discoveries of the past, this one has been followed by a Great Data Rush - a frenzy of entrepreneurial and scientific activity all hoping to benefit greatly from this new "oil". Sadly, we are also witnessing a repeat of our old colonial sins - from the exploitation of foreign lands and communities to extract "our" resources to the exploitation of immigrant workers in the gig economy to the willing surrender of all of our data to our new data overlords - it appears we have learned little from the harsh lessons of history.

The story of technology is generally seen as an overwhelmingly positive one. And for good reason - it has given humanity more than any other single activity. One could even argue, that technology is not just the best of our achievements - it is our *only* achievement and that Homo Sapiens - "the wise man" - would not even exist without technology. I certainly wouldn't be here writing this essay today without the help of just the antibiotics administered in my childhood alone. But when we look closer at the larger ramifications of all our inventions - then can we really call ourselves "wise"? Would a truly "wise man" do all this damage to our only home and all life on it and can there ever really exist any justification for all the destruction?

It would be easy to simply criticize technological progress, but that would also be intellectually lazy and just as myopic as constantly singing it's praises. We often use the word "challenge" when describing our interaction with technology - as in "GenAI is presenting new challenges to our education system". "Challenge" is a challenging word - it can mean anything from "a call to prove or justify something" to "exposure of the immune system to pathogenic organisms or antigens" and "a call to prove or justify something" - all seemingly valid definitions in the context of technology. But what if the most descriptive and constructive one is in fact "a call to someone to participate in a competitive situation"? What if technology is as much a challenger as it is our means to rise to the challenge? Like the person in front of you in a hundred meter dash - there to push you beyond your limits or to snatch the trophy if you fail. At the end of the day, maybe we should worry less about integrating Big Data and AI into our education system and more about making the learning experience more fulfilling and educational.

In the sage words of Henry David Thoreau: "our inventions are but improved means to an unimproved end". Maybe it's high time we begin improving our ends.

References

"AI Mistakes Are Way Weirder Than Human Mistakes IEEE Spectrum." 2025. https://spectrum.ieee.org/ai-mistakes-schneier.

Atkinson, David. 2024. "Putting GenAI on Notice: GenAI Exceptionalism and Contract Law." https://doi.org/10.2139/ssrn.4981332.

Cooper, A. Feder, and James Grimmelmann. 2025. "The Files Are in the Computer: Copyright, Memorization, and Generative AI," no. arXiv:2404.12590 (March). https://doi.org/10.48550/arXiv.2404.12590.

Eaton, Kit. 2025. "A Legal Win for Content Creators Could Change How AI Models Get Built." *Inc.* https://www.inc.com/kit-eaton/a-legal-win-for-content-creators-could-change-how-ai-models-get-built/91148118.

Emsley, Robin. 2023. "ChatGPT: These Are Not Hallucinations – They're Fabrications and Falsifications." *Schizophrenia* 9 (1): 1–2. https://doi.org/10.1038/s41537-023-00379-4.

Fine Licht, Karl de. 2024. "Generative Artificial Intelligence in Higher Education: Why the 'Banning Approach' to Student Use Is Sometimes Morally Justified." *Philosophy & Technology* 37 (3): 113. https://doi.org/10.1007/s13347-024-00799-9.

George, Dr A. Shaji. 2024. "When Trust Fails: Examining Systemic Risk in the Digital Economy from the 2024 CrowdStrike Outage." *Partners Universal Multidisciplinary Research Journal* 1 (22): 134–52. https://doi.org/10.5281/zenodo.12828222.

"Good Practices | Academic Information | TalTech 2020." 2020. https://taltech.ee/en/student/academic-information/good-practices.

"Google AI Search Tells Users to Glue Pizza and Eat Rocks 2024." 2024. https://www.bbc.com/news/articles/cd11gzejgz4o.

"How ChatGPT and Our Foundation Models Are Developed | OpenAI Help Center." 2025. https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed.

Jiang, Harry H., Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. 2023. "AI Art and Its Impact on Artists." In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 363–74. AIES '23. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3600211.3604681.

Johnston, Phillip, and Rozi Harris. n.d. "The Boeing 737 MAX Saga: Lessons for Software Organizations."

Lee, Hao-Ping (Hank), Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. "The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects from a Survey of Knowledge Workers." In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–22. Yokohama Japan: ACM. https://doi.org/10.1145/3706598.3713778.

Maleki, Negar, Balaji Padmanabhan, and Kaushik Dutta. 2024. "AI Hallucinations: A Misnomer Worth Clarifying." In *2024 IEEE Conference on Artificial Intelligence (CAI)*, 133–38.

https://doi.org/10.1109/CAI59869.2024.00033.

Milmo, Dan, and Dan Milmo Global technology editor. 2025. "Norwegian Files Complaint After ChatGPT Falsely Said He Had Murdered His Children." *The Guardian*, March. https://www.theguardian.com/technology/2025/mar/21/norwegian-files-complaint-after-chatgpt-falsely-said-he-had-murdered-his-children.

Rodriguez-Doncel, Victor, and Jaime Delgado. n.d. "Towards an Expression Language for Licensing Content in the Connected Semantic Web."

"When Data Breach Hits a Psychotherapy Clinic: The Vastaamo Case - Hadi Ghanbari, Kari Koskinen, 2024." 2024. https://journals.sagepub.com/doi/full/10.1177/20438869241258235.