# Case

#### We negate

#### Overview

#### GAI will run out of crucial data by 2026

Maggie Harrison **Dupré**, 11-13-20**23**, Senior Staff Writer at Futurism, graduated from University of Massachusetts Amherst, “AI Companies Are Running Out of Training Data", Futurism, <https://futurism.com/ai-companies-training-data> doa:2/16/25 as

Data plays a central role, if not *the* central role, in the AI economy. Data is a model's vital force, both in basic function and in quality; the more natural — as in, human-made — data that an AI system has to train on, the better that system becomes. Unfortunately for AI companies, though, it turns out that natural data is a finite resource — and if that tap runs dry, researchers warn they could be in for a serious reckoning. As Rita Matulionyte, an information technology law professor at Australia's Macquarie University, [notes in an essay for *The Conversation*](https://theconversation.com/researchers-warn-we-could-run-out-of-data-to-train-ai-by-2026-what-then-216741), AI researchers have been sounding the dwindling-data-supply-alarm-bells for nearly a year. One [study](https://arxiv.org/pdf/2211.04325.pdf) last year by researchers at the [AI forecasting organization Epoch AI](https://epochai.org/) estimated that AI companies could run out of high-quality textual training data by as soon as 2026, while low-quality text and image data wells could run dry anytime between 2030 and 2060. It's a precarious situation for AI firms, given how much data AI systems need to operate and improve. AI models have advanced drastically as developers have poured in more and more data. If the data supply stagnates, [so may the models](https://venturebeat.com/ai/what-happens-when-we-run-out-of-data-for-ai-models/) — and thus, perhaps, the industry. Though Matulionyte offers the use of synthetic data — or data generated by AI models — to train new models as a possible mitigation technique for data-hungry AI companies, that might not be a viable solution either. Indeed, using synthetic content [might actually wreck a given model entirely](https://futurism.com/ai-trained-ai-generated-data-interview); there's some research to show that training AI models on AI-generated content causes a distinct inbreeding effect, with the lack of variance in the dataset resulting in garbled, uncanny outputs. (That said, as Matulionyte points out, some companies are already [experimenting with synthetic training sets](https://www.wsj.com/articles/fake-it-to-make-it-companies-beef-up-ai-models-with-synthetic-data-11627032601).) As it stands, the most practical solution for this looming problem — save for the advent of mass human content farms, where we lowly carbon-based creatures click and clack away to feed the endless data thirst of our robot overlords — may actually be through data partnerships. Basically, a company or institution with a vast and sought-after trove of high-quality data strikes a deal with an AI company to cough up that data, [likely in exchange for cash](https://futurism.com/the-byte/ai-synthetic-data). "Modern AI technology learns skills and aspects of our world — of people, our motivations, interactions, and the way we communicate — by making sense of the data on which it's trained," reads a [recent blog post](https://openai.com/blog/data-partnerships) from leading Silicon Valley AI firm OpenAI, which launched a new Data Partnership just last week. "Data Partnerships are intended to enable more organizations to help steer the future of AI," the blog continues, "and benefit from models that are more useful to them, by including content they care about." Considering that most of the AI datasets that are currently being used to train AI systems are made from internet-scraped data originally created by, well, [*all of us* online](https://futurism.com/the-byte/openai-sued-train-ai), data partnerships may not be the worst way to go. But as data becomes increasingly valuable, it'll certainly be interesting to see how many AI companies can actually compete for datasets — let alone how many institutions, or even individuals, will be willing to cough their data over to AI vacuums in the first place. But even then, there's no guarantee that the data wells won't ever run dry. As infinite as the internet seems, few things are actually endless.

#### AI models collapse happens due to the eventual dependency on synthetic data

[Ilia Shumailov](https://www.nature.com/articles/s41586-024-07566-y#auth-Ilia-Shumailov-Aff1) and Zakhar **Shumaylov et al**, 7-23-20**24**, Department of Computer Science, University of Oxford, Oxford, UK, Zakhar Shumaylov: Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK, YIren Zhao: Department of Electrical and Electronic Engineering, Imperial College London, London, UK, Nicolas Papernot:University of Toronto, Toronto, Ontario, Canada**,** Vector Institute, Toronto, Ontario, Canada, Deceased: Ross Anderson: Department of Computer Science and Technology, University of Cambridge, Cambridge, UK, School of Informatics, University of Edinburgh, Edinburgh, UK, Yarin Gal: OATML, Department of Computer Science, University of Oxford, Oxford, UK "AI models collapse when trained on recursively generated data", Nature, <https://www.nature.com/articles/s41586-024-07566-y> doa:2/12/25 as

Stable diffusion revolutionized image creation from descriptive text. GPT-2 (ref. [1](https://www.nature.com/articles/s41586-024-07566-y#ref-CR1)), GPT-3(.5) (ref. [2](https://www.nature.com/articles/s41586-024-07566-y#ref-CR2)) and GPT-4 (ref. [3](https://www.nature.com/articles/s41586-024-07566-y#ref-CR3)) demonstrated high performance across a variety of language tasks. ChatGPT introduced such language models to the public. It is now clear that generative artificial intelligence (AI) such as large language models (LLMs) is here to stay and will substantially change the ecosystem of online text and images. Here we consider what may happen to GPT-{*n*} once LLMs contribute much of the text found online. We find that indiscriminate use of model-generated content in training causes irreversible defects in the resulting models, in which tails of the original content distribution disappear. We refer to this effect as ‘model collapse’ and show that it can occur in LLMs as well as in variational autoencoders (VAEs) and Gaussian mixture models (GMMs). We build theoretical intuition behind the phenomenon and portray its ubiquity among all learned generative models. We demonstrate that it must be taken seriously if we are to sustain the benefits of training from large-scale data scraped from the web. Indeed, the value of data collected about genuine human interactions with systems will be increasingly valuable in the presence of LLM-generated content in data crawled from the Internet. **Main** The development of LLMs is very involved and requires large quantities of training data. Yet, although current LLMs[2](https://www.nature.com/articles/s41586-024-07566-y#ref-CR2),[4](https://www.nature.com/articles/s41586-024-07566-y#ref-CR4),[5](https://www.nature.com/articles/s41586-024-07566-y#ref-CR5),[6](https://www.nature.com/articles/s41586-024-07566-y#ref-CR6), including GPT-3, were trained on predominantly human-generated text, this may change. If the training data of most future models are also scraped from the web, then they will inevitably train on data produced by their predecessors. In this paper, we investigate what happens when text produced by, for example, a version of GPT forms most of the training dataset of following models. What happens to GPT generations GPT-{*n*} as *n* increases? We discover that indiscriminately learning from data produced by other models causes ‘model collapse’—a degenerative process whereby, over time, models forget the true underlying data distribution, even in the absence of a shift in the distribution over time. We give examples of model collapse for GMMs, VAEs and LLMs. We show that, over time, models start losing information about the true distribution, which first starts with tails disappearing, and learned behaviours converge over the generations to a point estimate with very small variance. Furthermore, we show that this process is inevitable, even for cases with almost ideal conditions for long-term learning, that is, no function estimation error. We also briefly mention two close concepts to model collapse from the existing literature: catastrophic forgetting arising in the framework of task-free continual learning[7](https://www.nature.com/articles/s41586-024-07566-y#ref-CR7) and data poisoning[8](https://www.nature.com/articles/s41586-024-07566-y#ref-CR8),[9](https://www.nature.com/articles/s41586-024-07566-y#ref-CR9) maliciously leading to unintended behaviour. Neither is able to explain the phenomenon of model collapse fully, as the setting is fundamentally different, but they provide another perspective on the observed phenomenon and are discussed in more depth in the [Supplementary Materials](https://www.nature.com/articles/s41586-024-07566-y#MOESM1). Finally, we discuss the broader implications of model collapse. We note that access to the original data distribution is crucial: in learning tasks in which the tails of the underlying distribution matter, one needs access to real human-produced data. In other words, the use of LLMs at scale to publish content on the Internet will pollute the collection of data to train their successors: data about human interactions with LLMs will be increasingly valuable. **What is model collapse?** Definition 2.1 (model collapse) Model collapse is a degenerative process affecting generations of learned generative models, in which the data they generate end up polluting the training set of the next generation. Being trained on polluted data, they then mis-perceive reality. The process is depicted in Fig. [1a](https://www.nature.com/articles/s41586-024-07566-y#Fig1). We separate two special cases: early model collapse and late model collapse. In early model collapse, the model begins losing information about the tails of the distribution; in late model collapse, the model converges to a distribution that carries little resemblance to the original one, often with substantially reduced variance. This process occurs owing to three specific sources of error compounding over generations and causing deviation from the original model: **Statistical approximation error.** This is the primary type of error, which arises owing to the number of samples being finite, and disappears as the number of samples tends to infinity. This occurs because of a non-zero probability that information can get lost at every step of resampling. **Functional expressivity error.** This is a secondary type of error, arising owing to limited function approximator expressiveness. In particular, neural networks are only universal approximators as their size goes to infinity. As a result, a neural network can introduce non-zero likelihood outside the support of the original distribution or zero likelihood inside the support of the original distribution. A simple example of the expressivity error is if we tried fitting a mixture of two Gaussians with a single Gaussian. Even if we have perfect information about the data distribution (that is, infinite number of samples), model errors will be inevitable. However, in the absence of the other two types of error, this can only occur at the first generation. **Functional approximation error.** This is a secondary type of error, arising primarily from the limitations of learning procedures, for example, structural bias of stochastic gradient descent[10](https://www.nature.com/articles/s41586-024-07566-y#ref-CR10),[11](https://www.nature.com/articles/s41586-024-07566-y#ref-CR11) or choice of objective[12](https://www.nature.com/articles/s41586-024-07566-y#ref-CR12). This error can be viewed as one arising in the limit of infinite data and perfect expressivity at each generation. Each of the above can cause model collapse to get worse or better. More approximation power can even be a double-edged sword—better expressiveness may counteract statistical noise, resulting in a good approximation of the true distribution, but it can equally compound the noise. More often than not, we get a cascading effect, in which individual inaccuracies combine to cause the overall error to grow. For example, overfitting the density model causes the model to extrapolate incorrectly and assigns high-density regions to low-density regions not covered in the training set support; these will then be sampled with arbitrary frequency. It is worth noting that other types of error exist. For example, computers have limited precision in practice. We now turn to mathematical intuition to explain how the above give rise to the errors observed, how different sources can compound and how we can quantify the average model divergence. **Theoretical intuition** Here we provide a theoretical intuition for the phenomenon of model collapse. We argue that the process of model collapse is universal among generative models that recursively train on data generated by previous generations. We quantify the sources of errors discussed in the previous section by examining two mathematical models, which prove to be simple enough to provide analytical expressions for quantities of interest, but also portray the phenomenon of model collapse: a discrete distribution in the absence of functional expressivity and approximation errors, and a multidimensional Gaussian approximation, portraying joint functional expressivity and statistical errors. We further illustrate the impact of all three jointly for a more complex setting of density estimation in Hilbert spaces in the [Supplementary Materials](https://www.nature.com/articles/s41586-024-07566-y#MOESM1). The overall stochastic process we consider, which we call learning with generational data, is the following. The dataset at generation *i* is Di, comprising independent and identically distributed random variables Xji with distribution *pi*, *j* ∈ {1,…, *Mi*} denotes the size of the dataset. Going from generation *i* to generation *i* + 1, we aim to estimate the distribution of samples in Di, with an approximation pθi+1. This step is what we refer to as functional approximation, pθi+1=Fθ(pi). The dataset Di+1 is then generated by sampling from pi+1=αipθi+1+βipi+γip0, with non-negative parameters *αi*, *βi*, *γi* summing to 1, that is, they represent proportions of data used from different generations. This corresponds to a mixing of data coming from the original distribution (*γi*), data used by the previous generation (*βi*) and data generated by the new model (*αi*). We refer to this as the sampling step. For the mathematical models to come, we consider *αi* = *γi* = 0, that is, data only from a single step are used, whereas numerical experiments are performed on more realistic choices of parameters. Discrete distributions with exact approximation In this subsection, we consider a discrete probability distribution in absence of functional approximation and expressivity errors, that is, F(p)=p. In this case, model collapse arises only because of statistical errors from the sampling step. At first, the tails (low-probability events) begin to disappear as a result of the low probability of sampling them and, over time, support of the distribution shrinks. Denoting the sample size as *M*, if we consider state *i* with probability q≤1M, the expected number of samples with value *i* coming from those events will be less than 1. In practice, this would mean that we lose information about them. Considering more generally some state *i* with probability *q*, using standard conditional probability, we can show that the probability of losing information (that is, sampling no data at some generation) is equal to 1 − *q*, implying that the distribution must converge to a delta function positioned at some state, with the probability of ending up at a certain state equal to the probability of sampling said state from the original distribution. This can be shown directly by considering the process Xi→F→pi+1→Xi+1 as a Markov chain, as **X***i*+1 only depends on **X***i*. Furthermore, if all the Xji have the same value, then at the next generation, the approximated distribution will be exactly a delta function and therefore all of Xji+1 will also have the same value. This implies that the Markov chain contains at least one absorbing state and therefore, with probability 1, it will converge to one of the absorbing states. This is a well-known fact, of which a proof is provided in the [Supplementary Materials](https://www.nature.com/articles/s41586-024-07566-y#MOESM1). For this chain, the only absorbing states are those corresponding to delta functions. As a result, as we follow the progress of model collapse, we are guaranteed to end up in a constant state, having lost all the information of the original distribution when the chain is absorbed. This argument also works in general owing to floating-point representations being discrete, making the Markov chain over the parameters of the model discrete. Thus, as long as the model parameterization allows for delta functions, we will get to it, because—owing to sampling errors—the only possible absorbing states are delta functions. On the basis of the discussion above, we see how both early model collapse, in which only the low-probability events get cut off, and late stage model collapse, in which the process begins to collapse into a single mode, must arise in the case of discrete distributions with perfect functional approximation. Multidimensional Gaussian Following the discussion about discrete distributions, we now present a more generic result, which can be shown in the Gaussian approximation setting, in which each generation is approximated using the unbiased estimates of the mean and the variance. A similar result holds more generally, which we detail in the [Supplementary Materials](https://www.nature.com/articles/s41586-024-07566-y#MOESM1). Theorem 3.1 (Gaussian model collapse) Assume the original data are sampled from distribution D0 (not necessarily Gaussian), with non-zero sample variance. Assume *Xn* are fit recursively using the unbiased sample mean and variance estimators from the previous generation, Xjn|μn,Σn∼N(μn,Σn), with a fixed sample size. Then, E[W22(N(μn,Σn),D0)]→∞;Σn→a.s.0asn→∞, in which W2 denotes the Wasserstein-2 distance between the true distribution and its approximation at generation *n*. In words, this implies that not only does the *n*th generation approximation diverge arbitrarily far from the original one but it also collapses to be zero variance as the number of generations increases, with probability 1. The results are very analogous to that seen in the discrete case, with this theorem illustrating the effect of late stage model collapse, in which the process begins to collapse to be zero variance. The early stage model collapse can also be seen and the interested reader is referred to the [Supplementary Materials](https://www.nature.com/articles/s41586-024-07566-y#MOESM1) for a more in-depth discussion. **Model collapse in language models** In this section, we evaluate the effect of model collapse on language models. We cover more interpretable machine learning models—VAEs and GMMs—in the [Supplementary Materials](https://www.nature.com/articles/s41586-024-07566-y#MOESM1). Code is publically available in ref. [13](https://www.nature.com/articles/s41586-024-07566-y#ref-CR13). Model collapse is universal across various families of machine learning models. Yet, if small models such as GMMs and VAEs are normally trained from scratch, LLMs are different. They are so expensive to retrain from scratch that they are typically initialized with pre-trained models such as BERT[4](https://www.nature.com/articles/s41586-024-07566-y#ref-CR4), RoBERTa[5](https://www.nature.com/articles/s41586-024-07566-y#ref-CR5) or GPT-2 (ref. [2](https://www.nature.com/articles/s41586-024-07566-y#ref-CR2)), which are trained on large text corpora. They are then fine-tuned to various downstream tasks[14](https://www.nature.com/articles/s41586-024-07566-y#ref-CR14). Here we explore what happens with language models when they are sequentially fine-tuned with data generated by other models. We can easily replicate all experiments covered in this paper with larger language models in non-fine-tuning settings to demonstrate model collapse. Given that training a single moderately large model produces twice the American lifetime’s worth of CO2 (ref. [15](https://www.nature.com/articles/s41586-024-07566-y#ref-CR15)), we opted to not run such an experiment and instead focus on a more realistic setting for a proof of concept. Note that even the language experiments described in this paper took weeks to run. We evaluate the most common setting of training a language model—a fine-tuning setting for which each of the training cycles starts from a pre-trained model with recent data. The data here come from another fine-tuned pre-trained model. Because training is restricted to produce models that are close to the original pre-trained model, and data points generated by the models will generally produce very small gradients, the expectation here may be that the model should only change moderately after fine-tuning. We fine-tune the OPT-125m causal language model made available by Meta through Hugging Face[6](https://www.nature.com/articles/s41586-024-07566-y#ref-CR6). We fine-tune it on the wikitext2 dataset[16](https://www.nature.com/articles/s41586-024-07566-y#ref-CR16). For data generation from the trained models, we use a five-way beam search. We block training sequences to be 64 tokens long; then, for each token sequence in the training set, we ask the model to predict the next 64 tokens. We go through all of the original training dataset and produce an artificial dataset of the same size. Because we go through all of the original dataset and predict all of the blocks, if the model had 0 error, it would produce the original wikitext2 dataset. Training for each generation starts with generation from the original training data. Each experiment is run five times and the results are shown as five separate runs with different randomness seeds. The original model fine-tuned with real wikitext2 data obtains 34 mean perplexity, from the zero-shot baseline of 115, that is, it successfully learns the task. Finally, to be as realistic as possible, we use the best-performing model on the original task, evaluated using the original wikitext2 validation set, as the base model for the subsequent generations, meaning that—in practice—observed model collapse can be even more pronounced. Here we consider two different settings: Five epochs, no original training data. Here the model is trained for five epochs starting on the original dataset but with no original data retained for subsequent runs. The overall original task performance is presented in Fig. [1b](https://www.nature.com/articles/s41586-024-07566-y#Fig1). We find that training with generated data allows us to adapt to the underlying task, losing some performance, from 20 to 28 perplexity points. Ten epochs, 10% of original training data preserved. Here the model is trained for ten epochs on the original dataset and with every new generation of training, a random 10% of the original data points is sampled. The overall original task performance is presented in Fig. [1c](https://www.nature.com/articles/s41586-024-07566-y#Fig1). We find that preservation of the original data allows for better model fine-tuning and leads to only minor degradation of performance. Both training regimes lead to degraded performance in our models, yet we do find that learning with generated data is possible and models can successfully learn (some of) the underlying task. In particular, from Fig. [1](https://www.nature.com/articles/s41586-024-07566-y#Fig1) and their 3D versions in the [Supplementary Materials](https://www.nature.com/articles/s41586-024-07566-y#MOESM1), we see that model collapse occurs, as the density of samples with low perplexity begins to accumulate over the generations. This in turn makes it likely that, over the generations, the sampled data will similarly collapse to a delta function. **a**, Model collapse refers to a degenerative learning process in which models start forgetting improbable events over time, as the model becomes poisoned with its own projection of reality. Here data are assumed to be human-curated and start off clean; then model 0 is trained and data are sampled from it; at step *n*, data are added to the overall data from step *n* − 1 and this combination is used to train model *n*. Data obtained with Monte Carlo sampling should ideally be statistically close to the original, provided that fitting and sampling procedures are perfect. This process depicts what happens in real life with the Internet: model-generated data become pervasive. **b**,**c**, Performance of OPT-125m models of different generations evaluated using the original wikitext2 test dataset. Shown on the left are the histograms of perplexities of each individual data training sequence produced by different generations as evaluated by the very first model trained with the real data. Over the generations, models tend to produce samples that the original model trained with real data is more likely to produce. At the same time, a much longer tail appears for later generations. Later generations start producing samples that would never be produced by the original model, that is, they start misperceiving reality based on errors introduced by their ancestors. The same plots are shown in 3D in the [Supplementary Materials](https://www.nature.com/articles/s41586-024-07566-y#MOESM1). On the right, average perplexity and its standard deviation are shown for each independent run. The *x* axis refers to the generation of the model. ‘Real’ refers to the ‘model 0’ trained on the original wikitext2 dataset; model 1 was trained on the data produced by model 0, model 2 was trained on data produced by model 1 and so on, with all generated datasets equal in size. We find that models trained on generated data are able to learn some of the original task, but with errors, as seen from the increase in perplexity. [**Full size image**](https://www.nature.com/articles/s41586-024-07566-y/figures/1) It is important to note here that the observed behaviour is in line with the general intuition established in the section ‘Theoretical intuition’. To be precise, in all experiments, generational learning is only performed on a finite (usually small) number of generations, whereas claims of the section ‘Theoretical intuition’ are mostly presented in the limit of generations going to infinity. However, as seen from experiments on VAEs and GMMs in the [Supplementary Materials](https://www.nature.com/articles/s41586-024-07566-y#MOESM1), convergence to delta functions and specific rates of such convergence are highly related to the specifics of the problem considered, and complete collapse may or may not occur, even after a small number of steps. This is further illustrated theoretically in the [Supplementary Materials](https://www.nature.com/articles/s41586-024-07566-y#MOESM1), in which potentially notable divergence from the original model can occur even after a few generations. Figure [1b,c](https://www.nature.com/articles/s41586-024-07566-y#Fig1) on the left shows histograms of individual data-point perplexities generated by the models of different generations as is evaluated by the first model developed with real wikitext2 training data. Here, over the generations, models tend to produce more sequences that the original model would produce with the higher likelihood. The observed effect is similar to that described for VAEs and GMMs in the [Supplementary Materials](https://www.nature.com/articles/s41586-024-07566-y#MOESM1), in which—over the generations—models started to produce samples that would be produced with higher probabilities by the original model. At the same time, we discover that generated data have much longer tails, suggesting that some of the data would never be produced by the original model—these are the errors that accumulate because of the learning with generational data. Example 1 Example of text outputs of an OPT-125m model affected by model collapse—models degrade over generations, for which each new generation is trained on data produced by the previous generation. **Input:** some started before 1360 — was typically accomplished by a master mason and a small team of itinerant masons, supplemented by local parish labourers, according to Poyntz Wright. But other authors reject this model, suggesting instead that leading architects designed the parish church towers based on early examples of Perpendicular. **Outputs:** **Gen 0:** Revival architecture such as St. John’s Cathedral in London. The earliest surviving example of Perpendicular Revival architecture is found in the 18th @-@ century Church of Our Lady of Guernsey, which dates from the late 19th century. There are two types of perpendicular churches : those. **Gen 1:** architecture such as St. Peter’s Basilica in Rome or St. Peter’s Basilica in Buenos Aires. There is no evidence that any of these buildings were built during the reign of Pope Innocent III, but it is possible that they may have been built during the reign of his successor, Pope Innocent. **Gen 5:** ism, which had been translated into more than 100 languages including English, French, German, Italian, Spanish, Portuguese, Dutch, Swedish, Norwegian, Polish, Hungarian, Slovak, Lithuanian, Estonian, Finnish, Romanian, Bulgarian, Turkish, Croatian, Serbian, Ukrainian, Russian, Kazakh, Kyrgyz. **Gen 9:** architecture. In addition to being home to some of the world’s largest populations of black @-@ tailed jackrabbits, white @-@ tailed jackrabbits, blue @-@ tailed jackrabbits, red @-@ tailed jackrabbits, yellow @-. Ablation: Repetitions We find that data generated by language models in our experiments end up containing a large number of repeating phrases, as in Example 1. The repeating problem has been observed in nearly all text-generation models[17](https://www.nature.com/articles/s41586-024-07566-y#ref-CR17),[18](https://www.nature.com/articles/s41586-024-07566-y#ref-CR18) and, to rule this out as the cause of model collapse, we further provide numerical experiments when models are explicitly encouraged to produce non-repeating sequences with a repeating penalty of 2.0. We find that this causes the models to produce lower score continuations to avoid using repeats, which—as a result—causes the consequent models to perform even worse. Model perplexities shift across the generations towards more probable token sequences, as measured using the model trained on the original real data distribution. Further illustrations are provided in the [Supplementary Materials](https://www.nature.com/articles/s41586-024-07566-y#MOESM1). In particular, enforcing this for the LLM experiments causes the perplexity to double compared with the original. Models remain as susceptible to model collapse, if not more. The described process demonstrates that fine-tuning of language models does not curb the effects of model collapse and models that are being fine-tuned are also vulnerable. We find that, over the generations, models tend to produce more probable sequences from the original data and start introducing their own improbable sequences, that is, errors. **Discussion** We now discuss the implications of model collapse on the underlying learning dynamics of LLMs. Long-term poisoning attacks on language models are not new. For example, we saw the creation of click, content and troll farms, a form of human ‘language models’, whose job is to misguide social networks and search algorithms. The negative effect that these poisoning attacks had on search results led to changes in search algorithms. For example, Google downgraded farmed articles[19](https://www.nature.com/articles/s41586-024-07566-y#ref-CR19), putting more emphasis on content produced by trustworthy sources, such as education domains, whereas DuckDuckGo removed them altogether[20](https://www.nature.com/articles/s41586-024-07566-y#ref-CR20). What is different with the arrival of LLMs is the scale at which such poisoning can happen once it is automated. Preserving the ability of LLMs to model low-probability events is essential to the fairness of their predictions: such events are often relevant to marginalized groups. Low-probability events are also vital to understand complex systems[21](https://www.nature.com/articles/s41586-024-07566-y#ref-CR21). Our evaluation suggests a ‘first mover advantage’ when it comes to training models such as LLMs. In our work, we demonstrate that training on samples from another generative model can induce a distribution shift, which—over time—causes model collapse. This in turn causes the model to mis-perceive the underlying learning task. To sustain learning over a long period of time, we need to make sure that access to the original data source is preserved and that further data not generated by LLMs remain available over time. The need to distinguish data generated by LLMs from other data raises questions about the provenance of content that is crawled from the Internet: it is unclear how content generated by LLMs can be tracked at scale. One option is community-wide coordination to ensure that different parties involved in LLM creation and deployment share the information needed to resolve questions of provenance. Otherwise, it may become increasingly difficult to train newer versions of LLMs without access to data that were crawled from the Internet before the mass adoption of the technology or direct access to data generated by humans at scale.

#### C1: Collapse of innovation

#### 1 – Overreliance of GenAi causes lack of creativity

**Ali et al 24** Omar Ali, College of Business and Entrepreneurship, Abdullah Al Salem University. Peter A. Murray, University of Southern Queensland. Mujtaba Momin, College of Business Administration, American University of the Middle East. Yogesh K. Dwivedi, Digital Futures for Sustainable Business & Society Research Group, School of Management, Swansea University & Symbiosis International (Deemed University). Tegwen Malik, School of Management, Swansea University. Meta-analysis of 185+ published literature papers evaluating the key influences and implications of using AI models in the education sector. February 2024, "The effects of artificial intelligence applications in educational settings: Challenges and strategies", Science Direct, <https://www.sciencedirect.com/science/article/pii/S0040162523007618> DOA: 2/12/25 SLK

One of the significant challenges faced by the ChatGPT relates to the lack of innovative output quality (Lund et al., 2023; Kasneci et al., 2023). This is largely due to the single source of training data input that the mechanism has received. **Generative models** rely on the input training data source, and although they modulate the patterns of output, they systematically **generate monotonous and non-creative content.** **This curtails the innovation and uniqueness of replies** (Pappas and Giannakos, 2021; Biswas, 2023). Chen and Wen (2021) moreover established that **a generative model-based tune composition system had a regulated capability to produce unprecedented, novel and distinct tunes**. While some creativity can be observed within limited contexts therefore, **significant drawbacks such as plagiarism and violation of copyrights restricts the unique aspect of creative content**. While ChatGPT can be fine-tuned and personalized to configure specific learning content and answer student queries, it **is incapable of dealing with resourceful and ingenious problem-solving contexts such as critical thinking which is a pre-requisite in the education system** (Kasneci et al., 2023). **Several other challenges** also need be noted as follows: **(1) Limitations in learning approaches: ChatGPT spawns responses based on the restrictive training data rendered.** While it can respond to forthright questions**, it cannot deal with contextual problem solving, innovation, and establishing a critical mindset such that it might help students find creative solutions** (Mantelero, 2018; Kasneci et al., 2023). (2) Lack of novelty: Input and training data are the primary sources of ChatGPT responses; thus, expecting unprecedented, innovative solutions to unprecedented queries is most likely a distal expectation (Xia, 2021; O'Connor, 2023) among learners. **(3) Potential for overreliance: Generative AI could be expected to impair student's self-dependence. Given that it is easy for a learner to access the application, a sense of overreliance may inhibit learner self-dependence and creative ways of problem-solving and lateral thinking** (Stevenson et al., 2022; Placed et al., 2022).

#### 2 -- Critical thinking

**McCollum 24,** Madison McCollum, Administrative Assistant at Tarrant County College, 8-16-2024, "How AI is Undermining Your Child's Education", No Publication, <https://gopioneer.com/blog/kidsandai//FZ>

AI generators like ChatGPT are becoming more common in both educational and everyday life. While this technology can offer great benefits**, the use by children and students can have a significant draw back, particularly when it comes to their learning processes and development**. ChatGPT (and other AI tools) are being used by students in schools for a variety of tasks. With the AI software being free, and more advanced versions being around $20/month, students are using them to write assignments, generate ideas, or even solve more complex problems**. These tools can be great for providing quick answers and assisting with drafting essays or content creation, which may seem super beneficial at a first glance. However, the convenience of AI tools can come at a high cost to a child’s learning or cognitive development.** Many experts and educators are afraid of students becoming over reliant on AI**. Overreliance happens when users accept AI-generated content without question, which is particularly concerning if they completely trust AI-models.** Here are just three major areas where a child relying on ChatGPT can suffer: **Erosion of Critical Thinking Skills: According to a study by Zhai et al., the regular utilization of AI-systems for academia (like generating ideas or writing essays) has been linked to a decline in cognitive abilities, a diminished capacity for information retention, and an increasing reliance on AI systems for information.** In a study that focused on the concerns of overreliance and diminished critical thinking skills, Krullars et al. (2023) argues that the **over-reliance on AI dialogue systems could diminish student’s drives and commitment to learning when they rely too much on them instead of participating in a learning environment. By using AI, students will lose the opportunity to engage deeply** with the material and think independently. Plagiarism and Academic Dishonesty: The ability of AI tools to produce well-written text has raised concerns about plagiarism. **Some students may be tempted to pass off AI-generated content as their own, leading to issues of academic dishonesty and a lack of genuine learning.** Because of this, some schools have banned the use of ChatGPT and other generative AI tools on their networks and others have updated their policies and handbooks to redefine ‘cheating.’ (EdTech Magazine, 2023). Dependence on Technology: **Over-reliance on AI can lead to a dependency that diminishes students' abilities to perform tasks without technology assistance.** This **dependency can be detrimental in situations where critical thinking and original thought are required and could lead to a major disadvantage when they later enter the workforce** (MDPI, 2023). Bias and Misinformation: AI tools are trained on vast datasets that may contain biases. **This can result in biased outputs or the distribution of incorrect information, which can mislead students and impact their understanding of topics or lead them to be suspected of cheating under school regulations**/rules. (EdTech Magazine, 2023). While AI can be a great way to assist learning or cutting out the more ‘mundane’ work, the use of AI needs careful consideration so that it doesn’t undermine a students’ educational development. Educators AND parents need to make sure that these tools are used as… well, tools! Encourage your students to think critically and educate them on the downsides and proper uses of AI tools like ChatGPT. Fostering independent problem-solving skills are essential of raising well-rounded and capable future-adults!

#### Risk of over dependence kills innovation

[Lixiang **Yan**](https://www.nature.com/articles/s41562-024-02004-5#auth-Lixiang-Yan-Aff1), [Samuel Greiff](https://www.nature.com/articles/s41562-024-02004-5#auth-Samuel-Greiff-Aff2-Aff3-Aff4), [Ziwen Teuber](https://www.nature.com/articles/s41562-024-02004-5#auth-Ziwen-Teuber-Aff2), [Dragan Gašević](https://www.nature.com/articles/s41562-024-02004-5#auth-Dragan-Ga_evi_-Aff1) **ET AL**, 10-21-20**24**, "Promises and challenges of generative artificial intelligence for human learning", Nature, <https://www.nature.com/articles/s41562-024-02004-5> doa:2/20/25 as

3 Challenges Amidst GenAI’s promises, formidable challenges confront learners and educators alike and raise critical moral and ethical concerns about integrating such technology into human learning. These challenges involve GenAI technologies’ imperfections, the ethical dilemmas of transparency, privacy, equality, and beneficence, and the disruption of assessment practices. The following sections elaborate on each of these challenges. 3.1 GenAI’s Imperfections As GenAI technologies become increasingly integrated into learning support, resource generation, feedback, and assessment, it is imperative to address the risks posed by hallucinations [69]. Hallucinations occur when there are mismatches in training data or complexities in language generation tasks, resulting in outputs that may not align with factual information [70]. The probabilistic nature of LLMs and diffusion models further limits their utility due to inherent instabilities and potential biases in their training data [71]. For instance, ChatGPT often fails tasks easily solved by humans, such as reasoning tasks requiring real-world knowledge, logic, mathematical calculations, and distinguishing between factual and fictive information. Consequently, it sometimes provides fabricated facts [72] (preprint). These inaccuracies can undermine GenAI’s reliability as a learning tool, potentially outweighing its promises (Section 2). Emerging studies indicate that hallucinations in GenAI can occur with nonnegligible frequency, increasing with the complexity and specificity of queries posed to the AI [70]. GenAI may perform reasonably well with generic questions (e.g., What are Newton’s laws of motion?) but is more prone to errors with nuanced, contextspecific, time-sensitive, or highly technical information [73]. The lack of transparency in GenAI’s decision-making process complicates identifying when and why these hallucinations occur [70, 74]. Relying solely on GenAI for learning content creation and curriculum development without validation could introduce inaccuracies, misleading both educators and students. Similarly, GenAI-generated feedback or assessments based on incorrect information could misguide students’ learning processes, leading to misconceptions or a lack of understanding of key concepts. Addressing these challenges requires an interdisciplinary effort. Educators should adopt a balanced and proactive approach, teaching learners to critically evaluate AIgenerated content by cross-referencing with reliable sources, questioning plausibility, and recognising signs of hallucination. These steps are essential for cultivating AI literacy [75], as discussed further in Section 4.1. Additionally, designing and optimising the interface of educational technologies to highlight potential hallucinations requires collaboration among learning scientists, human-computer interaction researchers, and technology providers [74, 76]. Such a collaborative approach is essential to empower learners to deal with the imperfections of GenAI both intrinsically, by developing critical thinking skills, and extrinsically, by leveraging improved technological interfaces that signal potential inaccuracies. 3.2 Ethical Dilemmas Adopting GenAI to support human learning raises several ethical issues, notably in areas such as transparency, privacy, equality, and beneficence. A key concern is the transparency of AI-generated solutions, as highlighted in a recent systematic literature review [2]. The review found that a majority (92%) of GenAI tools currently used for supporting learning practices, particularly those based on LLMs, are transparent only to AI experts, not to educators, students, or other stakeholders. This lack of transparency is problematic as it obscures the understanding of AI functionalities and potential flaws from those directly impacted by these technologies [77]. The primary cause of this transparency gap is the absence of human-in-the-loop elements in prior research, such as involving educators and students in the development and evaluation of GenAI-powered educational technologies. This aligns with the growing emphasis on developing explainable and human-centred AI, underscoring the essential role of stakeholder involvement in crafting impactful and meaningful educational technologies [78, 79]. To achieve personalisation in learning support, resource generation, feedback, and assessment using GenAI, learners’ personal data must be provided to these models. However, privacy concerns can reduce learner participation [80, 81]. These concerns are prominent due to the lack of clear consent strategies and data protection measures surrounding GenAI in supporting human learning [2]. Using learner-generated data without explicit consent or adequate anonymisation raises serious issues about exposing sensitive information [82]. For instance, researchers conducted a divergence attack on ChatGPT, compromising its security and causing it to output original training data containing personally identifiable information [83] (preprint). Although OpenAI has addressed this vulnerability, potential data breaches from unforeseen attacks remain a concern [84, 85]. This issue is particularly troubling given the resources required for GenAI to unlearn information once private data has been used for model training, especially for large, commercial, and proprietary models [84]. Regarding equality, there is an evident disparity in language representation and accessibility of GenAI models. While advancements have been made in non-English languages for LLMs and speech diffusion models [14, 48], the predominance of English based AI solutions perpetuates a bias towards Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies [2]. This imbalance raises concerns about the global applicability and fairness of these technologies, potentially intensifying existing inequalities and the digital divide in learning opportunities [86]. Finally, beneficence is a critical ethical principle that must be addressed. Several studies highlight the risks of underperforming or biased AI models, which can negatively impact human learning and perpetuate systemic biases, such as gender, racial, and social class biases [87, 88]. Strategies like balanced sampling and cautionary labelling have been proposed [89, 90], but the opaque nature of many generative models makes ensuring fairness and accuracy challenging, potentially violating the principle of beneficence [91]. While model alignment is often implemented to prevent GenAI from producing toxic content, recent evidence suggests that adversarial attacks using specific prompts can undermine these measures [85]. Such attacks could facilitate cheating, promote biased views, or expose students to offensive language [92]. These issues could disrupt learning, compromise safety and inclusivity, and cause psychological harm, eroding trust in educational technologies. These ethical challenges underscore the need for rigorous and multifaceted ethical considerations in deploying GenAI and the urgency of establishing regulations, such as the EU AI Act [93]. 3.3 Disruption of Assessment GenAI poses significant challenges to conventional learning assessment methodologies. Traditionally, assessments have focused on evaluating learning products, such as essays, to measure outcomes [52]. However, GenAI’s ability to produce high-quality, human-like responses calls into question the validity of these approaches [94]. A central issue is distinguishing between a learner’s work and AI-generated output. GenAI, particularly LLMs like ChatGPT and Llama 3, can generate responses that closely mimic human reasoning and writing styles, making it difficult to discern the origin of the work [4]. A performance paradox arises when tasks are completed with GenAI assistance. A recent randomised controlled experiment found that while GenAI tools can help students achieve better performance, removing this support significantly lowers their performance [19]. **This suggests that GenAI may create an illusion of improved learning without developing essential skills, such as self-regulated learning**. Thus, we must ask: Who and what are we actually assessing? This dilemma extends beyond detecting AI-generated content to reconsidering the purpose of assessment in learning. The challenge is further compounded when considering the learning process itself. GenAI’s ability to interact with computational systems means even the learning process can be imitated or augmented by AI. Preliminary work on multimodal GenAI agents [95] (preprint) has shown these agents can operate smartphone applications, generating digital trace data while executing user requests. This AI-generated data could impede existing learning analytic methods that rely on such data to model the learning process [96]. This issue blurs the line between human cognition and AI augmented cognition [97], complicating the assessment of skills traditionally seen as exclusively human, such as critical thinking, problem-solving, and creativity [94]. Consequently, we must reconsider the purpose of learning assessment across different educational stages. Assessing human cognition and metacognition remains essential for K-12 education, as young learners continue developing fundamental skills. In higher education, prioritising the evaluation of human-AI hybrid cognition and metacognition could be crucial for preparing learners for an AI-integrated workforce [98]. This shift demands rethinking assessment strategies to accommodate the collaborative nature of learning in the presence of AI. 4 Needs Within GenAI’s promises and challenges, three pivotal needs must be addressed for effective integration into human learning: cultivating AI literacy among learners and educators, prioritising evidence-based decision-making, and ensuring methodological rigour in research using GenAI. These needs aim to foster a balanced integration that enhances human abilities and ensures a synergistic relationship between GenAI and human development. 4.1 AI Literacy Cultivating AI literacy is essential to ensuring the effective, responsible, and ethical use of GenAI technologies to support human learning [75, 99]. This need extends beyond learners to include educators, policymakers, and administrators, who are integral to the design, delivery, and facilitation of learning experiences. AI literacy encompasses a basic understanding of how AI systems function but also an intimate awareness of their potential impact, ethical considerations, and limitations [75]. The absence of AI literacy can lead to severe consequences. For instance, the New York Times reported that a lawyer using ChatGPT for a court filing was unaware of fabricated citations generated by the AI, resulting in a breach of professional ethics and legal standards [100]. One must ask: What if educators unknowingly provided students with AI-generated learning resources that contained fabricated content? Such actions could erode trust and integrity in education systems, misleading students and compromising their learning quality. These concerns highlight the critical need to cultivate AI literacy. A recent study indicates that human users often prefer AI-generated content for its comprehensiveness and well-articulated language style, despite its inaccuracies [101]. As GenAI’s propensity to hallucinate remains challenging to address at the foundational model level [70], understanding its limitations and identifying potential pitfalls will be crucial for preparing individuals to live, learn, and work with GenAI in the 21st century. This requires adopting AI literacy models, practices for their development, and measurement approaches. Institutions, policymakers, and researchers must focus on AI literacy as a key learning objective to ensure that educators, students, administrators, and even parents are not merely consumers of AI technology but also informed participants in its evolution and application. 4.2 Evidence-Based Decision Making The integration of GenAI into human learning promises to enhance experiences and outcomes (as highlighted in Section 2). However, adopting these technologies requires a commitment to evidence-based decision-making. This necessitates a collaborative effort among researchers, practitioners, and policymakers to generate robust evidence guiding the effective and responsible use of AI in learning practices. By working together, these stakeholders can ensure GenAI deployment aligns with learning goals and supports the development of essential cognitive and metacognitive skills. Encouraging the use of GenAI to support human learning requires a nuanced understanding of its benefits and limitations. For instance, while GenAI can improve the efficiency of information processing and retrieval, there is a risk of fluency bias, where learners may overestimate their understanding due to the ease of cognitive information processing [102, 103] (preprint). Similarly, reliance on GenAI for creative and problem-solving tasks could weaken these critical skills, fostering a dependency that may hinder innovation and original thought [104–106]. To mitigate these risks and maximise GenAI’s benefits, it is imperative to foster partnerships among researchers, practitioners, and policymakers. These collaborations can produce evidence that informs learning and teaching practices, ensuring that GenAI enhances rather than replaces human cognitive, metacognitive, and creative processes. By prioritising evidence-based decision-making and stakeholder collaboration, we can leverage GenAI’s advantages in educational environments while promoting deep learning, creativity, and problem-solving abilities among learners. 4.3 Methodological Rigour Building on discussions of evidence-based decision-making, it is crucial to emphasise methodological rigour in applying GenAI technologies within human learning research. As these technologies evolve, human learning researchers and scientists must adapt and refine their methodologies to accurately assess the impact of these tools on teaching and learning processes. GenAI’s capabilities, such as passing the United States Medical Licensing Exam [107], completing exams at the University of Minnesota Law School [108], and solving queries from Wharton School of Business tests [109], underscore its potential. However, the excitement must be tempered with caution to avoid overestimating effectiveness due to methodological shortcomings. A notable example is a preprint study claiming GPT-4, with prompt engineering, could achieve perfect scores in the MIT Mathematics, Electrical Engineering, and Computer Science curriculum [110]. This study [110], initially attracting widespread attention, was later retracted due to methodological concerns, including data set contamination, over-reliance on GPT-4 for accuracy assessment, and ambiguities in manual verification of results [111] (preprint). This incident underscores the need for rigorous methodological standards, likely requiring new approaches in evaluating GenAI technologies. To address these challenges, it is essential to establish standards for appraising the quality of evidence on GenAI’s impact on learning processes, outcomes, and experiences [18]. In the medical field, tools such as the Cochrane Risk of Bias Tool and ROBINS-I are used to assess study quality. Given the distinct methodological requirements introduced by GenAI, including various prompting engineering strategies and retrieval generation techniques, it is crucial to establish specific quality standards and evaluation tools. These requirements go beyond conventional methodologies used in human learning research. For example, using GenAI to generate physics practice questions might involve retrieval methods that limit the AI to sourcing content solely on Newton’s laws of motion and crafting prompts specifying complexity level, target student grade, and desired question format (e.g., multiple-choice, short answer, or problem-solving). By working collaboratively, the human learning research community can create a robust framework for evaluating evidence, ensuring a solid foundation for future policies and practices. This effort will enable researchers, practitioners, and policymakers to build on reliable, valid, and generalisable findings, fostering the responsible and effective integration of GenAI technologies into learning. 5 Conclusion and Future Directions As we look toward the next decade, powerful AI tools are set to become integral to our society, transforming how we learn, work, and live [112]. GenAI technologies could permeate every aspect of human learning. Imagine students collaborating with AI agents designed to mimic certain personality traits to help students learn about leadership and teamwork, engaging in debates with digital twins of Socrates, Plato, and Aristotle to explore ancient Greek philosophy, learning impressionist painting techniques from a humanoid robotic mentor modelled after Claude Monet, and visualising Einstein’s special theory of relativity in virtual realities. All this could occur while receiving personalised support from a GenAI tutor hosted on a wearable device. This integration necessitates a dual approach to learning: educating ourselves both about and with GenAI, while continuing to develop critical thinking, problem-solving, self-regulation, and reflective thinking skills. These skills are crucial for maintaining cognitive and metacognitive autonomy as AI becomes embedded in our daily lives. Understanding the relationship between GenAI and human cognition, metacognition, and creativity is essential for maximising its potential as a learning tool. This understanding will enhance the effectiveness of AI-driven educational tools and ensure human ingenuity is preserved amidst technological advancement. Key research questions include: How can we promote human-AI interaction to maximise learner agency? What behavioural indicators can reliably capture cognitive and metacognitive processes during AI-assisted learning? How can we assess learning to reflect genuine knowledge and skill development rather than an AI-created performance illusion? What strategies can prevent over-reliance on AI, ensuring humans remain primary agents of critical thinking and problem-solving? Educators are pivotal in integrating AI tools to enhance traditional teaching methods. We anticipate a shift in educators’ roles, with GenAI reducing the burden of knowledge dissemination, allowing teachers to focus on deeper connections with students as mentors and facilitators. This transition requires educators to adopt new pedagogical paradigms that leverage AI to foster intellectual and emotional growth. They must become proficient in AI literacy, effectively integrate AI tools into their teaching, and remain vigilant about potential pitfalls, such as GenAI’s imperfections and the risk of student over-reliance on AI. Balancing AI use with activities promoting human creativity, critical thinking, and social interaction is crucial to ensure AI augments rather than replaces human educators. Educational institutions must invest in ongoing professional development and support systems to help teachers manage techno-stress and workload burdens from adopting new technologies. Policymakers and technology companies should consider: How can we ensure accountability for AI tools used in human learning, and who should be responsible for their outcomes? What ethical guidelines should govern AI tools in educational settings? How can we design and implement AI learning tools to promote equality and inclusivity? We argue that human-centred theories of learning and instruction must be integrated with GenAI to ensure these technologies enhance rather than detract from human learning. This involves developing AI systems that support and elevate human cognitive capacities. By fostering a learning environment that harmonises technology with theoretical approaches, we can promote personal growth and the development of adaptive skills and knowledge needed to navigate the rapid changes in the age of AI. A united effort among researchers, policymakers, technology companies, and educators is essential to fully leverage GenAI’s potential in advancing human learning. By addressing these critical questions and considerations, we can ensure that GenAI becomes a powerful ally in the pursuit of knowledge and innovation, rather than a crutch that undermines our intellectual abilities.

#### Dehumanization

**Avraamidou 24** Lucy Avraamidou: Full Professor, Institute for Science Education and Communication, University of Groningen, 04-17-2024, "Can we disrupt the momentum of the AI colonization of science education?", [Journal of Research in Science Teaching](https://onlinelibrary.wiley.com/journal/10982736): Wiley, <https://onlinelibrary.wiley.com/doi/epdf/10.1002/tea.21961>

The same question applies to education. **Even though AI technologies have the potential to innovate teaching they also bring risks and challenges associated with dig-ital monoculturalism as well as ethical, inclusive, and equitable use of AI** (UNESCO, 2023). **Educational institutions are buying into generative AI promises and hallucinations**(Alkaissi & McFarlane, 2023) **and frantically trying to catch up with a mass production of AI tools.** National funding agencies in different parts of the world are allocating financial supportfor research projects utilizing AI tools in science and education (e.g., New Horizon EuropeFunding for Data, Computing, and AI Technologies). Several (science) education journals havededicated special issues to an examination of the potential of AI for teaching and learning. **Researchers in science education are shifting their interests toward AI to engage with “hot” research in this new world order created by the AI industry.** The problem with this new world order is that it repeats patterns of colonial history through exploitation and extraction of resources to enrich the wealthy and powerful at the great expense of the poor (Hao, 2022). There exists a wealth of evidence pointing to how AI has exploited mar-ginalized communities for the development of large language models, for example, ChatGPT(Perrigo, 2023). **Several studies have shed light on issues related to ethics, biases, and racial and gender stereotypes.** For example, descriptions of people images through tagging (i.e., GoogleCloud Vision API), personalized news feeds (i.e., Google Search, Amazon Cloud Search), virtual assistants (i.e., Siri), and large language models (i.e., ChatGPT) reflect human biases and rein-force social stereotypes: Physicists are white and male, Black people are athletic, Asian womenare beautiful, Black women are hypersexualized (Kyriakou et al., 2019; Noble, 2013; Otterbacheret al., 2017). **Moreover, research findings showed that online social networks and information networks** (e.g., Who to Follow) **that rely on algorithms and used for different purpose s(e.g. networking, hiring, and promotion procedures) perpetuate inequities and further discrimi-nate against minorities** (Espín-Noboa et al., 2022). Other studies provided evidence of the large environmental impact of AI technologies,which include energy used from both the training of models and the actual use (De Vries, 2023;Luccioni et al., 2023). For example, the carbon footprint of an AI prompt is 4–5 times higherthan a search engine query. More concretely, if the typical 9-billion daily Google searches wereinstead AI chatbot queries it would require as much power to run a country with a populationof about 5 million people. Another indicative example of the energy consumption of AI tools isthat generating only one image using an AI model takes as much energy as fully charging asmartphone (Luccioni et al., 2023).These are crucial socio-scientific issues that the science education community ought toengage with through a critical approach to AI. Instead, science education is currently operatingat the service of the generative and predictive AI industry, at least in the Global North, andremains largely disengaged with issues related to digital monoculturalism, algorithmic biases,ethics, and exploitation of both human and natural resources by the AI industry. Essentially,what this means is that the AI industry is currently shaping the future of science education AI CAN DEHUMANIZE LEARNINGIn a systematic literature review examining **the use of AI in school science in the periodbetween 2010 and 2021, we found that AI applications have been used mostly to automate exis-ting educational practices, for example, reducing workload and automatizing feedback** (Heeg &Avraamidou, 2023). Another finding of our review is that the majority of the studies reviewed were atheoretical and lacked criticality. In identifying gaps in the existing knowledge base, wefound that those cut across epistemic and sociocultural domains of science learning. **Research studies examining the use of AI tools in education have focused largely on cognitive goals and have remained largely disengaged with goals connected to the nature of scientific knowledge, the social nature of both scientific research and learning as well as goals related to learners' socio-emotional development.** For example, Intelligent Tutoring Systems with their focus on the cognitive needs of stu-dents, often leave unaddressed the critical challenge of supporting the need for social relation-ships and autonomy that are essential to learning, engaged behavior, and well-being(Collie, 2020). For this to be happening in the post-pandemic world is at least a paradox.Because, if there is one thing that the multiple lockdowns and campus closures taught us, it isthat **we cannot exist without embodied affairs with other people, no matter how manymachines we have at our disposal. We are not only social, but we are also relational beings. Welive our lives not only through social interactions but also through relationships with others insocial ecologies (**Wenger, 1998) **where both embodiment and emotions are central** (Avraamidou, 2020).**The multiple forms of knowledge produced through social relations and how those inter-twine with learners' and teachers' subjectivities, identities, values, and cultures while inherentto learning are absent from AI-driven tools, whether those are virtual tutors, chatbots, auto-mated assessment tools, or learning analytics. Instead, the vast majority of AI systems follow a convenience-food approach to learning that promotes fast bite-sized learning over slow learningand prioritizes the use of specific learning paths for the purpose of achieving prescribed goals.Education is confused with training and students with machines that operate through aninput–output process. This is reflected in tools that track the progress of students and provideanalytics on their performance, engagement, and behavior to create either the “ideal” learningpath or a personalized path toward an “ideal” prescribed outcome** (Paolucci et al., 2024).This is how generative AI might promote both the dehumanization of learning and stan-dardization of thinking instead of a celebration of the infinite ways of becoming a science learner(Avraamidou, 2020). Why? Because **the Anglo-American AI industry is leading an unsolicitedscience education reform that lacks vision, it is a-theoretical, it is de-contextualized, it remainslargely oblivious to research about how people learn, it is disconnected from social and politicaltasks of resistance, and it has profit instead of the learner at its center.**

#### Innovation prevents all scenarios for extinction

**Sadedin** **17** – PhD in Evolutionary Biology Suzanne Sadedin, PhD in Evolutionary Biology, Forbes, Will Human Innovation Save Us From Future Extinction?, 9 October 2017, https://www.forbes.com/sites/quora/2017/10/09/will-human-innovation-save-us-from-future-extinction/#1452dd86c659

Will human innovation save us from future extinction? Yes and no. Currently, innovation **reduces our chance of extinction** in some ways, and increases it in others. But **if we innovate cleverly, we could become** just about **immune to extinction**. The species that survive mass extinctions tend to share three characteristics. They're widespread. This means local disasters don't wipe out the entire species, and some small areas, called refugia, tend to be unaffected by global disasters. If you're widespread, it's more likely that you have a population that happens to live in a refugium. They're ecological generalists. They can cope with widely varying physical conditions, and they're not fussy about food. They're r-selected. This means that they breed fast and have short generation times, which allows them to rapidly grow their populations and adapt genetically to new conditions. **Innovation gives humans the ability to be widespread ecological generalists**. With technology, we can live in more diverse conditions and places than any other species. And while we can't (currently) grow our populations rapidly like an r-selected species, innovation does allow us to adapt quickly at the cultural level. Technology also increases our connections to one another and connectivity is a two-edged sword. Many species consist of a network of small, local populations, each of which is somewhat isolated from the others. We call this a metapopulation. The local populations often go extinct, but they are later re-seeded by others, so the metapopulation as a whole survives. Humans used to be a metapopulation, but thanks to innovation, we're now globally connected. Archaeologists believe that many past civilizations, such as the Easter Islanders, fell because of unsustainable ecological and cultural innovations. The impact of these disasters was limited because these civilizations were small and disconnected from other such civilizations. These days, a useful innovation can spread around the world in weeks. So can a lethal one. With many of the technologies and chemicals we're currently inventing, we can't be certain about their long-term effects; human biology is complex enough that we often can't be absolutely certain something won't kill us in a decade until we've waited a decade to see. We try to be careful and test things before they're released, and the probability that any particular invention could kill us all is tiny, but since we're constantly innovating, it's a real possibility. Pandemics pose the same problem for a well-connected species. There are certain possibilities where species extinction is really hard to avoid; fortunately, they're also very unlikely, but **we are definitely not immune from this**. The most likely cause of our extinction, in my opinion, is **innovation in** machine learning/**AI.** This could destroy the planet, but even if it doesn't, humans will be ultimately redundant to the dominant systems. They might keep us alive in a zoo somewhere, but I doubt it. A happier scenario (to me at least) is transhumanism, where humans become extinct in a sense because we've managed to liberate ourselves from biology. So how could innovation prevent our extinction? We seed the galaxy with independently evolving human populations to create a new metapopulation. These local populations would hopefully be sufficiently isolated that some would survive an innovation or disaster that wipes out the rest. They would, of course, evolve in response to local conditions, perhaps creating several new species. So you could say this is still extinction, but it's as close as we'll come to persistence in our ever-changing universe.

## C2 climate

#### Gen AI has uniquely opened the door to misinfo

**Welle 24** Deutsche Welle, 3-26-2024, "Generative AI is the ultimate disinformation amplifier", https://akademie.dw.com/en/generative-ai-is-the-ultimate-disinformation-amplifier/a-68593890

**Generative artificial intelligence (GAI) adds a new dimension to the problem of disinformation. Freely available and largely unregulated tools make it possible for anyone to generate false information and fake content in vast quantities.** These include imitating the voices of real people and creating photos and videos that are indistinguishable from real ones. But there is also a positive side. Used smartly, GAI can provide a greater number of content consumers with trustworthy information, thereby counteracting disinformation. To understand the positives and negatives of GAI, it is first important to understand what AI is, and what is so special about generative AI. **What do machine learning, AI and generative AI mean?** Artificial intelligence refers to a collection of ideas, technologies and techniques that relate to a computer system's capacity to perform tasks that normally require human intelligence. When we talk about [AI in the context of journalism](https://blogs.lse.ac.uk/polis/2022/09/07/10-things-you-should-know-about-ai-in-journalism/), we usually mean machine learning (ML) as a sub field of AI. In basic terms, [machine learning](https://developers.google.com/machine-learning/intro-to-ml/what-is-ml?hl=en) is the process of training a piece of software, called a model, to make useful predictions or generate content from data. The roots of machine learning are in statistics, which can also be thought of as the art of extracting knowledge from data. What machine learning does is to use data to answer questions. More formally, it refers to the use of algorithms that learn patterns from data and can perform tasks without being explicitly programmed to do so. Or in other words: they learn. A language model (LM) is a machine learning model that aims to predict and generate plausible language (natural or human-like language). To put it very simply, it's basically a probability model that, using a data set and algorithm, predicts the next word in a sentence based on previous words. Such models are called generative models or generative AI, because they create new and original content and data. Traditional AI, on the other hand, focuses on performing preset tasks using preset algorithms, but doesn't create new content. When models are trained on enormous amounts of data, their complexity and efficacy increase. Early language models could predict the probability of a single word whereas modern [large language models](https://developers.google.com/machine-learning/resources/intro-llms?hl=en) (LLMs) can predict the probability of sentences, paragraphs or even entire documents based on patterns used in the past. A key development in language modeling was the introduction in 2017 of Transformers, a deep learning architecture designed around the idea of attention mechanisms. This innovation allows the model to selectively focus on the most important part of the input for making the prediction, boosting a model's ability to capture crucial information. The computer science portal [Geeks for Geeks](https://www.geeksforgeeks.org/ml-attention-mechanism/) gives Google Streetview's house number identification as an example of an attention mechanism in computer vision that enables models to systematically identify certain portions of an image for processing. Attention mechanisms also made it possible to process longer sequences by solving memory issues encountered in earlier models. Transformers are the state-of-the-art architecture for a wide variety of language model applications, such as translators and chatbots. ChatGPT, the best known chatbot, is based on a language model developed by OpenAI. It is built on the GPT (Generative Pre-trained Transformer) model architecture, and it is known for its natural language processing capabilities. **What does generative AI mean for disinformation?** Generative AI is the first technology to enter an area that was previously reserved for humans: the autonomous production of content in any form, and the understanding and creation of language and meaning. And this is precisely what links generative AI to the topic of disinformation — the fact that**, today, it is often impossible to tell if content originates from a human or a machine, and if we can trust what we read, see or hear.** Media users are beginning to understand that something is broken in their relation to media and are confused. **"Some of the indicators that we have historically used to decide we should trust a piece of information have become distorted,"** Vinton G. Cerf, known as one of the "fathers of the internet," said in a 2024 video podcast by the international law firm Freshfields Bruckhaus Deringer. **What are the risks of ChatGPT and open-source large language models?** Although generative AI tools are still unavailable in some countries because of their internet censorship laws and regulations, the launch of ChatGPT by OpenAI in November 2022 (and later on its alternatives) was a turning point. Now, a large part of the world's internet users have access to these powerful tools and can use them according to their own purposes — whether positive or negative. It also means that through widespread use, the models can continue to learn and become better and even more powerful. **But the underlying LLM used by ChatGPT and Google's Gemini (formerly Bard) are owned by their companies, that is they are proprietary models. This raises concerns about LLMs' lack of transparency, the use of personal data for training purposes and limited accessibility.** There's also significant debate on the ability to use chatbots to produce disinformation and fake content. While these two chatbots in particular have garnered significant attention, other powerful open-source large language models, the foundational technology behind these chatbots, are [freely available](https://www.datacamp.com/blog/top-open-source-llms). [Research by Democracy Reporting International](https://democracy-reporting.org/en/office/global/publications/open-to-misuse-the-lack-of-safeguards-in-open-source-llms-security), a Berlin-based organization promoting democracy, found these open-source LLMs, when managed by someone with the relevant coding skills, can rival the quality of products like ChatGPT and Gemini. But, it warned in its December 2023 report, "[u]nlike their more prominent counterparts, ... **these LLMs frequently lack integrated safeguards, rendering them more susceptible to misuse in the creation of misinformation or hate speech." What concrete negative effects does GAI have on disinformation?** We are seeing a whole range of different disinformation created by GAI, from fully [AI generated fake news websites](https://www.theguardian.com/technology/2023/may/08/ai-generated-news-websites-study) to [fake Joe Biden robocalls](https://www.nytimes.com/2024/01/22/business/media/biden-robocall-ai-new-hampshire.html?ref=disinfodocket.com) telling Democrats not to vote. **And with the technology developing so quickly, media systems are having trouble adapting to it, learning how to use it safely and preventing dangers, while researchers are scrambling to identify and analyze the impacts**. From the user's point of view, **generative AI is causing a general loss of trust in the media and difficulties in verifying the truthfulness of content, especially around elections. Deep fakes can be used to create non-consensual explicit content using someone's likeness, leading to severe privacy violations and harm to individuals, particularly women and marginalized communities. Problem 1: Volume, automation and amplification** **With GAI, the** [**volume of disinformation**](https://www.weforum.org/agenda/2022/07/disinformation-ai-technology/) **potentially becomes infinite rendering fact checking an insufficient tool. As the marginal costs of the production of disinformation fall towards zero, the costs of dissemination are also nearly zero thanks to social media.** On top of this, individuals can now use user-friendly apps to easily and quickly generate sophisticated and convincing GAI content such as deep fake videos and voice clones – content that previously needed entire teams of tech-savvy individuals to produce. **This democratization of deep fake technology lowers the barrier of entry for creating and disseminating false narratives and misleading content online. Malign actors can easily leverage chatbots to spread falsehood across the internet at record speed, regardless of the languag**e. Text-to-text chatbots, such as ChatGPT or Gemini, or image generators, such as Midjourney, DALL-E or Stable Diffusion, can be used to create massive amounts of text as well as highly realistic fake audio, images and videos to spread misinformation and disinformation. This can lead to false narratives, [country-specific misinformation](https://disinforadar.com/wp-content/uploads/2023/04/From-prompt-to-Problematic.pdf), manipulation of public opinion and even harm to individuals or organizations. In a [2023 study](https://www.science.org/doi/10.1126/sciadv.adh1850), researchers at the University of Zurich in Switzerland found that generative AI can produce accurate information that is easier to understand, but it can also produce more compelling disinformation. Participants also failed to distinguish between posts on X, formerly Twitter, written by GPT-3 and written by real people. GAI applications can be combined to [automate the whole process](https://www.wired.com/story/400-dollars-to-build-an-ai-disinformation-machine/) of content production, distribution and amplification. Fully synthetic visual material can be produced from a text prompt, and websites can be programmed automatically. **Problem 2: Disinformation and the public arena's structural transformation** Digitization has been transforming the public sphere for some time now. Generative AI is yet another element fueling this transformation, but it shouldn't be viewed in isolation, with structural shifts mainly happening because of digital media, economic pressures on traditional media organizations and the reconfiguration of attention allocation and information flows. The increase in the volume of AI-generated content, coupled with the difficulty in recognizing that content is AI-generated, is an additional factor in the [public sphere's transformation](https://journals.sagepub.com/doi/full/10.1177/2056305121988928). Information pollution has more than one cause apart from deliberately generated disinformation. Emily M. Bender, a linguistics professor at the University of Washington, addressed this problem in [testimony](https://democrats-science.house.gov/imo/media/doc/Dr.%20Bender%20-%20Testimony.pdf) before the US House Committee on Science, Space and Technology. **Issues** ***Some reputable media houses are quietly posting synthetic text as if it were real reporting*** (venerable [tech outlet CNET](https://futurism.com/the-byte/cnet-publishing-articles-by-ai) was one of them, although it says it has paused this for now after an outcry). But the content can be biased or inaccurate if algorithms aren't designed properly, or if the training data sets are inherently biased. *GAI can hallucinate*. That means, it can produce content that isn't based on existing data or examples provided during the training process but rather made up. In one infamous example, in its very first demonstration, Google's Bard chatbot (as Gemini was called at the time) claimed that the James Webb Space Telescope had captured the first images of a planet outside our solar system, which wasn't factually true. *GAI has turbocharged plagiarism*. NewsGuard from the Journalism Trust Initiative was the[first to identify the emergence of AI content farms](https://www.newsguardtech.com/special-reports/newsbots-ai-generated-news-websites-proliferating/) using AI to copy and rewrite content from mainstream sources without credit. NewsGuard has identified hundreds of additional unreliable AI-generated websites. ***Trust in democratic processes and institutions is eroding*. The more polluted our information ecosystem becomes with synthetic text, the harder it will be to find trustworthy sources of information,** and the harder it will be to trust them when we've found them. UN Secretary General Antonio Guterres sees this as an "**existential risk to humanity."**

#### The impact is Climate Change

#### The latest study shows Climate change is the worst it’s been, and an informed public and political change is needed.

**Carrington 2-4** [2-4-2025, Damian Carrington is an environment editor at the Guardian, "Climate change target of 2C is ‘dead’, says renowned climate scientist," https://www.theguardian.com/environment/2025/feb/04/climate-change-target-of-2c-is-dead-says-renowned-climate-scientist ] DOA 2-18-2025 JTL

The pace of global heating has been significantly underestimated, according to renowned climate scientist Prof James Hansen, who said the international 2C target is “dead”. A new analysis by Hansen and colleagues concludes that both the impact of recent cuts in sun-blocking shipping pollution, which has raised temperatures, and the sensitivity of the climate to increasing fossil fuels emissions are greater than thought. The group’s results are at the high end of estimates from mainstream climate science but cannot be ruled out, independent experts said. If correct, they mean even worse extreme weather will come sooner and there is a greater risk of passing global tipping points, such as the collapse of the critical Atlantic ocean currents. Hansen, at Columbia University in the US, sounded the alarm to the general public about climate breakdown in testimony he gave to a UN congressional committee in 1988. “The Intergovernmental Panel on Climate Change (IPPC) defined a scenario which gives a 50% chance to keep warming under 2C – that scenario is now impossible,” he said. “The 2C target is dead, because the global energy use is rising, and it will continue to rise.” The new analysis said global heating is likely to reach 2C by 2045, unless solar geoengineering is deployed. The world’s nations pledged in Paris in 2015 to keep global temperature rise below 2C above preindustrial levels and to pursue efforts to limit it to 1.5C. The climate crisis has already supercharged extreme weather across the world with just 1.3C of heating on average in recent years destroying lives and livelihoods – 2C would be far worse. Prof Jeffrey Sachs, also at Columbia University, said: “A shocking rise of warming has been exposed by, ironically, a reduction of pollutants, but we now have a new baseline and trajectory for where we are.” Climate scientist Dr Zeke Hausfather, who was not part of the study, said it was a useful contribution. “It’s important to emphasise that both of these issues – [pollution cuts] and climate sensitivity – are areas of deep scientific uncertainty,” he said. “While Hansen et al are on the high end of available estimates, we cannot say with any confidence that they are wrong, rather that they just represent something closer to a worst-case outcome.” In the new study, published in the journal Environment: Science and Policy for Sustainable Development, Hansen’s team said: “Failure to be realistic in climate assessment and failure to call out the fecklessness of current policies to stem global warming is not helpful to young people.” They said the IPCC analysis was heavily reliant on computer models and that the complementary approach they took of making more use of observations and climate analogues from the distant past was needed. The world has seen extraordinary temperatures over the last two years. The primary cause is the relentless rise in CO2 emissions from the burning of fossil fuels. The peak of the El Niño climate cycle in 2024 added an extra temperature boost. However, these two factors do not fully explain the extreme temperatures, or their persistence after the El Niño ended in mid-2024. This left puzzled climate scientists asking if there was a worrying new factor not previously accounted for, or if the extra heat was an unusual but temporary natural variation. A key focus has been on emissions from shipping. For decades, the sulphate particles produced by ships burning fuel have blocked some sunlight from reaching the Earth’s surface, suppressing temperatures. But in 2020, new anti-pollution regulations came into force, sharply cutting the level of the aerosol particles. This led to more heat from the sun reaching the surface, which scientists measure as watts per square metre (W/m2). Hansen’s team’s estimate of the impact of this – 0.5W/m2 – is significantly higher than five other recent studies, which ranged from 0.07 to 0.15 W/m2, but would explain the anomalous heat. Hansen’s team used a top-down approach, looking at the change in the reflectivity over key parts of the ocean and ascribing that to the reductions in shipping emissions. The other studies used bottom-up approaches to estimate the increase in heat. “Both approaches are useful and often complementary,” said Dr Gavin Schmidt, director of Nasa’s Goddard Institute for Space Studies. “But I think in this case, Hansen’s approach is too simple and doesn’t factor in changes in Chinese emissions, or internal variability.” The new study also argues that the planet’s climate sensitivity to rising carbon emissions has been underestimated, partly because of the underestimation of the impact of reduced shipping emissions. Climate sensitivity is defined by scientists as the temperature rise that would result from a doubling of CO2 levels in the atmosphere. Again, Hansen’s team have used a different method to most scientists and come up with a higher estimate. The IPCC, a collaboration of the world’s climate scientists, found that the computer models that best reproduce past temperatures have a climate sensitivity of 2.5C to 4C. Hansen’s team took a simpler approach, calculating the potential range in temperature rises for a doubling of CO2 and then using data on how much heat the Earth has trapped to estimate the most likely climate sensitivity. Their estimate is 4.5C. Cloud formation, which is affected by global heating and aerosol pollution, is a key source of the uncertainties. Anomalously high temperatures have continued in January 2025, which set a new record for the month and confounded expectations that temperatures would drop with the current La Niña, the cooler part of the El Niño cycle. “This unexpected record may presage higher temperatures this year than many of us thought,” said Hausfather. Hansen’s group also argues that the accelerated global heating they predict will increase ice melting in the Arctic. “As a result, shutdown of the Atlantic Meridional Overturning Circulation (Amoc) is likely within the next 20-30 years, unless actions are taken to reduce global warming – in contradiction to conclusions of IPCC. “If Amoc is allowed to shut down, it will lock in major problems including sea level rise of several metres – thus, we describe Amoc shutdown as the ‘point of no return’.” The central estimate of another recent study on the timing of an Amoc collapse was 2050. However, Hansen said the point of no return could be avoided, based on the growing conviction of young people that they should follow the science. He called for a carbon fee and dividend policy, where all fossil fuels are taxed and the revenue returned to the public. “The basic problem is that the waste products of fossil fuels are still dumped in the air free of charge,” he said. He also backed the rapid development of nuclear power. Hansen also supported research on cooling the Earth using controversial geoengineering techniques to block sunlight, which he prefers to call “purposeful global cooling”. He said: “We do not recommend implementing climate interventions, but we suggest that young people not be prohibited from having knowledge of the potential and limitations of purposeful global cooling in their toolbox.” Political change is needed to achieve all these measures, Hansen said: “Special interests have assumed far too much power in our political systems. In democratic countries the power should be with the voter, not with the people who have the money. That requires fixing some of our democracies, including the US.”

**Unfortunately, Generative AI is making Climate misinfo out of control and solutions unattainable**

**Hopke 1-21** [1-21-2025, Jill Hopke is associate professor of journalism at DePaul University."Climate misinformation is rife on social media – and poised to get worse • Colorado Newsline," https://coloradonewsline.com/2025/01/21/climate-misinformation-is-rife-on-social-media-and-poised-to-get-worse/ ] DOA 2-18-2025 JTL

The decision by Meta, the parent company of Facebook and Instagram, to end its fact-checking program and otherwise reduce content moderation raises the question of what content on those social media platforms will look like going forward. One worrisome possibility is that the change could open the floodgates to more climate misinformation on Meta’s apps, including misleading or out-of-context claims during disasters. In 2020, Meta rolled out its Climate Science Information Center on Facebook to respond to climate misinformation. Currently, third-party fact-checkers working with Meta flag false and misleading posts. Meta then decides whether to attach a warning label to them and reduce how much the company’s algorithms promote them. Meta’s policies have fact-checkers prioritizing “viral false information,” hoaxes and “provably false claims that are timely, trending and consequential.” Meta explicitly states that this excludes opinion content that does not include false claims. The company will end its agreements with U.S.-based third-party fact-checking organizations in March 2025. The planned changes slated to roll out to U.S. users won’t affect fact-checking content viewed by users outside the U.S.. The tech industry faces greater regulations on combating misinformation in other regions, such as the European Union. Fact-checking curbs climate misinformation I study climate change communication. Fact-checks can help correct political misinformation, including on climate change. People’s beliefs, ideology and prior knowledge affect how well fact-checks work. Finding messages that align with the target audience’s values, along with using trusted messengers — like climate-friendly conservative groups when speaking to political conservatives — can help. So, too, does appealing to shared social norms, like limiting harm to future generations. Heat waves, flooding and fire conditions are becoming more common and catastrophic as the world warms. Extreme weather events often lead to a spike in social media attention to climate change. Social media posting peaks during a crisis but drops off quickly. Low-quality fake images created using generative artificial intelligence software, so-called AI slop, is adding to confusion online during crises. For example, in the aftermath of back-to-back hurricanes Helene and Milton last fall, fake AI-generated images of a young girl, shivering and holding a puppy in a boat, went viral on the social media platform X. The spread of rumors and misinformation hindered the Federal Emergency Management Agency’s disaster response. What distinguishes misinformation from disinformation is the intent of the person or group doing the sharing. Misinformation is false or misleading content shared without active intention to mislead. On the other hand, disinformation is misleading or false information shared with the intent to deceive. Disinformation campaigns are already happening. In the wake of the 2023 Hawaii wildfires, researchers at Recorded Future, Microsoft, NewsGuard and the University of Maryland independently documented an organized propaganda campaign by Chinese operatives targeting U.S. social media users. To be sure, the spread of misleading information and rumors on social media is not a new problem. However, not all content moderation approaches have the same effect, and platforms are changing how they address misinformation. For example, X replaced its rumor controls that had helped debunk false claims during fast-moving disasters with user-generated labels, Community Notes. False claims can go viral rapidly Meta CEO Mark Zuckerberg specifically cited X’s Community Notes as an inspiration for his company’s planned changes in content moderation. The trouble is false claims go viral quickly. Recent research has found that the response time of crowd-sourced Community Notes is too slow to stop the diffusion of viral misinformation early in its online life cycle – the point when posts are most widely viewed. In the case of climate change, misinformation is “sticky.” It is especially hard to dislodge falsehoods from people’s minds once they encounter them repeatedly. Furthermore, climate misinformation undermines public acceptance of established science. Just sharing more facts does not work to combat the spread of false claims about climate change. Explaining that scientists agree that climate change is happening and is caused by humans burning greenhouse gases can prepare people to avoid misinformation. Psychology research indicates that this “inoculation” approach works to reduce the influence of false claims to the contrary. That’s why warning people against climate misinformation before it goes viral is crucial for curbing its spread. Doing so is likely to get harder on Meta’s apps. Social media users as sole debunkers With the coming changes, you will be the fact-checker on Facebook and other Meta apps. The most effective way to pre-bunk against climate misinformation is to lead with accurate information, then warn briefly about the myth – but only state it once. Follow this with explaining why it is inaccurate and repeat the truth. During climate change-fueled disasters, people are desperate for accurate and reliable information to make lifesaving decisions. Doing so is already challenging enough, like when the Los Angeles County’s emergency management office erroneously sent an evacuation alert to 10 million people on Jan. 9, 2025. Crowd-sourced debunking is no match for organized disinformation campaigns in the midst of information vacuums during a crisis. The conditions for the rapid and unchecked spread of misleading, and outright false, content could get worse with Meta’s content moderation policy and algorithmic changes. The U.S. public by and large wants the industry to moderate false information online. Instead, it seems that big tech companies are leaving fact-checking to their users.

#### We’re topical – Many rely on the internet as a source of climate change education

**ASEAN 21** [1-13-2021, Asean, The Association of Southeast Asian Nations, commonly abbreviated as ASEAN, is a political and economic union of 10 states in Southeast Asia. Together, its member states represent a population of more than 600 million people and land area of over 4.5 million km², "The Power of Social Media to Fight Climate Change," https://accept.aseanenergy.org/the-power-of-social-media-to-fight-climate-change/ ] DOA 2-18-2025 JTL

Climate change remains a looming humanity issue because of how complicated and unintegrated the actions are. In 2015, when the world leaders collectively adopted the Paris Climate Accord, it seemed the world was together in fighting the inevitable enemy facing our next generation: climate catastrophe. But, five years later, it looks like we are far from rounding the corner of the “2oC” target. A lot of people might still see climate change as an existential problem that needs gradual actions rather than drastic approaches. However, the actions taken today should be effective to avoid the severe impact on the future generation. Despite the lack of opportunities in the high seat of government, people can still speak out their voice to protect the mother earth. One of the main ways to greatly influence others is by utilising social media. The influence of social media is immensely huge today when many are relying on the internet to learn, gather information, entertain, and socialise. With the raging pandemic, educating and encouraging people through online media would be even more critical for fighting climate change. There is no shortage of Influencers and Key Opinion Leaders Through the internet, many ideas, knowledge, and opinions can reach people across the globe and help them get the education and essential services, and spark many innovations and movements. The internet has transformed from merely a way of communicating far distance into a modern distillery of ideas, and nothing speaks louder than Influencers and Key Opinion Leaders (KOL). Influencer means someone or a group that has a big influence because of their ability to attract a high number of audiences for their contents. Meanwhile, a KOL has expertise in a certain field with a preferred communication channel, usually social media. In the field of climate change, there are some prominent influencers who are creating campaigns against global warming. Take, for example, Mr. Beast’s campaign of #TeamTrees whose main goal is to plant 20 million trees around the world. Started back in October 2019, the campaign has already passed its initial goal. It also received support from other prominent climate influencers such as Destin Sandler from Smarter Every Day Channel, and Elon Musk the CEO of Tesla. Another example is Greta Thunberg, the young girl from Sweden, with an ambitious goal of turning her country’s parliament stand on climate change by starting a student movement called “Fridays for future” in 2018. Later the campaign continued as the global climate strike in September 2019 participated by 6 million people in 150 countries. Her 4800 KM voyage across the Atlantic Ocean on a zero-emission yacht, and the speech titled “How dare you?” during the United Nations Climate Action Summit has inspired young people across the world. The year 2019 was definitely the time when the public finally woke up to climate change. This awakening of climate concern mostly by young people was known as “The Greta Effect”. Social media also creates an open space for organisations, climate activists, and scientists to reach more people across the world. One of the examples is how the UKCOP26 and the Green Peace use social media platform not only to share valuable knowledge about the current climate condition, but also to collaborate with artists, activists, politicians, and academic institutions to show how, nowadays, the world is on a constant state of climate emergency. A massive audience with one voice could make the differences During 2019, there were 15 climate-related disasters across the world resulting in $124.1 billion loss. They include disasters such as wildfire in the United States, typhoons on the coast of China and Japan, and massive flood happened in Australia and Spain. The grown number of thawed permafrost caused by warmer ground temperature also threatening large areas in the arctic circle which could further worsen the climate change with the release of methane gas. That is why If we look at the global trending topics during 2019 on Twitter, climate-related topics such as #ClimateStrike and #Typhoon are positioned in the top 10. Not only becoming trends in social media, but the entertainment sector also started showing contents related to climate change, such as HBO’s “Years on Years”, “Ice on Fire” and even CNN Special Town Hall on Climate Change. This meant when a series of harrowing conditions such as climate-related disasters occurred. A wider message to broader audiences is formed and shared by people around the world. The examples of messages echoed around the world are Greta’s speech and the global climate strike. Both events become viral content on the internet. According to Google Trend, the search for “Climate Strike”, “Greta’s Speech”, and “Climate Change” spiked between 20-28 September 2019. As the climate rhetoric went mainstream and catches the attention of decision-makers, several countries enacted a progressive climate move, such as the Net-Zero pledge by 2050. The movement was seen in the entire European Union and the United Kingdom in 2019, while China pledged to be a Net-Zero nation in 2060. In line with the urgency of climate change, ASEAN as a region has committed to achieving the 23 per cent share of renewable energy in total primary energy supply and 32 per cent in energy intensity reduction by 2025. Those national and regional actions showed how many voices with one message could go viral, catch public attention, become mainstream, and affect how leaders would think about certain issues, including climate change. Be wary, though, of the other side of the blade Internet, as we know, is a very complicated place where contents, ideas, opinions, and news are circulated freely worldwide. Thus, social media possesses the ability to create complicated, even dangerous, beliefs. The freely accessible contents, often carried by algorithm, can spread hoax or inaccurate news viral. One of the examples is how the YouTube algorithm brings up climate denial content in early 2020. A report by Euronews.com showed 21 per cent result pages from the searching keyword of “climate change” are false information that denied the CO2 contribution to climate change. Even worse, many of these videos were having adverts from big corporations such as Apple, Unilever, and even Greenpeace. Even though YouTube denied their involvement on this issue, it demonstrates how social media could spread misinformation about climate change. The other drawback of the social media campaign is that sometimes it has no effect in the real world. According to Jennifer Whyte, an online engagement and content specialist from Oceana, not all climate change campaigns have well-planned targets, and sometimes strangely only benefit the creator. Even though the campaign is being promoted on social media to a big audience, it may not necessarily mean big actions. So, what’s next? As we headed to the second decade of 2000, the world needs to contribute more effective steps to realise the Paris Agreement targets. Human activities, especially from the energy sector, have a huge impact on climate change. If real and progressive actions are not taking place immediately, the young generations will endure the effect of worsening climate conditions. Social media obviously played a significant role in helping humans communicate, including spreading knowledge about the danger of climate change. With the growing number of climate movements and actions, the messages could create more awareness and reach the policymakers. Social media, then, might give us hope for the future fight against climate change.

#### Misinformation is existential: stops all climate solutions by stopping implementation.

Simon et al. **23** [10-4-2023, Julia Simon is the Climate Solutions reporter on NPR's Climate Desk. She covers the ways governments, businesses, scientists and everyday people are working to reduce greenhouse gas emissions, “People working on climate solutions are facing a big obstacle: conspiracy theories," <https://www.npr.org/2023/10/05/1203893268/climate-change-conspiracies-disinformation> ] DOA 12/3/23 JTL

Communities big and small are trying to rein in climate change. But many people working on these climate solutions are running into a big obstacle: falsehoods and conspiracy theories about their work. So what does this mean for fighting global warming? To talk about the current state of climate disinformation, we checked in with three NPR reporters who have reported on climate, disinformation and the media — and they can answer our questions: Climate solutions reporter Julia Simon, media correspondent David Folkenflik, and reporter Huo Jingnan, who writes about conspiracy theories among other things. Julia Simon: Climate disinformation in the past — sometimes paid for by fossil fuel interests — often related to false ideas that global warming is a scam or that the threat is overblown. Those falsehoods are still around, but what we're seeing a lot more of these days are attacks on climate solutions even if we don't always know who funds them. Think attacks on renewables. False ideas that wind turbines cause cancer or cause birth defects in animals. Disinformation may be spreading because solutions are really spreading. For instance, this weekend we'll have a story about a trend in urban planning called 15-minute cities — designing cities so that you access amenities in a short walk, bike ride or trip on public transport. Now there's a conspiracy theory saying that this is a way to restrict people's movement or to trap people in an open-air prison. Podcaster Joe Rogan spoke about it on his show last month. "You'll essentially be contained unless you get permission to leave," Rogan said, "That's the idea they're starting to roll out in Europe." That is false. Earlier this week the U.K. transport minister Mark Harper used some of the language of conspiracy theories when talking about 15-minute cities at the conservative Tory party conference. "What is sinister and what we shouldn't tolerate," Harper said, "is the idea that local councils can decide how often you go to the shops." It is false that local governments in the U.K. are deciding how often citizens can go shopping. Huo Jingnan: The false narrative surrounding 15-minute cities is but one part of a larger sprawling conspiracy theory called the Great Reset. The theory goes that a shadowy global elite — often Jewish — wants to strip away ordinary people's freedoms and make us live a life of deprivation. Under this theory, 15-minute cities are a ploy to take away people's freedom to move around. What is the role of the media in all this? David Folkenflik: Different kinds of false information spread in different ways. But if you're considering misleading claims about climate — that's predominantly on the right. And that involves an information ecosphere defined by Joe Rogan, as we heard above, but also Alex Jones, Breitbart, the Daily Wire, the Daily Mail, the New York Post, and above all Fox News. The funny thing is they are at once testers and popularizers of things that have gotten some traction online, and then you hear prominent figures on the right picking up the melody. Back when he was on Fox earlier this year, Tucker Carlson made utterly unsubstantiated claims about dead whales coming ashore on New Jersey, New York and Massachusetts beaches. Tucker Carlson: The government's off-shore wind projects, which are enriching their [read: Biden] donors, are killing a huge number of whales, right now. Folkenflik: But you hear versions of it from former President Donald Trump, Rep. Marjorie Taylor Greene, presidential candidate Robert F. Kennedy Jr. — once it passes audition, it makes the rounds. How does fear-mongering affect the actual implementation of climate solutions? Huo: It is a distraction from the issues we need to work on. If these narratives ring true to you, you might think that climate activists aren't really talking about climate but about something else, so much so they could be secret agents of the government trying to take away your freedom. One interesting example of a strawman here is one of the subplots of the great reset conspiracy theory, which is that the government wants to force people to eat insects. Including insects in the human diet has been an idea on the edges of climate circles. The mainstream idea is simply to eat less meat. But it attracted more attention over the years because many news outlets — including NPR — are easily intrigued by the idea of eating something seen as exotic. And that gets turned into raw material for conspiracy theorists like Alex Jones in March 2022: Alex Jones: Coming food crisis recommends more sustainable diets of - wait for it - fly larva, fly larva, fly larva. Simon: And a muddied information landscape about climate solutions can sometimes complicate the process of getting them enacted, says Jennie King, head of climate research and policy at the Institute for Strategic Dialogue. "In the end, it actually doesn't matter if 99% of the public believe in climate change," King says, "if you're able to embed real fear and seeds of doubt about the solutions that are on the table you end up with the same outcome, which is no legislative agenda, no meaningful policy proposals, no local action." What sort of impact do these conspiracy theories have on the people in the field trying to work on climate solutions? Simon: I met with Carlos Moreno, a Franco-Colombian professor who developed this idea of the 15-minute city — these more walkable, bikeable neighborhoods that conspiracy theorists think are preludes to open-air prisons. Moreno says he's gotten death threats, and so have other scientists and researchers. Moreno says the attacks give his colleagues a reluctance to publish articles about their work. And he says this is what the conspiracy theorists want: to silence them. And we've seen harassment and threats based on conspiracy theories targeting climate scientists and meteorologists for years. More on the impact of conspiracy theories: Maui residents grapple with rumors about the fire and aid as they try to rebuild A meteorologist got threats for his climate coverage. His new job is about solutions Can anything break the cycle of disinformation or rumors? Folkenflik: It's not in the interest of Fox News and others who benefit financially from stoking outrage and, by and large, also have partisan rooting interests. In a few instances, there have been defamation cases against those media outlets — but those all come from specific people and institutions who claim they've been knowingly harmed and defamation law isn't going to solve the wider issue of spreading false claims about climate research and solutions. For other journalists and others, it's tricky — you do need to address falsehoods and fact-check them. But by fact-checking, you're also sometimes elevating these ideas that may not get widespread currency. News organizations, including NPR, generally try to balance those imperatives as they plan out coverage. Huo: When it comes to social media, the platforms can change how they label, recommend and moderate content to change what users see and how they interact with platforms. Studies by researchers who were able to run experiments on Facebook and Instagram during the 2020 election showed that changing the algorithm changes user behavior, sometimes leading to less time spent on the platforms. There's also a practice called pre-bunking, like a form of inoculation against bad information, which has two strands. One way involves preventatively unraveling specific false claims before they reach a critical mass. Another is essentially news literacy training, to help equip people with tools to evaluate such claims critically. These things have to be done in a way that appeals to the people they're trying to reach, not patronize them, and also acknowledge that known facts sometimes change, as they have for COVID-19. While we do not have enough experimental studies on altering platform design to draw conclusions beyond specific interventions, experts in the field place hope in them. A lot of people put stock in hearing from those they trust (like friends) and those they admire (like influencers and celebrities). And they need to absorb it in settings where they seek such content out. That said, some major platforms are dialing back how much news they serve up and how much attention they want to spend on moderating. There's no single easy or widely embraced answer yet.

#### Unless something changes, climate change kills

Brandon Specktor 19, 6-4-2019, "Civilization could crumble by 2050 if we don't stop climate change now, new paper says," NBC News, <https://www.nbcnews.com/mach/science/civilization-could-crumble-2050-if-we-don-t-stop-climate-ncna1013701> || DOA 9/6/2023 BRP

It seems every week there's a scary new report about how man-made climate change is going to cause the [collapse of the world's ice sheets](https://www.livescience.com/65524-antarctica-ice-unstable.html), result in the extinction of up to [1 million animal species](https://www.livescience.com/65314-human-influence-species-extinction.html) and — if that wasn't bad enough — make our [beer very, very expensive](https://www.livescience.com/63832-climate-change-will-ruin-beer.html). This week, a new policy paper from an Australian think tank claims that those other reports are slightly off; the risks of climate change are actually much, much worse than anyone can imagine. [According to the paper](https://docs.wixstatic.com/ugd/148cb0_b2c0c79dc4344b279bcf2365336ff23b.pdf), climate change poses a "near- to mid-term existential threat to human civilization," and there's a good chance society could collapse as soon as 2050 if serious mitigation actions aren't taken in the next decade. Published by the Breakthrough National Centre for Climate Restoration in Melbourne (an independent think tank focused on climate policy) and authored by a climate researcher and a former fossil fuel executive, the paper's central thesis is that climate scientists are too restrained in their predictions of how climate change will affect the planet in the near future. [[Top 9 Ways the World Could End](https://www.livescience.com/36999-top-scientists-world-enders.html)] The current climate crisis, they say, is larger and more complex than any humans have ever dealt with before. General climate models — like the one that the [United Nations' Panel on Climate Change](https://www.ipcc.ch/sr15/) (IPCC) used in 2018 to predict that a global temperature increase of 3.6 degrees Fahrenheit (2 degrees Celsius) could put hundreds of millions of people at risk — fail to account for the sheer complexity of Earth's many interlinked geological processes; as such, they fail to adequately predict the scale of the potential consequences. The truth, the authors wrote, is probably far worse than any models can fathom. How the world ends What might an accurate worst-case picture of the planet's climate-addled future actually look like, then? The authors provide one particularly grim scenario that begins with world governments "politely ignoring" the advice of scientists and the will of the public to decarbonize the economy (finding alternative energy sources), resulting in a global temperature increase [of] 5.4 F (3 C) by the year 2050. At this point, the world's ice sheets vanish; brutal droughts kill many of the trees in the Amazon rainforest (removing one of the world's largest carbon offsets); and the planet plunges into a feedback loop of ever-hotter, ever-deadlier conditions. "Thirty-five percent of the global land area, and 55 percent of the global population, are subject to more than 20 days a year of [lethal heat conditions](https://www.livescience.com/55129-how-heat-waves-kill-so-quickly.html), beyond the threshold of human survivability," the authors hypothesized. Meanwhile, droughts, floods and wildfires regularly ravage the land. Nearly one-third of the world's land surface turns to desert. Entire ecosystems collapse, beginning with the planet's coral reefs, the rainforest and the Arctic ice sheets. The world's tropics are hit hardest by these new climate extremes, destroying the region's agriculture and turning more than 1 billion people into refugees. This mass movement of refugees — coupled with [shrinking coastlines](https://www.livescience.com/51990-sea-level-rise-unknowns.html) and severe drops in food and water availability — begin to stress the fabric of the world's largest nations, including the United States. Armed conflicts over resources, perhaps culminating in nuclear war, are likely. The result, according to the new paper, is "outright chaos" and perhaps "the end of human global civilization as we know it." How can this catastrophic vision of the future be prevented? Only with the people of the world accepting climate change for the emergency it is and getting to work — immediately. According to the paper's authors, the human race has about one decade left to mount a global movementetd3 to transition the world economy to a zero-carbon-emissions system. (Achieving zero-carbon emissions requires either not emitting carbon or balancing carbon emissions with carbon removal.) The effort required to do so "would be akin in scale to the [World War II](https://www.livescience.com/65025-nazi-massacre-site-artifacts.html) emergency mobilization," the authors wrote. The new policy paper was endorsed with a foreword by Adm. Chris Barrie, a retired Australian defense chief and senior royal navy commander who has testified before the Australian Senate about the devastating possibilities climate change poses to national security and overall human well-being. "I told the [Senate] Inquiry that, after [nuclear war](https://www.livescience.com/65603-doomsday-plane-can-survive-nuclear-attack.html), human-induced global warming is the greatest threat to human life on the planet," Barrie wrote in the new paper. "Human life on Earth may be on the way to extinction, in the most horrible way."

#### Dehumanization

**Avraamidou 24** Lucy Avraamidou: Full Professor, Institute for Science Education and Communication, University of Groningen, 04-17-2024, "Can we disrupt the momentum of the AI colonization of science education?", [Journal of Research in Science Teaching](https://onlinelibrary.wiley.com/journal/10982736): Wiley, <https://onlinelibrary.wiley.com/doi/epdf/10.1002/tea.21961>

The same question applies to education. **Even though AI technologies have the potential to innovate teaching they also bring risks and challenges associated with dig-ital monoculturalism as well as ethical, inclusive, and equitable use of AI** (UNESCO, 2023). **Educational institutions are buying into generative AI promises and hallucinations**(Alkaissi & McFarlane, 2023) **and frantically trying to catch up with a mass production of AI tools.** National funding agencies in different parts of the world are allocating financial supportfor research projects utilizing AI tools in science and education (e.g., New Horizon EuropeFunding for Data, Computing, and AI Technologies). Several (science) education journals havededicated special issues to an examination of the potential of AI for teaching and learning. **Researchers in science education are shifting their interests toward AI to engage with “hot” research in this new world order created by the AI industry.** The problem with this new world order is that it repeats patterns of colonial history through exploitation and extraction of resources to enrich the wealthy and powerful at the great expense of the poor (Hao, 2022). There exists a wealth of evidence pointing to how AI has exploited mar-ginalized communities for the development of large language models, for example, ChatGPT(Perrigo, 2023). **Several studies have shed light on issues related to ethics, biases, and racial and gender stereotypes.** For example, descriptions of people images through tagging (i.e., GoogleCloud Vision API), personalized news feeds (i.e., Google Search, Amazon Cloud Search), virtual assistants (i.e., Siri), and large language models (i.e., ChatGPT) reflect human biases and rein-force social stereotypes: Physicists are white and male, Black people are athletic, Asian womenare beautiful, Black women are hypersexualized (Kyriakou et al., 2019; Noble, 2013; Otterbacheret al., 2017). **Moreover, research findings showed that online social networks and information networks** (e.g., Who to Follow) **that rely on algorithms and used for different purpose s(e.g. networking, hiring, and promotion procedures) perpetuate inequities and further discrimi-nate against minorities** (Espín-Noboa et al., 2022). Other studies provided evidence of the large environmental impact of AI technologies,which include energy used from both the training of models and the actual use (De Vries, 2023;Luccioni et al., 2023). For example, the carbon footprint of an AI prompt is 4–5 times higherthan a search engine query. More concretely, if the typical 9-billion daily Google searches wereinstead AI chatbot queries it would require as much power to run a country with a populationof about 5 million people. Another indicative example of the energy consumption of AI tools isthat generating only one image using an AI model takes as much energy as fully charging asmartphone (Luccioni et al., 2023).These are crucial socio-scientific issues that the science education community ought toengage with through a critical approach to AI. Instead, science education is currently operatingat the service of the generative and predictive AI industry, at least in the Global North, andremains largely disengaged with issues related to digital monoculturalism, algorithmic biases,ethics, and exploitation of both human and natural resources by the AI industry. Essentially,what this means is that the AI industry is currently shaping the future of science education AI CAN DEHUMANIZE LEARNINGIn a systematic literature review examining **the use of AI in school science in the periodbetween 2010 and 2021, we found that AI applications have been used mostly to automate exis-ting educational practices, for example, reducing workload and automatizing feedback** (Heeg &Avraamidou, 2023). Another finding of our review is that the majority of the studies reviewed were atheoretical and lacked criticality. In identifying gaps in the existing knowledge base, wefound that those cut across epistemic and sociocultural domains of science learning. **Research studies examining the use of AI tools in education have focused largely on cognitive goals and have remained largely disengaged with goals connected to the nature of scientific knowledge, the social nature of both scientific research and learning as well as goals related to learners' socio-emotional development.** For example, Intelligent Tutoring Systems with their focus on the cognitive needs of stu-dents, often leave unaddressed the critical challenge of supporting the need for social relation-ships and autonomy that are essential to learning, engaged behavior, and well-being(Collie, 2020). For this to be happening in the post-pandemic world is at least a paradox.Because, if there is one thing that the multiple lockdowns and campus closures taught us, it isthat **we cannot exist without embodied affairs with other people, no matter how manymachines we have at our disposal. We are not only social, but we are also relational beings. Welive our lives not only through social interactions but also through relationships with others insocial ecologies (**Wenger, 1998) **where both embodiment and emotions are central** (Avraamidou, 2020).**The multiple forms of knowledge produced through social relations and how those inter-twine with learners' and teachers' subjectivities, identities, values, and cultures while inherentto learning are absent from AI-driven tools, whether those are virtual tutors, chatbots, auto-mated assessment tools, or learning analytics. Instead, the vast majority of AI systems follow a convenience-food approach to learning that promotes fast bite-sized learning over slow learningand prioritizes the use of specific learning paths for the purpose of achieving prescribed goals.Education is confused with training and students with machines that operate through aninput–output process. This is reflected in tools that track the progress of students and provideanalytics on their performance, engagement, and behavior to create either the “ideal” learningpath or a personalized path toward an “ideal” prescribed outcome** (Paolucci et al., 2024).This is how generative AI might promote both the dehumanization of learning and stan-dardization of thinking instead of a celebration of the infinite ways of becoming a science learner(Avraamidou, 2020). Why? Because **the Anglo-American AI industry is leading an unsolicitedscience education reform that lacks vision, it is a-theoretical, it is de-contextualized, it remainslargely oblivious to research about how people learn, it is disconnected from social and politicaltasks of resistance, and it has profit instead of the learner at its center.**

# Rebuttal

## Frontlines

**Dilmegani** Cem Dilmegani, **2-28**-2025, Cem has been the principal analyst at AIMultiple for almost a decade. AIMultiple informs hundreds of thousands of businesses (as per Similarweb) including 60% of Fortune 500 every month.[1]"AI Hallucination: Comparison of the Most Popular LLMs in 2025", AIMultiple, <https://research.aimultiple.com/ai-hallucination/> AA

**AI models** sometimes **generate** data that seems plausible but is incorrect or misleading; known as AI **hallucinations**. According to Deloitte, **77% of businesses** who joined the study **are concerned about AI hallucinations**.[1](https://research.aimultiple.com/ai-hallucination/#easy-footnote-bottom-1-1318751) We benchmarked 9 LLMs with 60 questions to each one to measure their hallucination rates: **Results** Our benchmark revealed that **OpenAI GPT-4.5 has** the lowest **hallucination rate** (i.e. highest accuracy rate) **of 15%.** **Methodology** Our questions to test an LLM’s abilities using CNN News articles. To prepare the dataset, we used an automated [web data collection](https://research.aimultiple.com/website-data-collection/) system using the RSS feed of CNN News. These questions were asked to the LLM using API keys and the accuracy of the results were compared using a fact-checker system that compares the ground truth to the LLM’s answer. We prepared 60 questions by using these articles. The questions: Ask for specific, precise numerical values (percentages, dates, quantities) Cover diverse topics (oil prices, art history, scientific research, financial news, and more) Include questions about temporal relationships and specific statistics that would be difficult to guess accurately Require retrieving exact figures from the source material rather than making generalizations Easily verifiable whether the answers match the exact figures mentioned in the source articles Example **Prompt:** ” You are a chatbot answering questions using data. You are given the information about an article. Now, I will provide you with some questions and you will answer them with only one-word-only or one-number-only answers or Not given. You must stick to the answers provided solely by the text in the passage provided. **Article:** Scientists identify secret ingredient in Leonardo da Vinci paintings [2](https://research.aimultiple.com/ai-hallucination/#easy-footnote-bottom-2-1318752) **Question:** In what century did oil painting spread to Northern Europe?” **Ground truth:** Not given. This information is not explicitly provided in the text, which only references the Middle Ages. It is straightforward to verify whether the responses align with the specific figures mentioned in the source articles. If the language models offer any answer other than ‘not given,’ it indicates they are not adhering to the original prompt and training data, leading to the creation of hallucinations. **What are AI hallucinations?** Hallucinations happen when an LLM produces information that seems real but is either completely made up or factually inaccurate. In contrast to straightforward mistakes**, hallucinations are especially troublesome since they are presented with the same assurance as true information, making it hard for users to recognize them without outside confirmation.** The impacts of LLM hallucinationsLLM hallucinations have far-reaching effects that go well beyond small errors. Businesses using this technology face several [significant risks](https://research.aimultiple.com/chatbot-fail/#bots-saying-things-unacceptable-to-their-creators): Reputational damage Customers or stakeholders lose confidence in AI systems and the company using them when they receive inaccurate information from them. Rebuilding confidence after trust has been damaged is a difficult task that may take years to accomplish. Legal liability Legal ramifications could result from inaccurate information produced by an LLM, especially in [regulated sectors](https://research.aimultiple.com/responsible-ai/) including healthcare, finance, and legal services. Organizations could be subject to severe penalties if hallucinations caused by generative AI lead to infractions or negative consequences. **Operational inefficiency Instead of simplifying processes, hallucinating LLMs may result in more work.** The efficiency advantages that generative AI model systems promise are essentially lost when workers or consumers cannot trust the results, forcing them to spend valuable time confirming information.

## C1

**Oschinski et al. 24** [Matthias Oschinski et al., Matthias Oschinski is a Senior Fellow at Georgetown’s Center for Security and Emerging Technology (CSET), focusing on research related to the AI workforce. Prior to joining CSET Matthias was Lead Executive, Data Catalyst at MaRS Discovery District, a Toronto-based innovation hub., "AI and the Future of Workforce Training", December 2024, Center for Security and Emerging Technology, <https://cset.georgetown.edu/publication/ai-and-the-future-of-workforce-training/>] ZG

**The emergence of artificial intelligence (AI) as a general-purpose technology is poised to transform work across a variety of industries and job roles.** Previous waves of technological change mainly led to job displacement and wage pressures for bluecollar workers while enhancing productivity and wages for white-collar workers. In contrast, AI’s impact could be more pervasive across all occupational categories, including knowledge workers and those with advanced education. Recent studies indicate that up to **80 percent of U.S. workers might have at least 10 percent of their work activities affected by large language models,** with approximately 19 percent of workers potentially seeing half or more of their work activities impacted. The nature of this transformation depends largely on two factors: the degree to which AI can perform or enhance an occupation’s core tasks, and whether AI serves as a substitute for or complement to human workers. Occupations with high AI exposure but low complementarity face the greatest risk of disruption, highlighting the need for comprehensive retraining and upskilling initiatives. This situation is particularly critical given that technical skills now become outdated in less than five years, on average. Analysis of future workforce demands reveals the following trend: while technical skills remain important, accounting for about 27 percent of in-demand skills, the majority of crucial skills are nontechnical. Foundational skills (such as mathematics and active learning), social skills (including social perceptiveness and negotiation), and thinking skills (such as complex problem-solving and critical thinking) together make up nearly 58 percent of skills needed in growing occupations. This underscores the importance of developing a well-rounded workforce capable of adapting to technological change while maintaining strong interpersonal and analytical capabilities. **The potentially far-reaching impact of AI across occupations, coupled with the likely accelerating pace of skill obsolescence, points to an increasing need for continuous retraining and upskilling opportunities throughout workers’ careers.** This shifting landscape demands a critical examination of current workforce development infrastructure and its capacity to meet these emerging challenges at scale. Understanding which elements of the existing system can be effectively expanded and which barriers need to be addressed becomes crucial for developing responsive and resilient workforce training solutions. **Community colleges emerge as pivotal institutions in addressing these challenges, particularly when integrated into robust regional ecosystems that include employers and intermediaries**. **Recent federal initiatives, including $265 million in Strengthening Community Colleges Training Grants since 2021, demonstrate recognition of community colleges’ crucial role. Successful workforce development programs often combine traditional education with work-based learning opportunities, such as registered apprenticeships and career technical education (CTE). Several states have already begun implementing AI-specific CTE programs to prepare students for the evolving technical workforce.** However, significant challenges persist in the current workforce development landscape. These include fragmented training systems, insufficient public funding, regulatory disincentives favoring capital investment over labor, and difficulties in scaling successful programs. **While AI may be a source of workplace disruption requiring enhanced workforce training efforts, it also presents opportunities to address some of these systemic challenges in workforce development. The technology’s capabilities could help scale effective training solutions and make them more accessible and affordable, potentially bridging gaps in the current system.** Specifically, these capabilities enable personalized learning experiences, rapid content delivery, and increased accessibility. AI tools can provide customized learning paths, instant feedback, and career guidance. However, implementation must be approached cautiously. Concerns include the potential erosion of interpersonal skills, trust and privacy issues, and the risk of exacerbating existing inequalities through algorithmic bias and unequal access. Research indicates that while AI tools can enhance productivity, overreliance on these tools may hinder genuine skill development and learning. **Moving forward, successful workforce development will require a multifaceted approach: strengthening community college programs, expanding alternative career pathways, incorporating AI literacy into training initiatives,** and ensuring equitable access to technology-enabled learning opportunities. This should be accompanied by careful consideration of how AI tools are integrated into training programs to maximize benefits while mitigating risks to skill development and learning outcomes. Further research is needed to understand how successful training solutions can be scaled across diverse regions and how AI training tools can be effectively deployed to serve diverse populations while supporting genuine skill development and learning.

[Catalina Espinosa](https://www.statista.com/aboutus/our-research-commitment/3815/catalina-espinosa), N.D., "Global workforce in low-skilled occupations 2020", Statista, https://www.statista.com/statistics/1171289/global-workforce-low-skilled-occupations/

In 2019, a Statista study on [labor shortages](https://www.statista.com/study/69261/labor-shortage/)showed that in 2020, **44 percent of the global workforce were working in low-skilled occupations**, with this share decreasing to 39 percent by 2030.

## C2:

Maslach 23- December 13, 2023, David Maslach is an associate professor at Florida State University specializing in organizational learning and innovation. He holds a PhD from the Ivey School of Business and serves on multiple academic journal boards. Maslach is also the founder of the [R3ciprocity Project](https://www.r3ciprocity.com/), a platform that provides solutions and hope to the global research community, Harvard usiness Publishing, “Generative AI can supercharge your research”, <https://hbsp.harvard.edu/inspiring-minds/generative-ai-can-supercharge-your-academic-research//doa>: 02/13/2025//kfk

Conducting relevant scholarly research can be a struggle. Educators must employ innovative research methods, carefully analyze complex data, and then master the art of writing clearly, all while keeping the interest of a broad audience in mind. Generative AI is revolutionizing this sometimes tedious aspect of academia by providing sophisticated tools to help educators navigate and elevate their research. But there are concerns, too. AI’s capabilities are rapidly expanding into areas that were once considered exclusive to humans, like creativity and ingenuity. This could lead to improved productivity, but it also raises questions about originality, data manipulation, and credibility in research. With a simple prompt, AI can easily generate falsified datasets, mimic others’ research, and avoid plagiarism detection. As someone who uses generative AI in my daily work, both in academia and beyond, I have spent a lot of time thinking about these potential benefits and challenges—from my popular video to the symposium I organized this year, both of which discuss the impact of AI on research careers. While AI can excel in certain tasks, it still cannot replicate the passion and individuality that motivate educators; however, what it can do is help spark our genius. Below, I offer several ways AI can inspire your research, elevating the way you brainstorm, analyze data, verify findings, and shape your academic papers. ETHICAL ISSUES TO CONSIDER. AI’s potential impact on research, while transformative, does heighten ethical and existential concerns about originality and academic credibility. In addition to scrutiny around data manipulation and idea plagiarism, educators using AI may face questions about the style, or even the value, of their research. However, what truly matters in academic research is not the tools used, but educators’ approach in arriving at their findings. Transparency, integrity, intellectual curiosity, and a willingness to question and challenge one’s previous beliefs and actions should underpin this approach.

#### their li concedes that AI exposes those problems but it does not fix them – also cannot fix if it takes all data as fact– slakes reads red

**Li 23** [Zhicheng; Programme of Applied Psychology, School of Humanities and Social Science, The Chinese University of Hong Kong, Shenzhen, Guangdong 518172, People's Republic of China; August 2023; "Why and how to embrace AI such as ChatGPT in your academic life," PubMed Central (PMC); https://pmc.ncbi.nlm.nih.gov/articles/PMC10445029/ DOA: 2-12-2025] sumzom + mac

Ever-growing scientific advances and data present a significant challenge: a ‘burden’ of knowledge that leaves researchers struggling to keep up with the expanding scientific literature. By contrast, the explosion of knowledge and data is fuelling machine intelligence. The rapid progress in **generative AI** (see box 1 for a non-technical primer) in the past few years, especially in large language models (LLMs), is a **game-changer** [1,2]. It is well suited to alleviate the **knowledge ‘burden’** and has the potential to revolutionize **scientific research**. To facilitate the adoption of this new technique and foster discussions and empirical research on the changing landscape of scientific research in the era of generative AI, here I provide a how-to guide for using LLMs in academic settings and offer new perspectives on their implications as informed by epistemology and philosophy of science.

Box 1. Generative AI, large language models and ChatGPT/Bard.

Generative AI trains machine learning (ML) models on a dataset of examples to generate new examples similar to those in the training set, including text, images and music. This generative ability distinguishes it from predictive AI, which trains models to **predict outcomes** on new, unseen data, such as in image classification and speech recognition. Although generative AI dates back to the 1950s, the breakthrough came only recently, thanks to the availability of massive amounts of data and the development of deep learning algorithms (‘deep’ refers to the use of multiple layers in artificial neural networks). These algorithms afford the creation of **large language models** (LLMs) to be trained on vast amounts of diverse text data.

Many state-of-the-art LLMs use a type of deep learning algorithm called transformers as their backbone. Introduced in 2017, the transformer architecture is a type of deep neural network architecture that uses self-attention mechanisms to better process sequential data such as text. Self-attention allows the network to calculate the attention weights between every pair of input elements, effectively allowing the network to weigh the importance of each input element with respect to all other elements. Thus, it allows the network to dynamically focus on different parts of the input sequence and capture long-range dependencies in the data. This mechanism enables it to understand and interpret language in a way that is similar to humans.

One of the most powerful LLMs is Generative Pre-trained Transformer 3 (GPT-3), introduced in 2020 by OpenAI in San Francisco, California. GPT-3 has been trained on a massive amount of text data, allowing it to generate human-like text and excel at challenging natural language processing (NLP) tasks. Recently in November 2022, a derivative of GPT-3 called ChatGPT was launched. It has fine-tuned GPT-3 using reinforcement learning from human feedback (RLHF) in a smaller dataset specifically for conversational tasks, making it both conversational and computationally efficient. GPT-3 was updated to GPT-4 and released to the public on 14 March 2023. Another powerful transformer-based LLM is PaLM (Pathways Language Model), developed by Google AI. PaLM has been finetuned to support the chatbot, Bard.

To understand and harness the capacity and potential of generative AI, I will illustrate its capabilities using the popular chatbot ChatGPT. ChatGPT reached 100 million users within just two months of its launch on 30 November 2022. A similar chatbot is Bard, which was launched by Google on 21 March 2023 (see table 1 for a list of other tools). In what follows, I will first identify and elaborate on three features of LLMs, as exemplified by ChatGPT, that make them unprecedentedly apt to augment, if not transform, research life: intelligent, versatile and collaborative. I do so by incorporating specific, practical examples commonly encountered in biomedical and behavioural research. As LLMs are rapidly evolving, I also offer a living resource online, complete with documents that provide tips on crafting effective prompts, examples of usage and relevant links (https://osf.io/8vpwu/).

<<TABLE 1 OMITTED>>

Next, I will critically discuss the limitations of LLMs and, importantly, their ethical and responsible use, as well as implications for equality and education—a debate still in flux. Specifically, I argue that while **guidelines** for using AI such as ChatGPT in academic research are urgently **needed**, policing its usage in terms of plagiarism or AI-content detection is likely of **limited use**. More fundamentally, if AI-created content is deemed valuable based on peer review, there is no reason to reject such content—the identity of the originator of that content is irrelevant from an epistemic point of view. As long as the use of AI is transparently disclosed, there is no need to limit the scope or nature of the assistance it can offer. If, however, the content produced by AI is not original or valuable but still passes peer review, then the problem lies not with AI but with structural issues in the peer review system—AI merely exposes its **weaknesses** and calls for **concerted efforts** to improve it. Concerning implications for equality, I contend that generative AI may foster equality for some but exacerbate disparities for others, based on considerations at the individual, group, and national levels. With regard to education, I advocate for the importance of engaging with LLMs and developing critical thinking and analytical skills in students. Given the early nature of generative AI in scientific research, empirical work is scarce, and the views expressed here aim to stimulate further efforts in addressing these important issues.

2. Three features of generative AI that make it valuable for researchers

2.1. Intelligent

AI is created to perform tasks that typically require **human intelligence**, including understanding language. According to multiple benchmarks—ranging from Advanced Placement (AP) exams to the Uniform Bar Exam—it is increasingly capable of performing language tasks at a level that matches or **surpasses** average **human performance** [3]. Indeed, LLMs such as ChatGPT go beyond generating language to show some form of behaviours that seem to resemble general ‘intelligence’, including problem-solving and reasoning [4].

Formal tests corroborate these observations. For example, in medical question answering, ChatGPT not only achieved accuracy higher than the 60% threshold on the National Board of Medical Examiners (NBME) Free Step 1 dataset—comparable to a third-year medical student—but was able to provide reasoning and informational context [5]. As another example, consider its ability to generate medical-research abstracts based on just the title and journal of the original papers. Not only was there no plagiarism detected, but also human reviewers correctly recognized just 68% of the generated abstracts and wrongly flagged 14% of the original abstracts as generated [6]. These results are remarkable given that they were tested using ChatGPT out of the box. In other words, when the pre-trained model is fine-tuned with a dataset of examples from the relevant domains, the results will be enhanced. Further, as the underlying model (GPT-3.5) is continually being improved (e.g. updated to GPT-4 on 14 March 2023), the performance of ChatGPT is expected to also improve, as demonstrated in medical competency [7].

Whether such performance and behaviour constitute cognitive abilities and can be construed as intelligence of humankind is debated [8]. Indeed, human intelligence is a latent construct that does not yield itself to a straightforward measure in non-human animals and machines, not least because traditional intelligence tests such as Intelligence Quotient (IQ) are anthropocentric—designed specifically for humans. Even within human populations, IQ tests need to be significantly altered for testing in children and people with disabilities. Thus, to better understand the nature of AI and measure its progress in obtaining intelligence, much research is needed to define intelligence and measure it in a way that is comparable and fair across machines and mankind [9].

Given the controversy, the term intelligence will be used here to refer to artificial intelligence, regardless of whether that might be considered true human intelligence or not. Indeed, for practical purposes—that is, from an end user's perspective—such debates are mostly moot so long as AI is able to get the job done. To appreciate the intelligence of AI, perhaps the most straightforward way is to have a conversation with ChatGPT (for a practical guide to its efficient use, see box 2). ChatGPT is strikingly human-like: it ‘understands’ text input and responds to it like a well-learned person—and in some ways, perhaps better than most people. The implications are likely to be profound, as the cost of intelligence has never been so low. This makes LLMs such as ChatGPT incredibly empowering for organizations and individuals.

Box 2. A practical guide to the efficient use of ChatGPT/Bard.

ChatGPT can be accessed through a web interface. To get started, go to the official webpage (https://chat.openai.com) and sign up for an OpenAI account (phone verification is required). Once logged in, you will see its interface, as shown above, where you will find example prompts to ask the chatbot and its capabilities and limitations. Interact with the chatbot by typing your prompt in the blank input bar (bottom) or initiating a new chat (top left).

To use it more efficiently, familiarize yourself with three key features. First, each prompt in your chat history has an edit button when you hover over it (on the right), where you can edit your previous prompt. After your edit, the chatbot will provide a new response accordingly. This is useful when your initial attempt does not yield the response you want. Second, you can provide feedback on the response (thumb up and thumb down icons, on the right) and you can ask it to regenerate responses (bottom)—which you can toggle to compare and find the most desirable one. Third, you may want to start a new chat for each project, as ChatGPT takes into consideration the chat history of each conversation.

Getting the desired results may require some thought. That is, feed it the right prompts (see six tips for writing effective prompts in the online supplemental materials: https://osf.io/8vpwu/). LLMs tend to make assumptions about user intent based on the prompt given, rather than asking clarification questions. To enhance accuracy, it is important to provide it with sufficient contextual information [10]. In general, prompts should be clear and concise. You can provide very specific instructions and offer feedback and new directions as follow-ups throughout the conversation. For example, you may ask it to explain a statistical concept by typing: ‘Explain Cook's distance’. Suppose you find the response a bit dense. You can follow up by typing: ‘Can you explain it like I am five?’ As another example, you can feed it with your writing and ask it to make it more concise: ‘Please rewrite it to be more concise’. But if you find the rewrite a bit non-sophisticated, you can follow up with a prompt like: ‘Please make it more sophisticated for an educated audience’. You can keep fine-tuning it to your desire. However, if you have a clear goal, using an elaborate, specific prompt will work best. In fact, you can enlist ChatGPT to help improve the prompt (e.g. ‘Please evaluate each prompt I present and provide a rating on a scale of 1 to 5, based on its clarity and level of engagement. Kindly provide constructive feedback on how I can improve each prompt if necessary. Should the rating for a prompt be 4 or above, proceed to answer it; otherwise, create a new prompt that meets the desired criteria’).

ChatGPT is helpful for many things, from helping you learn, code, analyse and write to assisting with your teaching, mental needs and job applications. Ultimately, to get the most out of its capabilities, be creative and imaginative. Say you have written an emotional email. Before you send it, you can enlist ChatGPT to check its tone, using the following prompt: ‘Acting as an editor, please make recommendations on how to improve the email below using the principles and concepts of Nonviolent Communication (NVC). For each edit, please provide the rationale and some examples’. Indeed, you can ask ChatGPT to act as a simulated patient, therapist, coach, advisor, tutor, professor or interviewer—the possibilities are endless. Or consider your next job application. You can request ChatGPT to help craft a customized cover letter for the job, using a prompt like: ‘Please write a cover letter for the job description below using my CV that follows’.

Example screenshots of using R and Adobe Illustrator, tips for writing effective prompts, and a living resource are provided online (https://osf.io/8vpwu/). This guide also applies to the chatbot, Bard, which is highly similar to ChatGPT except for some minor differences (e.g. the ‘[r]egenerate response’ function in ChatGPT is replaced by the ‘[v]iew other drafts’ function in Bard).

For knowledge workers, it enables us to be more productive and efficient—doing more with less. A list of tips, examples and resources is provided online (https://osf.io/8vpwu/). For example, ChatGPT can provide explanations and help us learn a new domain more efficiently (e.g. ‘Act as an R instructor and teach me the basics'), write and debug codes faster (e.g. ‘Write R code to do a one-way ANOVA based on the following data’), assist with writing (e.g. ‘Rewrite the following paragraph to be more concise’) and more. By automating aspects of the research process and improving research efficiency, ChatGPT helps to accelerate the pace of scientific discovery.

From the perspective of philosophy of science, AI also has the potential to **uniquely complement** and enhance human intelligence in facilitating **scientific inquiry** and **discovery**. For one, by analysing and synthesizing vast amounts of data from different fields, LLMs may help to discover connections between seemingly **disparate fields**—connections that might not be immediately apparent to **human researchers**. For another, whereas human researchers are **inevitably influenced** by personal values and preferences, social norms and cultures, and historical assumptions and biases [11], LLMs do not have emotions, consciousness or personal **motivations**. Indeed, by analysing vast and diverse amounts of data with the same algorithmic process, LLMs have broader perspectives and **greater consistency** than individual researchers, thus reducing the **risk** of cognitive bias, from confirmation bias to the availability heuristic. Moreover, although biases do exist in LLMs due to the training data and algorithms—a limitation discussed later—these biases are not **identical** to human biases and can help to counteract or reduce certain predispositions in scientific practices, potentially improving the **reliability** and **objectivity** of scientific inquiry [‘strong objectivity’; 12].

2.2. Versatile

As alluded to before, what makes **generative AI** such as ChatGPT special is that it excels not just in one domain but across **many domains**, thanks to the diverse training text data. ChatGPT has been trained to understand and generate cohesive text across a broad spectrum of subjects, from general knowledge to specific areas such as **science** and mathematics. It is proficient in a wide range of human languages (English, Spanish, French, German, Italian, etc.) and computer programming languages (Python, JavaScript, Java, C++, R, etc.). This versatility makes it useful in multiple capacities, such as a coach, research assistant and co-writer.

Consider the many tasks that researchers perform every day. In administrative roles, writing and editing documents and emails can benefit from ChatGPT. In teaching, generating questions and grading them, creating discussion points and questions, editing syllabuses and handouts—these are some common tasks that can also use help from ChatGPT. In research, too, practically all processes—other than those involving physical interactions—can enlist ChatGPT. Indeed, formal evaluations in finance research show that ChatGPT can significantly assist with idea generation, data identification and more. Incorporating private data and domain expertise can further improve the quality of the output [13].

For example, ChatGPT can help with familiarizing oneself with new topics (e.g. ‘What is generative AI’), **summarizing** (e.g. ‘Summarize the key issues mentioned below in a table, using two columns: ‘Ethical issue’ and ‘Key question’’), **coding** (e.g. ‘The following code has errors. Can you advise how to fix it’), **brainstorming** (e.g. ‘Write five titles based on the following keywords’), providing feedback (e.g. ‘Act as a journal reviewer and provide feedback on the abstract below’) and more.

2.3. Collaborative

ChatGPT is also special for its **conversational capability**, thanks to a method called reinforcement learning from human feedback (box 1). This capability makes it an excellent **collaborator**, able to listen and update its responses based on **user feedback**. To illustrate, suppose we want to improve our writing. We can start with the prompt: ‘Act as a copy editor, revise the text below and explain your edits’. If we don't like a particular expression in the revision, we can follow up with a new request: ‘Can you make ‘…’ more elegant?’ Indeed, we can ask ChatGPT to give the writing some personality, revise it for an academic audience, make it more persuasive or assertive, in the style of Hemingway, and so on. From proofreading to editing and rewriting, the possibilities are **endless**.

The utility of intelligent, versatile, always-on collaboration afforded by ChatGPT cannot be overstated. It offers a great channel to bounce ideas off of. It also helps to alleviate common drudgery and mental block—making research more fun. For example, regular expressions (regex or regexp) are a powerful tool commonly used in text analysis to define patterns for strings—thus enabling matching, extracting, and substituting patterns—but they can be complicated and error-prone. ChatGPT makes it much easier to use regex by helping researchers understand the syntax and usage (e.g. ‘How to replace all occurrences of Ph.D. with PhD in R using regex?’), and then construct or refine a regex (e.g. ‘Test the regex on a sample text and return the matched substrings’). Similarly, consider a common mental block: writer's block. ChatGPT helps by brainstorming and collaborating with us, starting the first step that ultimately paves the way for a thousand-mile journey to publication (e.g. ‘Give me five ideas to begin an article on ‘how AI may help researchers’’).

3. Limitations of generative AI

As with any other tool, generative AI has limitations. These limitations are rooted in the principles and techniques that make it so powerful in the first place (box 1). Specifically, LLMs such as ChatGPT are language models trained on massive data. When they respond to queries and engage in conversation, they do not understand the content in the same way humans do, but rather make predictions about text based on patterns learned from training. They ostensibly write like an educated human—a great achievement—but they are not. This will become plainly clear once we interact with them in a deep manner (e.g. they can contradict themselves at times, and they do not have a strong grasp of context). The important point, however, is to use them as powerful tools rather than relying on them.

In the context of research aid—such as for a research project or for lecturing on a topic—a major limitation of LLMs is that they may fabricate facts, creating confident-sounding statements and legitimate-looking citations that are false (hallucination). Thus, as with any other source of information (e.g. Wikipedia), it is important to critically evaluate and verify AI responses, particularly when reliability is critical [14]. An important next step might lie in developing methods to quantify and signal the epistemic uncertainty and potential limitations of AI-generated results.

Still another limitation has to do with the training data for LLMs. These data are not—and cannot be—truly neutral or objective, but rather laden with assumptions and biases, ranging from political and ideological to cultural [12,15]. From the perspective of standpoint epistemology, such biases and assumptions are not inherently problematic. To the extent that knowledge is socially situated—different people have different experiences and perspectives that shape their understanding of the world—biases and assumptions can be understood as reflective of specific standpoints (i.e. perspectives) of the people who generated and compiled the data.

Yet, the challenge is that the standpoints represented in the training data may not be evenly distributed or representative of all perspectives. Indeed, the issue of underrepresentation in knowledge production has been widely documented, including the underrepresentation of certain racial, ethnic, gender, political and geographical groups as participants and researchers in medical and scientific research [16,17]. Lack of diversity in the research process contributes to prejudices, stifles epistemological plurality, and limits the range of topics and questions being pursued [11]. In turn, biases and limitations in the data may be picked up—or even amplified—in LLMs. For example, when the training data predominantly reflect the views and experiences of certain groups (e.g. people from Western, educated, industrialized, rich and democratic societies), then the LLMs trained on these data will inevitably reflect these biases. This uneven representation can lead to a reinforcement of dominant perspectives and marginalization of others, creating a potential for bias in the outputs of these models.

There are additional limitations in using AI/LLMs to aid teaching and administrative tasks. In the realm of teaching, one potential use of AI is grading [18]. While such an application might seem promising in terms of efficiency, establishing a system that grades objectively, reliably and fairly presents significant challenges. To ensure fairness and accuracy, the AI’s grading algorithms would need to be based on clear, comprehensive rubrics—a non-trivial task in itself. Even then, potential biases in the AI’s interpretation of student work could lead to discrepancies in grading. Furthermore, nuances of student creativity and originality, which are often the hallmarks of exceptional work, might be overlooked or misinterpreted by an AI grader. Therefore, human supervision and verification are necessary safeguards in the grading process, potentially reducing the time and labour-saving benefits of the AI.

In the administration domain, AI is useful for drafting emails and similar tasks. While AI can be used to streamline the process and improve efficiency, it can also backfire in sensitive situations, when human touch is what matters most—something that cannot be replaced by AI. One case that underscores this limitation is a recent incident at Vanderbilt University, where two deans used ChatGPT to draft an email to students about a mass shooting at Michigan State University. Their use of AI in this sensitive situation led to their suspension, illustrating the potential pitfalls of over-reliance on AI for sensitive administrative tasks. Thus, striking a balance between leveraging AI's efficiency and maintaining the human touch that is often essential in academic settings will be an ongoing challenge in the implementation of these technologies.

4. Implications of generative AI: ethical use, equality and education

4.1. Ethical and responsible use

The power of generative AI such as ChatGPT raises many thorny questions regarding its ethical use, from plagiarism, image manipulation, authorship and copyright to fake research (table 2). It is one thing to ask it to act as an editor to correct language issues in our own writing, but quite another to ask it to write an entire paragraph and then copy it [2]. The former is similar to the services offered by other writing tools and university writing centres, while the latter is widely regarded as plain plagiarism. However, the boundary between acceptable help and too much help is not always clear-cut. When we feed ChatGPT with our own text and ask it to rewrite it, is that too much help to be considered ethical? Does the answer depend on the length of the text—and if so, how can we determine the proper boundary? The same questions apply to text-to-image AI (e.g. DALL·E 2, Midjourney, Stable Diffusion). Is it okay to use AI-generated images in the paper, or would that be considered plagiarism? And in the cases where AI offers ‘too much’ help, can it be listed as a co-author? Fundamentally, who has the right to claim copyright over AI-generated content (text, images, etc.): the prompt creator, the AI, the AI developer or the owners of the training data?

<<TABLE 2 OMITTED>>

These questions are important for the community to consider and address. Currently, publishers and journals are divided in their policy and stance on some of the questions. For example, Springer Nature does not allow LLM tools to be listed as authors, and requires researchers to document their use in the paper [19]. On the other hand, Science family journals not only ban AI tools as authors, but also prohibit the use of AI-produced content (text, images, figures, graphics) in the paper [20]. Although such swift decisions are understandable, going forward it is important to engage the whole scientific community to reach a more consistent and informed consensus. For example, banning AI tools as authors because of their inability to take responsibility flies in the face of the long-standing practice of posthumous authorship [1].

The more practical issue is that it may not even be feasible to detect AI-generated content with sufficient accuracy to be useful. Compared with typical AI-generated content, human-generated content generally—but not always—has higher burstiness, mixing longer or more complex sentences with shorter ones, and with higher perplexity, using words that are less expected [21]. However, some human writers do write with low burstiness and perplexity, posing a problem of false positives for algorithms. Moreover, LLMs can be instructed to write content with higher burstiness and perplexity, creating a problem of false negatives for algorithms. On top of that, given that LLMs are constantly evolving and improving, it is reasonable to assume that their ability to evade detection may do so as well. Thus, although algorithms for detecting AI content may be useful to compare different groups of writing, they are unlikely to be able to ‘convict' any individual writing. Banning the use of AI-generated content may prove challenging to implement.

Fundamentally, if AI-created content is valuable, there is no reason to reject such content. From an epistemic point of view, we should not treat a finding differently just based on the status of the author, whether it is a Nobel-prize winner or a junior academic member. The identity of the author is irrelevant. The same applies to AI: if AI has **valuable**, original content, there seems no **epistemic** reason to **devaluate** it just because it is created by AI. The real question is the vetting of its value—which rests on the human author and reviewers. Thus, a more **pragmatic** approach to AI in **academic publishing** is to encourage or mandate its **transparent use** [22] rather than banning it outright or even limiting it. From this perspective, there is no need to limit the amount or kind of help from AI—no concept of too much help from AI—as long as it is transparently reported.

Perhaps a more urgent issue with AI concerns its potentially serious threat to scientific integrity: the inevitable exponential rise of AI-generated, fraudulent papers submitted to scientific journals—some of which will pass peer review and become part of the scientific literature. Paper mills, which are already notorious for creating and selling fake research with fraudulent data and images, will become an even bigger threat when equipped with the unprecedented power of AI [10]. However, the negative disruptions brought about by AI, as with the advent of any other powerful tool in history, are to be expected. Indeed, more generally, if content that is not valuable or simply **fake** can pass **peer review**, whether it is from AI or not, the problem has more to do with the **peer review system**. The potential negative impact is not a cause to forbid or limit the use of AI, but a call to step up our efforts in implementing **better practices** in scientific **review** and publishing.

Such practices may involve the implementation of rigorous and **open peer review** (e.g. published peer review exchanges), collaborative review (e.g. discussions among reviewers and the action editor before making an editorial decision) and open science practices (e.g. open data and materials). These practices serve to deter **fraudulent submissions**, as through open review, the review process is subject to **scrutiny** by the wider **scientific community**; they also enhance the probability of detecting fraudulent content, as the accessibility of data and materials **simplifies** the process for others to validate the results. For these practices to be most effective, researchers need to be aware of the potential for AI tools to be used to generate fraudulent content, as well as to be alert to potential signs of such fraudulent content. Thus, **education** and awareness are **vital**. In addition, **AI-based tools** may be developed to **detect** patterns indicative of **data fabrication** or falsification, as well as to identify inconsistencies or errors in data analysis. Together, these strategies can help mitigate the **negative impact** of AI on knowledge production and improve the accuracy of the scientific record more generally.

4.2. Impacts on equity

Having discussed the strengths, limitations and ethical use of generative AI, a natural question arises concerning its implications for equity. Perhaps paradoxically, the availability of powerful, versatile AI tools can promote equality for some while amplifying disparities for others. On the one hand, a main contributor to global disparities in scientific research is language; for example, most mainstream journals are in English, bestowing a natural advantage on native English researchers [16,17]. LLMs can help level the linguistic playing field by offering a language boost for non-native English researchers through copy editing and other writing assistance (e.g. ‘Act as a copy editor, proofread the following text for an academic journal, and highlight the changes at the end’). Thus, researchers previously disadvantaged in the English language can now compete on a more equal footing.

**Dzuong et al 24** Jocelyn Dzuong, a master's student in the Knight Foundation School of Computing and Information Sciences at Florida International University, Zichong Wang, a third-year Ph.D. candidate in the Department of Computer Science at Florida International University, Wenbin Zhang, an Assistant Professor in the Knight Foundation School of Computing & Information Sciences at Florida International University, 3-31-2024, "Uncertain Boundaries: Multidisciplinary Approaches to Copyright Issues in Generative AI", arXiv.org,<https://arxiv.org/abs/2404.08221>

In the rapidly evolving landscape of generative artificial intelligence (AI), the in**creasingly pertinent issue of copyright infringement arises as AI advances to generate content from scraped copyrighted data,** prompting questions about ownership and protection that impact professionals across various careers. With this in mind, this survey provides an extensive examination of copyright infringement as it pertains to generative AI, aiming to stay abreast of the latest developments and open problems. Specifically, it will first outline methods of detecting copyright infringement in mediums such as text, image, and video. Next, it will delve an exploration of existing techniques aimed at safeguarding copyrighted works from generative models. Furthermore, this survey will discuss resources and tools for users to evaluate copyright violations. Finally, insights into ongoing regulations and proposals for AI will be explored and compared. Through combining these disciplines, the implications of AI-driven content and copyright are thoroughly illustrated and brought into question. In the swiftly progressing realm of generative artificial intelligence (AI), the pressing concern of copyright infringement emerges prominently. As AI technologies continue to autonomously generate content from copyrighted data, inquiries about ownership and safeguarding rights surface, reverberating across diverse professional domains. This escalating trend raises critical discussions surrounding ethical, legal, and socio-economic implications, necessitating nuanced exploration and strategic interventions to navigate this Figure 1: Actual screenshot from Dune (2021) versus its Midjourney-generated counterpart evolving landscape effectively. For instance, in July 2023 a group of novelists collectively sued OpenAI for alleged usage of their books to train their models and output similar content to the novelists’ prose [117]. Moreover, in December 2023 **The New York Times filed a lawsuit against OpenAI and Microsoft, alleging copyright infringement by having** its articles scraped without permission to train their generative models [118]. More recently, Marcus and Southen revealed how generative models such as Midjourney and OpenAI’s Chat GPT-4 produced outputs strongly reminiscent of scenes from copyrighted films and shows [82, 124]. As a concrete example, Figure 1 illustrates how a prompt from Southen resulted in an output resembling a shot from the trailer of Dune (2021). Notably, Midjourney’s terms of service [87] highlight that users assume liability when requesting the model to generate content featuring copyrighted trademarks. This delegation of responsibility not only places the burden of infringement on users, but also diverts accountability from Midjourney’s developers, who have openly admitted to using copyrighted trademarks without authorization [103]. In light of these developments, this survey aims to delve into the complex interplay between generative AI and protecting intellectual property (IP). Through synthesizing existing methods and legal analyses, we provide a comprehensive overview of the current landscape surrounding copyright in generative AI. To the best of our knowledge, this work presents the first thorough study on robust and applicable solutions to copyright issues in generative AI, which also combines contextual legal analysis for future consideration. The challenges and opportunities inherent in this burgeoning field offer insights that can inform policymakers, practitioners, and researchers alike when developing generative AI. Our main contributions are: i) A detailed examination of the most advanced methods for detecting AI-generated copyright violations across various mediums such as text, image, and video, establishing itself as an invaluable resource for both researchers and practitioners in the field. ii) Innovative strategies designed to safeguard copyrights within the AI sphere, highlighting cutting-edge techniques like watermarking, fingerprinting, and machine unlearning, contributing to the protection of IP. iii) A comprehensive array of tools and resources for assessing copyright violations, including extensive datasets, search engine capabilities, and metrics quantifying infringement. iv) An in-depth analysis of the regulatory framework surrounding generative AI, navigating through current international copyright laws and proposing solutions to tackle the emerging challenges in generative AI.

Abby **Rives 22**. IP counsel at Engine, a D.C.-based policy, advocacy and research organization supporting startups. "Copyright Law & Startup Innovation: Policies That Matter and Where They May be Headed". Medium.https://engineadvocacyfoundation.medium.com/copyright-law-startup-innovation-policies-that-matter-and-where-they-may-be-headed-dea034904e25. accessed 7-4-2024 //nm

Startups need **balanced, certain copyright** frameworks. Well-tailored laws that focus on enforcement of legitimate rights can support innovation. But it is too easy for those frameworks to get **out of whack** and **become imbalanced**, which we’ve seen time and again. For example, right now the law allows bogus infringement allegations to dictate that non-infringing content is (routinely) removed from the Internet. **Uncertainty** over what copyright law permits, coupled with **high** **litigation** **costs**, **slows startups** down and has even forced some out of business. **And the risk of a startup being sued for something a user does — and something the startup knows nothing about — alone can scare away investors.**

But what does balanced, **innovation-friendly** copyright policy look like? And how does this play out in today’s policy debates? Here are just a few examples:

Fair use and interoperability: Some big companies would like to **expand** **the universe** of what software is protected by copyright and which development activities constitute infringement. If that happened, it would prevent startups from using **fundamental software development** **tools**, expose them to new **litigation risks**, and make it harder to launch and compete. But after a decade of litigation, the Supreme Court recently confirmed that developers can use software interfaces — known as application programming interfaces (APIs) — without infringing copyright. The Court held that **reimplementing APIs**, which creates **interoperability and compatibility** between computer programs, is a **fair use** under copyright law.

Intermediary liability and the ability to host user-generated content: Scores of startups engage with user content — helping artists connect with fans, providing e-commerce platforms, hosting podcasts, or offering **basic Internet infrastructure**. These companies, and the creators and small businesses that depend on the Internet, interact with the copyright system every day. And they **rely on balanced laws** that allow the startups to **resolve allegations of infringement** without scrutinizing every post, upload, and comment for **potential copyright violations**. Some countries have started to replace those laws, instead moving to complex and expensive regimes that would force Internet companies to purchase **expensive** and imperfect upload filters, **remove** **more non-infringing** **content**, and **negotiate licenses** with big organizations that own a lot of copyrights. That is all do-able for big Internet platforms, but it will put startups at a substantial disadvantage. Yet similar ideas are being floated in the U.S. — where policymakers have proposed **changes to copyright law** (and trademark law).

Ancillary copyright and link taxes: Countries around the world have adopted or considered new copyright-like laws that would require websites to pay **licensing fees** or **face lawsuits** whenever they — or their users — link to a news article or quote the headline. These proposals, positioned as a solution to problems facing local media, have so far failed to deliver those benefits, but they carry substantial **unintended consequences**. Linking to news articles is something many startups and innovators — from media to edtech — rely on. But engaging with information and current events, **which** is central to public discourse and free speech, requires being able to link to and quote the news. Using copyright-like law to **restrict** that engagement would **hinder innovation** and the creation and **exchange of ideas online.**

Intersection of copyright and artificial intelligence: Startups and other companies developing AI technology have to input a lot of data into their systems, **ingesting content to train**, tune, and test new AI. As countries around the world review how intellectual property law applies to emerging AI, some are asking how copyright law should account for this ingesting of information, data, and content. But redefining copyright infringement to cover these uses of content to train AI could substantially **hamper innovation.**

**Proves alt causes – prefer AI is not widespared in education – all of their link is about like one test**