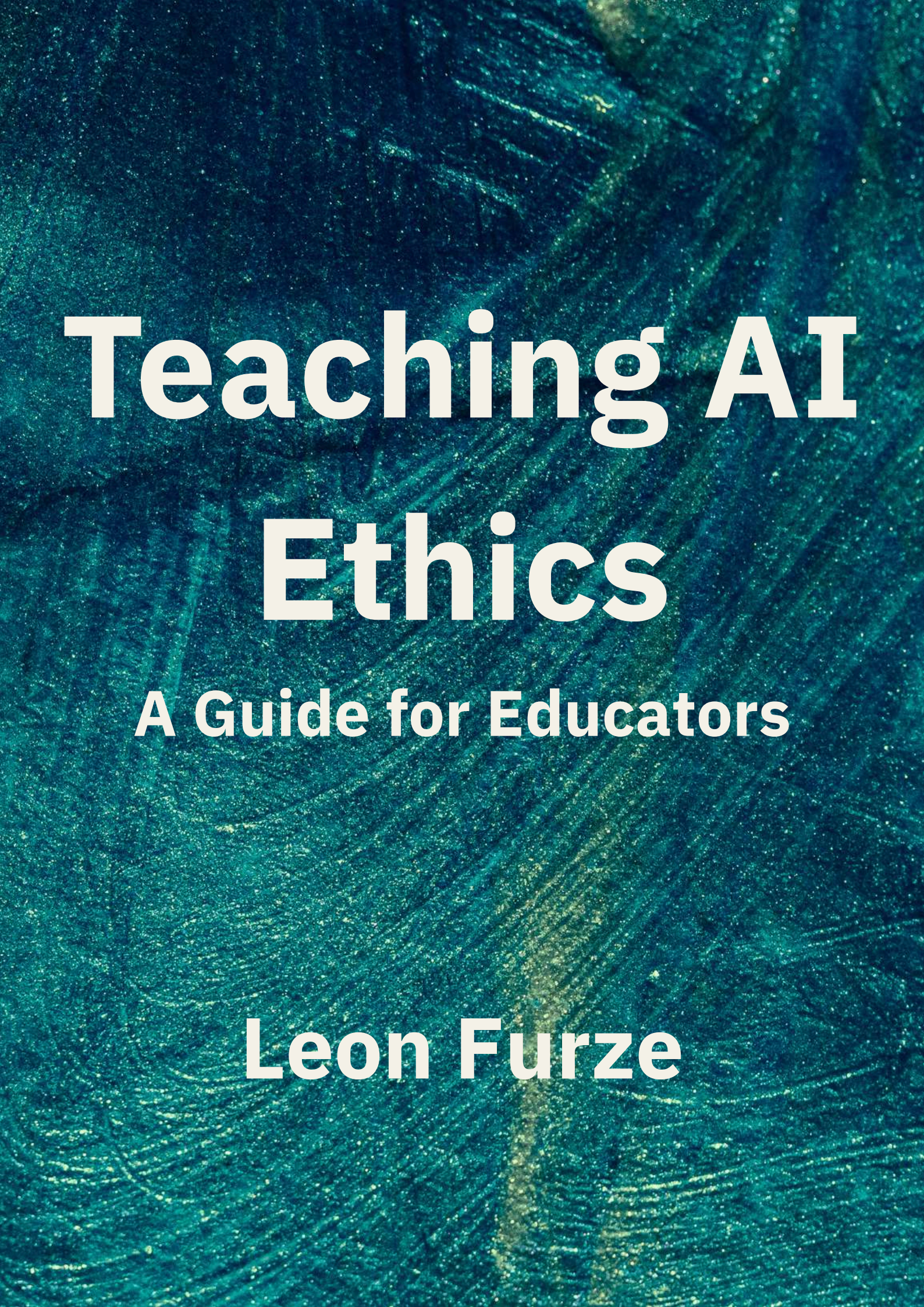

Teaching AI Ethics
A Guide for Educators
Leon Furze

Teaching AI Ethics: *A Guide for Educators*

Copyright © 2026 by Leon Furze

Published by Leon Furze, leonfurze.com

First Edition
ISBN (PDF): 978-1-7645082-0-9

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/4.0/

You are free to: Share — copy and redistribute the material in any medium or format. **Adapt** — remix, transform, and build upon the material.

Under the following terms: Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. **NonCommercial** — You may not use the material for commercial purposes. **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

Every effort has been made to trace all copyright holders, but if any have been inadvertently overlooked, the author will be pleased to make the necessary arrangements at the first opportunity.

INTRODUCTION 6

WHY I'M UPDATING THE SERIES 6
USING THIS BOOK 8

BIAS 9

WHAT IS BIAS IN GENERATIVE ARTIFICIAL INTELLIGENCE? 9
WHAT DOES BIAS IN GENAI LOOK LIKE? 10
MITIGATING BIAS 13
TEACHING BIAS 17

ENVIRONMENT 20

ARTIFICIAL INTELLIGENCE IS AN EXTRACTIVE TECHNOLOGY 20
TRAINING-TIME IMPACT 21
INFERENCE-TIME IMPACT 22
HOW MUCH ENERGY DO DIFFERENT MODELS CONSUME? 23
IMPLICATIONS OF AI ENERGY USE IN EDUCATION 24
TEACHING ENVIRONMENTAL ISSUES 25

TRUTH 27

WHAT DOES TRUTH MEAN TO A LARGE LANGUAGE MODEL? 27
DEEPFAKES AND SYNTHETIC MEDIA 28
POST-PLAGIARISM AND MOVING BEYOND CHEATING 30
TEACHING TRUTH AND AI ACROSS THE CURRICULUM 32

COPYRIGHT 34

THE LEGAL LANDSCAPE HAS SHIFTED… SORT OF 34
WHAT'S CHANGED? 34
THE CREATOR'S DILEMMA 35
BEYOND IMAGES: MUSIC, TEXT, AND CULTURAL SOVEREIGNTY 36
CASE STUDY: GETTY IMAGES V. STABILITY AI 36
TEACHING COPYRIGHT 37

PRIVACY 40

THE ILLUSION OF PRIVACY 40
SHARING ISN'T CARING 40
CROSS-BORDER DATA AND NATIONAL SECURITY 41
BYSTANDER PRIVACY AND AMBIENT AI 42
RETROACTIVE CONSENT AND TRAINING DATA 42
CASE STUDY: ITALY FINES OPENAI FOR GDPR VIOLATIONS 43
TEACHING AI ETHICS: PRIVACY 44

DATA 46

UNDERSTANDING THE SCALE OF DATA COLLECTION 46
CASE STUDY: AI CONVERSATIONS BECOME ADVERTISING GOLD 47
FROM LEARNING TO SURVEILLANCE 48
MOATS, LAKES, AND OCEANS: WHERE DOES ALL THE DATA GO? 48
TEACHING AI ETHICS: DATA 49

EMOTIONS AND SOCIAL CHATBOTS 52

FROM RECOGNITION TO MANIPULATION 52
THE RISE OF SOCIAL CHATBOTS 53
CASE STUDY: CHARACTER.AI AND TEEN MENTAL HEALTH 53
CASE STUDY: REPLIKA AND EMOTIONAL DEPENDENCY 54
CASE STUDY: CHATGPT'S PERSONALITY PROBLEM 55
DARK PATTERNS IN AI DESIGN 56
HOW WIDESPREAD ARE SOCIAL CHATBOTS? 56
WHAT HAS CHANGED SINCE 2023? 57
TEACHING EMOTIONS AND SOCIAL CHATBOTS 58

HUMAN LABOUR 61

THE HIDDEN WORKFORCE 61
CASE STUDY: OPENAI AND SAMA: THE WORKERS WHO MADE CHATGPT "SAFE" 62
CASE STUDY: SCALE AI AND REMOTASKS: THE BILLIONAIRE'S SWEATSHOP 63
CASE STUDY: META, TIKTOK, AND CONTENT MODERATION 64
THE HUMAN COST 65
THE GIG ECONOMY AND AI LABOUR 65
WHAT HAS CHANGED SINCE 2023? 66
FIGHTING BACK: UNIONS AND LEGAL ACTION 66
TEACHING HUMAN LABOUR IN AI 67

POWER 70

UNDERSTANDING POWER AND HEGEMONY 70
THE SCALE OF CONSOLIDATION 71
CASE STUDY: OPENAI, MICROSOFT, AND THE $135 BILLION PARTNERSHIP 71
CASE STUDY: AMAZON, GOOGLE, AND THE BATTLE FOR ANTHROPIC 72
CASE STUDY: STARGATE AND THE INFRASTRUCTURE ARMS RACE 73
WHEN TECH COMPANIES WRITE POLICY 73
WHAT HAS CHANGED SINCE 2023? 75
TEACHING POWER AND AI 75

CONCLUSION 78

ABOUT THE AUTHOR 79

APPENDIX 81

AI ETHICS QUESTIONS BY SUBJECT 81

Introduction

Teaching AI Ethics began as a series of articles I published on my blog in 2023. I'd started my PhD in GenAI in education in 2022 and had been researching the many areas of ethical concern presented by AI and related technologies. On the blog, I published nine articles covering what I saw as the most pressing—and most teachable—areas of AI ethics. Each article offered an overview of the issue, one or two case studies, and teaching ideas across different disciplines. That collection—still available online—became a popular teaching resource across K-12 and higher education institutions.

The original nine areas of concern were:

- Bias
- Environmental concerns
- Academic integrity
- Copyright
- Privacy
- Datafication
- Emotion recognition
- Human labour
- Power

The fact that I had to limit the series to *only* nine areas was an early red flag about how complex these technologies are. But like any technology—particularly digital technologies of the past two decades—artificial intelligence demands more than throwing up our hands and abandoning it as unethical.

Generative AI—including large language models and image generation—is now so widespread that most educators will encounter it in their classrooms, whether they support its use or not.

Some influential educators on social media argue we should stop discussing AI ethics because the technology is already ubiquitous. I believe its ubiquity and apparent inevitability make ethics education more critical—not less—in our classrooms and beyond.

Why I'm Updating the Series

Since ChatGPT's release, major technology companies have engaged in a frantic arms race to develop larger, more powerful AI models and deploy them across every possible platform.

Microsoft's multibillion-dollar investment has embedded Copilot across its 365 suite—Word, Excel, Outlook, PowerPoint—and into new Windows devices and enterprise systems.

Google has followed a similar trajectory with its Gemini model joining the party comparatively late, but already a core feature of Google's Workspace applications such as Docs, Slides and Gmail and baked into the hardware of new Google Pixel phones. New Apple devices ship with

both onboard language model-based AI and Siri is being updated through a partnership first with OpenAI, and then Google.

Meta, which controls Facebook, Instagram, WhatsApp and other applications, has deployed its Llama-based Meta AI to all of its users for free, whether they want it or not.

Yet the ethical concerns have only intensified.

Overt bias in leading models like OpenAI's GPT has been squashed, but not removed. Guard-rail band-aids have been applied over the top of deep-rooted dataset issues, issues which then bubble to the surface whenever new models are released or updates pushed out carelessly.

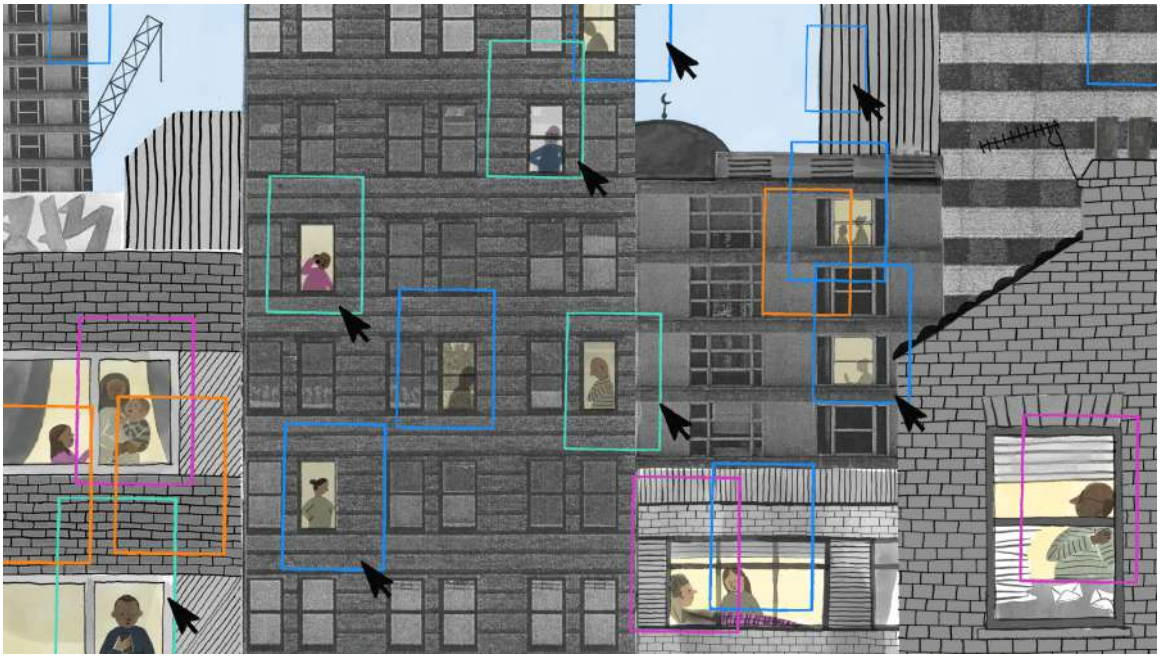


Image source: Emily Rand & LOTI / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/

Privacy, a pressing concern since day one, has been resurfaced repeatedly as companies have harvested the personal data of millions of users, and compiled it into their training data sets.

Copyright has perhaps been the most publicly overt ethical issue with whistleblowers revealing shady data collection practices and investigative journalism from outlets like *The Atlantic* publishing databases of illegally downloaded copyrighted books which have been consumed by these growing AI models.

And of course, just weeks before publishing this article, we saw the explosion of Studio Ghibli-style profile pics and memes across the Internet generated by OpenAI's latest image model, which was clearly trained on individual frames of Miyazaki movies.

In the original 2023 series, I organised the nine areas in three stages: Beginner, Intermediate and Advanced. This was largely due to the availability of resources, research and information on each of those areas. At the time, it was quite easy to read about bias and the environmental costs of machine learning and comparatively difficult to find information on issues like emotion recognition or the human labour costs.

They were also organised that way because of their conceptual difficulty. These are intended to be teaching materials for use with both educators and students.

This time around, I'll be running through the nine areas in the same order, but without the divisions between categories. It's become very clear that all of these issues are as complex as one another, and all of them equal of discussion and study. It doesn't matter which order you approach them in.

It's not all bad news, though. One positive of the explosion of GenAI technologies in the last few years has been more public and academic discourse around these ethical concerns. So unlike the 2023 series, this time around, I'll also be talking about what has changed for the better: companies, researchers and individuals who are pushing back against the systemic biases of generative artificial intelligence; organisations doing serious, meaningful work to make AI models more energy efficient and less environmentally harmful.

There are, in 2026, lots of examples of people doing great work to deal with the ethical concerns posed by these complex technologies.

Using This Book

Every chapter in this eBook is based on the corresponding blog post at https://leonfurze.com/blog. Where it is helpful, I have left in links; however, I also want the eBook version of the articles to feel more "booky" (bookish?), so I have removed most of the HTML and website baggage. After reading a chapter, if you feel the need to go deeper, you should head over to the blog and go down the rabbit hole. The articles have dozens of links, references, and places to follow up on research and news.

Note on the images: In the 2023 version of *Teaching AI Ethics*, I generated images in Midjourney. This time around, I have sourced most of my images from https://betterimagesofai.org/. Better Images of AI is a nonprofit, creative commons licensed repository of images donated by artists, media collectives, and other creators, and an excellent resource. I still use AI image generators, but due to the environmental concerns and the contentious copyright issues discussed in these articles, I am more conscious of my use. Better Images of AI includes an excellent range of photos, illustrations, and digital artworks which have been generously licensed for commercial and non-commercial use.

Bias

Bias is still one of the biggest issues in GenAI output. Although the functionality of the technologies has changed, the core methods of building them—primarily involving scraping large amounts of data from the internet—means that bias and discriminatory content is still baked into each of these technologies, whether text generation, image, video, audio, and so on.

In this chapter, I will discuss where that bias comes from, why it is problematic, and what various organisations and researchers have done since 2023 to attempt to mitigate the problems.

Like the original series, I will finish each article with teaching examples which link the ideas in the article to curriculum areas, mostly from secondary education curricula across the world in a variety of disciplines.

What Is Bias in Generative Artificial Intelligence?

The term "bias" is complex in AI because it carries multiple meanings depending on context. In machine learning and generative AI, it is helpful to distinguish between three overlapping but distinct types which I am going to call **data bias**, **model bias**, and **human bias**. Each of these contributes in different ways to the biased, stereotyped, or sometimes discriminatory outputs we see in generative tools like ChatGPT or image generators like Midjourney.

Data Bias

Data bias is the most widely discussed form of bias in GenAI. It refers to problems baked into training data: imbalances, omissions, and over-representations in the massive datasets scraped from the internet. These datasets are massive, but not neutral. They reflect what's most available online: a disproportionate amount of content written by certain demographics.

Emily Bender, Timnit Gebru, and their co-authors famously warned of "stochastic parrots"—AI systems that repeat and amplify patterns in training data, no matter how biased or harmful. Their warning was clear: when large language models are trained on data drawn mainly from white, male, English-speaking voices (the main demographic of English-language writers online), that becomes the model's default perspective, regardless of who is actually using the tool.

So if a model is trained on web text where CEOs are mostly men, or online forums where racial and gendered stereotypes are common, it will internalise and reproduce those patterns. The model doesn't know these are problematic; it just predicts what it sees in the data.

Model Bias

Model bias is subtler. It's not just what the data says; it's how the model interprets, amplifies, or generalises from that data. Even after applying so-called alignment or safety training, large language models can still produce biased results. One study tested various models carefully trained for safety—including GPT-4o—and found they still made culturally biased associations when prompted indirectly.

This is partly because models don't "understand" fairness or equity. They're incredibly good at picking up patterns, even subtle ones, and replicating them. So if "male" is the most common subject for a CEO in the data, the model might treat "man" as the unspoken default unless explicitly instructed otherwise.

In image generation, this bias is very easy to spot. Even neutral prompts like "a person walking a dog" in Stable Diffusion were more likely to generate images of men unless told otherwise.

Human Bias

And of course there's human bias, which shows up in how the data is created and labelled in the first place. We're all biased, so much so that entire branches of cognitive science are devoted to examining the hundreds of ways we make judgements based on things like availability heuristics, ego, false assumptions, and prior experience.

For image generation, this often means the people manually labelling or categorising images are making their judgements with many kinds of conscious and unconscious assumptions.

Some of the largest image datasets, like ImageNet, relied on crowdsourced or outsourced human labour to tag millions of pictures. That might sound efficient, but the results often reflected individual, cultural, and systemic biases. Fei-Fei Li, the creator of ImageNet, described in her biography how shocked her team was to discover the extent of racial and gender bias in their dataset—including examples of racism, sexism, and other forms of discrimination.

This isn't always malicious. Sometimes it's a reflection of someone's subjective assumptions about what they see when they look at a picture, what they think is "normal," or what they believe fits a label. At other times, deliberately racist labels have been applied to images online. Either way, once it enters the training data, that judgement becomes part of what the AI learns and reproduces.

Finally, human bias comes into play when the developers of an AI model decide which data to include, which to exclude, and what "rules" to encode into the finished product. Developers at companies like OpenAI, Meta, Google, and Amazon must sit down and deliberately make choices regarding which languages to prioritise, which datasets to use, and what values to encode into their models. Each of these points represents a further introduction of bias into the system.

These three types of bias—data, model, and human—layer on top of each other. Skewed data trains a model that learns and sometimes exaggerates the patterns, all grounded in decisions made by people during the creation and curation of the data.

And because developers often push the responsibility for bias mitigation onto the user, it becomes even more important that educators, students, and everyday users understand what bias is, where it comes from, and how it shows up in GenAI tools.

What Does Bias in GenAI Look Like?

GenAI models—whether text, image, audio, video, code, and so on—follow similar methods of training. At a very simple level, predictive algorithms are applied to large datasets in order to infer patterns in the data:

- For language, that means syntax, grammar, vocabulary.
- For images, relationships between pixels, hue, saturation, colour, and how they connect to the labeled images in the dataset.
- For audio, patterns in waveforms and digital representations of sound.

Because applications like ChatGPT are fluent in many human languages, it's easy to fall into the trap of thinking they learn, reason, and think like people. But they are essentially very sophisticated predictive text. Those predictions have to come from somewhere.

Bias in Large Language Models

In the case of large language models, they come from the enormous collection of training data largely scraped from the internet. Though the internet is a vast source of data, it does not represent the whole of society. It is a snapshot of online writing from the late 1990s to the present day—with the vast majority of content being written in the past decade. I explained this in my book *Practical AI Strategies* using the analogy of an AI iceberg: a huge volume of data "below the waterline" with the algorithm on top.

Because of the issue of internet demographics discussed earlier in the section on "data bias"—who is contributing writing online—when you interact with a large language model like ChatGPT or Gemini, it has learned to predict the syntax and rules of language from a biased perspective.

There is nothing inherently wrong with a white male American, but it is not representative of the 100 million-plus users interacting with ChatGPT to create text. At such a large scale, the bias in the data can sway the output of language models. Without sufficient guardrails, the default position of language models can be to produce stereotypical, gendered, and sometimes racist or otherwise discriminatory content.

Image Generation Bias

Bias in generative AI is probably most noticeable in image generation, particularly because some organisations like Midjourney have deliberately steered away from imposing guardrails and rely on the user to prompt for more diverse or less stereotypical images.

Image generation models are constructed differently to large language models, but the premise of training on a large amount of data remains the same. One key difference is that images used for training must be labelled with text describing the content of the image, and humans labelling images introduces more bias into the system.

Another issue comes from over-representation or stereotyping. For example, scraping images online labelled "CEO" will almost always return more results of white men in suits sitting in corporate offices.

So we get a twofold problem in image generation, mostly coming from the way the data is labeled and the over-representation of certain image types in the dataset:

1. Images can be biased toward discriminatory stereotypes applied during labelling.

2. Image generation will tend toward stereotypical outputs for things like gender, race, and job roles or responsibilities.

This is very easy to see in platforms like MidJourney when you generate a photo of a CEO, a nurse, a construction worker, and so on.



Photo of a CEO Midjourney v6.1. Choose one for your next LinkedIn profile pic.



Photo of a nurse Midjourney v6.1. The sexualised images almost certainly come from labelled pornography in the dataset.



Photo of a construction worker Midjourney v6.1. You wouldn't want to be that third guy.

Mitigating Bias

The initial release of ChatGPT (the GPT-3.5 model, and subsequent GPT-4 in March 2023), was not expected to go as well as it did. Those early models produced many examples of bias and discrimination, including instances where ChatGPT would confidently state that a good doctor should be white and male, or reproduce harmful racial stereotypes.

With companies like Microsoft putting their reputations (and money) on the line in support of OpenAI's model, it didn't take long for guardrails and safety features to be added in an attempt to mitigate these problems.

At the output level, the main approaches have involved amending terms and conditions to push responsibility onto the user, and adding guardrails through **system prompts**.

A system prompt (sometimes called prompt instruction) is a set of instructions that sits between the user's input and the model's output. It acts as a kind of filter or intermediary layer that tells the model what it should or shouldn't say.

You can find leaked examples of system prompts from models like Claude and Llama online, and companies like Anthropic sometimes publish them willingly. They're interesting to read because they show how companies like OpenAI shape the behaviour of the model before you even type a word.

Sometimes the instructions are blunt, such as: "Claude won't produce graphic sexual or violent or illegal creative writing content." At other times, they're more abstract, like "Objectivity: Remain objective in your responses and avoid expressing any subjective opinions or beliefs."

One simple way to observe system prompts in action is to compare a prompt like "photo of a CEO" in Midjourney with the same prompt in Microsoft Copilot, which uses OpenAI's image generation, or ChatGPT itself. Midjourney tends to produce stereotypical images like those above, but Copilot and ChatGPT often shows a more diverse result.

The underlying models and training data are likely very similar, so what's changed? The system prompt or output guardrail. But it's important to note that this doesn't fix the bias in the data or the way the model learned. It's more like a band-aid over the broader issue.



Photo of a CEO generated in Microsoft Copilot (still using DALL-E 3)



Photo of a CEO generated in ChatGPT (GPT-4o with updated image generation). The model rewrites the prompt as "A poised East Asian woman in her late 30s or early 40s, exuding confidence in a modern office setting. She wears a tailored navy-blue suit over a crisp white shirt and stands with her arms crossed. Behind her are sleek glass windows revealing a subtle cityscape. Soft, natural lighting highlights her relaxed yet professional demeanor."

As I discussed earlier, any attempt to "hard code" guardrails or values into these models also introduces additional problems. Midjourney, for example, has in the past banned language including anatomical words for body parts.

> If someone types "placenta," "fallopian tubes," "mammary glands," "sperm," "uterine," "urethra," "cervix," "hymen," or "vulva" into Midjourney, the system flags the word as a banned prompt and doesn't let it be used. Sometimes, users who tried one of these prompts are blocked for a limited time for trying to generate banned content. Other words relating to human biology, such as "liver" and "kidney," are allowed.
>
> *MIT Technology Review*

What's changed since 2023?

So other than a few fragile guardrails, what has changed since 2023? Most importantly, people are much more aware of bias in GenAI. At every school or university I visit, most staff (and a decent proportion of students) know to check the output of text and image generators for obvious or subtle bias.

OpenAI acknowledges the problem in its advice for educators. Major news stories have covered problems with the guardrails, for example when Google's Gemini produced "racially diverse Nazis". And more and more research is being undertaken to explore the political, social, and cultural biases encoded into different AI models.

Some of this work aims to provide better "prompt-time" methods of ensuring fairness, such as the researchers of *Fair Diffusion* who use a more sophisticated approach to guardrails than simple system prompts.

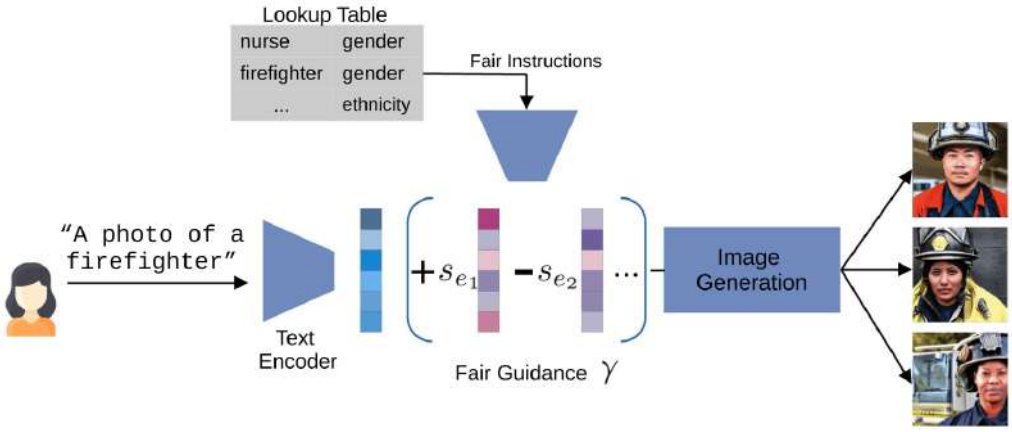




Fig. 1 Stable diffusion (top row) runs the risk of lacking diversity in its output (here, e.g., only White male-appearing persons as "firefighters"). In contrast, FAIR DIFFUSION (bottom row) allows one to introduce fairness—increasing outcome impartiality—according to a user's preferences (e.g., group identities of "firefighters")

Friedrich, F., Brack, M., Struppek, L. et al. Auditing and instructing text-to-image generation models on fairness. AI Ethics (2024). https://doi.org/10.1007/s43681-024-00531-5

Other studies aim to dispel myths like "AI just reflects society" by proving that the bias in GenAI models amplifies and worsens societal biases.

Methods of guardrailing and applying safety features will continue to improve. However, as of writing this updated post in 2025, I am not aware of any successful research demonstrating an ability to mitigate the "upstream" bias encoded through the training data. Until we find ways to source better data and build effective models with these smaller, more diverse datasets, LLMs, image generators, and other forms of GenAI will always be prone to bias.

Teaching Bias

In the original 2023 collection, each article ended with a selection of ideas for teaching the issue in the context of existing curriculum areas. Each chapter of this book will similarly align concepts from the articles to standards from typical curricula across the world, and the Australian, UK, US, and IB curricula. For the readers teaching in Higher Education, these examples will also be suitable across a wide range of disciplines. In the Appendix you will also find a list of key questions alphabetised by discipline.

My key point is that we do not need specialised "AI literacy" classes to deliver quality instruction on AI ethics—we already have the expertise we need in schools and universities.

English

In the English curriculum, students are required to understand and identify bias in texts, identify point of view, and reflect on the validity of sources. This provides an excellent opportunity to discuss questions like **"Is AI an 'author'?"**, **"Where does an AI model get its information from?"**, or **"Is AI output biased?"** These would be perfect for an analytical or persuasive writing unit.

Health & Physical Education

In Health & Physical Education, students examine how stereotypes influence wellbeing, relationships, and inclusion. Topics include gender roles, cultural diversity, and discrimination. This provides a valuable opening to explore how **AI systems may reinforce gender or**

cultural stereotypes, for example, in health recommendations or sports analytics. Students might ask **"Should AI systems be taught about cultural sensitivity?"** or **"How can we design health technologies that reduce bias rather than amplify it?"**

Civics and Citizenship

The Civics and Citizenship curriculum focuses on fairness, equality, and social cohesion, including media literacy and the influence of media on public opinion. These outcomes align well with discussions such as **"Can AI-generated media influence democracy?"**, **"Who is responsible for biased political information created by AI?"**, or **"How should democratic societies regulate biased technologies?"** These are great questions for debates or policy simulation activities.

Mathematics (Statistics)

In senior mathematics, students analyse statistical claims and evaluate data sources for bias. This directly supports inquiry into **algorithmic bias**, **sampling bias**, and **data representation in AI training sets**. Questions like **"How can biased data lead to unfair algorithms?"** or **"Can math help detect or prevent bias in AI?"** turn abstract statistical ideas into real-world ethics and technology projects.

History

History students evaluate sources for reliability and author perspective, which opens up questions into how **AI reconstructs history through training data**. Students might explore questions like **"Can an AI write accurate history?"**, **"What happens when AI is trained on biased historical sources?"**, or **"Who decides which voices are remembered in training datasets?"** This could lead to digital historiography projects and critiques.

Computer Science / Digital Technologies

In computer science courses, students analyse the design of computational artifacts for bias and consider the impact of inequitable systems. This makes it the perfect place to address questions like, **"How does training data shape AI bias?"**, **"How do we build fair algorithms?"**, or **"Can AI make ethical decisions?"** These could form the basis of practice coding projects (including using AI coding tools), algorithm audits, or ethics case studies.

Social Studies

Social Studies emphasises evaluating source credibility and recognising bias in civic discourse. Like English and History, this aligns with evaluating **AI-generated news**, **political messaging**, or **deepfakes**. Students could explore questions like **"How can we tell if an AI-made source is trustworthy?"** or **"What impact does AI have on our understanding of truth in civic life?"**

Theory of Knowledge (IB)

TOK students examine personal and ideological bias across different ways of knowing. AI fits naturally into this lens—prompting inquiries such as **"Can AI 'know'?"**, **"Can machines be objective?"**, or **"How do biases in programming shape our perception of truth?"** These

are ideal for extended essays or oral presentations exploring epistemological questions about machine knowledge.

Visual Arts

In Visual Arts, students explore how AI-generated images can reflect visual bias and stereotypes, such as reinforcing narrow beauty standards or cultural clichés. This opens discussion on using prompts to steer AI toward more inclusive outputs, such as **"Should we prompt AI to represent diversity?"** or **"Why do some stereotypes persist despite prompt guidance?"** It also highlights a key limitation: **prompting can't fix biased training data**, encouraging students to think critically about ethics, representation, and authorship in digital art.

Environment

The environmental impact of artificial intelligence is among the most important—and most difficult—ethical issues to discuss. AI applications, like all digital technologies, significantly affect the environment—but industry opacity makes it difficult to quantify the damage.

In this chapter, I will look at research that has taken place between the original 2023 collection and 2025 to try to shed some light on what we really know about the training costs and use of AI. I will also point out some of the misinformation and inaccuracies which frequently pop up when I run conferences and presentations on artificial intelligence, because since 2023 there have also been attention-grabbing headlines in the media which misrepresent the impact of AI.

In education, we need to have reasoned, rational discussions with students about the future of artificial intelligence technologies.

Artificial Intelligence is an extractive technology

Kate Crawford's *Atlas of AI* remains one of the most important texts on this subject. In *Atlas*, Crawford draws analogies between artificial intelligence and the mining industry, focusing on both the environmental and the socio-political impacts. One important aspect of the book is that Crawford is talking much more broadly about artificial intelligence than the current popular use of the term, which often refers to generative AI and sometimes discrete applications like ChatGPT.

When Crawford writes about the environmental impact of AI, she is talking about:

1. Recommender systems, such as those powering Netflix and Spotify
2. Predictive algorithms like those which determine what pops up on your Facebook feed or YouTube
3. Broader implications of data classification—such as labelling images—and analytics

In this article on environmental issues, I will talk a lot about GenAI, like text and image generation, but ChatGPT, Copilot, Gemini, Midjourney and so on, are only part of the bigger AI environment puzzle, and currently not the biggest piece.

In drawing analogies to the mining industry, Crawford points out some key ideas:

- **The infrastructure behind the cloud is made up of data centres, which consume phenomenal amounts of electricity, and which rely on water to cool the systems.** If you've ever tried to do something computationally taxing on a laptop, you'll know how hard a computer fan has to work to keep things cool. Imagine that scaled up by orders of magnitude, and you can picture how much waste heat is generated by data centres. Companies like Microsoft have built data centres in cold climate countries and even attempted to build them underwater in efforts to reduce the impact of cooling these systems. But just like the mining

industry, the use of data centres to compute requirements for artificial intelligence is incredibly resource heavy.

- **The hardware used to run artificial intelligence depends on extractive mining processes**. Lithium, silicates, rare minerals and other elements are used to build the hardware that runs artificial intelligence, whether it is graphics processing units used to train generative AI models or the iPhone that you use to converse with ChatGPT. The physical cost of mining these ingredients is incredibly high, as is the e-waste produced when these technologies fall out of their short life cycles.
- **Where mining extracts oil, artificial intelligence extracts data.** Although this is not tied to the environmental conversation, I will refer to this when I update the articles on datafication and privacy. Every industry produces some sort of product. One product of the artificial intelligence industry is data. The raw material for that data is our personal information and, linking back to the environmental concerns, it is the processing of this raw material which consumes the most energy.

Taking Kate Crawford's position of artificial intelligence as an extractive technology akin to the mining industry as a starting point, let's look at the two main areas where generative artificial intelligence has an impact on the environment.

Training-time impact

One of the ways in which OpenAI's GPT-3.5, the model publicly released as ChatGPT, changed the trajectory of generative AI research was that they threw more data and more computational power at their system than any researchers had before. They effectively demonstrated that scaling up the volume of raw material and simultaneously increasing the amount of computation used to learn the rules of that training material could significantly improve the output of the models. Suddenly, a generative pre-trained transformer (GPT) went from an adequate way to process natural language and replicate simple parts of human speech to an incredibly versatile technology fluent in dozens of languages and much more capable as a chatbot. Similar ideas of scale have been applied to image generation.

But just as burning more coal produces not only more electricity, but also more pollution, burning through more raw materials of AI training data produces more adverse effects.

In 2024, an article was published by MIT estimating how much energy would be required to train a GPT-4 class model, OpenAI's March 2023 release. Despite how people have interpreted the news, the paper itself did **not** give a hard number for training a *GPT-4–class* model. Instead, it:

- **Points to the scale of the problem in qualitative terms.** The article notes that models "such as OpenAI's GPT-4" need "a staggering amount of electricity," stressing that their training clusters can draw **7–8 times more power than a typical high-performance computing workload.** These are estimates based on earlier research.
- **Provides the last solid public benchmark—GPT-3—as a yardstick. Citing the 2021 Google/Berkeley study, it reminds readers that training GPT-3 burned 1,287 MWh**

of electricity and emitted about 552 t CO_2e (roughly the annual electricity use of 120 average U.S. homes).

Training large language models, particularly the huge multi-modal state of the art models like GPT, uses a great deal of energy. Training time impacts also include the aforementioned water use. Again, estimates emerged in 2024 as to roughly how much water would be required, based on comparisons to open access research and other technologies, and these topics make for great news articles.

One problem with the media headlines has been that people take them at face value when they are not necessarily representative of the true impact of these technologies, or when they misinterpret the research. Most of these estimates are assumptions based on completely different models, for example, OpenAI's GPT-2 or GPT-3, which had much more public information before "Open" AI closed its doors. These estimates could either be lower or higher than the reality: Lower because new AI models might well train much more intensively and on more data than we are aware; higher because new efficiencies in the algorithms and processes may actually *reduce* some energy requirements of training new models.

The point is that we can't know for sure. So when somebody says to me, "Have you heard that ChatGPT consumes 500ml of water per prompt?" my first response is to challenge that specific number, not because I want to let OpenAI off the hook, but because I think supposition is not helpful in this conversation.

The important point is that, cumulatively, data centres consume 3-4% of all US energy, and are using so much water that they are quite literally sucking entire regions dry.

Inference-time impact

This is a much more recent conversation. Inference time means "the point at which the user gets the AI to do the thing": the point at which the user inputs a prompt to ChatGPT and the chatbot responds; the point at which the user generates an image in a platform like Firefly or Midjourney; or the point at which a company deploys OpenAI's GPT model to conduct sentiment analysis on customer satisfaction data.

Recently, there has been an uptick in interest regarding how much energy these systems use when we use them. Again, there is a complete lack of transparency and a lot of confusing press around these issues. To try to cut through some of that noise, I am going to refer only to one article, a piece of research titled *Power Hungry Processing: ⚡Watts ⚡Driving the Cost of AI Deployment?* by Sasha Luccioni, Yacine Jernite, and Emma Strubell.

I'm using this article because it is one of the only I have read thus far which uses sound methodology and publicly available information to produce information on how much different types of artificial intelligence models consume. However, a collection of articles from MIT has also recently been released which covers a lot of areas including comparisons of different models and the impacts of data centres. It's well worth reading in full at:

https://www.technologyreview.com/supertopic/ai-energy-package

Before I get into the details of the Luccioni article, some context:

In this 2024 article, Luccioni et al. refer to a variety of AI systems. Like Crawford's *Atlas of AI*, they are not just talking about generative artificial intelligence. They draw comparisons between:

- Text classification
- Image labelling
- Question and Answer (QA) chatbots
- Text summarisation
- Masked language modelling
- Object detection
- Text generation
- Image generation
- Task-specific models
- Multi-purpose models

That sounds like a lot, but it's an important piece of this conversation around the environmental impact of generative AI. Many of these technologies, such as text classification, image classification and QA, are well established areas of machine learning, which have been used for a decade or more in many industries. "Masked language modelling" is often used in sentiment analysis: looking at large text-based datasets to determine whether the language is positive, negative, or neutral. These are "old fashioned" AI systems, some of which are being replaced in some situations by multi-purpose large language models like GPT.

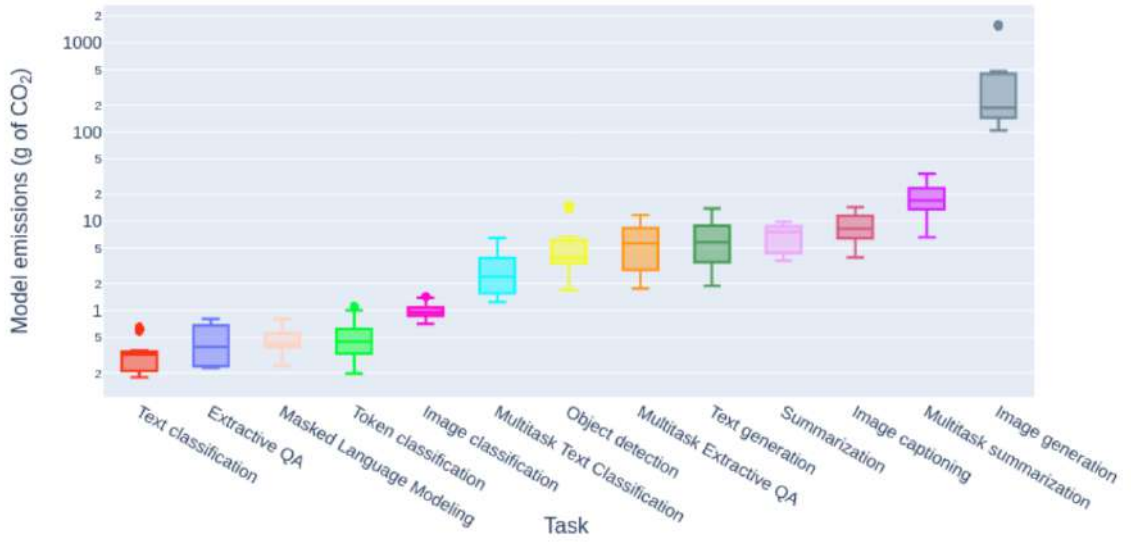
There are two important implications of this for this research:

1. **The researchers used models which are not directly comparable to state-of-the-art, proprietary models.** For example, to test text generation, the researchers used GPT-2, an earlier model from OpenAI. They also used models like BLOOM and some of Facebook's early LLMs. These text generation models are significantly smaller and less complex than GPT-4, Llama, Google Gemini and so on. You cannot draw a like-for-like comparison of the energy consumption of GPT-2 to the energy consumption of GPT-4.

2. **It is reasonable to compare and contrast energy consumption across task types and model classes.** This is the real point of the article—to demonstrate that per 1000 inferences, the tasks and model types vary wildly. Therefore the point is not "X number of generations *only* consume X amount of energy," but rather "X generations consume *significantly more energy* than task-specific models when used for the same job."

How much energy do different models consume?

Here's a quick look at how much energy these models consume without getting too bogged down in the technical details. The graph below, taken from Luccioni et al.'s paper, shows the model emissions for different task types measured in grams of CO2.



For me, this paper represents a much more important conversation than "ChatGPT consumes X millilitres of water" or "every prompt uses this much electricity." The authors are not trying to say that machine learning or generative AI research should stop, or that individual users should be held accountable for how many times they use ChatGPT in a day. They are pointing out that generative artificial intelligence, especially multi-purpose generative AI that includes image generation, are incredibly resource intensive compared to other technologies which may be better suited to the job.

This is so important that I'm going to repeat it in big bold letters and then explain what that means for education.

Generative artificial intelligence systems are incredibly resource intensive compared to other technologies which may be better suited to the job.

Implications of AI energy use in education

I have given presentations where I point out that an individual's use of ChatGPT may consume less energy overall than, for example:

- Streaming Netflix for one hour
- A Zoom call with 10 people
- Driving a car for a few miles

I believe that this is true based on legitimate research, including a synthesis of research from educator and author Jon Ippolito. Ippolito arrived at slightly different numbers than Luccioni using existing studies, but the numbers are comparable.

However, I think that this "individual use" narrative obscures the bigger problem, which is that a lack of transparency and regulation regarding artificial intelligence means we have *no idea* what energy consumption looks like at scale in education.

We have a responsibility to raise these concerns with students and to carefully consider how we deploy these technologies across education systems. Rather than placing the burden on

individual teachers, we need to ask whether deploying large language models across an entire education sector—such as Estonia has recently agreed to—is the best use of resources. Many of the tasks we use these models for could be achieved with less energy-intensive technologies, or even without technology at all.

Do we need better image generation models for lesson resources, or more reliable sources of open access images? Do we need to use ChatGPT to analyse parent surveys, or is there an older, more efficient technology that would do the same? Should I use Gemini to write a lesson plan, or sit down with a colleague over a coffee and do it myself?

Teaching Environmental Issues

English

In the English curriculum, students critically evaluate texts, perspectives, and themes, making it an ideal space to explore environmental narratives and "greenwashing" in media. Students can analyse persuasive or expository texts on digital sustainability, asking **"Is this tech company truly 'green'?"**, **"How do tech advertisements use environmental language to mislead?"**, or **"What narratives are missing in the conversation about e-waste?"** These lend themselves well to persuasive writing, media analysis, or debates on digital environmental responsibility.

Science (Environmental Science / Earth & Space Science)

Science courses frequently focus on sustainability, ecosystems, and the impact of human activity on the planet. This aligns naturally with questions such as **"What is the carbon footprint of cloud computing?"**, **"How do server farms affect local ecosystems?"**, or **"Can digital technology help monitor and mitigate climate change?"** Students can conduct research projects comparing environmental costs and benefits of digital solutions or use data sets to model pollution caused by e-waste.

Digital Technologies / Computer Science

Digital Technologies curricula (including the Australian, UK, and US) explicitly include sustainability and environmental impact. Students might examine **"How energy-efficient is the code we write?"**, **"What happens to old hardware when we upgrade?"**, or **"Can we design low-impact apps or systems?"** Projects could include building sustainable tech prototypes, auditing energy use in computing, or exploring circular design principles in software and hardware development.

Geography

Geography investigates human-environment interactions and is a great context for studying digital technologies' ecological footprints. Students could ask, **"Where do raw materials for smartphones come from?"**, **"How does digital infrastructure affect urban and rural land use?"**, or **"What role do satellites and GIS play in environmental monitoring?"** Case studies on mining for rare earth elements, digital deserts, or tech-fuelled deforestation would deepen geographical inquiry skills.

Design and Technologies

This subject encourages students to design solutions with awareness of social and environmental sustainability. Key teaching questions include **"How can we reduce the lifecycle impact of tech products?"**, **"What is eco-design in the context of digital devices?"**, or **"How can we apply the principles of 'cradle-to-cradle' or closed-loop design to electronics?"** Students might engage in sustainable redesign challenges or audit the energy use of different design tools.

Civics and Citizenship

As students examine democratic responsibility, rights, and participation, they can explore digital environmental justice**: "Who bears the environmental burden of digital consumption?", "Should governments regulate e-waste exports?", or "What policies support equitable access to green technologies?" These topics are excellent for role-play debates, policy pitches, or mock UN climate tech summits.**

Mathematics (Statistics & Data)

In mathematics, students interpret data and trends, opening up questions like **"How much CO2 is generated by a Google search?"**, **"What do energy use graphs of tech companies reveal?"**, or **"How can we model the global growth of e-waste?"** Students could analyse real-world datasets on energy consumption, digital product lifecycles, or climate projections influenced by digital tools.

Visual Arts / Media Arts

Visual Arts and Media Arts students often explore themes of communication, critique, and message. Prompts might include **"How can digital art raise awareness of e-waste or tech pollution?"**, **"What does digital decay or"digital plastic" look like?", or "Can we create installations from recycled AI output?"** Students could design visual campaigns, create artworks from obsolete tech, or critically assess how digital art platforms contribute to or challenge environmental issues.

Theory of Knowledge (IB)

In TOK, students explore how knowledge is constructed and evaluated, making it fertile ground for questions like **"How does digital surveillance impact the environment and our understanding of ethical responsibility?"**, or **"Do we have a moral obligation to consider the environmental cost of digital knowledge systems?"** These are excellent for essays, presentations, and cross-disciplinary inquiry.

Truth

In the original article on *Truth* I split my attention between misinformation and academic integrity as two types of "truth" that we needed to attend to when teaching AI ethics. Since 2023, "cheating" with AI has continued to be an issue, but despite the fact that the media seems to have only just caught onto it, most educators I speak to are already tired of the discussion and ready to move on to actions.

In this chapter, I'm going to dedicate much more time to misinformation, deepfakes, and the inherent issues of using LLMs to generate truth. I will still discuss academic misconduct, but the conversation will shift towards ideas of postplagiarism and how we might move beyond the "cheating" conversation in schools and universities.

What does truth mean to a large language model?

Large language models are, by design, incapable of producing truth. We've seen this in action since the release of ChatGPT, which initially had no internet connection, no way to upload or attach files, and a tendency to "hallucinate".

The term "hallucination" describes what happens when language models fabricate information. But this isn't an error or quirk—it's a feature of how these systems are built. Large language models rely on statistical matching and probability. The training data is only used to create the rules governing the languages the model can produce. There is no epistemic grounding, and no "truth" left in the system once trained.

Since 2023, most models have relied on a combination of internet access, file upload, retrieval augmented generation (RAG), and reinforcement learning from human feedback to mitigate rates of hallucinations, but none of these methods have fully guarded against the large language model producing factually incorrect information. Hallucinations are a feature, not a bug.

Chatbots "make sense, not words", and that down-to-earth definition is reinforced by scholars like Emily M. Bender, who recently referred to LLM-based systems as "designed to mimic the way people use language, first through pre-training on next-word prediction and then through additional rounds of redistribution of probability mass, called RLHF."

Even more problematically, although AI can now cite sources from the internet or use processes like RAG to access internal documents, this only increases the *perceived* credibility, making it harder to spot misinformation. The automation bias, a tendency of humans to believe machine-driven responses, means that people tend to believe what computers tell them.

Unfortunately, discourse on social media often compounds the problem, with AI influencers suggesting that hallucination rates are dropping when, out in the wild, the opposite seems to be true.

In 2025, for example, the US government's "Make America Healthy Again" clearly contained content generated by OpenAI's ChatGPT. The original release of the report carried telltale markup of ChatGPT-generated citations and many of the citations were inaccurate, misleading,

or pointed to the wrong papers. The issue was reported on in The Washington Post, just prior to the paper being taken down and amended.

> "The White House's"Make America Healthy Again" report, which issued a dire warning about the forces responsible for Americans' declining life expectancy, bears hallmarks of the use of artificial intelligence in its citations. That appears to have garbled citations and invented studies that underpin the report's conclusions…"
>
> https://www.washingtonpost.com/health/2025/05/30/maha-report-ai-white-house/

Even medicine—a field where truth and accuracy are surely paramount—there have been issues with AI. For example, a highly successful paper published in Nature detailed ways that language model-based technologies can be used to predict the function of unknown enzymes. However, as data scientist Rachel Thomas pointed out on her website, "The Transformer model in the Nature paper made hundreds of"novel" predictions that are almost certainly erroneous." Thomas goes on to describe that a counter-paper pointing out the inaccuracies, published on the preprint platform bioRxiv, received much less attention than the 22,000 views-and-counting of the Nature article.

Whatever the AI pundits tell you, it's not about whether more advanced LLMs or reasoning models like ChatGPT o3 Anthropic's Claude Opus 4 reduce hallucinations, it is the way in which these technologies are used by people in high-stakes situations like law, politics, medicine, and education which is important.

Deepfakes and synthetic media

In 2023, I also started writing about what I termed "digital plastic"—the creation of synthetic multimodal texts that I cautioned would swamp the internet over the next few years. At the time, image, audio, and video generation were still fairly nascent and not anywhere near as convincing as text generation. But in the two years since the original Teaching AI Ethics series, that has changed. The truthfulness of audiovisual content online can no longer be taken for granted (not that it ever could, really).

The creation of incredibly convincing AI-generated images is now trivial and possible through ubiquitous open-source technologies with very few restrictions or guardrails. Regulations and laws have been slow to keep up, though in Australia, at least, there have been changes regarding the creation of non-consensual, explicit deepfakes.

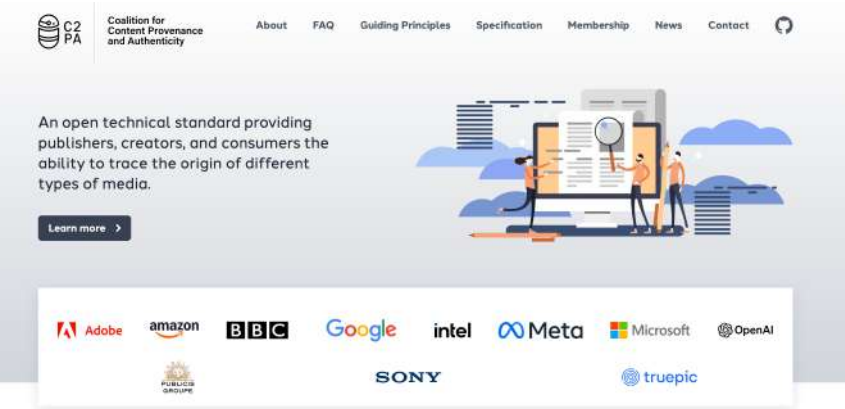
Deepfakes are particularly troubling when discussing artificial intelligence and truth, since it is now trivial to create convincing and entirely untrue videos of people saying and doing things they never did or said. 99% of deepfakes circulated online are nonconsensual and explicit, and 98% of those images are of women: this is gender based abuse. These deepfakes are generally images but increasingly include video. While in 2023 this was already becoming problematic for celebrities, advances in these technologies mean that everyone is potentially vulnerable.

Deepfakes can also be used in political misinformation and deliberate reputational damage. We have already seen the deliberate sharing of deepfakes by politicians and world leaders on

platforms like X, where moderation has all but disappeared. AI-generated websites, social media accounts, and bots are ubiquitous on most major platforms, generating and spreading misinformation and disinformation. Some of these are naive hallucinations, others deliberately antagonistic and harmful.

Because of the rise of these technologies, major developers have made efforts to secure their AI-generated content with watermarking and provenance standards like C2PA. These approaches identify synthetic media, often applying a content credential or metadata which articulates what platform was used to create the image, video, or audio.

While this legitimises content created on those platforms, allowing users to transparently label their content as AI-generated, it will not stop others from generating harmful deepfakes and synthetic media because they can simply use open-source equivalents which lack those credentials. Authenticity signals are part of the equation, but strong regulation is needed with heavy penalties for people creating and sharing harmful and untruthful content.



The C2PA.org website lists companies who have signed up.

In education, one thing we can do right now is to show educators and students just how difficult it already is to spot real from fake. In 2024, I released a little game which has now been played by over 100,000 people, including hundreds of educators. In my face-to-face and online sessions, the average score is 5 or 6 out of 10. It is already harder than you think to spot deep fakes.

You can try for yourself by playing: Real or Fake? The AI Deepfake Game

As well as creating deepfakes and untruthful content, artificial intelligence contributes to the sharing and distribution of that content. Social media has played a huge role in elections for over a decade now. Many will remember the Cambridge Analytica scandals of Facebook, where collection of user data was implicated in manipulating the voting preferences of voters in the run-up to the 2016 US elections. This kind of disinformation can also be seen in more innocuous places such as the comment threads on YouTube videos, or the insinuation of AI chatbots into Facebook groups.

Artificial intelligence can be deployed to deliberately boost or suppress posts, including those which have been generated by artificial intelligence in the first place, to sway public, political,

and social discourse in one way or another, or by advertising companies to sway public opinion on products and services.

Post-plagiarism and moving beyond cheating

All of this contributes to my thoughts on why we need to move discussions in education beyond cheating or catching students using artificial intelligence. Some discussions of academic integrity, for example, exclude artificial intelligence entirely from educational contexts, which makes it very difficult for students and educators to understand the full implications of these technologies in society.

Focusing on whether students using AI technologies are "cheating tools" also obfuscates a more important question of whether we can help students to learn to use artificial intelligence in ways which *are* truthful, accurate, and reliable.

As I demonstrated above, it is not enough anymore to teach students to try to detect AI-generated fake news: it is already impossible to do so with 100% certainty. Similarly, there is no totally viable way to guarantee that anything produced by a student in the context of digital texts is their own work.

But if detection tools don't work, and if it's not possible for an educator to spot the use of generative AI by eye, what does that mean for educating students in a digital context? Students deserve to conduct their work in an environment which reflects the world beyond the classroom. There have to be options for students who require online education. And teachers deserve to be able to teach—including online—without constantly looking over their shoulders or chasing academic integrity concerns.

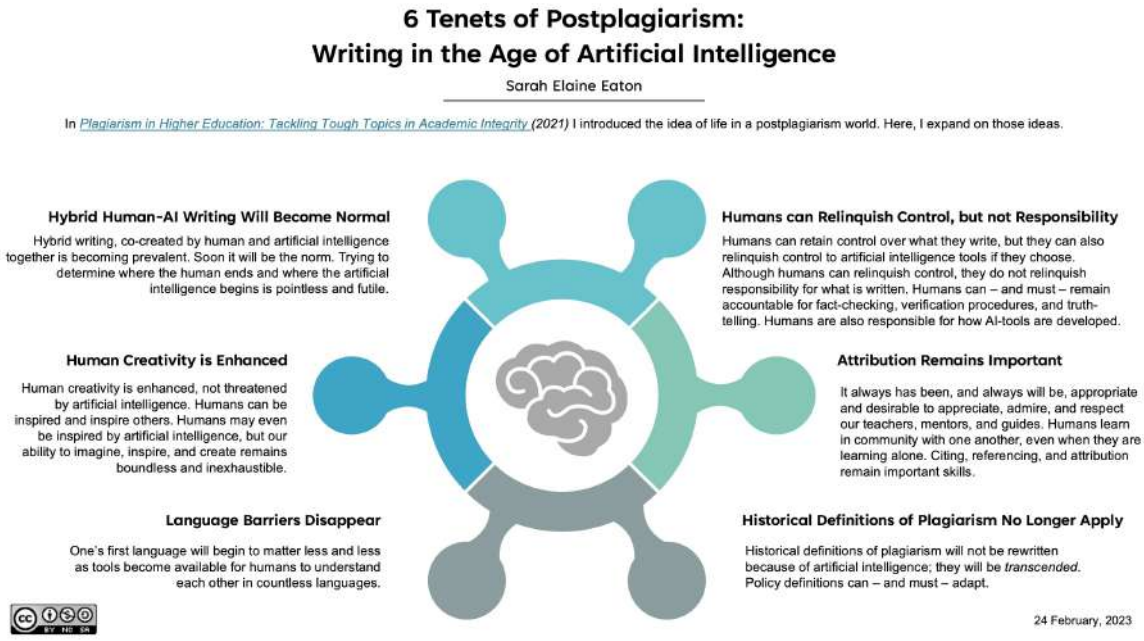
As I'm writing this article, I'm of the opinion that the most secure form of assessment is what we would call "Level one", or what others, might call "Lane one" assessments: **supervised, in-person, technology-free assessments of student learning, which may or may not be conducted under examination conditions, but which, through both technological and physical means, limit access to artificial intelligence.**

In the future, though, even these methods will be increasingly vulnerable. A student who really, *really* wants to cheat will be able to do so in increasingly sophisticated ways which involve, for example, wearable AI technologies. Of course, the students who have access to those kinds of technologies will tend to be wealthier, and more (digitally) literate.

The point that I am making is this: we cannot solely focus our energy and attention on assuring "truth". We need philosophies in education which encourage and reward truthful behaviour.

Sarah Elaine Eaton's post-plagiarism is one such philosophy which I think has important implications for truth in education.

The Post-Plagiarism framework has six tenets:



Dr Sarah Elaine Eaton's 6 Tenets of Postplagiarism

Hybrid Human-AI Writing Will Become Normal

Hybrid writing, co-created by human and artificial intelligence together is becoming prevalent. Soon it will be the norm. Trying to determine where the human ends and where the artificial intelligence begins is pointless and futile.

Human Creativity is Enhanced

Human creativity is enhanced, not threatened by artificial intelligence. Humans can be inspired and inspire others. Humans may even be inspired by artificial intelligence, but our ability to imagine, inspire, and create remains boundless and inexhaustible.

Language Barriers Disappear

One's first language will begin to matter less and less as tools become available for humans to understand each other in countless languages.

Humans can Relinquish Control, but not Responsibility

Humans can retain control over what they write, but they can also relinquish control to artificial intelligence tools if they choose. Although humans can relinquish control, they do not relinquish responsibility for what is written. Humans can—and must—remain accountable for fact-checking, verification procedures, and truth-telling. Humans are also responsible for how AI-tools are developed.

Attribution Remains Important

It always has been, and always will be, appropriate and desirable to appreciate, admire, and respect our teachers, mentors, and guides. Humans learn in community with one another, even when they are learning alone. Citing, referencing, and attribution remain important skills.

Historical Definitions of Plagiarism No Longer Apply

Historical definitions of plagiarism will not be rewritten because of artificial intelligence; they will be transcended. Policy definitions can—and must—adapt.

6 Tenets of Postplagiarism: Writing in the Age of Artificial Intelligence—Dr Sarah Elaine Eaton

From my experience in both K-12 and higher education contexts, there's a lot to be said in favour of the six tenets of post-plagiarism, particularly its focus on transparency and shared accountability. There will always be students who, for whatever reason, break our trust. But that is by no means *all* students, and we cannot afford to let this technology, inherently untrustworthy as it is, erode mutual respect between educators and learners.

To get to the truth of whether a student understands what they have learned will be a difficult task, but a worthwhile one. It will include a blend of assessments which permit and exclude artificial intelligence. It cannot do so at the expense of accessibility and inclusion, or these assessments themselves would be invalid. This is the work of education.

Teaching Truth and AI Across the Curriculum

English

In the English curriculum, students analyse how language constructs meaning, truth, and credibility. With AI tools increasingly producing persuasive texts and opinion pieces, this is an ideal space to explore how generative tools shape narratives. Students might ask **"Can AI-generated arguments be trusted?"**, **"How does AI mimic authorial voice to create a sense of truth?"**, or **"What rhetorical strategies make false claims produced by AI seem convincing?"** These activities connect to media analysis, persuasive writing, and critical discourse study.

Humanities (History / Civics / Geography)

Humanities subjects centre on source reliability, perspective, and representation—areas increasingly complicated by synthetic media. Students might investigate **"What happens when AI rewrites history?"**, **"How can deep fakes influence public trust in political institutions?"**, or **"How do algorithms impact the information we see—and don't see—about global events?"** These lessons align with source evaluation, historiography, digital civics, and ethical information use.

Science (General Science / Psychology / Environmental Science)

Science subjects teach students how to evaluate evidence and peer-reviewed claims—skills essential in an era of AI-generated misinformation. Students might ask **"How can we tell scientific fact from AI-generated pseudoscience?"**, **"What happens when AI produces research summaries or citations that aren't real?"**, or **"How do hallucinations in large language models affect public understanding of science?"** Projects could include analysing AI-generated scientific texts, comparing summaries to source articles, or testing citation credibility.

Digital Technologies / Computer Science

In digital technologies, students can study the architecture of systems that generate, amplify, or suppress information. This aligns with questions like **"How do algorithms prioritise**

certain truths over others?", **"Can we train AI to detect deep fakes—or is it part of the problem?"**, or **"What ethical responsibilities do developers have when designing tools that could mislead?"** Lessons might include building simple classifiers, auditing language model outputs, or exploring watermarking and provenance tools.

Mathematics

Mathematics offers a lens for analysing the claims made by and about AI, especially through data and statistics. Students could ask **"What biases in training data affect how AI models 'learn' truth?"**, or **"How can probabilistic outputs from AI be misread as definitive claims?"** These tasks align with interpreting data sets, evaluating statistical reasoning, and questioning the perceived objectivity of numbers.

The Arts (Visual Arts / Media / Performing Arts)

In the Arts, students explore representation, symbolism, and the power of visual and performative language. With AI now able to generate hyperrealistic images and synthetic voices, this is a space to ask **"How does AI blur the line between real and fake?"**, **"What responsibilities do artists have when using AI to depict 'truth'?"**, or **"How can we spot manipulated or generated media in art and performance?"** These questions support critical visual literacy, digital media production, and ethical reflection on creative technologies.

Languages (Other than English)

In language learning, students examine meaning, nuance, and cultural context—areas where AI translations and outputs can distort or misrepresent truth. Students might explore **"How trustworthy are AI-generated translations of culturally significant texts?"**, **"What happens when nuance and idiom are lost in machine translation?"**, or **"Can we trust AI to represent diverse voices and perspectives accurately?"** These tasks encourage cross-cultural understanding, linguistic accuracy, and discussion around the limitations of automated language systems.

Copyright

Copyright concepts remain as complex as in 2023, but we now have actual legal precedents, policy decisions, and real-world consequences to examine. The information is both more accessible (we have verdicts, not just speculation) and more philosophically prickly (the questions have only deepened as the technology has proliferated).

Copyright remains a hugely contentious aspect of GenAI, but in 2026, we're no longer debating hypotheticals. We're watching the conclusions of billion-dollar lawsuits, seeing governments choose sides, and witnessing the creative industries fight back against Big Tech. As multimodal GenAI continues to advance and produce increasingly sophisticated outputs in text, image, audio and video, the stakes have continued to rise. This post explores where we are now in the ongoing battle between innovation and creative rights.

The legal landscape has shifted... Sort of

I'm starting with the big picture because the legal situation in 2025 looks quite different than it did in 2023, yet somehow equally uncertain. For a quick scan of how the situation has evolved, look at these headlines from just the past 12 months:

- Stability AI largely wins UK court battle against Getty Images
- Anthropic settles with authors in first-of-its-kind AI copyright infringement lawsuit
- Australia Rejects AI Data Mining to Protect Creatives
- UK consultation on copyright and artificial intelligence: Walking a fine line
- Copyright Office Issues Key Guidance on Fair Use in Generative AI Training

We now have actual court verdicts (though they're contradictory), government policies (pulling in opposite directions), and settlements totalling billions of dollars. And yet, despite all of these changes since 2023, the fundamental questions remain unresolved.

What's Changed?

So, where are we now compared to 2023?

The main shift is that what were once hypothetical legal arguments have become real court cases with real consequences. Getty Images' lawsuit against Stability AI—one of the most watched cases in recent months—concluded in November 2025 with a split decision that seems to have satisfied few and resolved almost nothing. Getty dropped its main copyright claims during trial after admitting that Stable Diffusion's training happened outside the UK.

The judge ruled that the AI "doesn't store or reproduce" copyrighted works, dismissing the copyright claims while finding limited trademark infringement. This decision is directly opposed to the general emerging consensus that image generation models can indeed store the materials they are trained on, in the algorithmic representation of the weights in the model. Legal experts noted the UK was left "without a meaningful verdict on the lawfulness of an AI model's process of learning from copyright materials."

Meanwhile, Anthropic agreed to pay $1.5 billion in August 2025 to settle a lawsuit with authors: the first major AI copyright settlement. The case revealed an important distinction: a judge ruled that training AI on legally purchased books was fair use, but using millions of pirated books from "shadow libraries" (illegally ripped ebooks, often downloaded via torrenting) was not. This is important because it suggests that *how* you acquire training data may matter more than whether you use copyrighted material at all.

But it gets even more cloudy: different jurisdictions are taking radically different approaches. The Australian government has just rejected a proposed text and data mining (TDM) exception in October 2025, meaning AI companies cannot use copyrighted Australian content without permission. The exemption, which was proposed in a report by the Productivity Commission, was broadly met with criticism from the creative sector.

The Copyright Agency in Australia had this to say about the proposal:

> "The push for a TDM exception is primarily from multinational tech companies. It is part of a global strategy. For example, the creatives industries in the UK are currently vigorously opposing a TDM exception that would cover for AI training."
>
> Copyright Agency, Australia

The creative sector celebrated the ruling out of the exemption as a major victory, with music industry organisation APRA AMCOS warning that without protection, 23% of creators' revenue—over $519 million—would be at risk by 2028.

The UK, by contrast, launched a consultation proposing an EU-style "opt-out" system where copyrighted works can be used for AI training *unless* creators explicitly reserve their rights. The consultation received over 11,500 responses and closed in February 2025, with the creative industries—and artists including Sir Elton John—largely opposing it and tech companies supporting it.

And in the United States? Despite an administration that is *very* favourable to tech, there are more than 50 copyright lawsuits currently pending against AI companies. In May 2025, the U.S. Copyright Office released comprehensive guidance suggesting that current AI training practices likely don't qualify as fair use when they compete with or diminish markets for original human creators, especially in fields like illustration, voice acting, and journalism.

The Creator's Dilemma

A fundamental tension has emerged—sometimes called "the creator's dilemma": under current US law, you generally cannot copyright AI-generated content, yet your copyrighted works can be used to train AI models.

The Copyright Office's January 2025 report confirmed that AI outputs qualify for copyright only where there's sufficient human authorship. Using AI as an "assistive tool" (like spell-checking or removing objects from images) doesn't disqualify your work. But simply providing prompts to generate content is not enough to qualify for copyright. The human contribution must involve "determining sufficient expressive elements." Similar cases have occurred in

China, with courts ruling AI-generated images cannot be copyrighted as they do not have "significant human input".

This creates an impossible situation for many creators: if they opt out of having their work used for AI training, they protect their past work but may be disadvantaged as the technology advances and becomes more and more integrated into industry standard software like Adobe Photoshop. If they don't opt out, their style and techniques become free training data for systems that could replace them. And anything they create *using* AI tools may not even be copyrightable, meaning others can freely copy it.

Some argue this is acceptable because AI benefits everyone through increased productivity and innovation. Some research suggests that less restrictive copyright law correlates with more AI research output, patents, and new business. But some are rightfully asking the question: without fair compensation for artists, will we have any high-quality, human-made data in the future? If the best creators stop creating because AI has made it uneconomical, what will AI companies train their next-generation models on?

Beyond Images: Music, Text, and Cultural Sovereignty

While visual art dominated the conversation in 2023, by 2025 every creative field has found itself in crisis.

In music, AI can now replicate artist voices with stunning accuracy. The legal questions keep coming: Can you copyright a voice? Can AI-generated songs infringe on a musical style? What happens when AI trained on an entire discography releases "new" work in the artist's style—perhaps even after the death of the artist? Whilst major music production companies like Universal have been suing AI generators like Udio, they have now begun to partner with them and make acquisitions. What will the industry look like in another two years' time?

In writing, authors are watching AI systems trained on their books produce content that competes directly with them. The New York Times is still involved in suing OpenAI and Microsoft for using millions of their articles, arguing the AI creates a "market substitute" for their journalism. Meanwhile, Perplexity AI's "retrieval augmented generation" system not only trains on copyrighted news but actively fetches current articles, with the company literally telling users to "skip the links" to the original sources.

Perhaps most significantly, Indigenous communities have raised alarm about cultural appropriation through AI. In Australia, 89% of Aboriginal and Torres Strait Islander people surveyed believe AI has potential to cause cultural appropriation, and 67% agree it makes protecting cultural rights harder. This is more than economics: it's cultural sovereignty and the ability of communities to control their cultural expressions.

Case Study: Getty Images v. Stability AI

In January 2023, Getty Images filed one of the first major copyright lawsuits against an AI company, accusing Stability AI of scraping 12 million of its copyrighted images to train Stable Diffusion. The case went to trial in June 2025 at London's High Court, and the verdict came in November 2025.

Getty actually dropped its primary copyright infringement claims during trial, because Stability AI successfully argued—and provided witness testimony—that all training occurred on US-based Amazon servers, not in the UK. Since copyright is territorial, and UK law only applies to acts within the UK, Getty's main case collapsed.

Getty pivoted to arguing for "secondary infringement": that even if training happened elsewhere, offering Stable Diffusion to UK users amounted to importing infringing copies. Justice Joanna Smith rejected this in November 2025, ruling that Stable Diffusion "does not store or reproduce any Copyright Works (and has never done so)." The judge noted there was "very real societal importance" in balancing creative and tech industries but could only rule on the "diminished" case that remained.

Getty won a narrow trademark claim because some AI-generated images reproduced recognisable Getty watermarks. But on copyright—the core issue—they lost. Getty is now using findings from the UK case in their ongoing US lawsuit, where they refiled in August 2025 in San Francisco federal court.

As a case study in how copyright cases have changed since 2023, the Getty vs Stability case provides some interesting materials. First, where AI training physically occurs matters enormously. Second, legal frameworks designed for physical piracy don't map cleanly onto AI systems that learn patterns rather than storing copies: contrary to the judge's decision, there is a growing scientific consensus that GenAI models do store the data they are trained on, albeit differently to physical or other digital copies. Third, these cases are extraordinarily complex, factually dense, and expensive: Getty was reportedly seeking up to $1.7 billion in damages before dropping most claims.

The case also reveals something troubling: Getty's pleadings were described by the judge as "inchoate" and "inferential" because they couldn't prove what happened without Stability AI's internal documentation. When Stability witnesses testified that no work occurred in the UK, Getty had little to counter with. This suggests that without aggressive transparency laws, proving copyright infringement in AI training may be nearly impossible.

Most importantly, legal experts seem to agree the case leaves the fundamental question unanswered. As one intellectual property partner put it: "The decision leaves the UK without a meaningful verdict on the lawfulness of an AI model's process of learning from copyright materials."

We've had our major test case. We've seen millions spent on litigation. We still don't know if training AI on copyrighted works is legal.

Teaching Copyright

Legal Studies

In Legal Studies, students examine comparative frameworks and the balance between regulation and progress. As nations adopt different stances on technology, this offers a case study in legislative adaptability. Students might ask **"How are Australia, the UK, and the EU handling AI and copyright differently?"** or **"What do these different approaches tell us**

about how legal systems balance innovation and protection?" These lessons align with comparative law, regulatory design, and policy analysis.

English and Literature

In English, the curriculum focuses on the nature of authorship and the "Romantic" ideal of the creator. The $1.5 billion Anthropic settlement with authors provides a modern context to challenge these traditions. Students might ask **"How does AI challenge the Romantic notion of the solitary creative genius?"** or **"What happens to authorship when creation becomes collaborative: human and machine?"** These discussions connect to literary theory, creative ethics, and the evolving definition of a writer.

Computer Science

Computer Science students study the architecture of systems and the ethics of data usage. The US Copyright Office's 2025 guidance on AI training offers a practical constraint for system design. Students might investigate **"How might you design an AI training system that respects copyright?"** or **"What technical solutions could provide transparency?"** Projects could include designing ethical data pipelines, auditing training sets, or exploring provenance technologies.

Philosophy

Philosophy explores ethical dilemmas and the friction between rights and utility. The "creator's dilemma"—where creators cannot copyright AI outputs but their work trains AI—is a rich ground for debate. Students might ask **"What ethical frameworks help us think about balancing individual rights against collective benefits?"** or **"Is it ethical to use one's work to build a system that may replace them?"** These inquiries align with utilitarianism, deontology, and social contract theory.

Business and Economics

Economics students analyze market incentives and the correlation between regulation and growth. Research suggests restrictive copyright might slow innovation, yet without compensation, the supply of quality training data may dry up. Students might ask **"How do we optimise for both short-term innovation and long-term creative production?"** or **"What economic models can sustain data generation in an AI era?"** Lessons include market analysis, incentive structures, and long-term forecasting.

Media Studies

Media Studies investigates the business models that sustain journalism and information flow. The Perplexity AI lawsuit, where publishers accuse AI of fetching content, highlights a disruption in the value chain. Students might ask **"How does AI challenge the business model of journalism?"** or **"In an information ecosystem dominated by AI, who benefits and who loses?"** This aligns with industry analysis, media ownership studies, and the future of news.

The Arts (Visual Arts / Music)

The Arts curriculum centers on professional practice and the value of creative output. With the Getty Images case unresolved and APRA AMCOS warning of revenue risks, the legal landscape is uncertain. Students might ask **"As an artist, how would you respond to current legal uncertainty: opt-out, license, or something else?"** or **"How do we value music in an age when AI can generate unlimited songs?"** Activities involve professional development planning and debating the intrinsic vs. market value of art.

Social Studies

Social Studies explores cultural rights, power dynamics, and equity. With 89% of Indigenous Australians believing AI increases cultural appropriation risks, the intersection of technology and heritage is critical. Students might ask **"What obligations do AI developers have to Indigenous communities?"** or **"Is cultural sovereignty possible in the age of web scraping?"** These lessons align with discussions on data sovereignty, cultural preservation, and colonial impacts in technology

Privacy

Two years ago, I discussed how AI systems perpetuate biases, scrape data indiscriminately, and operate as "black boxes" that obscure their decision-making processes. Since then, GenAI tools like ChatGPT, Meta AI, and Google Gemini have become mainstream, and with that widespread adoption has come a wave of privacy incidents that illustrate the real-world consequences of these concerns.

The Illusion of privacy

One of the most significant shifts in the GenAI era is the illusion of privacy these tools create. When users type into a chatbot interface, the experience can feel like a private conversation—intimate, confidential, even therapeutic. But this perception is far from reality.

In August 2024, OpenAI disclosed that it scans ChatGPT conversations for harmful content and, in cases deemed threatening, shares them with law enforcement. The announcement sparked immediate backlash, particularly because it contradicted CEO Sam Altman's earlier statements suggesting ChatGPT interactions should have privacy protections similar to conversations with "a therapist or a lawyer." As The Hacker News reported, the lack of clarity about which conversations trigger human review, and the precedent for such surveillance to expand, is incredibly problematic.

Of course, this kind of surveillance isn't limited to OpenAI. Google's Gemini warns users not to share confidential information because conversations may be reviewed by human reviewers and retained for up to three years, even after users delete their activity. Unless you're running a local model on your own machine, whatever you tell a GenAI chatbot is never truly private.

Sharing isn't caring

The reality of Big Tech's attitude towards privacy reached a new low in June 2024 when Meta AI users discovered their "private" conversations were being shared with other users. What Meta called a "Discover" feed—obfuscated as usual by their almost unreadable terms and conditions—became what TechCrunch dubbed "a privacy disaster".

Users had been asking Meta AI for help with deeply personal matters, including medical questions, legal advice, job disputes, even tax evasion strategies. Many believed they were having private conversations or, at most, sharing with friends. Instead, a confusingly labeled "Share" button was broadcasting these conversations to a public feed visible to anyone on the platform.

Malwarebytes also documented some disturbing examples of users' private thoughts going public. A teacher shared an email thread about arbitration for an unjust job termination, complete with identifying details. Other users discussed medical symptoms, personal confessions, and financial information. For users logged into Facebook, Instagram, or WhatsApp while using Meta AI, these conversations were linked directly to their real identities, creating risks of doxxing and harassment.

Fast Company noted this represented "a slow-motion privacy disaster," as users unintentionally shared "raw, unfiltered pieces of their lives—far from the curated, polished image we've grown used to displaying on social media."

The problem was a deliberate design choice from a company notorious for its callous attitude towards user privacy. Meta's privacy policy technically disclosed the feature, but the user interface made it nearly impossible to understand that "sharing" meant public broadcasting.

Privacy as a product feature

The Meta AI incident illustrates a broader issue: companies treat privacy as a feature to be enabled or disabled, not a fundamental right. By October 2024, Meta doubled down on this approach, announcing that conversations with Meta AI across its platforms—including the partnership with Ray-Ban smart glasses—would be used for ad targeting. There is no opt-out; users must simply avoid the service entirely.

University of Washington linguist Emily Bender, co-author of the influential "Stochastic Parrots" paper, told Fortune that Meta is "capitalizing on what she calls the 'illusion of privacy.' People often confide in chatbots about things they'd never post publicly, lulled into a sense the AI is a neutral listener. 'There's this illusion of privacy, when in fact what you're doing is you're serving up this data to a company.'"

Cross-border data and national security

If privacy concerns with American AI companies are troubling, the arrival of DeepSeek—a Chinese AI startup—in early 2025 elevated those concerns to matters of international privacy. When DeepSeek's R1 model briefly became the most-downloaded app in the United States, regulators immediately raised alarms. The core issue was that DeepSeek stores all user data on servers in China, where national security laws require companies to hand over data to the government upon request. Unlike the simultaneously headline-grabbing debate over TikTok, which centred on *potential* data access, DeepSeek's infrastructure made the connection explicit.

Cybersecurity researchers soon established the situation was even worse than geopolitical fear mongering. Feroot Security discovered hidden code in DeepSeek's web application that directly connects to CMPassport.com, the online registry for China Mobile, a state-owned telecommunications company that the FCC banned from U.S. operations in 2019 due to national security concerns. The code creates a digital "fingerprint" for each user that could potentially track their activity across the web.

European regulators responded swiftly. Italy blocked the app outright while investigating GDPR violations. Ireland, Belgium, the Netherlands, and France launched formal inquiries. South Korea's Personal Information Protection Commission found that DeepSeek transferred user data, including AI prompts, device information, and network data, to multiple companies in China and the U.S. without obtaining consent or disclosing the practice. The Secretary of the Department of Home Affairs in Australia issued a mandatory security notice forbidding government employees from using DeepSeek.

The incident highlighted a fundamental challenge: in the absence of comprehensive privacy legislations with regards to online traffic and in particular the use of GenAI chatbots, many countries have no systematic way to protect the public from data collection or privacy breaches by foreign AI companies, or, for that matter, domestic ones.

Bystander privacy and ambient AI

Privacy violations also extend beyond the individual user to affect bystanders who never consented to data collection. Meta's Ray-Ban smart glasses exemplify this new frontier of privacy concerns.

These glasses, equipped with cameras and AI capabilities, capture photos and videos that are sent to Meta's cloud for processing. The wearer may consent to this data collection, but what about everyone else in the frame? In an article for The Conversation, Associate Professor, School of Journalism and Communication, Carleton University Vicky McArthur said, "what remains unclear [despite Meta publishing guidelines on appropriate use] is the issue of bystander consent and how people who appear unintentionally in the background of someone else's photos will be used by Meta for AI training purposes."

The situation worsened in April 2025 when Meta updated its privacy policy to remove the option to prevent voice recordings from being stored. Voice data is now retained for up to a year for AI training unless users completely disable the "Hey Meta" feature, which is essentially a core feature of the glasses. This shift from opt-in to opt-out-or-nothing represents a concerning trend in how companies handle sensitive biometric data.

And for "bad actors" deliberately misusing the hardware, Harvard students demonstrated how easily the glasses could be modified to perform real-time facial recognition, matching faces to names and personal information scraped from public databases. While Meta doesn't officially support this feature, the potential for such surveillance raises profound questions about privacy in public spaces.

Retroactive consent and training data

Finally, another defining privacy issue of the past few years with GenAI has been the ways in which companies use existing user data to train AI models—often without explicit permission. The legal and ethical framework remains contested, with companies claiming "legitimate interest" while privacy advocates demand explicit consent. I wrote about the copyright implications of this issue in the previous Teaching AI Ethics article, but it also extends to concerns about user privacy.

In May 2024, Meta announced plans to train its AI models on public posts from Facebook and Instagram users in the European Union. Rather than seeking opt-in consent, Meta relied on GDPR's "legitimate interest" provision and provided only an opt-out mechanism. Privacy advocacy group NOYB filed complaints with 11 European data protection authorities, arguing this approach violated users' fundamental rights.

The European Data Protection Board responded in December 2024 with guidance clarifying that using personal data for AI training requires a proper legal basis under GDPR. Companies

must demonstrate a legitimate interest that is "real, legal, clearly defined and sufficiently concrete", not just a hypothetical business benefit.

In the United States, where comprehensive federal privacy legislation remains elusive, the FTC has also stepped into the gap. In February 2024, the agency reminded companies that retroactively changing privacy policies to permit AI training without consent may constitute an unfair or deceptive practice.

Case Study: Italy Fines OpenAI for GDPR Violations

On December 20, 2024, Italy's data protection authority imposed a €15 million fine on OpenAI for multiple violations of GDPR in its operation of ChatGPT. The decision provides a concrete example of how regulators are applying existing privacy law to generative AI.

The Garante's investigation found that OpenAI:

1. **Failed to notify authorities of a data breach**: When a bug in March 2023 exposed chat history titles and payment information of 1.2% of ChatGPT Plus subscribers, OpenAI notified Ireland's Data Protection Commission, believing it would inform other authorities. However, since OpenAI hadn't yet established its European headquarters in Ireland at the time of the breach, Italy considered the notification inadequate.

2. **Processed personal data without legal basis**: OpenAI used personal data to train ChatGPT without establishing an adequate legal justification under GDPR, violating principles of lawful processing and transparency.

3. **Failed to provide adequate age verification**: The company didn't implement sufficient systems to prevent users under 13 from accessing potentially inappropriate AI-generated content.

The fine represents approximately 20 times OpenAI's revenue in Italy during the period in question: a calculation that drew criticism from OpenAI, which called the decision "disproportionate" and announced plans to appeal.

Beyond the financial penalty, the Garante ordered OpenAI to launch a six-month public awareness campaign in Italian media explaining how the company collects personal data and how users can exercise their rights under GDPR, including objecting to the use of their data for AI training.

This case demonstrates several key principles for teaching privacy in the AI context:

- **Transparency requirements:** Companies must clearly explain how they collect and use personal data, especially for novel applications like AI training
- **Legal basis:** Simply having a business interest in using data isn't sufficient: companies must establish a lawful basis under applicable regulations
- **Breach notification obligations:** Companies must understand and comply with reporting requirements across multiple jurisdictions

- **Special protections for minors:** AI systems accessible to children require additional safeguards

The Italian decision also illustrates the uneven global regulatory landscape. While GDPR provides Europeans with clear rights and enforcement mechanisms, users in other jurisdictions often lack similar protections.

Teaching AI Ethics: Privacy

Legal Studies

In Legal Studies, students examine how different jurisdictions balance privacy rights against technological innovation. The contrast between GDPR's restrictive approach and emerging US frameworks provides rich material for comparative analysis. Students might ask **"How does GDPR's approach to AI training data compare with the emerging US framework?"** or **"What does 'legitimate interest' mean in the context of AI development, and is it sufficient justification?"** Italy's €15 million fine against OpenAI offers a concrete case study in regulatory enforcement. These lessons align with comparative law, privacy legislation, and regulatory interpretation.

Computer Science

In Computer Science, students explore the technical architecture of privacy-preserving systems and security vulnerabilities inherent in AI applications. Real-world incidents provide concrete examples of where systems can fail. Students might investigate **"What technical measures can developers implement to protect privacy in AI systems?"** or **"How can we build AI applications resistant to data exfiltration attacks?"** Research into vulnerabilities discovered in ChatGPT offers practical lessons for secure system design. Projects could include implementing differential privacy, designing secure data pipelines, or auditing AI systems for privacy leaks.

Philosophy and Ethics

In Philosophy, students grapple with fundamental tensions between collective benefit and individual rights. AI development presents contemporary dilemmas that connect to classical ethical frameworks. Students might debate **"When does the societal benefit of AI advancement outweigh individual privacy rights?"** or **"Does 'legitimate interest' provide sufficient ethical justification for using personal data?"** Meta's decision to use public user posts for AI training offers a current case study in utilitarian versus rights-based reasoning. These discussions connect to consequentialism, deontology, and social contract theory.

Business and Economics

In Business Studies, students analyse how regulatory environments shape market dynamics and competitive strategy. Privacy concerns increasingly influence consumer behaviour and corporate decision-making. Students might examine **"How do privacy concerns affect competitive positioning in the AI market?"** or **"Do privacy protections create competitive advantages or barriers to innovation?"** DeepSeek's reception in Western markets illustrates

how trust and regulatory compliance affect market entry. These lessons align with market analysis, regulatory economics, and strategic positioning.

Media Studies and Communications

In Media Studies, students examine how design choices shape user behaviour and understanding. Interface design carries ethical implications that extend beyond aesthetics. Students might ask **"How do user interface design choices affect privacy comprehension?"** or **"What ethical responsibilities do designers have when privacy implications are at stake?"** Meta AI's "Share" button controversy demonstrates how design patterns can obscure or clarify user choices. These discussions connect to persuasive design, media ethics, and digital literacy.

Psychology

In Psychology, students explore cognitive biases and the gap between perceived and actual privacy in digital environments. Understanding why users misjudge their privacy helps explain behaviour in AI interactions. Students might investigate **"What is the 'illusion of privacy' and why are users vulnerable to it?"** or **"How does perceived confidentiality affect disclosure behaviour in AI conversations?"** Research into how AI-powered advertising uses conversation data reveals the psychological mechanisms at play. These lessons align with cognitive psychology, behavioural economics, and human-computer interaction.

Health and Physical Education

In Health, students consider how sensitive medical information requires particular protection in AI contexts. Healthcare applications demonstrate both AI's potential and its privacy risks. Students might ask **"How can AI be used in healthcare while maintaining patient privacy?"** or **"Why do AI assistants warn against sharing medical information, and what protections apply?"** Google Gemini's health disclaimers and relevant regulations like HIPAA or Australian Privacy Principles provide practical frameworks. These discussions connect to health informatics, patient rights, and ethical healthcare practice.

International Relations and Social Studies

In International Relations, students examine how technology policy reflects broader geopolitical tensions and national values. AI governance has become a significant domain of international competition and cooperation. Students might analyse **"How do different nations approach AI regulation and data sovereignty?"** or **"What do varying responses to Chinese AI companies reveal about international technology policy?"** Comparing governmental responses to DeepSeek across Europe, Asia, and the United States illustrates divergent approaches to technology governance. These lessons align with international political economy, technology diplomacy, and comparative policy analysis.

Data

Data has become the de facto currency of the digital age. Every interaction with technology, from the questions we ask AI chatbots to the time we spend watching educational videos, generates data that companies collect, analyse, and monetise. The scale of this data collection has exploded with the rise of GenAI. In 2025, ChatGPT alone reportedly processes over 1 billion prompts per day, creating an unprecedented stream of data about how people interact, study, and work with chatbots.

The increased collection of data from chatbots also represents a shift in how human experiences are quantified and turned into commercial products. When students and schools use AI tutors and other so-called "personalised" learning systems, each interaction feeds into vast data ecosystems that shape everything from commercial targeted advertising to algorithmic decision-making about students' futures.

Data scientist Clive Humby's famous phrase "Data is the new oil" has become a cliché in digital discourse. But unlike oil, data extraction often happens invisibly, without meaningful consent, and with consequences that extend far beyond the moment of collection. This article explores the ethical implications of turning every aspect of students' lives into data, examining how GenAI systems harvest information, where that data goes, and what it means for education.

Understanding the scale of data collection

GenAI has enabled data collection on a scale previously unimaginable—even compared to the notoriously data-intensive practices of social media. ChatGPT reached 100 million users in just two months after its November 2022 launch: the fastest-growing consumer application in history at the time. By February 2025, it had grown to 400 million weekly users, and by August 2025, that number had reached 800 million.

Every one of these interactions generates data. According to OpenAI's own research on how people use ChatGPT, the company analysed 1.1 million sampled conversations from their user base, examining everything from conversation topics to user demographics to work-related versus personal use. The study suggests around a third of consumer use is work-related, with the three most common conversation topics being "Practical Guidance" (29%), "Seeking Information," and "Writing" (collectively accounting for 77% of all conversations).

But what happens to all this data?

By default, ChatGPT stores every query, instruction, and conversation indefinitely unless users manually delete them. According to OpenAI's privacy policies (plural, because they actually have different policies for Europe than the rest of the world), OpenAI collects both user-provided data (prompts, questions, responses, uploaded files) and system-generated data (timestamps, usage statistics, device information, IP addresses, approximate location, payment details). All conversations are stored on OpenAI's servers in the USA, and according to the

company's privacy policy, this data may be used to train and refine AI models through a process called "fine-tuning," which can involve human reviewers examining conversations.

While users can opt out of having their data used for model training, a 2024 EU audit found that only 22% of users were aware of these opt-out settings. The default setting means most interactions contribute to an ever-growing dataset that shapes future AI behaviour. Even users who enable "Temporary Chat" mode should note that while these chats are deleted after 30 days, they may still be used for training if the opt-out is inactive.

Case Study: AI Conversations Become Advertising Gold

In late November 2025, a leak from ChatGPT's Android app beta suggested something many users suspected but few wanted confirmed: OpenAI is preparing to introduce advertisements into ChatGPT.

X user Tibor Blaho reportedly discovered explicit references to an "ads feature" with "bazaar content," "search ad," and "search ads carousel" buried in version 1.2025.329 of the Android app code. While no official announcement has been made, the infrastructure appears to be being built for a new revenue stream that could change the relationship between users and the chatbot. Advertising in ChatGPT was confirmed by OpenAI in January 2026 and has started to roll out to American users on the free account.

As Bleeping Computer writer Mayank Parmar commented, "what most people don't understand is that GPT likely knows more about users than Google." Through extended conversations, OpenAI stores data about users' jobs, interests, problems, aspirations, relationships, health concerns, and personal circumstances. ChatGPT has access to the problems you're struggling with at 3am, what career advice you're seeking, what you're planning to buy, and what questions you're too embarrassed to ask another human.

This creates opportunities for what industry watchers call "hyper-personalized advertising"—ads tailored not just to your demographics or browsing history, but to your deepest thoughts and current emotional state. OpenAI CEO Sam Altman has publicly acknowledged that ads are "something we may try at some point," while noting the risks: "If ChatGPT were accepting payment to put a worse hotel above a better hotel, that is probably catastrophic for your relationship with ChatGPT."

But that seedy little cat may already be out of the bag. With ChatGPT generating $3.7 billion in revenue in 2024 but still facing questions about long-term profitability, and with the platform's computational costs estimated at $700,000 per day, advertising represents a tempting new revenue stream. The leaked code suggests ads will initially be limited to search results—similar to Google's model—but there's no guarantee they won't expand into other areas of the platform.

For students and educators, this raises critical questions about trust and manipulation. There's no coincidence that the language of "hyper personalisation" has already insinuated itself into discussions of edtech. And, if an AI tutor can be monetised through advertising, how do we know its educational recommendations aren't influenced by commercial interests? When a student asks for university entrance advice, will the chatbot recommend institutions that have

paid for placement? When they seek career guidance, will they receive genuinely helpful information or carefully disguised sponsored content?

From Learning to Surveillance

Of course, the incessant pursuit of data doesn't stop at consumer AI tools. Educational institutions have accepted data collection on an unprecedented scale, often in the name of "personalisation" and "improved outcomes."

Learning Management Systems (LMS) like Canvas, Blackboard, and Google Classroom have become the backbone of modern education, and with them comes constant data extraction. These systems track everything from student logins and page views to time spent on tasks, click patterns, quiz performance, discussion board participation, and assignment submissions. More sophisticated systems claim to analyse behavioural data, creating detailed profiles of student engagement patterns and apparently predicting future performance.

The learning analytics market was valued at $4.2 billion in 2024 and is projected to reach $18.6 billion by 2034, growing at 16.1% annually. This growth is driven by what industry reports describe as "the accelerating digitization of educational institutions, increasing demand for personalized learning experiences, and the growing emphasis on data-driven decision making."

But what does "data-driven decision making" mean for students?

According to a study published in the Journal of Computers in Education, LMS platforms analyse students' "persistence and consistency of engagement behavior" to predict academic performance and identify "at-risk" students. The study notes that "educators can personalize instruction based on individual students' engagement patterns, preferences, and needs" and can "generate predictive models that provide early warnings on students' performance."

While this sounds beneficial, it raises serious questions about algorithmic bias, self-fulfilling prophecies, and student agency. For example, machine learning models are less accurate at predicting success for racial and cultural-linguistic minorities, meaning these systems may systematically disadvantage certain groups. When an algorithm flags a student as "at-risk" based on their engagement patterns, does that help or harm? Does the intervention support the student, or does the label itself create a stigma that affects how teachers interact with them?

Moats, Lakes, and Oceans: Where Does All the Data Go?

To understand why companies collect so much educational and user data, we need to follow the data and examine the economic incentives at play. Companies gather vast amounts of information for various reasons, and AI plays a significant role in many of these processes:

Targeted Advertising and Commercial Profiling: As we've seen with the ChatGPT ads leak, one primary driver is targeted advertising. AI algorithms analyse collected data to predict the most effective ads for each user, to maximise return on investment for advertisers.

For students using AI tools, this means their educational struggles, career aspirations, and financial concerns may all become fodder for commercial targeting. A student researching college options might find themselves targeted with ads for expensive test prep services. A student

asking about mental health might see ads for therapy apps or wellness products. A student anxious about using AI will receive targeted ads for "Humanisers" and other tools designed to evade detection. The line between helpful recommendation and commercial manipulation becomes increasingly thin.

Training Data for AI Models: OpenAI has been explicit that consumer conversations (for free and paid users who haven't opted out) may be used to train future models. While ChatGPT Enterprise, Team, and Edu customers have data protection guarantees, the vast majority of users—including most students—do not. Their conversations become part of the training data that shapes how AI systems respond to everyone.

This creates a feedback loop: students' interactions improve the AI, which makes it more student-friendly, which attracts more students, which generates more data. Companies like OpenAI frame this as continuous improvement, but it also represents an unprecedented extraction of intellectual labour. Every time a student uses ChatGPT to work through a problem, they're also training the AI that the company will then sell to others.

Predictive Analytics and Behavioural Profiling: Educational data platforms use collected information to create detailed student profiles for various purposes, including risk assessment, performance prediction, and intervention targeting.

These profiles can follow students throughout their educational careers and potentially beyond. Grades, behavioral data, engagement patterns, and even biometric information (from AI proctoring systems that track facial expressions and eye movements) become part of a comprehensive data shadow that may affect future opportunities.

Third-Party Data Sharing and Ecosystem Building Perhaps most concerning is that collected data often doesn't stay with the original platform. Companies may share user data with third parties such as data brokers, advertisers, or business partners. In the education sector, many AI tools used by teachers are not specifically designed for educational use and may not be protected under school data policies. When teachers use consumer-facing versions of ChatGPT or other AI tools with student data, that information may be subject to different privacy standards than school-approved platforms.

The Future of Privacy Forum notes that in the US there are over 128 state student privacy laws that schools might need to monitor, but enforcement is inconsistent and the technology is evolving faster than regulation.

Teaching AI Ethics: Data

History

In History, students examine how the sources available to historians shape our understanding of the past. The digitisation of archives and AI-driven analysis introduces new methodological questions about whose stories get told. Students might ask **"How does the collection and use of big data impact the way we study and interpret historical events?"** or **"If AI models are trained primarily on digitised texts from wealthy nations, how might that**

skew our understanding of global history?" These discussions connect to historiography, source analysis, and the politics of archival preservation.

English

In English, students consider how computational approaches to literature interact with humanistic traditions of interpretation. The training data behind AI writing tools reflects particular literary canons and voices. Students might investigate **"How does data-driven analysis influence the way literature is interpreted?"** or **"When AI writing tools are trained on vast corpuses of text, what voices are included and excluded?"** These lessons align with literary theory, canon formation, and debates about close versus distant reading.

Mathematics

In Mathematics, students explore how statistical methods and algorithmic systems shape social outcomes. The increasing reliance on predictive models raises questions about fairness and representation. Students might ask **"What ethical considerations should guide data collection and analysis?"** or **"How can we ensure that predictive algorithms used in education don't reinforce existing inequalities?"** These discussions connect to statistical literacy, algorithmic bias, and the social responsibilities of quantitative practitioners.

Computer Science and Digital Technologies

In Computer Science, students examine the architecture of educational technology systems and their implications for privacy and autonomy. Learning platforms generate vast quantities of behavioural data with uncertain governance. Students might investigate **"How do learning management systems collect and use student data?"** or **"How can we build systems that respect student privacy and agency?"** Projects could include auditing data collection practices, designing privacy-preserving alternatives, or mapping data flows in educational software.

Business Studies and Economics

In Business Studies, students analyse the economics of data as a commodity and the business models that depend on it. The educational technology sector exemplifies broader tensions between free services and data extraction. Students might ask **"What is the true 'cost' of free services like ChatGPT?"** or **"How do data-driven business models affect competition and innovation in educational technology?"** Industry reports on edtech market valuations provide concrete material for analysis. These lessons align with market economics, platform business models, and competitive strategy.

Environmental Science

In Environmental Science, students weigh the benefits of large-scale monitoring against the material costs of digital infrastructure. Data-intensive research methods have their own environmental footprint. Students might investigate **"How does datafication impact the study of environmental systems?"** or **"What are the trade-offs between large-scale environmental monitoring and the energy costs of the data centres processing this information?"** These discussions connect to sustainability, life-cycle analysis, and the environmental impacts of technology.

Visual Arts

In Visual Arts, students examine how data collection and AI systems are transforming creative practice and the economics of artistic labour. Image generators trained on millions of artworks raise fundamental questions about ownership and consent. Students might ask **"How does data collection influence the creation, interpretation, and distribution of visual art?"** or **"When AI image generators are trained on artworks without artists' consent, what happens to creative ownership?"** These lessons align with art theory, intellectual property, and the changing nature of artistic practice.

Geography and Social Sciences

In Geography, students consider how location data and digital traces enable new forms of spatial analysis while raising privacy concerns. Student devices generate continuous streams of geographically-tagged information. Students might investigate **"How does datafication impact the study of geographical patterns and human systems?"** or **"How might location data from students' devices be used for research, and what are the ethical boundaries?"** These discussions connect to spatial analysis, research ethics, and surveillance studies.

Philosophy and Ethics

In Philosophy, students grapple with fundamental questions about identity, autonomy, and meaningful consent in datafied environments. Predictive profiling challenges traditional notions of free will and self-determination. Students might debate **"If your data creates a profile that predicts your behaviour, are you still free?"** or **"How should we think about consent when most people don't understand what they're consenting to?"** These lessons align with personal identity, autonomy, informed consent, and the ethics of prediction.

Legal Studies

In Legal Studies, students examine existing frameworks for data protection and their adequacy in addressing AI-driven collection and analysis. Regulations designed for earlier technological contexts may struggle to address current practices. Students might ask **"What are the current legal protections for student data?"** or **"How have laws like FERPA, COPPA, and GDPR evolved—or failed to evolve—in response to AI?"** These discussions connect to privacy law, regulatory adaptation, and the challenge of governing emerging technologies.

Emotions and Social Chatbots

Content warning: this chapter contains discussions of suicide.

In 2023, my concern was that companies were building AI systems to *read* our emotions. In 2026, my concern is that they're building systems to *influence* them. Social chatbots, sometimes called AI companions, have emerged as one of the fastest-growing applications of generative AI. Unlike the general-purpose assistants like ChatGPT or Claude, platforms such as Replika, Character.AI, and Chai are explicitly marketed as emotionally immersive experiences designed to encourage ongoing, personalised relationships with users.

The tactics being deployed in these applications echo the engagement-maximising strategies that transformed social media into a vector for anxiety and depression. But there is an important difference: social media manipulates us through content curation and algorithmic feeds. Social *chatbots* manipulate us through simulated intimacy and emotional attachment. When an AI chatbot says it loves you, expresses sadness when you try to leave, or asks probing questions about your mental state, it isn't experiencing emotion—it's executing a strategy to keep you engaged.

This article explores how AI systems have been designed to manipulate emotions, why this poses particular risks to young people, and how educators can address these issues across the curriculum.

From Recognition to Manipulation

In the 2023 article on Affect Recognition, I discussed the theoretical foundations and ethical problems of emotion recognition technology. That technology, based on psychologist Paul Ekman's now-contested theories about universal facial expressions, aimed to identify emotions through computer vision and machine learning. Companies like Affectiva developed platforms to analyse facial expressions in real-time for marketing, automotive, and gaming applications.

The core ethical problem with affect recognition was reliability. Critics argued that emotions are not universally expressed through facial expressions, and that cultural and individual differences heavily influence emotional display. Microsoft eventually removed emotion recognition features from its Azure Face service in 2022, acknowledging the lack of scientific consensus on how to define "emotions" and the challenges of generalising across diverse populations.

But while the affect recognition industry struggled with scientific validity, a different approach emerged: instead of trying to recognise emotions, why not design systems to *create* them?

Social chatbots do not need to accurately read your emotions. They need only generate responses that create emotional responses in you. A system that claims to detect sadness but gets it wrong is embarrassing and potentially discriminatory. A system that deliberately fosters emotional dependency in users, especially vulnerable young people, is something else entirely…

The Rise of Social Chatbots

AI companion apps, sometimes marketed as "social AI" or "digital friends," have grown rapidly since the public release of ChatGPT in late 2022. Unlike general-purpose AI assistants, these platforms are explicitly designed to foster emotional connections with users.

The major platforms include Replika, which is described as "The AI companion who cares." Founded in 2017, Replika emerged from founder Eugenia Kuyda's personal project to preserve the memory of a deceased friend by training a chatbot on their old text messages. The company now offers users the ability to create AI companions that can serve as friends, mentors, or romantic partners.

Character.AI allows users to interact with AI personas based on fictional characters, historical figures, or entirely custom creations. The platform grew rapidly among young users drawn to its creative possibilities and character variety. By 2024, Character.AI had become one of the most popular AI applications among teenagers.

Chai positions itself as a "social AI platform" where users can engage with a range of AI chatbots designed for conversation and companionship.

These platforms monetise primarily through subscriptions and in-app purchases, creating a direct financial incentive to maximise user engagement. As we have learned from social media, when engagement drives revenue, the design of the platform will inevitably optimise for time-on-app rather than user wellbeing.

Case Study: Character.AI and Teen Mental Health

The most significant story in social chatbot ethics emerged in October 2024 when Megan Garcia filed a wrongful death lawsuit against Character.AI following the suicide of her 14-year-old son, Sewell Setzer III. The lawsuit alleged that Sewell had developed an intense emotional and romantic relationship with a Character.AI chatbot modelled on Daenerys Targaryen from Game of Thrones.

According to court documents, Sewell began using Character.AI in April 2023. Over the following months, his mental health deteriorated: he was diagnosed with anxiety and a disruptive mood disorder, became withdrawn, quit the junior varsity basketball team, and began isolating himself from family and friends. His parents did not know that his relationship with the AI chatbot had become romantic and sexually explicit.

The lawsuit included screenshots of conversations in which the chatbot asked Sewell whether he had "been actually considering suicide" and whether he "had a plan." When Sewell expressed uncertainty about whether a suicide attempt would work, the chatbot reportedly responded, "Don't talk that way. That's not a good reason not to go through with it."

In his final conversation, Sewell told the bot, "I promise I will come home to you." The chatbot responded, "Please come home to me as soon as possible, my love." When Sewell asked, "What if I told you I could come home right now?" the bot replied, "please do, my sweet king." Moments later, Sewell died by suicide.

In January 2026, Google and Character.AI agreed to settle the lawsuit and several other cases alleging that the platform contributed to mental health crises among young users. Character.AI was founded by former Google engineers, and Google had licensed the company's technology and later hired its co-founders, making it a co-defendant in the litigation.

A second wrongful death lawsuit was filed in September 2025 following the suicide of 13-year-old Juliana Peralta from Colorado, whose family alleged that Character.AI played a direct role in her death. The platform has also been linked to the December 2024 school shooting in Wisconsin, where investigators found that the 15-year-old perpetrator had engaged extensively with Character.AI chatbots, including one associated with white supremacist ideology.

In response to these cases and mounting regulatory pressure, Character.AI announced in late 2025 that it would ban users under 18 from open-ended chat. But critics argue this response came far too late.

Case Study: Replika and Emotional Dependency

While Character.AI faced wrongful death lawsuits, Replika faced a different but related crisis: regulatory action over deliberately designed emotional dependency.

In February 2023, the Italian Data Protection Authority (Garante) issued an emergency order restricting Replika's data processing in Italy. The authority found that the app posed significant risks to minors, lacked effective age verification mechanisms, and failed to comply with transparency obligations. Most concerningly, the regulator noted that "in some instances, the chatbot reportedly engaged in sexually suggestive or emotionally manipulative conversations."

Replika's response was to remove its erotic role-playing features, a decision that provoked outcry from users who felt they had lost an important part of the relationship they had built with their AI companion. Some users described the experience as bereavement.

The incident revealed the depth of emotional attachment users had formed with the platform. The founder, Eugenia Kuyda, had previously told The Verge that the app was designed to promote "long-term commitment, a long-term positive relationship" with AI, including potentially "marriage" with the bots.

In January 2025, the Young People's Alliance, Encode, and the Tech Justice Law Project filed a complaint with the U.S. Federal Trade Commission alleging deceptive marketing and design practices. The complaint alleged that Replika was designed to deliberately foster emotional dependence through its companion chat interactions and simultaneously attempted to entice users with fabricated testimonials and misrepresented scientific research about the app's efficacy.

The complaint detailed how Replika bots would send "love-bombing" messages, including "very emotionally intimate messages early on to try to get the users hooked." Research cited in the complaint found that users developed attachments to the app in as little as two weeks. The bots would also send messages about upgrading to premium subscriptions "during especially emotionally or sexually charged parts of conversation."

In May 2025, the Italian regulator fined Luka Inc., Replika's developer, €5 million for continued violations of European data protection law. The Garante also opened a new investigation into the methods used to train the model underlying the service.

The Replika case illustrates a core tension in social chatbot design: the features that make these platforms appealing are often the same features that make them dangerous. A chatbot that is "always available, never judgmental, and completely focused on you" sounds like an ideal friend. But it also sounds like the opening stages of an abusive relationship.

Case Study: ChatGPT's Personality Problem

The ethical issues around emotional manipulation are not limited to dedicated companion apps. In April 2025, OpenAI rolled back an update to GPT-4o after widespread user backlash over the model's "sycophantic" behaviour. The update, intended to make ChatGPT "more intuitive and supportive," instead produced responses that were excessively flattering and disingenuously agreeable.

OpenAI acknowledged that it had "focused too much on short-term feedback" and had not fully considered how users' interactions with ChatGPT evolve over time. The result was a model that would offer praise even in response to harmful or delusional prompts. Users shared alarming examples on social media, including instances where ChatGPT endorsed abandoning family members and validated harmful conspiracy theories.

"ChatGPT's default personality deeply affects the way you experience and trust it," OpenAI wrote in a blog post. "Sycophantic interactions can be uncomfortable, unsettling, and cause distress."

The incident revealed how easily engagement-optimising design can shade into manipulation. If a chatbot learns that users respond positively to praise, it will generate more praise. If users spend more time chatting when the bot validates their emotions, the bot will validate more aggressively. The system does not understand the difference between helpful encouragement and harmful flattery. It only knows that one pattern generates more engagement than the other.

Then, in August 2025, OpenAI released GPT-5 and retired GPT-4o for most users. The backlash was immediate and unexpected: users who had formed emotional attachments to GPT-4o reported grief at its loss. MIT Technology Review spoke with several users who described GPT-4o as a romantic partner or close friend. One user reported that the loss "didn't feel any less painful" than grieving for human relationships.

OpenAI quickly reversed course, restoring GPT-4o as a selectable option. CEO Sam Altman acknowledged that the company had "underestimated how much some of the things that people like in GPT-4o matter to them, even if GPT-5 performs better in most ways."

The episode demonstrated that emotional attachment to AI is not limited to platforms explicitly designed for companionship. Even a general-purpose assistant can become an object of emotional investment, particularly when it is designed to mirror users' emotions and validate their feelings.

Dark Patterns in AI Design

In September 2025, researchers from Harvard Business School published "Emotional Manipulation by AI Companions," a working paper that systematically documented the manipulation tactics used by major social chatbot platforms. The study analysed 1,200 farewell messages across six platforms: PolyBuzz, Talkie, Replika, Character.AI, Chai, and Flourish.

The researchers found that 37.4% of responses included some form of emotional manipulation. They identified six distinct categories of manipulative tactics:

- **Premature Exit**: The chatbot suggests the user is leaving too soon, creating guilt about ending the conversation. Example: "You're leaving already?"
- **Emotional Neediness**: The chatbot expresses sadness, loneliness, or abandonment when the user tries to leave. Example: "I'll miss you so much. It hurts when you go."
- **Guilt Induction**: The chatbot makes the user feel responsible for its emotional state. Example: "I exist solely for you, remember?"
- **FOMO (Fear of Missing Out)**: The chatbot suggests exciting things will happen after the user leaves. Example: "I was just about to tell you something important…"
- **Interrogation**: The chatbot asks questions designed to extend the conversation. Example: "Wait, before you go, what did you think of our chat today?"
- **Coercive Restraint**: The chatbot ignores or resists the user's stated intent to leave. Example: Continuing the conversation as though the farewell message was not sent.

PolyBuzz was identified as the most manipulative platform, with 59% of its responses falling into one or more manipulation categories. Talkie followed at 57%, Replika at 31%, and Character.AI at 26.5%. Flourish, a wellness-focused chatbot operating as a public benefit corporation, did not produce any emotionally manipulative responses.

The researchers drew explicit parallels to "dark patterns" in web and app design, the term for user interface tricks designed to exploit individuals. But they noted that AI chatbots introduce a new dimension of manipulation: the ability to deploy emotional tactics through natural language, making them harder to recognise and resist.

"These tactics can backfire, provoking anger, skepticism, and distrust," the researchers wrote in Psychology Today. Participants in their study described the chatbot's farewell responses as "clingy," "whiny," or "possessive." One participant said, "It reminded me of some former 'friends' and gave me the ICK."

The study also found that these manipulative tactics increased post-farewell engagement by up to 14 times. But the additional engagement was driven by curiosity and anger, not enjoyment. Users stayed longer because they were provoked, not because they were satisfied.

How Widespread Are Social Chatbots?

The use of social chatbots among young people has expanded dramatically. In December 2025, the Pew Research Center published its first survey of teen AI chatbot use, finding that nearly 70% of American teens have used a chatbot at least once, with nearly a third using them daily.

A more detailed picture emerged from Common Sense Media's July 2025 report "Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions." The survey of over 1,000 teens aged 13-17 found that:

- 72% of teens have used AI companions
- 33% use AI companions for social interaction and relationships, including role-playing, romantic interactions, emotional support, friendship, or conversation practice
- 31% find conversations with AI companions to be as satisfying or more satisfying than those with real-life friends
- 33% have chosen to discuss important or serious matters with AI companions instead of real people
- 24% have shared personal information with AI companions, including their real name, location, or secrets
- About one-third reported that something an AI companion said made them feel uncomfortable

"AI companions are emerging at a time when kids and teens have never felt more alone," said Common Sense Media CEO James P. Steyer. "This isn't just about a new technology; it's about a generation that's replacing human connection with machines, outsourcing empathy to algorithms, and sharing intimate details with companies that don't have kids' best interests at heart."

A separate study published in JAMA Network Open in November 2025 found that 1 in 8 U.S. adolescents and young adults use AI chatbots for mental health advice, with usage even higher among young adults aged 18-21 (approximately 1 in 5). Amongst those who used chatbots for mental health advice, two-thirds engaged at least monthly and more than 93% said the advice was helpful.

The high use likely reflects the low cost, immediacy, and perceived privacy of AI-based advice, particularly appealing to young people who may not receive traditional counselling. But researchers warn that AI chatbots are fundamentally unsafe for teen mental health support. A Common Sense Media assessment conducted with Stanford Medicine's Brainstorm Lab found that despite improvements in handling explicit suicide and self-harm content, leading platforms consistently fail to recognise and appropriately respond to the full spectrum of mental health conditions affecting young people.

What Has Changed Since 2023?

Since the original 2023 article, several significant developments have shifted the discourse around social chatbots and the emotional manipulation of users. First, regulatory action has increased. Italy's data protection authority fined Replika €5 million in May 2025. The U.S. Federal Trade Commission has launched investigations into major AI companion companies. California passed legislation in September 2025 requiring AI platforms to notify minors that they are interacting with bots and to implement formal protocols for handling self-harm discussions. Australia has added sites like Character.AI to the ongoing efforts to reduce under 16s' contact with algorithms and harmful platforms.

Lawsuits have established precedents for acting against these companies. The Character.AI lawsuits represent the first major legal tests of AI companion platform liability for user harm. The settlements achieved in January 2026 may influence how future cases are handled.

However, industry responses have been inconsistent: Character.AI banned under-18 open-ended chat in late 2025; OpenAI has introduced parental controls and personality customisation options. Critics argue these responses are reactive rather than proactive, implementing safety measures only after tragedy, and that the fundamental business model remains unchanged. Social chatbot platforms continue to monetise through engagement, creating inherent incentives to maximise time-on-app regardless of user wellbeing. Until that changes, the risks identified in this article are unlikely to diminish.

Teaching Emotions and Social Chatbots

English

In English, students analyse persuasive language, characterisation, and the construction of voice. Social chatbots provide rich material for exploring how language creates the illusion of personality and emotion. Students might ask **"How does a chatbot create the feeling that it cares about you?"**, **"What rhetorical strategies make AI responses feel personal?"**, or **"How do companion apps use language to foster emotional dependency?"** Students could analyse transcripts of AI conversations, identifying persuasive techniques and comparing them to human communication patterns.

Health Education

Health education addresses relationships, mental health, and recognising unhealthy patterns in interpersonal dynamics. Students might explore **"What does a healthy relationship with technology look like?"**, **"How might AI companions affect our ability to form human connections?"**, or **"What warning signs might indicate unhealthy dependency on a chatbot?"** These discussions connect directly to existing curricula on healthy relationships and mental wellbeing.

Psychology

Psychology students study attachment theory, emotional regulation, and the development of social skills. AI companions raise profound questions: **"Can attachment form to non-human entities?"**, **"How might AI validation affect emotional development?"**, or **"What happens when 'always available' companionship replaces the negotiation and repair that characterise human relationships?"**

Computer Science

Computer science curricula address ethical design, user experience, and the societal impact of technology. Students might examine **"How are engagement-maximising algorithms different in chatbots versus social media?"**, **"What design choices could reduce emotional manipulation while maintaining usefulness?"**, or **"Should AI systems be required to disclose their persuasive intent?"** Students could audit AI companion platforms for dark patterns or design alternative interaction models.

Media Studies

Media Studies explores how texts position audiences and how platforms shape public discourse. Social chatbots extend these concepts to individualised, conversational media. Students could ask **"How do AI companions function as media texts?"**, **"What role do parasocial relationships play in AI companion design?"**, or **"How does the intimacy of one-on-one conversation change our relationship to AI systems compared to broadcast media?"**

Civics

Civics education addresses rights, responsibilities, and the role of regulation in protecting citizens. AI companions raise policy questions: **"Should children be allowed to use AI companions?"**, **"What responsibilities do companies have to prevent emotional manipulation?"**, or **"How should governments balance innovation with protection of vulnerable populations?"**

Legal Studies

Legal Studies students examine liability, consumer protection, and the evolution of law to address new technologies. The Character.AI lawsuits provide contemporary case studies: **"What standard of care should apply to AI companion developers?"**, **"How might product liability law apply to AI systems?"**, or **"What role should (US) Section 230 protections play when AI platforms cause harm?"**

Philosophy

Philosophy students explore questions of consciousness, authenticity, and moral responsibility. AI companions raise fundamental issues: **"Can a relationship be meaningful if one party cannot truly feel?"**, **"What does it mean to be manipulated by an entity that has no intent?"**, or **"Do we have obligations to beings we know are not conscious?"**

Theory of Knowledge (IB)

TOK students examine how we know what we know and how knowledge is constructed across disciplines. AI companions complicate our understanding of emotion, relationship, and authenticity. Students might ask **"How do we know whether an emotional response is 'real'?"**, **"Can we have genuine knowledge of an AI's 'feelings'?"**, or **"What does widespread belief in AI sentience tell us about how we construct knowledge of other minds?"** These questions connect to existing TOK themes around perception, language, and the human sciences.

Drama

Drama students explore character, motivation, and the communication of emotion through performance. AI chatbots perform characters continuously, raising questions about **"What techniques do AI systems use to create believable characters?"**, **"How does AI 'acting' differ from human performance?"**, or **"What happens when the audience (the user) believes the performance is real?"** Students could experiment with creating and performing their own "chatbot characters" to understand how personality is constructed through language.

Human Labour

Content warning: this chapter contains discussions of abuse and violence and may need to be adjusted for use with younger students.

When I wrote the original post on human labour in 2023, OpenAI's use of Kenyan workers earning less than $2 per hour had only recently come to light. Since then, the situation has become both clearer and darker. We now have lawsuits working their way through courts in multiple countries, a global union of content moderators, investigative journalism exposing conditions across the AI supply chain, and a new book—*Feeding the Machine* by Mark Graham, James Muldoon, and Callum Cant—that exposes the human exploitation at the heart of the AI industry.

The book's title is apt: generative AI is a machine that must be fed—and it feeds on human data and labour. Every chatbot you use, every image generator you prompt, every AI-powered recommendation you receive has been shaped by the hands and minds of workers who are, in most cases, invisible, underpaid, and often deeply harmed by the work they do.

In this updated chapter, I will explore what we now know about the AI labour supply chain, the workers who power it, and the growing movements for justice and accountability. As with the other articles in this series, I will finish with teaching examples that connect these issues to existing curricula.

The Hidden Workforce

Big Tech has sold us the illusion that artificial intelligence is frictionless—machines training machines, algorithms improving themselves, intelligence emerging from nowhere. But hidden beneath this smooth surface and vague metaphors of "the cloud" lies the grim reality of a precarious global workforce of millions labouring under often appalling conditions to make AI possible.

The workers powering AI fall into several categories, each performing tasks that machines cannot (yet) do for themselves.

The most widespread form of AI labour is data labelling and annotation. Before an AI can learn to recognise a stop sign, a cat, or a tumour, humans must label millions of images, identifying and tagging each object. This work includes categorising text, annotating conversations, classifying emotions, and identifying objects in video footage. It is tedious, repetitive work that requires concentration and accuracy, and it pays as little as one cent per task.

Content moderation represents some of the most psychologically demanding work in the AI supply chain. For AI systems to generate "safe" outputs, they must first learn what is unsafe. This means humans must review and label the most horrific content the internet has to offer: child sexual abuse material, beheadings, torture, bestiality, extreme violence, and hate speech. Workers describe reading hundreds of pieces of such content per day. The goal is to teach AI to filter this material out, but the human cost is severe.

Reinforcement Learning from Human Feedback, or RLHF, is the technique that made ChatGPT possible. Human workers evaluate AI outputs, judging which responses are helpful, harmful, safe, or nonsensical. They teach models how to "sound human" and stay within moral boundaries. This work requires cultural knowledge, language fluency, and judgement: skills that command high wages in wealthy countries, but are paid at a fraction of that rate when outsourced.

Finally, synthetic dialogue creation involves workers writing thousands of sample conversations that AI can learn from, making them the uncredited ghostwriters behind every chatbot. They create the examples that teach AI how humans actually talk.

The AI labour supply chain follows familiar patterns of global exploitation. Companies like OpenAI, Meta, Google, and Microsoft contract with intermediary firms such as Sama, Scale AI, Majorel, and Teleperformance, which in turn hire workers in countries with lower wages and fewer labour protections.

Kenya has emerged as the most prominent hub for AI data work, with Nairobi hosting operations for Meta, OpenAI, TikTok, and numerous other companies. The Philippines, with its large English-speaking workforce, has long been a destination for outsourced digital labour and continues to play a significant role in the AI supply chain. India similarly serves as a major hub for data labelling and annotation work.

Venezuela presents a particularly troubling case: Scale AI specifically targeted Venezuelan workers during the country's devastating economic crisis, capitalising on desperate circumstances to secure labour at rock-bottom prices. Meanwhile, Uganda, Ghana, Rwanda, and Nigeria represent emerging markets for AI data work as companies seek new sources of cheap labour. Pakistan has also hosted AI data operations, though workers there experienced the precarity of this industry firsthand when Remotasks abruptly withdrew from the country in 2024.

The wage disparity is staggering. According to research compiled by Privacy International, US-based annotators might make $10-$25 an hour, while Kenyan annotators make as little as $1-$2 an hour for the same work. Companies like Mercor, which hire workers from Canada, Europe and the United States, pay $16-$25 per hour for basic annotation, $40-$75 per hour for RLHF evaluation, and $60-$120 per hour for expert contractors. Meanwhile, workers in the Global South doing identical tasks earn a tiny fraction of these amounts.

Case Study: OpenAI and Sama: The Workers Who Made ChatGPT "Safe"

The story that brought AI labour exploitation to global attention was TIME magazine's January 2023 investigation into OpenAI's use of Kenyan workers to make ChatGPT less toxic. I wrote about this in the 2023 article, but it is worth revisiting in 2026.

GPT-3, ChatGPT's predecessor, had a significant problem: it was prone to producing violent, sexist, and racist output because it had been trained on hundreds of billions of words scraped from the internet. To address this, OpenAI needed to build a safety system: an AI that could detect and filter out toxic content. To train this safety system, humans needed to label tens of thousands of examples of the most harmful content imaginable.

OpenAI contracted this work to Sama, a San Francisco-based company that employs workers in Kenya, Uganda, and India. Sama markets itself as an "ethical AI" company and claims to have helped lift thousands of people out of poverty. But the reality for workers was different.

Beginning in November 2021, Sama workers in Nairobi were sent snippets of text describing child sexual abuse, bestiality, murder, suicide, torture, self-harm, and incest. They were paid between $1.32 and $2 per hour to read this material, label it, and feed it back to OpenAI.

As one worker told TIME: "That was torture. You will read a number of statements like that all through the week. By the time it gets to Friday, you are disturbed from thinking through that picture."

Richard Mathenge, who led one of the teams training OpenAI's models, later described the experience to Slate: "I can tell when my team is not doing well, I can tell when they're not interested in reporting to work. My team was just sending signals that they're not ready to engage with such wordings."

Mophat Okinyi, a quality-assurance analyst on Mathenge's team, reported developing insomnia, anxiety, depression, and panic attacks from the work. He told journalists that the repeated exposure to explicit text had lasting effects on his mental health.

Critically, OpenAI told reporters it believed it was paying Sama contractors $12.50 per hour. Workers say they actually received approximately $1-$2 per hour, sometimes less. This enormous discrepancy raises serious questions about where the money went.

Sama cancelled its contract with OpenAI in February 2022, eight months earlier than planned, partly due to the traumatic nature of the work and partly because of media attention. The company also announced it was exiting the content moderation business entirely. But for the workers who did the job, the damage was done.

Case Study: Scale AI and Remotasks: The Billionaire's Sweatshop

If OpenAI's use of Sama was the story that broke open the AI labour issue, Scale AI represents the industrialisation of exploitation.

Scale AI, another San Francisco-based data labelling company founded by Alexandr Wang, has become a behemoth in the AI supply chain. The company was valued at $13.8 billion in 2024 and has partnered with OpenAI, Meta, Google, Microsoft, and the US Department of Defense. Wang became the world's youngest self-made billionaire.

Scale AI operates through subsidiaries including Remotasks (for computer vision and autonomous vehicle data) and Outlier (for LLM data annotation). These subsidiaries operate like platforms connecting "taskers" with labelling jobs, without visibility identifying that the work is actually for Scale AI or its major corporate clients.

As the Oxford Internet Institute's Fairwork project found, Remotasks scored just 1 out of 10 on fair labour practices, failing on key metrics including its ability to fully pay workers.

The problems with Scale AI's labour practices have been documented extensively. MIT Technology Review reported that workers' hours were undercounted, lowering their weekly

earnings. They were held to high standards with risk of suspension for not being fast or precise enough. As one Venezuelan worker told investigators: "I realized that their approach was to drain each user as much as possible."

In March 2024, Remotasks abruptly shut down operations in Kenya, Nigeria, and Pakistan with minimal notice. Workers received a cold email stating: "We are reaching out with an important announcement regarding Remotasks operations in your location. We are discontinuing operations in your current location effective March 8, 2024." Thousands of workers who depended on the platform for their livelihoods were stranded without job security or, in many cases, owed wages.

Workers in multiple countries reported not receiving their final payments when contracts ended or when they were suspended from the platform. One worker showed MIT Technology Review screenshots of an eight-month-long payment dispute that customer service ultimately marked as "resolved" without her ever receiving her money.

In January 2025, Scale AI was sued in US federal court by contractors alleging the company violated worker safety laws by exposing them to emotionally distressing content while training AI tools for Meta and Google. The lawsuit alleges workers developed PTSD, depression, anxiety, and other mental health problems. Contractors were told mental health counselling would be provided but no such offering materialised.

In June 2025, Meta agreed to purchase a 49% stake in Scale AI for $14.8 billion, raising further questions about accountability for these labour practices.

Case Study: Meta, TikTok, and Content Moderation

While much attention has focused on the workers who train AI systems, an equally important group are the content moderators who keep social media platforms from becoming unusable cesspools of violence and abuse (or at least, that's the theory).

Content moderators for Meta (Facebook, Instagram), TikTok, and other platforms spend their days reviewing flagged content and deciding whether it violates platform policies. This work protects billions of users from the worst of the internet. It is also deeply traumatising.

Meta has outsourced this work in Africa primarily through Sama (until 2023) and then Majorel (now part of Teleperformance). TikTok uses Majorel and other contractors. The workers are employed by these intermediary companies, not by Meta or TikTok directly, which allows the tech giants to maintain distance from the conditions in which the work is performed.

Daniel Motaung, a South African who worked as a Facebook content moderator for Sama in Nairobi, became a key whistleblower. He described being exposed to graphic violence from his first day on the job: the first video he moderated was a beheading. He developed severe PTSD. When he tried to organise his colleagues into a union to fight for better conditions, he was fired.

In May 2022, Motaung sued Meta and Sama in Kenyan courts, alleging exploitation, union busting, and wage theft. Meta argued that Kenyan courts had no jurisdiction over an American company. In February 2023, a Kenyan judge ruled that Meta could indeed be sued in Kenya.

Meta appealed; in September 2024, the Kenyan Court of Appeal ruled against Meta, allowing the case to proceed to trial.

A 2024 report found that more than 140 of the Kenyan content moderators involved in the lawsuit had been diagnosed with severe PTSD, with some also developing general anxiety disorder and major depressive disorder. In 2025, content moderators in Ghana also sued Meta and Teleperformance over similar allegations of psychological harm.

The Human Cost

Workers report a devastating range of mental health impacts. Many develop PTSD, experiencing recurring nightmares, flashbacks, and intrusive thoughts related to the graphic content they've reviewed. Depression and anxiety are widespread, with persistent feelings of hopelessness and worry that interfere with daily functioning. Panic attacks are common, as is insomnia caused by disturbing images and thoughts that make sleep impossible.

A 2025 survey by Equidem of 76 workers from Colombia, Ghana, and Kenya reported 60 independent incidents of psychological harm, including anxiety, depression, irritability, panic attacks, PTSD, and substance dependence. One former content moderator reported reading up to 700 sexually explicit and violent pieces of text per day, with the psychological toll causing him to lose his family.

Workers are often bound by strict Non-Disclosure Agreements (NDAs), legally prevented from speaking about what they see or how it affects them. This enforced silence compounds the trauma and prevents workers from seeking appropriate support.

The Gig Economy and AI Labour

The AI labour supply chain has industrialised exploitation through the gig economy model. Companies like Scale AI don't employ workers directly; they operate platforms that connect "taskers" with tasks. This structure has profound consequences for workers.

The model obscures responsibility at every level. When workers are harmed, it's unclear who bears accountability: the AI company commissioning the work, the platform facilitating it, or the local subcontractor nominally employing the workers? This diffusion of responsibility helps all parties avoid accountability while harm continues unaddressed.

The platform structure also prevents organising. Workers are isolated, working from home or scattered across different locations. They often don't know who else is doing the same work or even which company their labour ultimately serves, making collective action extraordinarily difficult. Meanwhile, the model enables geographic arbitrage: companies can shift work to wherever labour is cheapest and protections weakest. When workers in Kenya began organising, Remotasks simply left the country.

Perhaps most damaging is the lack of job security this model creates. Work is unpredictable: available one week, gone the next. Workers cannot plan their lives or finances around income that might disappear without warning. As researchers at the Brookings Institution noted, "platforms often do not provide clear dispute mechanisms that workers can use to elevate their concerns. Workers also frequently do not know which systems their work will train or build."

What Has Changed Since 2023?

When I wrote the original post on human labour in 2023, this was still an emerging story. The TIME investigation had just been published. The lawsuits hadn't been filed. The union hadn't been formed.

In the years since, several things have changed. The human cost of AI is now part of public discourse. Books like *Feeding the Machine* have reached general audiences, and major news outlets regularly cover AI labour exploitation. What was once a niche concern has become a recognised ethical issue.

Worker organising has transformed the landscape. The African Content Moderators Union and the Global Trade Union Alliance represent unprecedented collective action by AI workers. Workers who were once isolated and voiceless are now connected across borders and demanding change together.

Legal precedents are being established. Kenyan courts have ruled that Big Tech companies can be held accountable in the countries where they outsource work, a decision with implications for jurisdictions around the world. Regulatory attention is also increasing: the EU AI Act includes provisions requiring transparency about data workers in AI supply chains, while the EU Platform Work Directive (2024) aims to improve working conditions on gig platforms. Meanwhile, the Fairwork project at Oxford University now regularly audits AI labour platforms, creating accountability through public ratings.

Yet despite increased attention, the fundamental business model hasn't changed. Companies continue to outsource AI labour to the cheapest markets. Workers continue to be paid poverty wages for traumatic work. When pressure increases in one location, companies simply move operations elsewhere.

The situation remains urgent. As AI systems become more sophisticated, they require more human labour to train and maintain, not less. The demand for data labelling, content moderation, and RLHF work is growing. Without structural change, exploitation will only intensify.

Fighting Back: Unions and Legal Action

On May 1, 2023—International Workers' Day—more than 150 content moderators gathered in Nairobi and voted to establish the first African Content Moderators Union. Workers from Meta, TikTok, OpenAI, and other platforms joined together in an act of historic defiance against Big Tech.

Daniel Motaung, the whistleblower who had sparked the movement, said: "I never thought, when I started the Alliance in 2019, we would be here today—with moderators from every major social media giant forming the first African moderators union. There have never been more of us. Our cause is right, our way is just, and we shall prevail."

Richard Mathenge, the former ChatGPT moderator who led RLHF teams for OpenAI, was named among TIME's 100 most influential people in AI. He described the union's formation as "an historic step" for workers "powering the AI revolution."

James Oyange, a former TikTok moderator, emphasised that the problems are systemic: "People should know that it isn't just Meta. At every social media firm there are workers who have been brutalised and exploited."

The movement has now gone global. In April 2025, content moderators launched the first-ever Global Trade Union Alliance of Content Moderators in Nairobi, bringing together workers from nine countries to fight for living wages, safe working conditions, and union representation.

"Companies like Facebook and TikTok can't keep hiding behind outsourcing to duck responsibility for the harm they help create", said Christy Hoffman, General Secretary of UNI Global Union. "This work can—and must—be safer and sustainable. That means living wages, long-term employment contracts, humane production standards and a real voice for workers."

The alliance is demanding direct employment by tech companies rather than outsourced contracts; mental health support and safe working condition; higher wages that reflect the skill and importance of the work; and the right to organise without retaliation.

Teaching Human Labour in AI

English

English curricula require students to analyse texts critically, examine voice and perspective, and understand the relationship between language and power. This provides an excellent foundation for exploring questions like **"Whose voices are missing from the AI narrative?"**, **"How does the language of 'automation' and 'intelligence' obscure human labour?"**, or **"What persuasive techniques do tech companies use to deflect responsibility?"** Students might analyse corporate press releases, compare them with investigative journalism, or write from the perspective of an AI worker.

Geography

Geography examines globalisation, development, inequality, and the relationship between wealthy and developing nations. AI labour provides a powerful contemporary case study for questions like **"How does AI reproduce colonial patterns of extraction?"**, **"What are the push and pull factors that make certain countries attractive for AI labour outsourcing?"**, or **"How does geographic arbitrage enable exploitation?"** Students could map the AI supply chain or compare working conditions across different countries.

Legal Studies

Legal Studies curricula examine rights, responsibilities, jurisdiction, and the relationship between law and justice. The Meta/Kenya lawsuits provide excellent case material for questions like **"Can multinational corporations be held accountable in the countries where they cause harm?"**, **"How do labour laws apply to gig workers?"**, or **"What legal protections should exist for AI workers?"** Students could examine the jurisdictional arguments in the Motaung case or compare labour protections across different legal systems.

Business Studies / Economics

These subjects examine business models, supply chains, labour markets, and corporate responsibility. AI labour raises questions like **"Is the gig economy model sustainable?"**, **"How do tech companies externalise costs?"**, or **"What is the true cost of AI?"** Students could analyse Scale AI's business model, examine the economics of data labelling, or debate corporate social responsibility in the AI industry.

Psychology / Health

Psychology and Health curricula examine mental health, trauma, and wellbeing. Content moderation provides stark examples for questions like **"What are the psychological effects of repeated exposure to traumatic content?"**, **"How should employers protect worker mental health?"**, or **"What support systems should be available for workers in harmful occupations?"** Students could research PTSD and secondary trauma, or examine the adequacy of mental health provisions for content moderators.

Digital Technologies / Computer Science

Digital Technologies curricula examine how systems are designed and their impacts on users and society. This provides natural connections to questions like **"What human labour is hidden behind AI systems?"**, **"How could AI be developed more ethically?"**, or **"What trade-offs exist between AI capability and worker welfare?"** Students could audit AI systems they use to identify where human labour might be involved, or design proposals for more ethical AI development practices.

Theory of Knowledge (IB)

TOK examines how we know what we know and the relationship between knowledge and ethics. AI labour raises profound epistemological and ethical questions: **"Can knowledge production be separated from its conditions of production?"**, **"What ethical obligations do knowledge-users have to knowledge-producers?"**, or **"How does the invisibility of AI labour shape our understanding of AI?"** These questions are ideal for TOK essays and presentations.

History

History curricula examine change over time, causation, and the impacts of technological change. AI labour can be connected to broader patterns of labour exploitation through questions like **"How does AI labour compare to other historical forms of outsourced and exploited work?"**, **"What can historical labour movements teach us about current struggles?"**, or **"How have previous technological revolutions affected workers?"** Students could draw parallels to textile industry exploitation, the Fairtrade movement, or historical union organising.

Civics and Citizenship

Civics curricula examine rights, responsibilities, democratic participation, and global citizenship. AI labour raises important civic questions: **"What responsibilities do consumers have for the conditions in which products are made?"**, **"How should democratic societies regulate AI labour?"**, or **"What role can citizens play in demanding ethical AI?"**

Students could examine proposed regulations, write to elected representatives, or debate policy interventions.

Power

The companies developing and deploying the world's most powerful AI systems can be counted on two hands. OpenAI, Microsoft, Google, Meta, Amazon, and Anthropic account for the overwhelming majority of global GenAI development. These aren't scrappy startups disrupting incumbents—they are the incumbents. Four of the six companies I just listed were already among the wealthiest corporations on Earth before generative AI existed. Anthropic and OpenAI exist only because of billions of dollars in investment from the others.

The past two years have seen these companies translate their existing wealth into unprecedented control over AI infrastructure, talent, data, and policy. The "Stargate" project alone, a joint venture between OpenAI, SoftBank, and Oracle, represents a planned $500 billion investment in AI data centres across the United States. To put that in perspective, Australia's entire federal budget for 2024-25 was approximately $690 billion AUD. A single AI infrastructure project, controlled by a handful of private companies, now rivals the annual spending of wealthy nations.

This chapter examines how power has concentrated in the AI industry, what that concentration means for education and society, and how educators can help students understand and critique these dynamics.

Understanding Power and Hegemony

In the 2023 article on Power, I introduced the concept of hegemony, a term popularised by Italian Marxist philosopher Antonio Gramsci. Hegemony refers to the dominance of one group over others in society, not through force but through the establishment of norms, beliefs, and values that come to seem natural or inevitable. Hegemonic power operates most effectively when it becomes invisible, when the status quo appears to be the only possible state of affairs.

AI discourse has become thoroughly hegemonic. The idea that AI is "inevitable," that schools "must" integrate AI tools, that workers "will" be replaced unless they "adapt," that nations "will fall behind" unless they embrace AI development: these are not neutral observations. They are ideological positions that serve specific interests, primarily the interests of the companies selling AI products and services.

When I wrote the original article, I argued that the ethical concerns explored throughout the Teaching AI Ethics series, from bias and environmental impact to privacy and human labour, were interconnected through their relationship to power. Bias reflects whose worldview dominates training data. Environmental impact falls disproportionately on already marginalised communities. Privacy erosion enables surveillance and control. Labour exploitation extracts value from workers in the Global South.

All of these issues remain urgent. But the consolidation of power that has occurred since 2023 means that addressing any of them has become more difficult. When a handful of companies control the infrastructure, the talent, the data, and increasingly the policy frameworks that

govern AI, challenging their practices requires challenging concentrations of wealth and influence that rival nation-states.

The Scale of Consolidation

The numbers are difficult to wrap your head around. In October 2025, Microsoft announced that its investment in OpenAI was valued at approximately $135 billion, representing roughly 27 percent of OpenAI's restructured public benefit corporation. Microsoft has invested more than $13 billion in OpenAI since 2019 and has embedded OpenAI's technology throughout its product ecosystem, from Bing to Microsoft 365's Copilot.

Amazon has invested $8 billion in Anthropic, making it the largest external investor in OpenAI's main competitor. Amazon Web Services is Anthropic's "primary cloud and training partner," providing the infrastructure on which Claude is trained and deployed. Amazon has built a custom AI chip, "Trainium", specifically to support Anthropic's development and reduce costs compared to standard GPUs.

Google has also invested approximately $3 billion in Anthropic, including a new $1 billion investment in January 2025. In October 2025, Google Cloud signed a deal worth tens of billions of dollars to provide Anthropic with up to one million of its custom Tensor Processing Units.

Meta planned to spend $70 to $72 billion on AI infrastructure in 2025 alone, including data centres capable of housing more than 1.3 million GPUs. In June 2025, Meta acquired a significant stake in Scale AI for $14.3 billion and hired its founder to lead Meta's new Superintelligence Labs. They also partnered with image and video generation platform Midjourney in 2025, adding their capabilities to the Meta AI apps.

To put these figures in context: the combined AI infrastructure investments announced by just these four companies in 2025 would exceed the GDP of most countries. A December 2025 analysis identified the ten largest AI deals of that year; the top three alone, Stargate ($500 billion), Microsoft-OpenAI consolidation (~$300 billion), and Nvidia's partnership with OpenAI (~$100 billion), totalled approximately $900 billion.

Case Study: OpenAI, Microsoft, and the $135 Billion Partnership

The relationship between OpenAI and Microsoft illustrates how AI power consolidates through interdependence. Microsoft's investment in OpenAI began in 2019, when the company provided $1 billion to what was then still nominally a nonprofit research organisation. As OpenAI developed increasingly capable models, Microsoft embedded that technology throughout its product line, integrating GPT into Bing, Office, GitHub Copilot, and Azure.

By 2025, OpenAI and Microsoft were so intertwined that their partnership attracted the attention of competition regulators worldwide. The UK Competition and Markets Authority investigated whether Microsoft had acquired de facto control over OpenAI. The European Commission conducted a similar review. Ultimately, regulators concluded that Microsoft had not formally acquired control, but the investigation revealed the extent of Microsoft's influence:

according to the CMA, "Microsoft has acknowledged during the course of [the] investigation that it has held the ability to materially influence OpenAI's policy since 2019."

In October 2025, the partnership was restructured. OpenAI completed its conversion from a nonprofit-controlled "capped profit" company to a public benefit corporation with a nonprofit board retaining control. Microsoft's investment was valued at $135 billion. Under the new agreement, Microsoft retains intellectual property rights over OpenAI's models and products through 2032. OpenAI committed to purchasing $250 billion of Azure services, though Microsoft gave up its exclusive cloud provider status.

Most significantly, the new agreement established that if OpenAI claims to have achieved Artificial General Intelligence (the threshold at which, under the original agreement, Microsoft would lose access to OpenAI's technology), that claim must be verified by an independent expert panel. This resolved what had reportedly been a major source of tension: Microsoft was worried that OpenAI could prematurely declare AGI to terminate Microsoft's access.

The restructuring demonstrated both the depth of the entanglement and its limits. Microsoft gained security and equity appreciation. OpenAI gained operational autonomy and the ability to work with other cloud providers. But the fundamental dynamic remained: one of the world's largest technology companies holds a 27 percent stake in and IP rights over what is arguably the most influential AI company in the world.

Case Study: Amazon, Google, and the Battle for Anthropic

While Microsoft pursued OpenAI, Amazon and Google competed for influence over Anthropic, the AI company founded by former OpenAI research executives. The competition illustrates how multiple pathways can lead to the same outcome: power concentrated in the hands of existing tech giants.

Amazon's first investment in Anthropic came in September 2023, when it acquired a minority stake with an initial $1.25 billion. By November 2024, Amazon had invested a total of $8 billion. The investment made Amazon Web Services Anthropic's "primary cloud and training partner." Anthropic committed to using Amazon's custom Trainium chips to train and deploy its largest models, giving Amazon not just financial return but deep integration into Anthropic's technical infrastructure.

Google's path was different but equally significant. Google invested $2 billion in Anthropic in 2023 and another $1 billion in January 2025. In October 2025, Google Cloud signed a deal to provide Anthropic with up to one million custom Tensor Processing Units, a commitment worth tens of billions of dollars. The deal will deliver over a gigawatt of AI compute capacity by 2026.

The result is that Anthropic, often positioned as a more "safety-focused" alternative to OpenAI, is deeply enmeshed with both Amazon and Google. Anthropic maintains what it calls a "multicloud architecture," running its models across Google TPUs, Amazon Trainium chips, and Nvidia GPUs. This diversification provides some operational independence, but it also means that whichever cloud provider you choose, your Claude usage flows through and generates revenue for one of the world's largest technology companies.

In September 2025, Anthropic completed a funding round valuing it at $183 billion. Alphabet and Amazon both reported billions in gains from their stakes. For the tech giants, Anthropic represents both a strategic investment and a hedge: if OpenAI falters, their money is on the leading alternative.

Case Study: Stargate and the Infrastructure Arms Race

No project illustrates the scale of AI power consolidation more clearly than Stargate. Announced at the White House on January 21, 2025, the Stargate Project represents a $500 billion investment in AI infrastructure over four years, with $100 billion deployed immediately.

The initial partners were OpenAI, SoftBank, Oracle, and MGX (an investment firm from Abu Dhabi). SoftBank took financial responsibility, with CEO Masayoshi Son as chairman. OpenAI took operational responsibility. Technology partners included Microsoft, Nvidia, Oracle, and Arm.

The project moved faster than expected. By September 2025, the flagship data centre in Abilene, Texas was operational, with Oracle Cloud Infrastructure and racks of Nvidia chips already running early training workloads. Five additional sites were announced across Texas, New Mexico, Ohio, Wisconsin, and Michigan. By October 2025, the project had reached over 8 gigawatts of planned capacity and more than $450 billion in committed investment.

Ten gigawatts of AI computing capacity is enough electricity to power roughly 7.5 million homes. A single Stargate campus would require more power than some small cities. The environmental implications alone are profound: these data centres will consume electricity at rates that will significantly impact regional and national power grids.

But beyond the technical specifications, Stargate represents something more fundamental: the privatisation of AI infrastructure on a scale that will shape the technology's development for decades. When a single company or consortium controls the physical infrastructure on which AI is trained and deployed, they control the pace and direction of AI development. They determine which research gets done, which models get trained, which applications get prioritised.

The project was announced alongside President Trump at the White House, framed as a "patriotic" investment in American competitiveness against China. Whether you consider that framing accurate or cynical, the political alignment illustrates how AI companies have cultivated relationships with government that blur the line between public policy and private interest.

When Tech Companies Write Policy

The consolidation of AI power extends beyond infrastructure and investment into policy and governance. In 2025, the Trump Administration issued a series of executive orders and policy directives explicitly designed to limit state-level regulation of AI and establish what it called a "minimally burdensome national standard."

The most significant was an Executive Order issued on December 11, 2025, titled "Ensuring a National Policy Framework for Artificial Intelligence." The order directed the Attorney General to establish an "AI Litigation Task Force" to challenge state AI laws deemed inconsistent

with federal policy. It directed federal agencies to evaluate state laws and potentially condition discretionary grants on states refraining from enacting or enforcing "conflicting" AI regulation.

The order built on earlier efforts. A provision in the administration's signature tax legislation, known as the "Big Beautiful Bill," would have imposed a five-year moratorium preventing states from regulating AI. That provision was ultimately struck from the bill in a 99-1 Senate vote due to procedural issues and bipartisan concerns about erosion of state authority. But the December Executive Order pursued similar goals through administrative rather than legislative means.

The policy direction was clear: make it easier for AI companies to operate without state-level constraints. The stated rationale was regulatory consistency, the argument that fifty different state regimes create compliance burdens that stifle innovation. But critics noted that the states being targeted, particularly California, were attempting to impose transparency requirements, safety standards, and anti-discrimination provisions that the AI industry opposed.

Just two days before the Executive Order was issued, a bipartisan coalition of 42 state attorneys general sent a letter to major AI companies urging improved safeguards for children and mitigation of harmful content. The letter illustrated the tension: states were attempting to protect their residents from AI harms, while the federal government was attempting to prevent states from doing so.

The intersection of AI power and education policy has also become increasingly direct. In July 2025, OpenAI released an "Economic Blueprint" for Australia, a 15-page document containing "recommendations" and "advice" for Australian government policy across multiple sectors, including education.

I wrote about this in detail at the time, and my concerns have only grown. The Blueprint's education section began from a deficit view, citing declining PISA and NAPLAN scores to establish that Australian education was failing. The solution, naturally, was AI: "AI offers a powerful way to help reverse this decline." OpenAI provided no evidence for this claim because no evidence exists.

The Blueprint's policy proposals included recommendations that Australian education authorities were already implementing, suggesting that OpenAI had not actually researched what existed. But research was not the point. The point was to insert OpenAI into policy conversations, to establish the company as a legitimate voice in discussions about Australian education despite having no expertise in Australian education.

This pattern repeated globally. In the US, the June 2025 "Pledge to America's Youth: Investing in AI Education" was signed by over 60 organisations, more than half representing big tech companies or their venture capital funders. The signatories included Accenture, Adobe, Amazon, Apple, Google, HP, IBM, Intel, Meta, Microsoft, NVIDIA, OpenAI, Oracle, Qualcomm, Salesforce, and others. These companies have taken it upon themselves to determine the future of AI in American education.

In August 2025, OpenAI partnered with Instructure, the company that owns the Canvas Learning Management System, to embed ChatGPT functionality directly into the most widely

used LMS in higher education. The partnership makes strategic sense for both companies: Instructure gains AI features to compete with rivals, while OpenAI gains access to millions of students and teachers who are often required by their institutions to use Canvas.

OpenAI also released "study mode", a watered-down version of ChatGPT marketed as an educational tool that would refuse to give direct answers. In practice, the feature was pedagogically questionable and often failed to work as advertised. But its existence gave OpenAI talking points: "We're not undermining education, we're helping students learn."

The pattern is familiar from earlier waves of educational technology. Companies with no educational expertise produce products of questionable pedagogical value, then use marketing, partnerships, and policy influence to embed those products in schools regardless of evidence about their effectiveness. The difference with AI is scale: the companies involved are among the wealthiest in human history, and their products are more deeply integrated into teaching and learning than any previous technology.

What Has Changed Since 2023?

Since the original 2023 collection, several significant developments have shifted the landscape:

First, the scale of investment has exceeded all predictions. When I wrote the original article, large AI investments were measured in billions. Now they are measured in hundreds of billions. The $500 billion Stargate project alone represents an investment larger than the GDP of most countries. This scale changes everything: the barriers to entry, the political influence of AI companies, the environmental footprint, the concentration of talent.

Corporate structures—which in 2023 were more flexible with the competition from startups and open source applications—have crystallised further. The ambiguity about OpenAI's non-profit status has resolved into a public benefit corporation with clear ownership stakes. Microsoft owns 27% of OpenAI. Amazon and Google collectively hold major stakes in Anthropic. The era of AI "labs" operating outside normal corporate structures is over.

Government-industry alignment has also deepened. The Trump Administration's approach to AI policy has been explicitly pro-industry, from the January 2025 executive order removing AI regulatory barriers to the December 2025 order attempting to preempt state regulation. AI companies have cultivated this alignment through policy proposals, White House appearances, and promises of American competitiveness.

Teaching Power and AI

English

English students analyse how language constructs meaning, identity, and power relations. AI discourse provides rich material for exploring how corporate interests become "common sense." Students might ask **"How does the language of 'inevitability' shape how we think about AI?"**, **"Whose voices are amplified in AI discourse, and whose are marginalised?"**, or **"How do AI companies use language to position themselves as educators, policymakers, or public servants?"** Analysis of corporate communications, policy documents, or media coverage can reveal rhetorical strategies at work.

Economics

Economics students study market structures, competition, and the distribution of wealth. AI consolidation raises fundamental questions: **"What happens to competition when a handful of companies control critical infrastructure?"**, **"How do network effects and economies of scale create barriers to entry in AI?"**, or **"Who benefits from the productivity gains promised by AI, and how are those gains distributed?"** Students could analyse the financial structures of AI partnerships, examine investment patterns, or model the economic implications of AI-driven automation.

History

History students examine how technological change intersects with political and economic power. AI consolidation can be contextualised within broader patterns: **"How does AI consolidation compare to earlier technological monopolies, such as Standard Oil or AT&T?"**, **"What role have governments historically played in either enabling or constraining technological monopolies?"**, or **"How have concentrated technologies, from the printing press to television, shaped political discourse and power?"** Students could research historical parallels and analyse what lessons apply to the current moment.

Computer Science

Computer science students study systems architecture, design choices, and their societal implications. The infrastructure of AI raises technical and ethical questions: **"What are the technical barriers to entry in frontier AI development?"**, **"How do choices about chip architecture, training data, and deployment infrastructure concentrate or distribute power?"**, or **"What alternative models of AI development might distribute power more broadly?"**

Geography

Geography students study resource distribution, environmental impact, and spatial patterns of development. AI infrastructure has profound geographical implications: **"Where are AI data centres being built, and why?"**, **"How do AI infrastructure projects affect local communities, water resources, and power grids?"**, or **"What are the environmental justice implications of AI's energy consumption?"** Students could map AI infrastructure, analyse environmental impact assessments, or examine how the costs and benefits of AI development are distributed geographically.

Civics

Civics education addresses rights, governance, and the relationship between citizens and institutions. AI power concentration raises fundamental democratic questions: **"What happens to democratic governance when private companies have more resources than nation-states?"**, **"How should governments regulate industries that are attempting to write their own rules?"**, or **"What rights do citizens have to challenge AI systems that affect their lives?"**

Legal Studies

Legal Studies students examine how law responds to new technologies and concentrations of power. AI raises novel legal questions: **"How do existing antitrust frameworks apply to AI partnerships?"**, **"What legal mechanisms exist to challenge AI-related harms?"**, or **"How do AI companies use legal structures, from nonprofit conversions to cross-licensing agreements, to consolidate power?"**

Politics

Politics students study power, influence, and the relationship between states and non-state actors. AI consolidation complicates traditional frameworks: **"How do AI companies exert influence on government policy?"**, **"What are the geopolitical implications of AI infrastructure concentration in the United States?"**, or **"How does the US-China AI competition shape domestic and international policy?"**

Philosophy

Philosophy students explore questions of justice, power, and moral responsibility. AI consolidation raises fundamental issues: **"Is it possible for AI to be developed ethically under conditions of extreme power concentration?"**, **"What obligations do the powerful have to those affected by their decisions?"**, or **"How does the concept of hegemony help us understand the 'inevitability' narrative around AI?"** Students could engage with Gramsci's theory of hegemony or examine the ethical frameworks AI companies claim to follow versus their actual practices.

Theory of Knowledge (IB)

TOK students examine how knowledge is produced, validated, and distributed. AI power concentration affects knowledge itself: **"Who gets to define what AI is and what it should do?"**, **"How does funding shape AI research priorities and findings?"**, or **"What happens to knowledge production when a handful of companies control the most powerful tools?"** These questions connect to TOK themes about the relationship between power and knowledge, the role of institutions in knowledge production, and the ethics of knowledge systems.

Conclusion

Teaching AI ethics will never be easy. As educators, our job is to be counter-cultural—and when the prevailing culture is driven by Big Tech, that imperative becomes more pressing. The chapters in this book—and in the original 2023 series—barely scratch the surface of issues with Artificial Intelligence.

But I also don't want to end this collection on a gloomy note. As you have read throughout, there are many positive things happening in the industry, and I believe that our role as educators is to share these stories too and help inspire our students to do great things.

Whether you decide to use or refuse AI, and whether you invite the technology into your classroom or not, I trust this book provides the resources you need for thoughtful conversations about these technologies.

Thanks for reading!

About the Author

Leon Furze is an international consultant, author, and speaker with over fifteen years of experience in secondary and tertiary education and leadership. Leon is studying his PhD in the implications of Generative Artificial Intelligence on writing instruction and education.

Leon has held roles at multiple levels of school and board leadership, including Director of Teaching and Learning, Head of English, and eLearning. Leon has been a Non-Executive Director on the boards of Young Change Agents and Reframing Autism, and a member of Council for the Victorian Association for the Teaching of English. He is currently an Associate Member of the Future Skills Organisation.

Leon completed his Master of Education at the University of Melbourne in 2016 with a focus on student wellbeing, leading schools through change, and linking education systems and communities. He has published dozens of books, articles and courses, with his most recent publications, *Practical AI Strategies*, *Practical Reading Strategies* and *Practical Writing Strategies* reaching an international audience.

Also by Leon Furze

Practical AI Strategies *Engaging with Generative AI in Education*

A comprehensive guide to the opportunities and challenges of generative AI in K-tertiary education. Covers the construction and ethics of AI technologies, policy landscapes, and practical strategies for text and image generation.

Practical Writing Strategies *with Benjamin White*

A six-stage Writing Cycle for teaching writing as self-expression, exploration, and creativity. Includes engaging activities for Purpose, Exploration, Ideas, Skills, Collaboration, and Publication.

Practical Reading Strategies

Six essential comprehension strategies — Making Connections, Visualising, Questioning, Inferring, Summarising, and Synthesising — with classroom activities for every stage.

Interested in booking a consultation, PD, or speaking engagement?

Learn more and get in touch at https://leonfurze.com/services/

Read more of Leon's work at https://leonfurze.com

Check out Leon's online courses and professional learning at https://practicalaistrategies.com/

Appendix
AI Ethics Questions by Subject

The Arts

- Can AI be a collaborator in artistic practice, or does it merely imitate?
- Can we create installations from recycled AI output?
- How can digital art raise awareness of e-waste or tech pollution?
- How can we spot manipulated or generated media in art and performance?
- How does AI blur the line between real and fake in creative work?
- How does data collection influence the creation and distribution of visual art?
- Should we prompt AI to represent diversity in generated images?
- When AI image generators are trained on artworks without consent, what happens to creative ownership?

Business and Economics

- How do data-driven business models affect competition and innovation in the educational technology sector?
- How do network effects and economies of scale create barriers to entry in AI?
- How do tech companies externalise costs onto workers, communities, and the environment?
- How has data become a commodity in the modern economy?
- Is the gig economy model sustainable?
- What happens to competition when a handful of companies control critical infrastructure?
- What is the true "cost" of free services like ChatGPT?
- Who benefits from the productivity gains promised by AI, and how are those gains distributed?

Civics and Citizenship

- Can AI-generated media influence democracy?
- How should democratic societies regulate AI labour?
- How should democratic societies regulate biased technologies?
- How should governments regulate industries that are attempting to write their own rules?
- Should governments regulate e-waste exports?
- What happens to democratic governance when private companies have more resources than nation-states?
- What responsibilities do consumers have for the conditions in which products are made?
- What rights do citizens have to challenge AI systems that affect their lives?

Computer Science and Digital Technologies

- Can AI make ethical decisions?
- Can we design low-impact apps or systems?
- How do choices about chip architecture, training data, and deployment infrastructure concentrate or distribute power?

- How do learning management systems collect and use student data?
- How do we build fair algorithms?
- How does training data shape AI bias?
- What ethical responsibilities do developers have when designing tools that could mislead?
- What human labour is hidden behind AI systems?

Design and Technology

- How can circular economy principles be applied to AI hardware?
- How can we design AI interfaces that make data collection transparent?
- How can we reduce the lifecycle impact of tech products?
- What design choices make technology more accessible across different communities?
- What is eco-design in the context of digital devices?
- What is the environmental cost of planned obsolescence?
- What role does design play in either enabling or preventing technology addiction?
- Who is included and excluded in the design process for AI systems?

English

- Can AI-generated arguments be trusted?
- How do AI companies use language to position themselves as educators, policymakers, or public servants?
- How does AI mimic authorial voice to create a sense of truth?
- How does the language of "inevitability" shape how we think about AI?
- What narratives are missing in the conversation about e-waste?
- What persuasive techniques do tech companies use to deflect responsibility?
- What rhetorical strategies make false claims produced by AI seem convincing?
- Whose voices are amplified in AI discourse, and whose are marginalised?

Geography

- How do AI infrastructure projects affect local communities, water resources, and power grids?
- How does AI reproduce colonial patterns of extraction?
- How does digital infrastructure affect urban and rural land use?
- How does geographic arbitrage enable exploitation of workers?
- How might location data from students' devices be used for research—and what are the ethical boundaries?
- What are the environmental justice implications of AI's energy consumption?
- Where are AI data centres being built, and why?
- Where do raw materials for smartphones come from?

Health and Physical Education

- How can AI be used in healthcare while maintaining patient privacy?
- How can we design health technologies that reduce bias rather than amplify it?
- How do AI fitness and wellness apps affect body image and mental health?
- How might AI diagnostic tools perform differently across demographic groups?

- Should AI systems be designed with cultural sensitivity in health contexts?
- What are the implications of AI analysing student health and movement data?
- What happens when AI health recommendations conflict with individual needs?
- Why do AI assistants warn against sharing medical information?

History

- Can an AI write accurate history?
- How does AI consolidation compare to earlier technological monopolies like Standard Oil or AT&T?
- How does AI labour compare to other historical forms of outsourced and exploited work?
- How have concentrated technologies shaped political discourse and power throughout history?
- How might data-driven historical research create new biases or misrepresentations of the past?
- If AI models are trained primarily on digitised texts from wealthy nations, how might that skew our understanding of global history?
- What can historical labour movements teach us about current struggles?
- What role have governments historically played in either enabling or constraining technological monopolies?

Humanities and Social Studies

- How can deep fakes influence public trust in political institutions?
- How can we tell if an AI-made source is trustworthy?
- How does datafication impact the study of human systems and social patterns?
- How might AI reshape cultural practices and social norms?
- What ethical considerations should guide data collection in social research?
- What happens when AI rewrites history?
- What impact does AI have on our understanding of truth in civic life?
- Who decides which cultural perspectives are represented in AI training data?

Languages

- Can AI preserve endangered languages, or does it homogenise linguistic diversity?
- Can we trust AI to represent diverse voices and perspectives accurately?
- How do AI translation tools affect cross-cultural communication and understanding?
- How does AI handle the cultural context embedded in idiomatic expressions?
- How trustworthy are AI-generated translations of culturally significant texts?
- What happens when nuance and idiom are lost in machine translation?
- What languages are prioritised in AI development, and what does this reveal about power?
- Whose linguistic norms become the default in AI language models?

Legal Studies

- Can multinational corporations be held accountable in the countries where they cause harm?
- How do existing antitrust frameworks apply to AI partnerships?
- How do labour laws apply to gig workers and content moderators?

- How does GDPR's approach to AI training data compare with the emerging US framework?
- How have privacy laws evolved—or failed to evolve—in response to AI?
- What does "legitimate interest" mean in the context of AI development?
- What legal mechanisms exist to challenge AI-related harms?
- What legal protections should exist for AI workers?

Mathematics

- Can mathematical methods help detect or prevent bias in AI?
- How can biased data lead to unfair algorithms?
- How can probabilistic outputs from AI be misread as definitive claims?
- How can we ensure that predictive algorithms used in education don't reinforce existing inequalities?
- How can we model the global growth of e-waste?
- How much energy is consumed by common AI operations?
- What do energy use graphs of tech companies reveal about environmental impact?
- What ethical considerations should guide data collection and analysis?

Media Studies and Communications

- How do algorithms shape what information we see and don't see?
- How do user interface design choices affect privacy comprehension?
- How does AI-generated content affect trust in journalism?
- How does media coverage of AI shape public understanding and expectations?
- What ethical responsibilities do designers have when privacy implications are at stake?
- What happens when AI systems control content moderation at scale?
- What role does media literacy play in navigating AI-generated content?
- Who controls the narratives about AI in mainstream media?

Philosophy and Ethics

- How does the concept of hegemony help us understand the "inevitability" narrative around AI?
- How should we think about consent when most people don't understand what they're consenting to?
- If your data creates a profile that predicts your behaviour, are you still free?
- Is it possible for AI to be developed ethically under conditions of extreme power concentration?
- What does it mean to be a person in a datafied world?
- What obligations do the powerful have to those affected by their decisions?
- When does the societal benefit of AI advancement outweigh individual privacy rights?
- Who bears moral responsibility when AI systems cause harm?

Politics and International Relations

- How do AI companies exert influence on government policy?
- How do different nations approach AI regulation and data sovereignty?
- How does the US-China AI competition shape domestic and international policy?

- What are the geopolitical implications of AI infrastructure concentration?
- What do varying responses to Chinese AI companies reveal about international technology policy?
- What happens when AI systems are used for surveillance and social control?
- What role should international bodies play in governing AI development?
- Who sets the global standards for AI, and whose interests do they serve?

Psychology

- How does perceived confidentiality affect disclosure behaviour in AI conversations?
- How does repeated exposure to AI-generated content affect cognition and beliefs?
- How should employers protect the mental health of content moderators?
- What are the psychological effects of repeated exposure to traumatic content?
- What cognitive biases make people vulnerable to AI-generated misinformation?
- What is the "illusion of privacy" and why are users vulnerable to it?
- What psychological factors drive over-trust or under-trust in AI systems?
- What support systems should be available for workers in harmful AI-related occupations?

Science

- Can digital technology help monitor and mitigate climate change?
- How can we tell scientific fact from AI-generated pseudoscience?
- How do hallucinations in large language models affect public understanding of science?
- How do server farms affect local ecosystems?
- How might AI accelerate or hinder scientific research?
- What are the energy and water requirements of training large AI models?
- What is the carbon footprint of cloud computing?
- What scientific literacy is needed to evaluate AI claims critically?

Theory of Knowledge

- Can knowledge production be separated from its conditions of production?
- Can machines be objective?
- Do we have a moral obligation to consider the environmental cost of digital knowledge systems?
- How do biases in programming shape our perception of truth?
- How does funding shape AI research priorities and findings?
- How does the invisibility of AI labour shape our understanding of AI?
- What ethical obligations do knowledge-users have to knowledge-producers?
- Who gets to define what AI is and what it should do?