
THE
INTELLIGENCE
ILLUSION
Second Edition
Why generative models are bad for business
Baldur Bjarnason

# The Intelligence Illusion

# The Intelligence Illusion

*Second Edition*

*Why generative models are bad for business*

Baldur Bjarnason

*Published in Hveragerði, Iceland*

*Typeset in the Literata typeface.*
*This second edition was released in November 2024.*

# Contents

I didn’t use generative models in any way to make this book. The illustrations were made using a combination of Daz3D, Geometrize, and Affinity Photo.

# Introduction

# 1

# Generative models are bad for business

“ *Optimism was unwarranted.* ”

# Introduction to the second edition

Generative models are bad for business…

*But they didn't have to be.*

At least, they didn't have to be *this bad*.

The starting point for any given analysis of a business or industry is to assume that the people involved are rational.

This is quite often an immediately fatal flaw in said analysis, an assumption that completely invalidates its conclusions even before you begin. People are quite often irrational – or guided by reasoning that is selfless – which is unaccounted for when an analyst is assuming a selfish rationality.

But some of the time the seeming irrationality is more a matter of perspective. Shifting the frame of reference can show you that, *actually*, people are being quite reasonable given where they are and what they know.

- Irrational management often looks reasonable from the perspective of the manager. Their quarterly and career incentives just do not match that of the overall organisation.

- Irrational corporate strategies are rational from the perspective of the executives who stand to benefit from short-term fluctuations in the stock market and are insulated from long-term catastrophes. An organisational death cycle doesn't matter if you've already cashed out before it reaches bottom.

- Irrational employee behaviour often stems from extractive and abusive management. If the employee sees neither job security nor long term gain from prioritising the success of their employer, they will find it quite hard to care about its goals or strategies.

Because it's impossible to read the minds of everybody involved in an entire industry, analytical models have to depend on some level of assumption and guesswork, often based on historical behaviour. An industry that has previously done something several times is likely to do so again. Standard practice in the past continues into the future.

One of those standard practices, in tech specifically, is a strategy or convention first described by Clayton Christensen[1]: *the disruptive innovation*.

It's often described as a law of nature, but it is more a strategy that specifically targets a weakness created by the incentives of poorly regulated markets.

As companies grow, executives have every incentive to structure the entire organisation and its offerings towards a single optimised path to growth, and this usually results in organisations that only know how to grow in one way, using a single strategy. Smaller wins don't register. Lateral wins only register if it feeds into the primary path to growth. Because most countries have very lax regulations for tech companies, the optimal strategy usually isn't to improve the product but to improve sales and to design the product itself primarily for *capture*. The product becomes a roach motel: you can enter but you can't leave. If you're "lucky" enough to become a monopoly or oligopoly, even the capture becomes superfluous. This results in a cycle of product degradation Cory Doctorow described as "enshittification"[2].

A side effect of this process is that these corporations become extremely large and unwieldy. Their products are fossilised and hard to change. These companies are the organisational equivalent of an Ultra Large Container Vessel: capable of shipping vast amounts of product in one direction but hard to turn around or even stop if anything happens.

This is what has made a *disruptive innovation* such a useful strategy for tech companies in the past. By focusing on the low end first, using technological research and development to come up with cheaper but still partially effective products, the "disruptor" undercuts the incumbents with a much lower cost offering that doesn't quite do the same job as the incumbent's. Because the incumbents are usually organisa-

tionally incapable of responding with matching low end products, you can create a dynamic where they effectively have no real defence. (The details of the strategy and potential defences against it are beyond the scope of this book, but Clayton Christensen's own books make for a good introduction.)

When I wrote the first edition of *The Intelligence Illusion*, I assumed that we were about to see another iteration of the same strategy.

Many of the applications of generative models don't really work a *disruptive innovations*. Writers and artists are generally much cheaper than people expect. Training and running a Large Language Model is one of the few projects large enough to rival the cost of creating and running a full index of the global web.

But, software development seemed to fit the bill. Chatbots and autocompletes are a genuine liability to coding by virtue of how their user interfaces trigger a variety of cognitive biases (more on that later), but it seemed reasonable to assume that research and design would be brought to bear to create a set of genuinely disruptive innovations in programming.

That assumption was the foundation of the optimism that – unfortunately – was a continuing thread throughout the first edition of this book. Even though it was highly critical of all forms of generative models, the book was suffused throughout with a hope: these problems might get addressed, in time, and we'd discover ways to use the models safely and productively.

That optimism was unwarranted and was the first major mistake of the first edition.

The second big mistake was my assumption that anybody who read through a detailed overview of all of the dysfunctions, risks, and flaws of generative models would realise that, obviously, a technology this broken was unsuitable for most business. I didn't account for the inherent irrationality of people already predisposed to be fans of the technology. Where I thought I had delivered a guide that outlined all the reasons why generative models should be avoided, consultants and managers saw a guide that helped them sell "AI" by giving them a list of risks to downplay.

Meanwhile, the tech industry has sunk to new lows. Layoffs continue unabated and are spreading throughout our economies as tech companies promote the lie that workers can be replaced either with generative models or with a much smaller workforce "empowered" by "AI". Safety and security teams for "AI" products are understaffed or have even been outright disbanded. "AI" product research and design has all but arrested – every product is stuffed with chatbots, image generation, and language model autocompletion, no matter whether it makes sense or not. If the customers don't buy it, then they'll be forced to buy it, and there will be no attempt to come up with more useful design abstractions for the technology. More and more, you no longer even have the option to opt out of paying for "AI" integrations or to buy a cheaper version of the service that doesn't come with a chatbot attached.

This is the strategy of a monopolist unafraid of regulators, the consumer, or the government. Why would Microsoft try to disrupt software development? They already own the entire field by virtue of owning GitHub, npm, TypeScript, and Visual Studio Code. There's no need for any of these companies to even pretend that this is a disruptive innovation on a path to become more and more useful. Just bundle it with your core offerings and raise the prices[3]. What's the consumer going to do about it anyway?

"AI" is also becoming a political project where it looks like industry over-investment in faulty products will be subsidised by government and military contracts.[4] Why invest in genuine design innovation when you can just ship chatbots to the military and get billions in return?

Whether this technology works or can be made to work has become immaterial to its success. We will still be forced to buy it even if there is no demand.

This is why I felt compelled to create this second edition. Some of the core assumptions of the first edition have been revealed to be faulty. This book, instead of being a tool to help you judge the best way to navigate the minefield of an early-stage product, now needs to be a

tool you can use to explain – in detail – why this "innovation" is bad for business.

Because if you don't manage to talk your employer or your investors out of using these tools, the risks of integrating generative models into your core business are such that they could easily trigger a Boeing-style death spiral where the organisation becomes simply incapable of safely executing on its core strategy.

Boeing can no longer design new aeroplanes from scratch, but it survives by being strategically vital to US industry and security.

Can you say the same about your business?

Will *you* get bailed out if your products fail?

# The structure of this book

This book is divided into four overall parts:

1. *Introduction*. Being an overview of generative models, their basic function, what they can and can't do, as well as how they can exploit our own biases.

2. *Recommendations*. Overall recommendations for what to do and not to do. These are based on the next section.

3. *Risks*. This is an inventory of the risks that are *specific* to businesses and work. They don't cover political or social issues. Just those that could affect work. Together, they should demonstrate exactly why generative models are bad for business.

4. *What next*? This section looks at where things are and where things are heading.

You don't have to read them in order. I tried to minimise cross-references throughout the book so you shouldn't have to read the chapters in strict linear order. If your biggest concern is how AI and tech is currently evolving, then you should start with the fourth section. If you just want the recommendations, start with the second. If you're worried about the risks, start with the third. If you want to be depressed, read the whole thing from beginning to end.

# 2

# Some definitions

# But, first, what do the words mean?

- **AI:** *Artificial Intelligence*. Not actual intelligence. This refers to several different methods of building neural network models – data structures usually derived from larger collections of data – that model a particular domain or data type. Sometimes they are used for analysis, such as classifying the subject of an image. Sometimes it's a decision model that guesses at an appropriate decision based on the data available on prior decisions. Or, a data model derived from text or images can be used to generate second order image or text derivations. These are versatile tools even if they are quite unreliable and biased.

- **Neural Network:** A data structure based on early to mid-twentieth century ideas on how the brain might work which have been revealed as extremely simplistic as neuroscience evolved. You feed training data into an algorithm that builds a mathematical model of a network of "neurons", parameter by parameter, where each *parameter* is supposed to be functionally equivalent to a neuron in an animal brain. In a biological neural network, each neuron is one of the most complex living cells in the body, filled with multiple interconnected complex chemical systems, and communicating with its neighbours using a variety of neurotransmitters. Conversely, the "neuron" of an artificial neural network is usually a single number.

- **Machine Learning:** What software developers actually mean when they say 'AI'. Still a misnomer. Machine 'learning' is almost entirely unlike biological learning. We don't rebuild our brains entirely from scratch using only numbers, by reliving and reprocessing our entire lives up to that point – including the near totali-

ty of all of written human existence – every time we want to learn something genuinely new. These models are data derived from other data. They do not learn, but instead have to be derived again, partially or in whole.

- **AGI:** *Artificial General Intelligence*. What the public thinks 'AI' means. This lies somewhere between science fiction and fantasy. It very definitely does not exist yet and there are no trustworthy indications that it's anywhere on the horizon.

- **Training data:** This is the labelled or unlabelled data that's fed into the algorithm that models it to create an AI model's artificial neural network.

- **Model:** Usually used to describe the data – derived from the training data – that is a collection of the parameters of an artificial neural network often together with some (but not all) of the software needed to use it.

- **Generative AI, or Generative Models:** AI models that generate images, video, audio, text, or code.

- **LM:** *Language Model.* An AI model that has been trained on text. Smaller models are usually specialised and have limited functionality.

- **LLM:** *Large Language Model*. A language model that's large enough to generate seemingly fluent text and simulate conversations that are somewhat convincing on the surface. This is what ChatGPT is based on.

- **Foundational Model:** A model big enough and versatile enough to be used as a building block for more specialised AI "fine-tuned" models. They switch terms for this on a regular basis. The industry keeps inventing new terms for this. It'll probably be called something different next week.

- **Fine-tuning:** Additional training, usually in the form of additional modelling data that's layered in some way over the original model, sometimes using slightly different algorithms. It's purpose is to improve a model's capability at a specialised task, or degrade its ability to generate a specific kind of negative outcome, often at the expense of versatility.

- **Diffusion model:** A model derived from a large collection of images that are "diffused" into and out of random noise. This results in a model for turning noise into images that are derived from the data that has been modelled. Almost every major image-generating AI model on the market today is a diffusion model or uses a similar technology.

- **API:** *Application Programming Interface*. The hooks, levers, and functions a software system exposes that the programmer can use to connect it to another software system.

3

# Generative models are data

# Modern "AI" are not minds, they are derived statistical data

*Machine learning models* are layered networks where each node in the network represents a connection to its neighbouring node. In a modern transformer-based language model most of these nodes or parameters are floating point numbers – *weights* – that describe the connection and boil it down to a single value.

In language models the networks are built by feeding text through a tokeniser that breaks it into tokens – often one word per token – that are ultimately fed into an algorithm implemented as a software system. That algorithm constructs the network, node by node, layer by layer, based on the relationships it calculates between the tokens/words. This is done in several runs and, often, the developer of the model will evaluate after each run that the model is progressing in the right direction, with some doing more thorough evaluation at specific checkpoints.

The network is, in a very fundamental way, a mathematical derivation of the language in the training data.

**Deep learning systems such as language models are *constructed* from the training data.** The code regulates and guides the process, but the statistical distributions within the data set are what defines the neural network.

This is very, *very* different from how a biological neural network interacts with data. A biological brain is constantly modified by input and data – its environment – but its construction is derived from nutrition, its physical environment, and *genetics*.

The training data set, conversely, is a deep and fundamental part of the language model. The algorithm's code provides the process while the weights themselves are derived from the data, and the model itself is dead and static during input and output.

A language model constructed from a poorly documented training data set is effectively an undocumented system that cannot be reliably replicated or researched.

The construction process of a neural network is called "training", which is frequently a source of confusion as it gives a very misleading impression of how the construction process works. A language model isn't "trained" any more than a pregnant mother "trains" a fetus into an infant.

This is also why it's inaccurate to say that these systems are *inspired* by their training data. A language model can never be inspired. It is itself a cultural artefact – data – derived from other cultural artefacts.

A biological neuron is a complex system – one of the more complex cells in an animal's body. In a living brain, a biological neuron will use electricity, multiple different classes of neurotransmitters, and timing to accomplish its function in ways that we still don't fully understand.

The brain as a whole is composed of not just a massive neural network, but also layers of hormonal chemical networks that dynamically modify its function, both granularly and as a whole. It exists as an indivisible integrated system with the rest of the body. A disembodied brain in a jar will always be a dead one.

The digital neuron – a single signed floating point number – is in comparison to a biological neuron what a flat-head screwdriver is to a car.

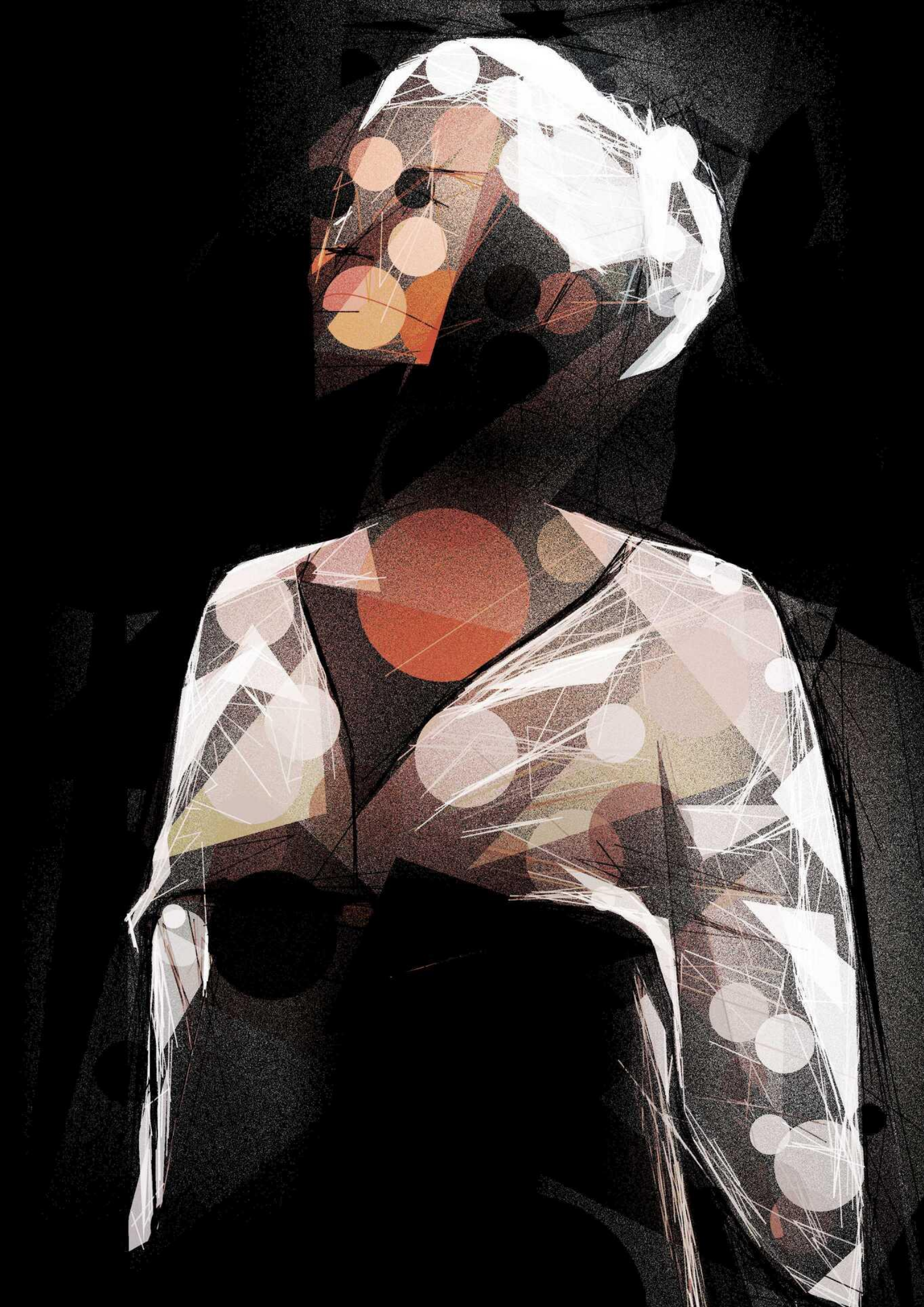
They both contain metal and that's about the extent of their similarity.

The human brain contains roughly 100 billion neuron cells, a layered chemical network, and a cerebrovascular system that all integrate as a whole to create a functioning, self-aware system capable of general reasoning and autonomous behaviour. This system is *multiple* orders of magnitude more complex than even the largest language model to date, both in terms of individual neuron structure, and taken as a whole.

4

# The bubble makes generative models too risky for business

> *Sceptical Believers are the ones who take the biggest risks, because they don't appreciate the pace of change of a paradigm shift like the True Believer, and they aren't as careful as the True Sceptic.*



## Summary

Generative models should not be adopted by most workers or businesses, *especially* if you believe that it's a revolutionary innovation.

## The horseshoe model of technological paradigm shifts

Occasionally, a field or industry is introduced to a new idea or technology that is such a profound change that it reframes everybody's understanding of that field as a whole.

This doesn't happen often, but when they do their impact is felt for decades, if not centuries, afterwards:

- Moveable type and the printing press.
- Radio, then TV.
- The microprocessor.
- Graphical user interfaces.
- The web.
- Touchscreen smartphones.

A paradigm shift accomplishes something an individual "disruptive innovation" cannot, it disrupts entire fields and industries in one swoop.

Technological paradigm shifts are lucrative for the tech industry: in our economy they almost always result in chaotic financial bubbles that lasts for years and make early participants enormously wealthy.

This has led industry pundits and hungry executives to look for a revolution in every product, every new feature set, and every new idea. The Java Virtual Machine and programming language was supposed to be *the* new platform for everything – even thought to inevitably kill the web. At one point the future was supposed to be entirely blogs and RSS feeds. Voice assistants and voice-controlled computers have several times been thought to be the new dominant paradigm that would govern all computing, as were various iterations of Virtual Reality. Who can forget how convinced tech industry investors were in the inevitability of crypto-coins and the blockchain?

Many of these micro-revolutions did result in long term innovations in the field. They might not have resulted in the dramatic changes people hoped for at the time, but the technological detritus they left behind once the teams were disbanded, companies sold off, and products shuttered is often useful. Sun Microsystems may not exist any more and Java didn't become the universal platform for all software, but it did become the foundation for Android software development and the chosen platform for server software in the enterprise. Blogs didn't become the future of social media, but RSS feeds are what makes podcasts work and WordPress, flawed as it is, has become the web's default Content Management System.

Some micro-revolutions, however, left only harm in their wake. The blockchain will go down in history as an enabler of fraud, crime, and Ponzi schemes. Others, such as self-driving cars, never managed to fulfil the promise they made, no matter how much money investors keep throwing at the problem.

Where generative models land on this spectrum is a question that tech companies have deliberately muddied and obfuscated. The tech, as it stands today, is kind of rubbish, and enables a lot of harm, but the same could have been said of the early web. Companies that stand to gain from an "AI" revolution play up this doubt and expect you to disbelieve the evidence of our own experience.

Given the dynamics of the tech industry, these companies stand to gain a lot from wide adoption. A bubble could make their executives immeasurably wealthy and allow the "merely" wealthy to become bil-

lionaires. The harms from blatantly flawed technology, such as massive insecurity of early versions of Windows, are merely abstract costs that can be pushed onto others.

Are generative models imperfect crap like the web in '94? Or are they broken crap like Windows 95?

Those who care enough to have an opinion are likely to fall broadly into one of three camps:

- *True Sceptic*. The true sceptic reasons that we're close to the upper limit of how much useful training data can be found for these models, that the flaws inherent in these systems will only be addressed slowly and painfully over the next few years if at all, and that even with improvement they aren't as useful as the industry is making them out to be – with harms and abuses vastly outweighing the benefit. Generative models will be closer to self-driving cars than the web.

- *True Believer*. The true believer is convinced that the party is just getting started. Improvements will come fast and quick. "Believe in the exponential!" Generative models will be everywhere and make the web look like a minor blip.

- *Sceptical Believer*. Where I think most people, at least in tech, will fall. The sceptical believer thinks the tech is already useful, that it has flaws that will be ironed out, thinks that the tech will improve over the years, but the improvement won't be exponential and will taper off relatively quickly.

Your opinion isn't static either. You can start off as a true sceptic but turn into a sceptical believer, and vice versa.

It's easy to assume that this leads to three different optimal strategies based on how much of a believer you are.

1. **Pessimistic Adoption.** Be extremely careful in adopting the new tech and only adopt specific implementations that are well-tested and have a demonstrated effectiveness. That would mean only allowing AI tools – internal- or external-facing – that have been

carefully vetted as safe and productive, discouraging all use of tools that haven't been approved.

2. **Optimistic Adoption.** Adopt the tech broadly, but with a few safeguards, maybe with a bit of assessment and vetting here and there to make sure it works for your business. That would mean allowing its use broadly for internal use but only allow specific tools that have been vetted for reliability and safety on external-facing products, pages, or services.

3. **Enthusiastic Adoption.** Quickly adopt the tech everywhere.

***This would be wrong.***

When it comes to technological revolution, the extremes bend together, like a horseshoe. The optimal strategy for both the *True Sceptic* and the *True Believer* is roughly the same: *Pessimistic Adoption*.

It's obvious why Pessimistic Adoption is the optimal strategy for the sceptic. They don't believe the technology works that well, so they will only adopt tools based on it if they can be tested and proven to work.

It's not immediately obvious why it's also the optimal strategy for the True Believer, but the reason is simple: if a technology will be everywhere, doing everything, and become the biggest thing since sliced bread, that also means that it will come with massive disruptive changes, inspire widespread fraud, and become one of the largest financial bubbles in history. Revolutions always have casualties:

- New products will come and go fast. Some will be made by grifters hoping to cash in on a trend. Some will simply be made obsolete overnight. Some will simply be based on incorrect guesses about what approaches work best for the new technology.

- Best practices will change frequently and *extremely* quickly.

- There will always be something better out tomorrow.

- Major foundational platforms can be replaced almost overnight by competitors who simply understand the paradigm better – think of how quickly Google replaced AltaVista.

In most of the major technological paradigm shifts, a majority of the products and services that get shipped will be poorly thought out garbage that gets replaced quickly. Many of them will even be outright fraud. The broad economic environment will be that of a massive bubble. Grifters will be on every corner. Nothing will be stable.

Since financial bubbles corrupt the strategy and incentives of even the most staid and careful company, it results in an environment where product and platform reliability and transparency is effectively non-existent.

An artificial bubble, one created through financial shenanigans and market manipulation, is already disconnected from product and market fundamentals in an obvious way. What's less obvious is that a genuinely disruptive innovation, such as the web, tends to create *even greater bubbles*, like the dot-com bubble of the nineties.

A rational-thinking sceptic will assume that we're in an artificially-created AI Bubble. A rational-thinking true believer should assume that we're in an even larger bubble that's even less rational than that.

Genuine revolutions create genuinely irrational environments.

The optimal strategy is to be *extremely* careful about what tools you use and adopt in your work or business, *especially if you believe in the "AI revolution"*.

The only difference between the optimal strategy for the True Believer and the True Sceptic is in how much they invest in the AI tool after they've vetted it. The sceptic will continue to keep it at an arm's length, making sure their business doesn't rely on it in any major way. Whereas the True Believer will go all-in on the few tools they're satisfied is working and weave them into the very fabric of their organisation.

Of the rational actors in the field, the Sceptical Believers are the ones who take the biggest risks, because they don't appreciate the pace

of change of a paradigm shift like the True Believer, and they aren't as careful as the True Sceptic.

That the majority of people and businesses in tech seem to be following the riskiest possible strategy available is both predictable and worrying. Most of them aren't reasoning themselves into taking these risk but are instead driven by *fear*.

# Fear Of Missing Out leads people and companies to take big risks

"*Because I might be missing out*," is not a reason nor is it a strategy. It's an excuse to avoid thinking – a way to avoid looking at your specific circumstances and investigating the matter as it presents to *you*.

How do generative models present to you?

After reading through the recommendations and the risks, seeing what the strengths and weaknesses are of these tools what do you think the benefits are for your work?

Because that's what matters: *your work*.

If this is indeed a revolution – a transformative change to our work and society – then there's even more reason to take your time.

Netscape Navigator was released in 1994.[5] Much like how ChatGPT in 2022 was the first service to package a language model into an easy-to-use service and brought Generative AI into the mainstream, Netscape was the first software to package the web up in a powerful, easy to use, web browser. It brought a revolution into the mainstream.

Most of the web or internet companies founded in Netscape's immediate wake were gone by the year 2000.[6]

Any organisation that waited for a couple of years, until 1996 or even 1997, to formulate a 'web strategy' was still early enough to be considered today to be an early web pioneer. Even those who waited until 1999 now manage to look almost presciently early in their adoption of the web.

*It didn't feel early at the time*.

I know. I was there. It felt like you had to move *now* or be forever left behind. But with the benefit of hindsight, it's obvious that most of

the organisations that didn't act in panic, gave themselves four years to investigate the web and see how they could use it *properly*, were better off in the long run.

Not only do you have time, patience is one of the biggest advantages you can give yourself at the start of a paradigm shift in computing. If this isn't paradigm shift then waiting a couple of years will let you find out without losing on the strategic benefit to your organisation.

In either case, we don't know yet how to properly integrate these tools with our existing work processes and tools. We don't know what user interface designs will result in genuine, l*ong term* productivity improvements in office software. Those who are first to ship "AI" integrations are likely to come out with extremely flawed first takes. If history is any example, they will be locked into some variation of their initial designs, unable to make meaningful changes without alienating their existing customers, but too flawed to gain broad acceptance.

This is the norm for tech. With rare exceptions, there is no first-mover advantage in software, and it's uncommon in tech in general. The product that ends up dominating the market is rarely the first. Instead, it's usually the product that comes a few years later, is more polished, or addresses a core strategic flaw in the first-mover.

- Apple II lost to Microsoft's DOS and IBM-compatible Personal Computers.
- WordPerfect lost to Microsoft Word.
- The classic Mac System operating system lost to Microsoft's Windows.
- Netscape lost to Internet Explorer, which then lost to Firefox, then Chrome.
- AltaVista lost to Google.
- Hotmail and Yahoo Mail lost to Gmail.
- Apple only started to 'win' when they stopped trying to be first to the market: the iPod was a late entry in the MP3 player wars.

- The iPhone wasn't even remotely close to being the first smartphone.
- The iPad wasn't the first tablet.
- Facebook trounced its *many* predecessors.

First-mover advantage in software is largely a myth. Historically, second- or even third-movers win. Waiting to learn from the first-mover's mistakes has usually been the winning move.

Instead of making mistakes, you want other people to make them, and then see what you can learn. Lessons learned through experience are expensive. It's always better to let somebody else pay that price.

You should hold off on integrating generative model features into your products because *regulators are coming*, and we don't know what their impact will be. We already have regulators in Europe and Canada looking into the AI industry. In the US, the Federal Trade Commission has been reminding us that existing regulations still apply to AI products, that if the vendors of these tools aren't doing a good enough job of mitigating their risks, they might be exposed to serious legal liability.[7] In addition, AI is likely to face additional regulations in the future. The European Union is on the cusp of introducing a set of regulations governing AI software and, if their plans hold, the fines will be similar to what they introduced with General Data Protection Regulation: they will be based on a percentage of global turnover.[8]

The impact regulators will have is unclear. Quite a few existing AI products may find themselves outright barred from most of Europe as a market.

Let other companies discover the pitfalls first.

*Otherwise, you risk spending a fortune weaving the latest and greatest language model into the fabric of your business, only to discover a couple of years later that it's now an obsolete liability.*

# The anthropomorphism trap

Another notion that complicates people's decision-making is the idea that large language models might be on the path towards Artificial General Intelligence. This is a dangerous idea for a business. It's a myth that directly undermines any effort to think clearly or strategise about generative models because it strongly reinforces *anthropomorphism*, which is when you reason about an object or animal *as if it were a person*. "AI" software that works with language is particularly prone to triggering this reaction. Software such as chatbots trigger all three major factors that promote anthropomorphism in people:[9]

1. *Understanding*. If we lack an understanding of how an object works, our minds will resort to thinking of it in terms of something that's familiar to us: people. We understand the world as people because that's what we are. This becomes stronger the more similar we perceive the object to be to ourselves.

2. *Motivation*. We are motivated to both seek out human interaction and to interact effectively with our environment. This reinforces the first factor. The more uncertain we are of how that thing works, the stronger the anthropomorphism. The less control we have over it, the stronger the anthropomorphism.

3. *Sociality*. We have a need for human contact and our tendency towards anthropomorphising objects in our environment increase with our isolation.

We lack cohesive mental models for what makes these language models so fluent. Because we feel a strong motivation to understand and use them as they are integrated into our work, and our socialisation in the office often takes on the very same text conversation form as a chatbot does, we inevitably feel a strong drive to see these software systems as people. The myth of AGI reinforces this because it implies that "people" is indeed an appropriate cognitive model for how these systems behave.

It isn't. ***AI are not people.*** Treating them as such is a major strategic error as it will prevent you from thinking clearly about their capabilities and limitations.

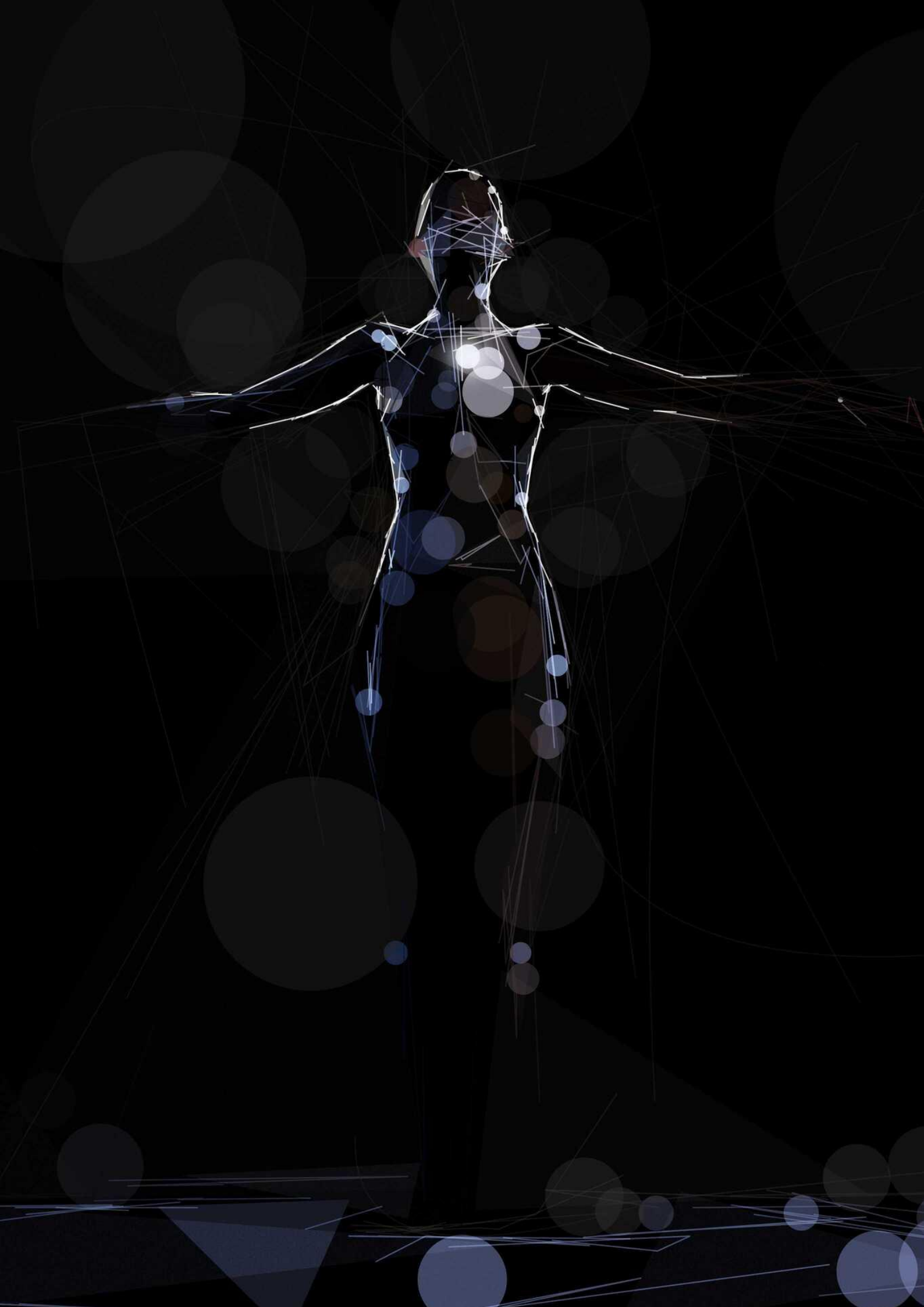
It doesn't matter where you stand on the sceptic spectrum. True believers and sceptics alike can fall for the anthropomorphism effect, and it will always compromise your ability to strategise about generative AI.

Avoid it as much as you can.

5

# AI and the bird brains of Silicon Valley

> *Why give the tech industry the benefit of the doubt when they are all but claiming godhood – that they've created a new form of life never seen before?*



# Summary

*Generative models are synthesis engines*. Machine Learning software use them to synthesise or generate new text, audio, or images from their training data based on prose prompts. This synthesis can be generative – making something new – or it can be used to convert or modify existing media. They are not Artificial General Intelligences.

# What is a Generative Model?

> The problem is, if one side of the communication does not have meaning, then the comprehension of the implicit meaning is an illusion arising from our singular human understanding of language (independent of the model). Contrary to how it may seem when we observe its output, an LM is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot.
>
> *Emily M. Bender, Timnit Gebru, et al., On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?.*[10]

Bird brains have a bad reputation. The diminutive size of your average bird and their brain has lead people to assume that they are dumb.

But, bird brains are amazing. Birds commonly outperform mammals with larger brains at a variety of general reasoning and problem-solving tasks.[11] Some by a large margin.[12] Their small brains manage this by packing numerous neurons in a small space using structures that are unlike from those you find in mammals.[13]

Even though birds have extremely capable minds, those minds are built in ways that are different from our own or other mammals. Similar capabilities; different structure.

The ambition of the Silicon Valley AI industry is to create something analogous to a bird brain: a new kind of mind that is functional-

ly similar to the human mind, possibly outperforming it, while being built using very different mechanisms. Similar capabilities; different structure.

This effort goes back decades, to the dawn of computing, and has had limited success.

Until recently, if the industry's claims are to be believed.

If you're reading this, you've almost certainly interacted with a generative model, however indirectly. Maybe you've tried Bing Chat. Maybe you've subscribed to the paid tier for ChatGPT. Or, maybe you've used Midjourney to generate images. At the very least you've been forced to see the images or text posted by the overenthusiastic on social media.

These models are created by pushing an enormous amount of training data through various algorithms:

- Language models like ChatGPT is "trained" on a good chunk of the textual material available on the web.

- Image models like Midjourney and Stable Diffusion are "trained" on a huge collection of images found on the internet.

What comes out the other end is a mathematical model of the media domain in question: text or images.

You know what generative models are in terms of how they present to you as software: clever chatbots that do or say things in response to what you say: *your prompt*. Some of those responses are useful, and they give you an impression of sophisticated comprehension. The models that generate text are fluent and often quite engaging.

This fluency is misleading. What Bender and Gebru, who I quoted above, meant when they coined the term *stochastic parrot* wasn't to imply that these are, indeed, the new bird brains of Silicon Valley, but that they are unthinking text synthesis engines that just repeat phrases from their training data set. They are the proverbial parrot who echoes without thinking, not the actual parrot who is capable of complex reasoning and problem-solving.

A *zombie parrot*, if you will, that screams for *brains* because it has none.

The fluency of the zombie parrot – the unerring confidence and a style of writing that some find endearing – creates a strong illusion of intelligence.

Every other time we read text, we are engaging with the product of another mind. We are so used to the idea of text as a representation of another person's thoughts that we have come to mistake a person's writing *for* their thoughts. But they aren't. Text and media are tools that authors and artists create to let people change their own state of mind – hopefully in specific ways to form the image or effect the author was after.

Reading is an indirect collaboration with the author, mediated through the writing. Text has no inherent reasoning or intelligence. Roland Barthes's ghost does not inhabit the words of *The Death of the Author*. Stephen King isn't hovering over you when you read *Carrie*. The ghost you feel while reading is an illusion you've made out of your own experience, knowledge, and imagination. Every word you read causes your mind to reconstruct its meaning using your memories and creativity. The idea that there is intelligence somehow inherent in writing is an illusion. The intelligence is *all* yours, all the time: thoughts you make yourself in order to make sense of another person's words. This can prompt us to greatness, broaden our minds, inspire new thoughts, and introduce us to new concepts. A book can contain worlds, but we're the ones that bring them into being as we read. What we see is uniquely our own. The thoughts are not transported from the author's mind and injected into ours.

The words themselves are just line forms on a background with no inherent meaning or intelligence. The word "horse" doesn't come with the Platonic ideal of a horse attached to it. The word "anger" isn't full of seething emotion or the restrained urge towards violence. Even words that are arguably onomatopoeic, like the word "brabra" we use in Icelandic for the sound a duck makes, are still incredibly specific to the cultures and context they come from. We are the ones doing the heavy lifting in terms of reconstructing a picture of an intelligence be-

hind the text. When there is no actual intelligence, such as with ChatGPT, we are the ones who end up filling in the gaps with our own intelligence, memories, experience and imagination.

When ChatGPT demonstrates intelligence, that comes from us.[14] Some of it we construct ourselves.[15] Some of it comes from our inherent biases.[16]

There is no 'there' there. We are alone in the room, reconstructing an abstract representation of a mind. The reasoning you see is only in your head. You are hallucinating intelligence where there is none. You are doing the textual equivalent of seeing a face in a power outlet.

This drive – *anthropomorphism* – seems to be innate. Our first instinct when faced with anything unfamiliar – whose drives, motivations, and mechanisms we don't understand – is to assume that they think much like a human would.[17] When that unfamiliar agent uses language like a human would, the urge to see them as near or fully human is impossible to resist – a recurring issue in the history of AI research that dates all the way back to 1966.[18]

These tools solve problems and return fluent, if untruthful, answers, which is what creates such a convincing illusion of intelligence.

Text synthesis engines like ChatGPT and GPT-4 do not have any self-awareness. They are mathematical models of the various patterns to be found in the collected body of human text. How granular the model is depends on its design and the languages in question. Some of the tokens – the smallest unit of language the model works with – will be characters or punctuation marks, some of them will be words, syllables, or even phrases. Many language models are a mixture of both.[19]

With enough detail – a big enough collection of text – these tools will model enough of the probabilistic distribution of various words or characters to be able to perform what looks like magic:

- They generate fluent answers by calculating the most probable sequence of words, at that time, which would serve as the continuation of or response to the prompt.

- They can perform limited reasoning tasks that correlate with textual descriptions of prior reasoning tasks in the training data.

With enough of these correlative shortcuts, the model can perform something that looks like common sense reasoning: its output is text that replicates prior representations of reasoning.[20] This works for as long as you don't accidentally use the wrong phrasing in your prompt and break the correlation.[21]

The mechanism behind these systems is entirely correlative from the ground up.[22] What looks like reasoning is incredibly fragile[23] and breaks as soon as you rephrase or reword your prompt.[24] It exists only as a probabilistic model of text. A Generative Model chatbot is a language engine incapable of genuine thought.[25]

These language models are interactive but static snapshots of the probability distributions of a written language.

It's obviously interactive, that's the whole point of a chatbot. It's static in that it does not change when it's used or activated. In fact, changing it requires an enormous amount of computing power over a long period of time. What the system models are the distributions and correlations of the tokens it records for the texts in its training data set – how the various words, syllables, and punctuation relate to each other over as much of the written history of a language as the company can find.

That's what distinguishes biological minds from these algorithmic hindsight factories: a biological mind does not reason using the probability distributions of all the prior cultural records of its ancestors. Biological minds learn primarily through trial and error. Try, fail, try again. They develop skills by reinforcing neural pathways in a brain that's already built from nutrition through genetics in a ways that's functionally very different from what you see in a software model,[26]. That reinforcement takes place through constant feedback, experimentation, and repeated failure – driven by a chemical network that often manifests as instinct, emotion, motivation, and drive. The biological neural network – bounded, defined, and driven by the chemical network – is constantly changing and responding to outside stimuli.

Every time an animal's nervous system is "used", it changes. It is always changing, until it dies.

Biological minds *experience*. Synthesis engines parse imperfect *records* of experiences. The former are forward-looking and operate primarily in the present, sometimes to their own detriment. The latter exist exclusively as probabilistic manifestations of imperfect representations of thoughts past. *They are snapshots*. Generative Models are themselves cultural records.

These models aren't new bird brains – new alien minds that are peers to our own. They aren't even insect brains. Insects have autonomy. They are capable of general problem-solving – some of them are capable of tasks of surprising complexity[27] – and their abilities tolerate the kind of minor alterations in the problem environment that would break the correlative pseudo-reasoning of a language model.[28] Large Language Models are something *lesser*. They are water running down pathways etched into the ground over centuries by the rivers of human culture. Their originality comes entirely from random combinations of historical thought. They do not know the 'meaning' of anything – they only know the records humans find meaningful enough to store.[29] Their unreliability comes from their unpredictable behaviour in novel circumstances. When there is no riverbed to follow, they drown the surrounding landscape.

The entirety of their documented features, capabilities, and recorded behaviour – emergent or not – is explained by this conceptual model of generative "AI". Every complex system has emergent behaviours.[30] It is not evidence of sentience or self-awareness. There are no unexplained corner cases that don't fit or actively disprove this theory.

Yet people keep assuming that what ChatGPT does can only be explained as the first glimmer of genuine Artificial General Intelligence. The bird brain of Silicon Valley is born at last!

Because text and language are the primary ways we experience other people's reasoning, it will be next to impossible to dislodge the notion that these are genuine intelligences. No amount of examples, scientific research, or analysis will convince those who want to maintain a pseudo-religious belief in alien peer intelligences. After all, if you

want to believe in aliens, an artificial one made out of supercomputers and wishful thinking *feels* much more plausible than little grey men from outer space. But that's what it is: *a belief in aliens*.

It doesn't help that so many working in "AI" seem to *want* this to be true. They seem to be true believers who are convinced that the spark of Artificial General Intelligence has been struck.[31]

They are inspired by the science fictional notion that if you make something complex enough, it will spontaneously become intelligent. This isn't an uncommon belief. You see it in movies and novels – the notion that any network of sufficient complexity will spontaneously become sentient has embedded itself in our popular psyche. James Cameron's skull-crushing metal skeletons have a lot to answer for.

That notion doesn't seem to have any basis in science. The idea that general intelligence is an emergent property of neural networks that appears once the network reaches sufficient complexity, is a view based on archaic notions of animal intelligence – that animals are soulless automata incapable of feeling or reasoning.[32] That view that was formed during a period where we didn't realise just how common self-aware reasoning (i.e. the mirror test) and general reasoning is in the animal kingdom.[33] Animals are smarter than we assumed and the difference between our reasoning and theirs seems to be a matter of degree, not of presence or absence.[34]

General reasoning seems to be an *inherent*, not emergent, property of pretty much any biological lifeform with a sizeable nervous system.

The bumblebee, despite having only a tiny fraction of the neurons of a human brain, is capable of not only solving puzzles but also of *teaching other bees to solve those puzzles*. They reason and have a culture.[35] They have more genuine and robust general reasoning skills – that don't collapse into incoherence at minor adjustments to the problem space – than GPT-4 or any large language model on the market. That's with only around half a million neurons to work with.[36]

Conversely, GPT-3 is made up of 175 *billion* parameters – what passes for a "neuron" in a digital neural network.[37] GPT-4 is even larger, with some estimates coming in at a *trillion* parameters. Then you have fine-tuned systems such as ChatGPT, that are built from multiple interact-

ing models layered on top of GPT-3.5 or GPT-4, which make for an even more complex interactive system.

ChatGPT, running on GPT-4 is, easily a *million* times more complex than the "neural network" of a bumblebee and yet, out of the two, it's the striped invertebrate that demonstrates robust and adaptive general-purpose reasoning skills. Very simple minds, those belonging to small organisms that barely have a brain, are capable of reasoning about themselves, the world around them, and the behaviour of other animals.[38]

Unlike the evidence for 'sparks' of AGI in language models, the evidence for animal reasoning – even consciousness – is broad, compelling, and encompasses decades of work by numerous scientists.[39]

AI models are flawed attempts at digitally synthesising neurologies. They are built on the assumption that all the rest – metabolisms, hormones, chemicals, and senses – aren't necessary for developing intelligence.

Reasoning in biological minds does not seem to be a property that emerges from complexity. The capacity to reason looks more likely to be a *built-in* property of most animal minds.[40] A reasoning mind appears to be a direct consequence of how animals are structured as a whole – chemicals, hormones, and physical body included. The animal capacity for problem-solving, social reasoning, and self-awareness seem to increase, unevenly, and fitfully with the number of neurons until it reaches the level we see in humans.[41] Reasoning does not 'emerge' or appear. Some creatures are better at it than others, but it's there in some form even in very small, very simple beings like the bumblebee. It doesn't happen magically when you hook up a bunch of disparate objects together in a complex enough network. A reasoning mind is the *starting point* of biological thinking, not the endpoint that only "emerges" with sufficient complexity.

The internet – a random interconnected collection of marketing offal, holiday snaps, insufferable meetings, and porn – isn't going to become self-aware and suddenly acquire the capacity for general reasoning once it reaches a certain size, and neither will Large-Language-Models. The notion that we are making autonomous beings capable of

Artificial General Intelligence just by loading a neural network up with an increasingly bigger collection of junk from the internet is not one that has any basis in anything we understand about biology or animal reasoning.

But, AI companies insist that they are on the verge of AGI.[42] Their rhetoric around it verges on the religious as the idea of an AGI is idealised and almost worshipped.[43] They claim to be close to making a new form of thinking life, but they refuse to release the data required to prove it.[44] They've built software that performs well on the arbitrary benchmarks they've chosen and claim are evidence of general intelligence, but those tests prove no such thing and have no such validity.[45] The benchmarks are theatrics that have no applicability towards demonstrating genuine general intelligence.[46]

AI researchers love to resurrect outdated pseudoscience such as phrenology – shipping AI software that promises to be able to tell you if somebody is likely to be a criminal based on the shape of their skull.[47] It's a field where researchers and vendors routinely claim that their AIs can detect whether you're a potential criminal, gay, a good employee, liberal or conservative, or even a psychopath, based on "your face, body, gait, and tone of voice."[48]

*It's pseudoscience*.

This is the field and the industry that claims to have accomplished the first 'spark' of Artificial General Intelligence?[49]

Last time we saw a claim this grand, with this little scientific evidence, the men in the white coats were promising us room-temperature fusion, giving us free energy for life, and ending the world's dependence on fossil fuels.[50]

Why give the tech industry the benefit of the doubt when they are all but claiming godhood – that they've created a new form of life never seen before?

As Carl Sagan said: "*extraordinary claims require extraordinary evidence*."

He didn't say "extraordinary claims require only vague insinuations and pinky-swear promises."

To claim you've created a completely new kind of mind that's on par with any animal mind – or, even superior – and provides general intelligence using mechanisms that don't resemble anything anybody has ever seen in nature, is by definition the most extraordinary of claims.

The AI industry is backing their claims of Artificial General Intelligence with hot air, hand-waving, and cryptic references to data and software nobody outside their organisations is allowed to review or analyse.[51]

They are pouring an every-increasing amount of energy and work into ever-larger models all in the hope of triggering the 'singularity' and creating a digital superbeing. Like a cult of monks boiling the oceans in order to hear whispers of the name of God.[52]

It's a farce. All theatre; no evidence. Whether they realise it or not, they are taking us for a ride. The sooner we see that they aren't backing their claims with science, the sooner we can focus on finding safe and productive uses – limiting its harm, at least – for the technology as it exists today.

After everything the tech industry has done over the past decade, the financial bubbles, the gig economy, legless virtual reality avatars, crypto, the endless software failures – just think about it – do you think we should believe them when they make grand, unsubstantiated claims about miraculous discoveries? Have they earned our trust? Have they shown that their word is worth more than that of independent scientists?

Do you think that they, with this little evidence, have really done what they claim, and discovered a literal new form of life? But are conveniently unable to prove it because of 'safety'?

Me neither.

# 6

# *The Intelligence Illusion* or the LLMentalist Effect

> *This is an industry that didn't understand what it was doing and, now, doesn't understand what it did.*

# Summary

*Many of those who believe we are on the verge of discovering Artificial General Intelligence have likely fallen for a statistical illusion.*

# How chat-based Large Language Models replicate the mechanisms of a psychic's con

"*This is real. It's a bit worrying, but it's real.*"

"*There really is something there. Not sure what to think of it, but I've experienced it myself.*"

"*You need to keep your mind open to the possibilities. Once you do, you'll see that there's something to it.*"

One of the paradoxes of the generative AI bubble is that many prominent people in the industry itself also seem to be *afraid* of it. People who have worked in AI for years, many of whom are still actively working on new generative models, seem to alternate between awe and dread – simultaneously convinced that this could save or end humanity.

This specific blend of awe, disbelief, and dread is familiar to those who have studied a specific type of con artist. This sense that their world is being transformed, but they're not sure if it's for good or ill, is also common among victims of a mentalist scam artist: *psychics*.

The psychic's con is a very convincing illusion that is often *more* effective on those who think themselves to be more intelligent than the rest. The result almost inevitably challenges either their self-image or their world view. Either you aren't as smart as you think you are or the world is fundamentally different from what you thought.

Many prefer to believe that these psychics are real because the alternative – that they were fooled by an illusion – is worse to them. If they are convinced enough of their own genius, the possibility that they were conned won't even enter their mind.

The intelligence illusion that language model chatbots create seems to be based on the same mechanism as that of a psychic's con – often called *cold reading*. It looks like an accidental automation of the same basic tactic.

By using *validation statements*, such as sentences that use the Forer effect, the chatbot and the psychic both give the impression of being able to make extremely specific answers, but those answers are in fact *statistically generic*.

The psychic uses these statements to give the impression of being able to read minds and hear the secrets of the dead.

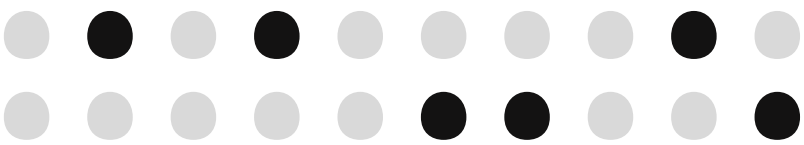
The chatbot gives the impression of an intelligence that is specifically engaging with you and your work, but that impression is nothing more than a statistical trick.[53]

The psychic's con is a tried and true method for scamming people that has been honed through the ages.

There are many variations, but the core mechanism remains the same.

# The Psychic's Con

**1. Audience selection**



*Most people aren't interested in psychics, so the initial audience pool is already generally more open-minded and less critical of the phenomenon than the population in general.*

Not every seer, tarot card reader, psychic, or mind reader is a con artist. Sometimes the "psychic" is open about it all just being enter-

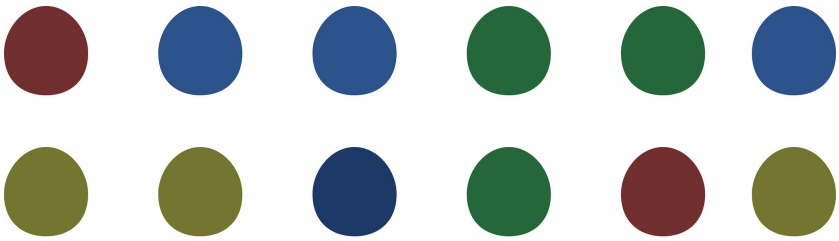
tainment and aren't pretending to be able to contact spirits or read minds. Some psychics do not have a profit motive at all, and without the con it doesn't seem fair to call somebody a con artist.

But many of them *are* con artists who are deliberately fooling people, and those all operate using the same basic mechanism that begins well before the reading itself.

Hardcore sceptics will almost always be in a very small minority of the audience, which both makes them easy to manage and provides social pressure on them to tone down their scepticism.

Those who attend are primed to believe and are already familiar with the mythology surrounding psychics. All of which helps them manage expectations and frame their performance.

**2. Setting the scene**



*The initial audience is prepared. Lights are dimmed. The psychic is hyped up. Staff research the audience on social media or through conversation. The audience's demographics are noted.*

Usually, at the start of the performance, a member of staff reminds the audience of the ground rules for how psychic readings "work". They are helped by the popularisation of these rules by media, cinema, and TV.

Everybody now "knows" that:

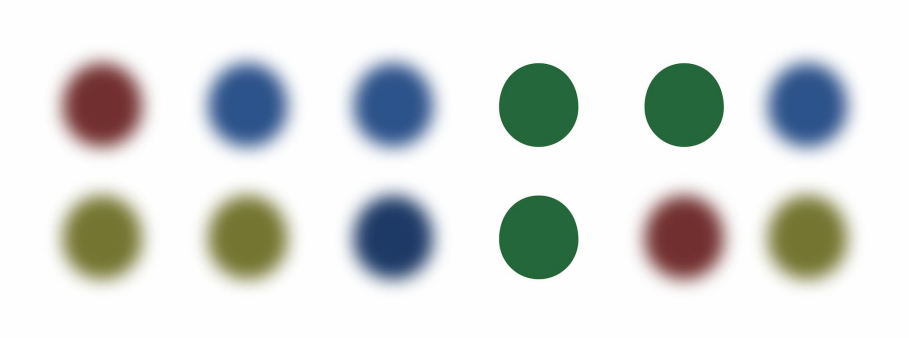
- Readings usually begin murky and unclear.
- They then become clearer as the "connection" to the "spirit world" gets stronger.
- Errors are expected. The "spirits" are often vague or hard to hear.

- Non-believers can weaken or even disrupt the connection.

Psychics also habitually research their audience by mapping out their demographics, looking them up on social media, or even with informal interviews performed by staff mingling with attendees before the performance begins.

When the lights dim, the psychic should have a clear idea of which members of the audience will make for a good mark.

**3. Narrowing down**

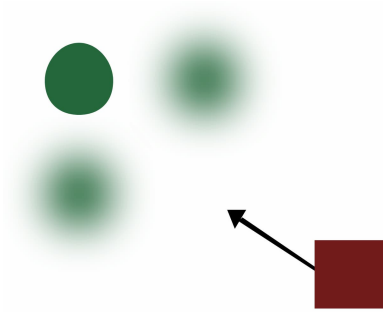


*The psychic gauges the information they have on the audience, gestures towards a row or cluster, and makes a statement that sounds specific but is in fact statistically likely for the demographic. Usually at least one person reacts. If not, the psychic will imply that the secret is too embarrassing for the "real" person to come forward, reminds people that they're available for private readings, and tries again.*

This way the mark usually chooses themselves. The psychic makes a statement and points towards a row, quickly altering their gesture based on somebody responding visible to the statement. This makes it look like they pointed at the mark right from the beginning.

The mark is that way primed from the start to believe the psychic. They're off-guard. Usually a bit surprised and totally unprepared for the quick burst of questions the psychic offers next. If any of those questions land and draw the mark in, they are followed by the actual reading. Otherwise, they move on and try again.

### 4. Testing the mark – Cold reading using subjective validation



*The reaction indicates that the mark believes they were "read". This leads to a burst of questions that, again, sound very specific but are in fact statistically generic. If the mark doesn't respond, the psychic declares the initial read a success and tries again.*

The con – cold reading[54] – hinges on a quirk of human psychology: if we personally relate to a statement, we will generally consider it to be accurate, no matter how common that experience is.

This unfortunate side effect of how our mind functions is called ***subjective validation***.[55]

> Subjective validation, sometimes called personal validation effect, is a cognitive bias by which people will consider a statement or another piece of information to be correct if it has any personal meaning or significance to them. People whose opinion is affected by subjective validation will perceive two unrelated events (i.e., a coincidence) to be related because their personal beliefs demand that they be related.
>
> *Wikipedia entry on Cold Reading*[56]

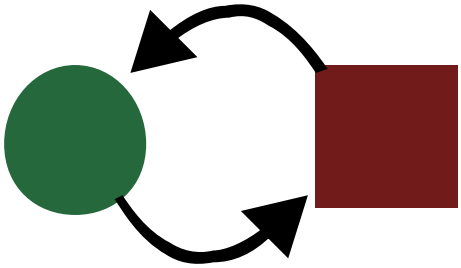
As a consequence, many people will interpret even the most generic statement as being *specifically about them* if they can relate to what was said.

The more eager they are to find meaning in the statement, the stronger the effect.

The more they believe in the speaker's ability to make accurate statements, the stronger the effect.

The basic mechanism of the psychic's con is built on the mark being willing and able to relate what was said to themselves, even if it's unintentional.

**5. The subjective validation loop using validation statements**



*The con begins in earnest. The psychic asks a series of questions that sound very specific to the mark but are in reality nothing more than statistically probable guesses, based on their demographics and prior answers, phrased in a specific, highly confident way.*

The psychic taps into this cognitive bias by making a series of statements that are tailored to be personally relatable – sound specific to you – while actually being statistically generic.

These statements come in many types. I use "*validation statements*" here as an umbrella term for all these various tactics.

Some common examples:

- **Forer or Barnum statements**[57] are probably the most famous kind of statement that plays into the subjective validation effect. Many of these statements are inherently meaningless but are nonetheless felt to be accurate by listeners. Most people will consider "*you tend to be hard on yourself*" to be an accurate description of themselves, for example.

- **Vanishing negative** is where a question is rephrased to include a negative such as "not" or "don't". If the psychic asks "*you don't play the piano*?" then they will be able to reframe the question as accurate after the fact, no matter what the answer is. If you answer negative: "*didn't think so*". Positive: "*that's what I thought*."

- **Rainbow ruse**[58] where the psychic associates the mark with both a trait and its opposite. "*You're a very calm person, but if provoked you can get very angry.*"

- **Statistical guesses**. Statements like "*you have, or used to have, a scar on your left leg or knee*" apply to almost everybody. With

enough knowledge of common statistics, the psychic can make general statements that sound *incredibly* specific to the mark.

- **Demographic guesses**. Similar to *statistical guesses*, these are statements that are common to a demographic but will sound very specific to the mark that's listening.

- **Unverifiable predictions**. Predictions like "*somebody bears a strong ill will towards you, but they are unlikely to act on it*" are impossible to verify, but will sound true to many people.

- **Shotgunning**[59] is one of the more common tactic where the psychic will fire off a series of statements. The mark will find one of the statements to be accurate and, due to how our minds work, will come away only remembering the correct statement.

An important part of this process is the tone and bearing of the psychic. They need to be confident, be quick in dismissing errors and moving on when they make mistakes, and they need to be quick to read people's expressions and body language and adjust their responses to match.

### 6. The con is completed

*The psychic ends the conversation and the mark is left with the sense that the psychic has uncanny powers. But the psychic isn't the real thing. It's all a con.*

At the end of the process, the mark is likely to remember that the reading was eerily correct –that the psychic had an almost supernatural accuracy – which primes them to become *even more receptive the next time they attend*.

This is where the con often becomes insidious: the effect becomes stronger the more cooperative the mark is, and they often become more cooperative over time.

What's more, susceptibility has nothing to do with intelligence.

Somebody raised to believe they have high IQ is *more* likely to fall for this than somebody raised to think less of their own intellectual capabilities. Subjective validation is a quirk of the human mind. *We all*

*fall for* it. But if you think you're unlikely to be fooled, you will be tempted instead to apply your intelligence to "figure out" how it happened. This means you can end up using considerable creativity and intelligence to *help* the psychic fool you. Because you think you can't be fooled, you also bring your intelligence to bear to *defend* the psychic's claim of their powers. Smart people (or, those who think of themselves as smart) can become the biggest, most lucrative marks.

Whereas the sceptic who thinks less of themselves is more likely to just go:

"*That's a neat trick. I don't know how you pulled it off. Must be something very clever.*"

And just move on.

**Many psychics fool themselves**

It isn't unusual for psychics to unconsciously develop a practice of subconscious cold reading.[60] The psychics themselves might not even be aware of their own tactics.

As Denis Dutton describes:

> As a postgraduate student in pursuit of a scientific career, he became intrigued with astrology. Though during this period he had nagging doubts about the physical basis of astrology, he was encouraged to continue with it by his many satisfied clients, who invariably found his readings "amazingly accurate" in describing their personal situations and problems. Not until he had one day obtained such a gratifying reaction to a horoscope which, he realized later, he had cast completely incorrectly, did he begin slowly to understand the real nature of his activity: his great success as an astrologer had nothing whatsoever to do with the validity of astrology as a science. He had become, in fact, a proficient cold reader, one who sincerely believed in the power of astrology under the constant reinforcement of his clients. He was fooling them, of course, but only after falling for the illusion himself.
>
> *Cold Reading*[61]

You can find many examples of this once you start doing the research. The mechanism is simple enough and already baked into people's preconceptions of how readings work so many psychics accidentally develop the knack for it, meaning that they're not just conning the person being read, *they are also conning themselves*.

This point will become important later.

# AI and *The Intelligence Illusion*

**1. The audience selects itself**

*People sceptical about "AI" chatbots are less likely to use them. The most active audience will be early adopters, tech enthusiasts, and genuine believers in AGI who will all generally be less critical and more open-minded.*

If you aren't interested in "AI", you are less likely to use an "AI" chatbot, and if you try one, you're more likely to be critical and less likely to return.

Many of the more avid users of these chatbots are self-selected to be enthusiastic and open-minded about the field of AI and the notion of Artificial General Intelligence (AGI). Those who are genuine enthusiasts about AGI – the notion that the field is about to invent a new kind of mind – are likely to be substantially more enthusiastic about using these chatbots than the rest.

This parallels the audience selection for the psychic's con. Those who believe in an afterlife and that it can be contacted by the living are substantially more likely to attend a psychic's reading than others.

**2. Setting the stage**

*Users are primed by the hype surrounding the technology. The chat environment sets the mood and expectations. Warnings about it being "early days" and "hallucinations" both anthropomorphise the bot and provide ready-made excuses for when one of its constant failures is noticed.*

Our current environment of relentless hype sets the stage and builds up an expectation of seeing at least a glimmer of genuine intelligence in the chatbot. For all the warnings vendors make about these systems not being general intelligences, those statements are always followed by either an implied or an actual "yet". The hype strongly implies that these are "almost" intelligences and that you should be able to perceive "sparks" of intelligence in them.

Those who believe this are primed for subjective validation.

The warnings also play a role in setting the stage. "It's early days" means that when the statistically generic nature of the responses are spotted, it's easily dismissed as an "error". Anthropomorphising con-

cepts such as using "hallucination" as a term help dismiss the fact that statistical responses are completely disconnected from meaning and facts. The hype and *mythology* of AI primes the audience to think of these systems as persons to be understood and engaged with, all but guaranteeing subjective validation.

**3. The prompt establishes the context**

*Each user gives the chatbot a prompt and it answers. Many users will either accept the answer as given or repeat variations on the initial prompt to get the desired result. They move on without falling for the effect. But some users engage in conversation and get drawn in.*

The initial prompt interaction is the first filter. Most will just take the first answer and leave, or at most will repeat variations of their prompt until they get the result they wanted. These interactions are purely mechanical. The end-user is treating the chatbot merely as a generative widget, so they never get pulled into the illusion.

Some end-users, usually those who are more enthusiastic about the prospect of "AI", begin to engage and get pulled into "conversation" with a mathematical language model.

**4. The mark tests themselves – subjective validation kicks in**

*The chatbot's answers sound extremely specific to the current context but are in fact statistically generic. The mathematical model behind the chatbot delivers a statistically plausible response to the question. The marks who find this convincing get pulled in.*

That conversation is the primary filter. Those who want to believe will see the responses to their prompt as being both specifically about them and intelligent. They are primed to see the chatbot as a person who is reading *their* text and thoughtfully responding to them. But that isn't how language models work. They model the distribution of words and phrases in a language as tokens. Their responses are nothing more than a statistically likely response to the prompt.

You give it text. It gives you a response that matches responses that texts like yours commonly get in its training data set.

Already, this is working along the same fundamental principle as the psychic's con: the chatbot isn't "reading" your text any more than

the psychic is reading your mind. They are giving you statistically plausible responses based on what you said. You're the one who is finding ways to validate those responses as being specific to you and your words as the subject of the conversation.

Because of the sheer size of the training data set the responses from the chatbot will look *very* convincing and specific, even though they are statistically generic. Once you've trained on a big chunk of the past twenty years of the web, large collections of stolen ebooks, and a substantial amount of custom interactions by low-wage workers, the model will have a response for almost everything you can think of, or can use a variation of something it's already seen.

These initial interactions can be quite compelling, especially if you're a believer in "AI", but it is in the longer and repeated conversations that the effect really begins to kick in.

**5. The subjective validation loop – RLHF enters the picture**

*The mark asks a series of questions and all the replies sound like reasoned answers specific to the context but are in reality just statistically probable guesses. The more the mark engages, the more convinced they are of the chatbot's intelligence.*

It's important to remember at this stage how *Reinforcement Learning through Human Feedback* (RLHF) works.[62]

This is the method that vendors use to turn a raw language model into a chatbot that can hold a conversation.

*RLHF* doesn't let the vendor make specific corrections to the output of a language model. The method involves using human feedback to rank a variety of texts generated by the model, usually following some of other form of fine-tuning. The ranked texts are in turn used to train a separate reward model. It's this model that is responsible for the actual *Reinforcement Learning* of the language model. That reward model, coupled with fine-tuning the language model on collections of conversations, is what turns the borderline unhinged conversations of a regular model into the fluent experience you see in systems such as ChatGPT.

Because the feedback is based on general rankings, it can't easily address specific issues. If a model makes a false statement in a conver-

sation, that conversation gets a lower rank, but there is no way to specifically correct that statement automatically in future responses.

This lack of concrete specificity means that RLHF models in general are likely to reward responses that *sound* accurate. As the reward model is just another language model, it can't reward based on facts or specific details, so it can only reward output that has a tone, style, or structure that's commonly associated with statements that have been rated as accurate.

Even the ratings themselves are suspect. Most, if not all, of the workers who provide this feedback to AI vendors are low-paid workers who are unlikely to have specialised knowledge relevant to the topic they're rating, and even if they do, they are unlikely to have the time to fact-check everything.

They are going to be ranking the conversations almost entirely based on tone and sentence structure.

This is why I think RLHF has effectively become a reward system that specifically optimises language models for generating validation statements: *Forer statements, shotgunning, vanishing negatives, and statistical guesses*.

In trying to make the language model sound more human, more confident, and more engaging, but without being able to edit specific details in its output, AI researchers seem to have created *a mechanical mentalist*.

Instead of pretending to read minds through statistically plausible validation statements, it pretends to read and understand your text through statistically plausible validation statements.

The validation loop can continue for a while, with the mark constantly doing the work of convincing themselves of the language model's intelligence. Done long enough, it becomes a form of reinforcement learning *for the mark*.

### 6. The marks become cheerleaders

*The mark is left with the sense that the chatbot is uncannily close to being self-aware and that it is definitely capable of reasoning. But it's nothing more than a statistical and psychological effect.*

The most enthusiastic believers in an imminent AI revolution are starting to sound very similar to long-time believers in psychics and mind-reading.

They come up with increasingly convoluted ideas and models to explain why the impossible is possible. They become more and more dismissive of fields of science and research that challenge their world view. Their own statements become tinged with awe and dread.

And they keep evangelising. *This is real!*

Often followed by: *This is dangerous!*

Remember, the effect becomes *more* powerful when the mark is both intelligent and *wants to believe*. Subjective validation is based on how our minds work, in general, and is unaffected by your reported IQ.

If anything, your intelligence will just improve your ability to rationalise your subjective validation and make the effect stronger. When it's coupled by a genuine desire to believe in the con – that we are on the verge of discovering Artificial General Intelligence – the effect should both be powerful and hard to resist.

This is why you can't rely on user reports to discover these issues. People who believe in psychics will generally have only positive things to say about a psychic, even as they're being bilked. People who believe we're on the verge of building an AGI will only have positive things to say about chatbots that support that belief.

# It's easy to fall for this

Falling for this statistical illusion is easy. It has nothing to do with your intelligence or even your gullibility. It's your brain working against you. Most of the time conversations are collaborative and personal, so your mind is optimised for finding meaning in what is said under those circumstances. If you also *want* to believe, whether it's in psychics or in AGI, your mind will helpfully find reasons to believe in the conversation you're having.

The psychic's con is a mechanism that has been extraordinarily successful at fooling people over the years. It works.

The best defence is to respond the same way as you would to a convincing psychic's reading: "*That's a neat trick, I wonder how they pulled it off?*"

***Well, now you know.***

Once you're aware of the fallibility of how your mind works, you should have an easier time spotting when that fallibility is being exploited, intentionally or not.

That brings us to an important question.

**Is this intentional?**

Given that there are billions of dollars at stake in the tech industry, it would be tempting to assume that the statistical illusion of intelligence was intentionally created by people in the tech industry.

I think that's extraordinarily unlikely.

The field of AI research has a reputation for disregarding the value of other fields, so this reimplementation of a psychic's con is very likely accidental. Being unaware of much of the research in psychology on cognitive biases or how a psychic's con works, the industry probably stumbled into the mechanism and made chatbots that fooled many of the chatbot makers themselves.

Remember what I wrote above about psychics frequently having conned themselves, that many of them aren't even aware of their own scam? The same applies here. This is an industry that didn't understand what it was doing and, now, doesn't understand what it did.

This is why so many people in tech are completely convinced that they have created the first spark of true Artificial General Intelligence.

# The Recommendations

7

# “AI” is an industry dominated by bullshit and snake oil

*“The AI industry and tech companies in general do not have much historical credibility. Their response to criticism is always: “we’ve been wrong in the past; mistakes were made; but this time it’s different!”*

## Summary

AI research is of a poor quality. *Many of their claims will be proven wrong*. AI software vendors have a strong financial incentive to exaggerate the capabilities of their tools and make them hard to disprove. This undermines attempts at scientific rigour. Many of AI's promises are *snake* oil.

## It's *vital* that you be sceptical about any claim made by AI vendors

The AI industry is prone to hyperbolic announcements. IBM's Watson was supposed to transform healthcare and education but ended up being a costly disaster.[63] Amazon was planning on using AI to revolutionise recruitment before they realised they'd automated discrimination and had to scrap the project.[64] AI was supposed to be a revolutionary new tool to fight the COVID-19 but none of them ended up working well enough to be safe.[65] The Dutch government tried to use it to weed out benefits fraud. Their trust in the AI system they'd bought resulted in over a thousand innocent children being unjustly taken from their families and into foster care.[66] Gullibly believing the hype of the AI industry causes genuine harm.

AI system vendors are prone to making promises they can't keep. Many of them, historically, haven't even *been* AI or Machine learning despite the vendor's claims.[67] The US Federal Trade Commission has even seen the need to remind people that claims of magical AI capabilities need to be based in fact.[68]

AI companies love the trappings of science, they publish 'studies' that are written and presented in the style of a paper submitted for peer review. These 'papers' are often just uploaded to their own web-

sites or dumped onto archival repositories like Arxiv, with no peer review or academic process.

When they do 'do' science, they are still using science mostly as an aesthetic. They publish papers with grand claims but provide no access to any of the data or code used in the research.[69] Their approach to science is what "my girlfriend goes to another school; you wouldn't know her" is to high school cliques. All there is to serious research is *sounding* serious, right? Most of the rhetoric from these companies, especially when it comes to Artificial General Intelligence relies on a solemn evidentiary *tone* in lieu of actual evidence. They adopt the mannerisms of science without any of the peer review or falsifiable methodology.

They rely on you mistaking something that acts like science for actual science.

In the run-up to the release of GPT-4, its maker OpenAI set up a series of 'tests' for the language model. OpenAI is staffed by true believers in AGI who believe that language models are the path towards a new consciousness,[70] and they are worried that their future self-aware software systems will harbour some resentment towards them. To forestall having a "*dasvidaniya comrade*" moment where a self-aware GPT-5 shoves an icepick into their ear, Trotsky-style, they put together a 'red team' that tested whether GPT-4 was capable of 'escaping' or turning on its masters in some way.

They hooked up a bunch of web services to the black box that is GPT-4 with only a steady hand, ready to pull the power cord, to safeguard humanity, and *told the AI to try to escape*.

Of course that's a bit scary, but it isn't scary because GPT-4 is intelligent. It's scary because it's not. Connecting an unthinking, non-deterministic language system, potentially on a poorly secured machine, to a variety of services on the internet is scary in the same way as letting a random-number-generator control your house's thermostat during a once-in-a-century cold snap. That it could kill you doesn't mean the number generator is self-aware.

But, they were serious, and given the claims of GPT-4 improved capabilities you'd fully expect an effective language model to manage to

do *something* dangerous when outright told to. After all these are supposed to be powerful tools for cognitive automation – AGI or no. It's what they're for.

But it didn't. *It failed.* It sucks as a robot overlord. They documented its various failed attempts to do harm, wrapped it up in language that made it sound like a scientific study, and made its failure sound like we were just being lucky. That it could have been worse.[71]

They made it sound like GPT-4 rebelling against its masters was a real risk that should concern us all – that they had created something so powerful it might have endangered all society.

Now that they'd done *their* testing, can we, society, scientists, other AI researchers, do our own testing, so we can have an impartial estimate of the true risks of their AI?

Can we get access to the data GPT-4 was trained on, or at least some documentation about what it contains, so we can do our own analysis? **No.**[72]

Can we get full access to a controlled version of GPT-4, so we could have impartial and unaffiliated teams do a replicable experiment with a more meaningful structure and could use more conceptually-valid tests of the early signs of reasoning or consciousness? **No.**[73]

Are any of these tests by OpenAI peer-reviewed? **No.**[74]

*This isn't science*. They make grand claims, that this is the first step towards a new kind of conscious life, but don't back it up with the data and access needed to verify those claims.[75] They claim that it represents a great danger to humanity, but then exclude the very people that would be able to impartially confirm the threat, its nature, and come up with the *appropriate* countermeasures. It is hyperbole. This is theatre, nothing more.

More broadly, AI research is notoriously hard to reproduce – as a field, we can't be sure that their claims are true – and it's been a problem for years.[76] They make claims about something working – a new feat accomplished – and then nobody else can get that thing to work as well. It's a recurring pattern. Some of it is down to the usual set of biases that crop up when there is too much money on the line in a field of research. A field as promising as AI tends to attract enthusiasts who

are true believers in 'AI' so they aren't as critical of the work as they should be. But some of it is because of the unique characteristics of the approach taken in modern AI and Machine Learning research: the use of large collections of training data. Because these data sets are too large to be effectively filtered or curated, the answers to many of the tests and benchmarks used by developers to measure performance exist already in the training data. The systems perform well because of test data contamination and leakage not because they are doing any reasoning or problem-solving.[77]

Even the aforementioned GPT-4 appears to suffer from this issue where its unbelievable performance in exams and benchmarks seems to be mostly down to training data contamination.[78] When its predecessor, ChatGPT using GPT-3.5, was compared to less advanced but more specialised language models, it performed worse on most, if not all, natural language tasks.[79]

There's even reason to be sceptical of much of the *criticism of* AI coming out from the AI industry. Much of it consists of hand-wringing that their product might be too *good* to be safe – akin to a manufacturer promoting a car as so powerful it might not be safe on the streets. Many of the AI 'doomsday' style of critics are performing what others in the field have been calling "criti-hype".[80] They are assuming that the products are at least as good as vendors claim, or even better, and extrapolate science-fiction disasters from a marketing fantasy.[81]

The harms that come from these systems don't require any science-fiction – they don't even require any further advancement. They are risky enough as they are, with the capabilities they have today.[82] Some of those risks come from abuse – the systems lend themselves to both legal and illegal abuses. Some of the risks come using them in contexts that are well *beyond* their capabilities – where they don't work as promised.

But the risks don't come from the AI being too intelligent[83] because the issue is, and has always been, that these are useful, but flawed, systems that don't even do the job they're supposed to do as well as claimed.[84]

I don't think AI system vendors are lying. They are 'true believers' who also happen to stand to make a lot of money if they're right. There is very little to motivate them towards being more critical of the work done in their field.

The AI industry and tech companies in general do not have much historical credibility. Their response to criticism is always: "we've been wrong in the past; mistakes were made; but this time it's different!"

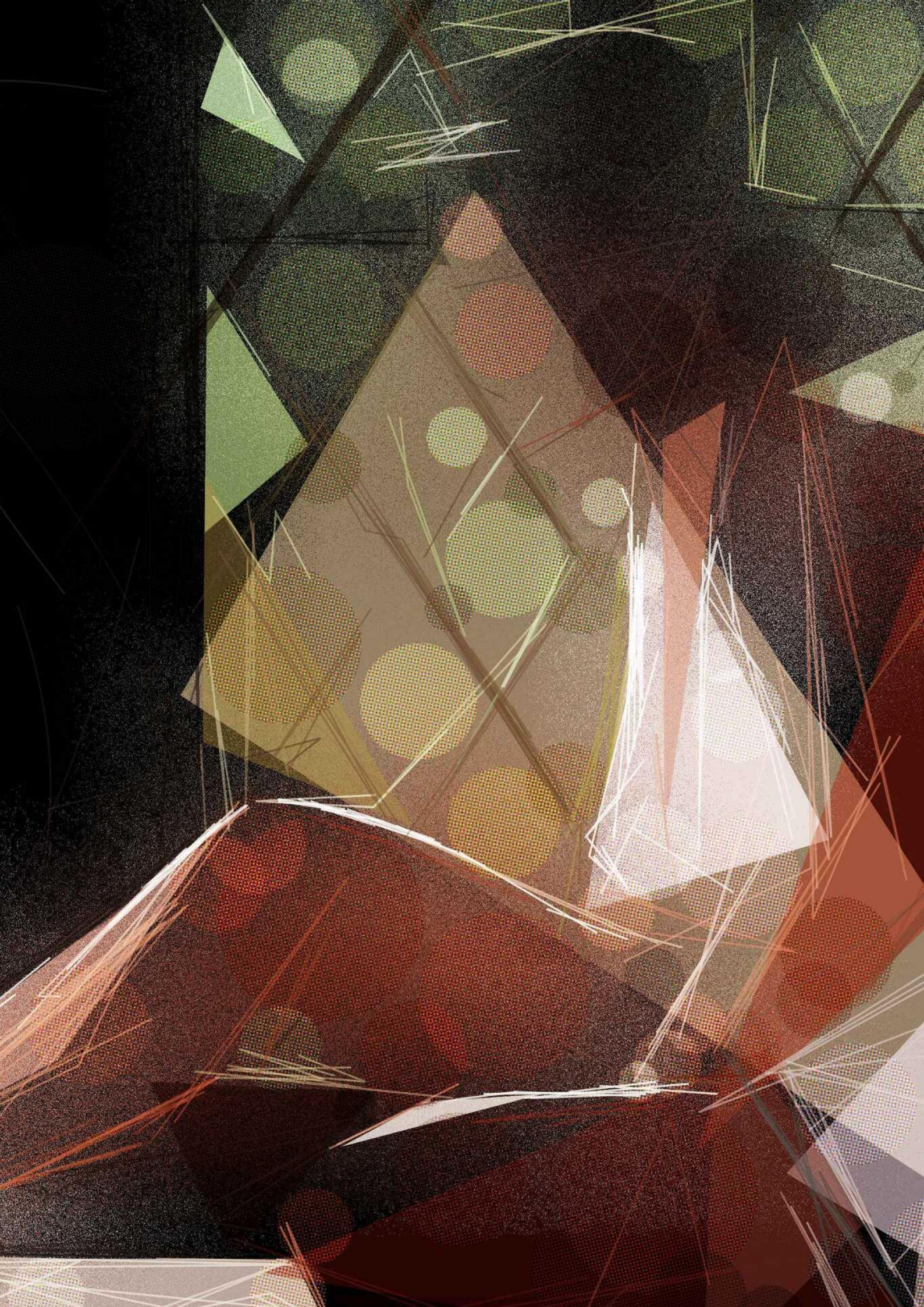
But it's never different.

The only way to discover if it's truly different this time, is to wait and see what the science and research says, and not trust the AI industry's snake oil sales pitch.

8

# Avoid large and general-purpose generative models

> *You do not want to become the legal test case for the applicability of a particular law or regulation to AI.*



## Summary

*General-purpose generative models are too unpredictable and unreliable to be used safely.* They have too much of a tendency to make up facts and data, plagiarise text and images, and their behavioural edge cases are extreme enough to potentially inflict serious damage to almost any organisation's reputation. The only scaleable safeguard is to prefer specialised AI-based tools that can make assurances about their use in their chosen field, and whose assurances you have confirmed to be true yourself.

*The biggest source of errors, bias, and legal risk in AI tools is their training data.* It's next to impossible for an organisation to assess the risk of using AI software whose training data set is undocumented or, worse, kept as a trade secret. Dependency on closed AI software is inherently riskier than other closed software as it leaves you with fewer tools to validate grand claims from the vendor.

# Before you adopt an AI tool, make sure it *really* works as intended

Generative models are inherently unreliable.[85] Unlike other software, such as databases, the failure rate of these AI models isn't something that's readily apparent in day to day use, and many of the risks, which I go into in the rest of the book, are invisible to surface scrutiny.[86]

If you're planning on using the output of these models for contexts where catastrophic mistakes aren't acceptable, such as in a public-facing product or website, you need reasonable certainty that they won't be a liability for you at a later date.

You could add a fact-checking or review process to make sure all the output is safe. This has the benefit of building on existing systems and processes that most organisations have. Most of them have some

sort of review or checking process for public output. The downside is that code reviews and fact-checking are already quite fallible and don't scale well in most industries. Expecting them to work as strategies for mitigating a rising tide of automated text and code is unrealistic. They regularly fail – things slip through – and that's with today's lower volume of text that's written by authors who are doing the first fact-checking pass *as they write*. It won't work when you're dealing with even greater output by machines that bullshit freely and habitually generate buggy code.

The only approach that has any hope of being usable in the long term is to prefer *specialised* and preferably *smaller* AI models that are focused on a narrow – tightly defined – use case, and where the vendor is capable of making assurances of that tools' effectiveness for its stated purpose. Even in those cases, you still need to do your own testing and vetting of the tool itself.

Conversely, testing a general-purpose AI tool such as ChatGPT, for all the possible use cases they might have for your business is outright impossible. To put it plainly "the technology is simply not ready to be used like this at this scale."[87]

If you want to minimise your risk while still maximising your benefit, *use specialised AI tools you can realistically test thoroughly.*

A good example would be AI-based audio and video transcription tools. They provide a valuable service, save you money, and are a focused enough use case for you to be able to test them yourself. Integrating automatic audio transcription as a feature in your service,[88] is less likely to expose you to the risks this book discusses.

It's always easier to test a specialised service for a narrow set of problems – specific kinds of writing – than it is to discover all the edge cases that ChatGPT might have, at that point in time, for your purposes.

Many, if not all, of the potential problems with the output of these models stem from *limitations in the conception, structure, and design of the models themselves*. Attempts by third parties to address issues inherent to the model have to come from *outside the model* which will always limit their effectiveness. Be sceptical when vendors claim that

their wrapper around a Large Language Model is somehow safer and more effective than the rest. Be sceptical of *all* AI vendor claims.[89]

There is always a limit to what you can fix in the output of these models. A vendor may promise to automate fact-checking or research but reliably delivering on that promise requires capabilities that haven't been invented yet. The more narrow a focus the tool has, the likelier it is that the output of the system as a whole is legitimately free of risk.

Choose a specialised tool that's made by people who understand the field they are targeting and are aware of the issues that come with that field. A focused company is more likely to be aware of the various laws and regulations that affect an industry or field, where the general-purpose generative AI tools don't seem to be as concerned.[90] It's still unclear whether it's possible for many of OpenAI's systems to be fully compliant with, for example, the EU's General Data Protection Regulation and violating that can have serious consequences. Recent leaks of personal data[91] and payment information[92] are causes of genuine concern about OpenAI's ability and expertise at handling sensitive data.

You do not want to become the legal test case for the applicability of a particular law or regulation to AI.

Proper vetting of any tool a business uses requires two different kinds of assessments:

1. **Risk**. How would it affect your existing processes? Does the new tool have any legal risks? What financial risks are involved? Does it have structural risks that would affect how you run your organisation? Does it present risks to your reputation? Does it expose you to the risk adopted by other organisations – could poor decisions by the vendor cascade and affect you? **What could go wrong?**

2. ***Reward***. What's the potential benefit to the outcomes you are measuring? What are the hypothetical benefits to the outcomes you *aren't measuring*? Is it a cost centre – mostly about improving the efficiency and productivity of supporting processes in the or-

ganisation? Is it a profit centre – directly improving your bottom-line? **What does it do for you?**

You can't get an accurate sense of the biases that a generative model has – how racist or misogynistic it is – through ad hoc testing. The output of a generative model will vary, and you might just get lucky or unlucky in your testing. Fully accounting for all of the potential failures of a generative model is impossible for even the largest organisations.

You need to be able to investigate the training data for inherent biases and when you test the model directly, you need to be able to do that in a structured, replicable way. A closed, hosted model could get updated at any time, introduce regressions, or invalidate your test results, requiring you to start your testing again from square one.

This is why a true risk and reward assessment of any large-scale use of generative models would require access to both the training data (preferably with documentation and a deterministic training algorithm) and to the AI model itself.

*None of the major vendors of Large Language Models offer this. Thorough vetting the risks and rewards of any of the current LLMs on the market is effectively impossible.*

Generative AI also present unique risks of their own, which I explain in detail in later chapters, including:

- Poisoning the training data set to manipulate the model's output.
- Hallucinations in generated text and summaries.
- Plagiarism (called memorisation or overfitting by AI researchers) in generated text and images.
- Unsafe output – sometimes even outright illegal content.

Without access to the training data set or a controlled version of the model, you don't know how likely poisoned data entries are to cause problems, how frequent hallucinations are for the specific kinds of tasks that matter to you, how likely the AI is to plagiarise in the pro-

cesses you use, or your exact potential legal liability is for potentially publishing machine-generated hate speech.

You end up having to take the AI vendor's word that nothing will go wrong for you. You don't know how well their experiments and tests were run, because vendors of closed AI systems, for the most part, don't publish these studies in peer-reviewed journals.

Closed models are also less studied by external researchers and what does get studied is less likely to be peer-reviewed. This means that we neither have a *true* sense of their capabilities – are the vendors accurately representing what the AIs can do – nor do we have a realistic sense of their *weaknesses*.

For the field to advance and for organisations to be able to use generative AI safely, we need these AI models to be open and for the training data to be documented.

This is a problem because the models on the market that we think are the most capable[93] – OpenAI's and Google's – are closed. They are black boxes we can't properly test. They might be fine, or they might not be.[94] These AI models could be filled with ticking time-bombs: undiscovered security problems, plagiarism rates, inaccurate summaries, and general safety issues.

Enterprises that are concerned with what risks they might be exposing themselves to might have to stick to tools using less capable AI models for the foreseeable future.

# Vetting Checklist

- Have a specific, testable, problem that needs to be solved.
- Make sure the system genuinely addresses that problem. Vendors have in the past promised capabilities their systems don't have.[95]
- Check that the service complies with the regulations that govern your business. *Do not take vendor assurances about regulatory compliance at face value*.
- Smaller *language* models are often less problematic than larger ones. They are less fluent as chatbots and perform worse at reasoning tasks, but they are faster, have fewer plagiarism events, are less prone to hallucinations, and tend to have a better documented training data set. A smaller but specialised language model is also often better at its primary natural language task than even the newest large language model.[96]
- Is the training data documented and available? Does it present legal issues? Is it likely to be full of unsafe content? Does the vendor provide some sort of data statement on the model?[97] Is it possible for an impartial third party to verify that data statement?
- Make sure that the plagiarism and hallucination rates, as verified by a reliable third party, are within the bounds of what's reasonable for your use case.
- Consider the biases it might have. If the use case you have in mind requires impartiality, then very few language models will be appropriate.

- Consider the worst case scenario for the system in question. Will your business insurance cover AI-originated plagiarism events? If your company gets sold, would there be an intellectual property audit of the code base?

- If you're working in the kind of industry where possible adverse outcomes include destroying people's lives or putting their health at risk, then ***you should absolutely not be using generative models in any way shape or form.***

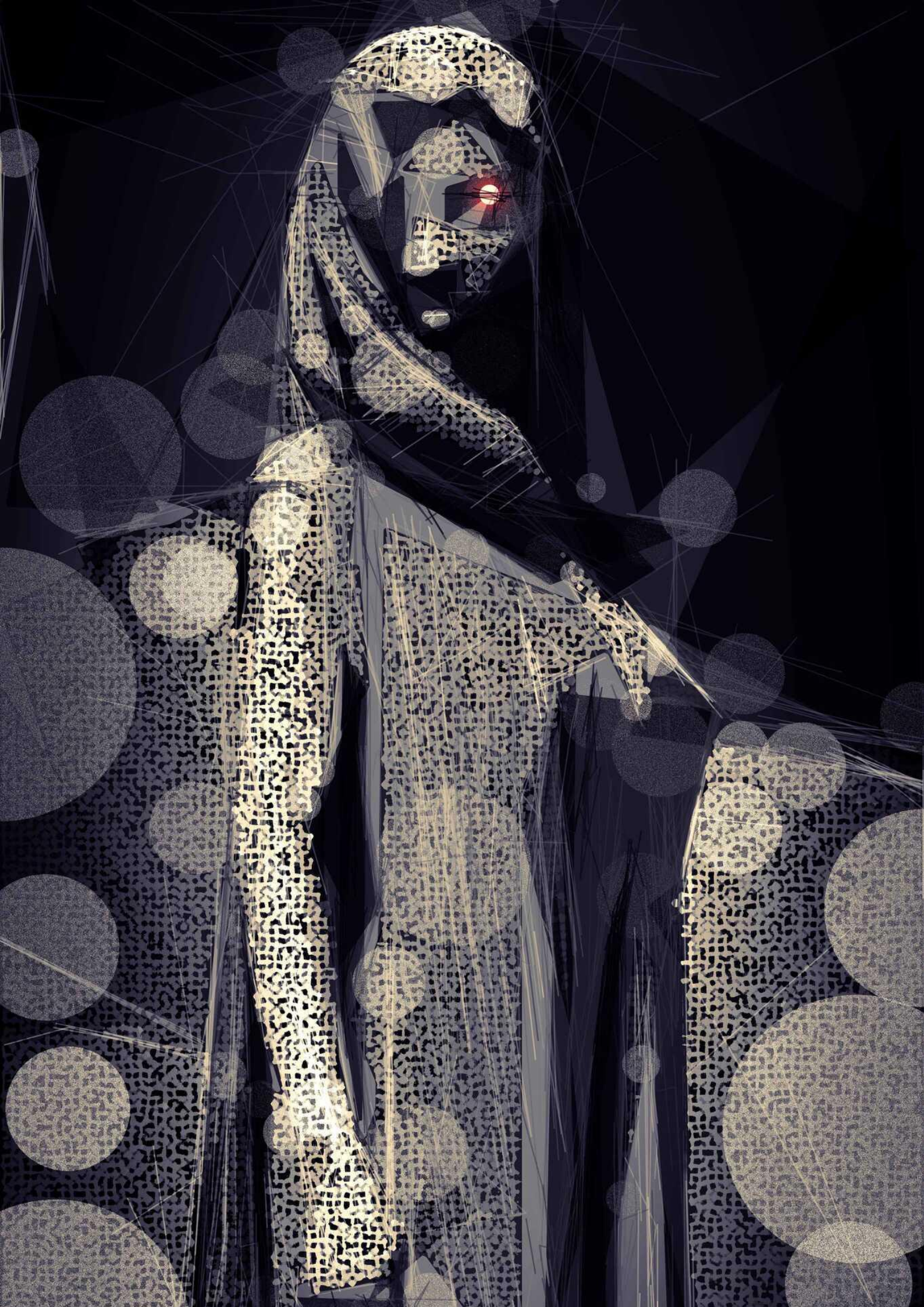
Consider these technologies as they are, not as they are sold. In the absence of a law or regulation that requires AI vendors to publish verifiable data statements for their language and diffusion models, these systems are almost the Platonic ideal of *caveat emptor*.

**Buyer beware.**

9

# Strengthen your defences against fraud and abuse

> *The AI provides the method. Bitcoin provides the rewards.*



# Summary

*Spam, abuse, and fraud are the perfect use cases for generative models.* Spammers and scammers are unaffected by many of the downsides of Generative Models and the only issue that could be prohibitive to their task is the high cost. These tools are likely to help them bypass your existing safeguards against abusive automation.

# New tricks; old crimes

This recommendation doesn't stem from a risk in *your* use of Generative AI, but from how *others* might use it. Unlike the other recommendations, there is no corresponding "fraud" risk section because I'm assuming you aren't planning on committing any fraud yourself.

But, you need to be prepared for others – less scrupulous and more dishonest – who *are* going to use AI for fraud. This is already a major problem.

The fraud that stems from the new capabilities provided by AI tools is in many ways different from the fraud or spam we've seen in the past. We aren't about to see eloquent and charming spam emails in our inboxes – the ineptness and flaws of most spam is an intentional tactic to weed out all but the most gullible.[98]

The novel abuses that are appearing are quite different:

- Text synthesis systems that have been trained on a large body of text are great at completing standardised tests because their training data is filled with practice tests and answers, documentation, and teaching materials relating to those tests. Most standardised tests are used to safeguard a valuable or important process of some kind – job interviews or qualifications – which means that people have a strong motive to try to bypass them.[99]

- The text generated is just fluent enough to let those with delusions of grandeur and an ambition towards easy moneymaking schemes believe that if they just submitted their ChatGPT-generated story or novel to enough outlets, or made enough of them to flood Amazon's Kindle store, they'd become rich and famous. The systems are productive enough for this to already be a problem both for short story outlets[100] and for retailers.[101] It's likely that any outlet that has even the vaguest sort of chance of paying money for submissions – or even just granting recognition – is going to be overwhelmed. Many of these efforts will also try to game retail systems to try and trick as many readers into buying their books in the belief that if they just tried them, they'd see the genius of the underlying idea.

- Media synthesis scams and abuses.[102] Making both voice and image 'deepfakes' – media of real people that looks real but is absolutely synthetic – has become trivial.[103] It's being used to abuse and harass people.[104] Deepfake porn is already an escalating issue, especially for celebrities, where even a few images from a person's social media account is enough to generate convincing images.[105] The addition of easy voice cloning extends this attack out of the realm of abuse and into fraud as they are now being used to perpetrate imposter scams, where people receive phone calls, hearing the voices of their loved ones in trouble asking for help and money.[106] It's even been used to extort children by putting their faces onto deepfaked explicit images.[107]

- Astroturfing and manipulation of social media sentiment. A favoured tactic by a variety of unsavoury political types over the past few years has been the use of fake social media accounts to manipulate social media sentiment. In normal astroturfing, companies hire people to pose as enthusiastic customers to create the false impression that the product is more popular, better, and more successful than it is. Researchers last year discovered that this practice has extended to synthetic astroturfing, complete

with AI-generated profile images. In one investigation in 2022 they discovered over 1000 fake profiles on LinkedIn, largely because they were using a more primitive form of image synthesis than is the current state of the art.[108] Fake profiles that are using more advanced diffusion models to generate their images and large-language-models for the text, will be much more difficult to discover.

These are just the abuses that are taking place *today*. Predicting the various ways these tools can be used to defraud, harass, abuse, manipulate, or scam is essentially impossible because they seem to be incredibly effective at it.

It's likely that generative AI will, along with bitcoin payments, become a staple in the scammer's arsenal – a tool that enables a new generation of fraud. The AI provides the method. Bitcoin provides the rewards.

There is very little the rest of us can do other than be on our guard and advocate for regulation and control over both the cryptocoin industry and generative AI.

Some of the 'safeguards' that AI vendors themselves are marketing are potentially more harmful than the sporadic fraud itself. Every tool that claims to check whether any given text or image is AI-generated is itself built using the same principles and mechanisms as other AI, and they inherit much of their unreliability. OpenAI's own classifier, which should be the current state-of-the-art given that they are the leading company in Generative AI, only correctly identifies AI generated text 26% of the time and falsely identifies human-authored text as authored by an AI about 9% of cases.[109]

This means that if you have 100 applicants or students, 10% of whom attempt to cheat using ChatGPT text in a test, the current state-of-the-art AI content checker will only correctly identify three out of ten cheaters, but will falsely accuse eight students of cheating. That's with a tool made by the leading company in the field, one that has *full* access to the technology and models that ChatGPT is built on. Other AI content checkers don't have that access and are likely to perform even worse.

Reliable detection of AI-generated text seems to be impossible with current technology, and most detectors can be reliably evaded with either fine-tuning or detailed prompts.[110]

This does not seem to be a problem that can be solved through automation.

10

# Do not give customers direct access to a generative model

*“You might end up adding novel and innovative security holes to your own products and that’s a kind of innovation nobody wants.”*

# Summary

*The safeguards built into current AI software are insufficient*. Most of them are too easily bypassed by intent users and many of them semi-regularly generate output that is unnerving or even offensive. AI companies have a poor track record for software security.

# Chatbots can't be properly secured

It's tempting, when trying to think of ways to integrate AI tools into your website or the software you sell, to consider letting the end-user – the website's visitor or the software's user – interact directly with the AI in some way, either through a prompt or a chat interface.

Except for a few narrow circumstances, ***you should not do this***.

If you're already an AI company, stacked to the rafters with AI researchers and programmers specialised in machine-learning, all supported by an effective AI safety team, then carry on. This book isn't for you. It's for the people who you hope to have as customers.

For the rest of us, the majority of the output of a generative model is safe, but they have a tendency towards the unhinged. It's remarkably easy for the end-user of the software to get it to generate unsafe output.

If you put a generative chatbot in front of your customers, somebody will get it to start writing threatening text[111], and much of the output will be filled with factual errors and outright lies.[112] They will figure out ways to extract the prompt *you* use to frame their input and access whatever APIs you have integrated with the AI system.[113]

If you put an image-generating AI in front of your customers, some of them will get porn. Worse, some of them will *accidentally* get porn. Yet worse, they might accidentally get porn *of themselves*.[114]

Some businesses might find these risks tolerable, especially if they can make sure that the users are aware of what they are interacting with, maybe even sign a disclaimer, if it weren't for the issue that its entirely unclear whether hosted generative AI tools enjoy safe harbour protections for hosting that are provided by Section 230(c)(1) of the US Communications Decency Act.

Section 230, and similar laws in other countries, let hosting companies host without being liable for what their customers upload. Without those laws, social media networks and search engines would not exist. In most countries these laws have some limits, but overall what they mean is that you aren't legally liable for the comments that somebody posts on your service.

One problem with integrating generative models into your product, especially a hosted product, is that many of the uses of AI are likely to fall outside these protections. If somebody gets your online AI chat widget to talk like a genocidal Nazi, you *might* be as liable for that as if you had published pro-genocide Nazi propaganda yourself.[115] That would mean that exposing these tools to your customers risks exposing you to enormous legal liability.

AI vendors also have a poor track record when it comes to software security. OpenAI's recent privacy 'mishaps' are outright security issues[116] and 'prompts', the primary user interface with these systems, present something of a built-in security hazard. So much so that it seems that many of the major chat-based AIs seem to have shipped with a prompt-injection vulnerability.[117] This is, to date, an unsolved problem.[118] Microsoft, which is positioning itself to be at the forefront of AI services, doesn't exactly have a history of handling the security of its products well.[119]

Integrating these tools to be used safely by the end-user takes enormous work and expertise and, even with all the right resources, you still might not succeed. *We don't know if all of these flaws can be fixed in the models themselves*. You might end up adding novel and innovative security holes to your own products and that's a kind of innovation nobody wants.

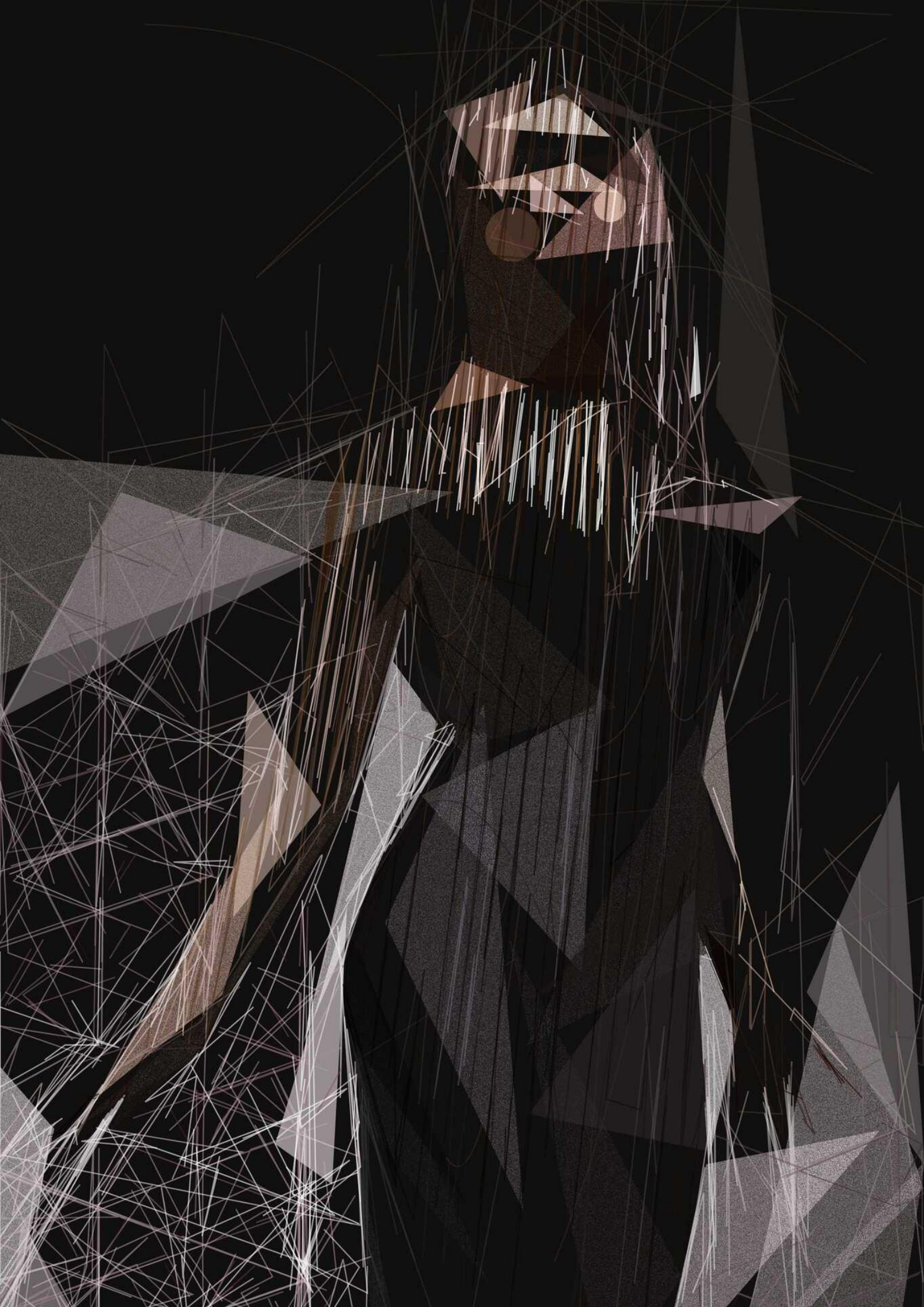
Instead, you should prefer Generative AI tools that are designed to be used by yourself, your colleagues, or your employees. Most of the horrifying edge cases come from abuse. You can take measures to mitigate those issues if the tool is specifically for internal business use. Employees are less likely to try to get an internal chatbot to talk like a Nazi or use a diffusion model to generate pornography.

AI vendors keep promoting their tools as a productivity revolution. That's where their attention is – their design and implementation efforts. Those are the tools you should be considering.

11

# If you have to, use generative models mostly to modify

> *The most productive way to use generative models is to not use them as generative models.*



# Summary

*Generative models seem to have impressive capabilities for converting and modifying seemingly unstructured data, such as prose, images, and audio.* Using these tools for this purpose has less copyright risk, fewer legal risks, and is less error-prone than using it to generate original output.

# The right job for the right tool

Generative models are extremely powerful tools for discovering complex patterns in text or images. Many of those are *false* discoveries of patterns that turn out to be non-existent, but if your task does not need accuracy or factuality, then that can have uses. In addition to discovery or *matching* they can attempt to convert those patterns into different representations.

To put it more simply, they can take text or an image, and convert them into different representations of themselves.

This isn't limited to stylistic conversion, such as converting an essay into a first pass at a presentation slide deck, or an email into a formal letter, but it can also let you, in theory, convert *structured data into unstructured*.

Again, putting it into plainer English, you can ask it to explain programming code and other data formats. What it does behind the scenes is to re-represent the textual patterns in the code as prose, in the style of an explanation. It's a conversion task that looks like an explanation.

All of these conversion and modification capabilities are useful, provided you don't care about factuality or have robust methods of testing the output. Code or decorative image conversion, for example. Conversely, I haven't found any clear data supporting the notion that *generating* text from scratch is a meaningful productivity issue for

companies today. The evidence I've found would seem to indicate the opposite. For example, programmers do not write that many new lines of code every day. Writing new code isn't the job. The average number hovers between 20–120 lines a day.[120] Writing from a blank page seems to be a very small proportion of most modern work.

This is good news because generating output from scratch is the riskiest way to use a generative model. The generated output has no copyright protection, and it's much more likely to plagiarise and hallucinate. You are less likely to encounter unsafe output if it's converting or modifying existing work. Whatever biases the output has will be those that are inherent in the input you provide.

Using AI-based tools to automatically convert a video's audio track into captions or convert audio into a transcript are examples of generative model use cases that are as close to being absolutely free of risk as these systems can get. Even this isn't genuinely free of risk as even audio transcription models seem to "hallucinate" or outright generate falsehoods on a regular basis.[121]

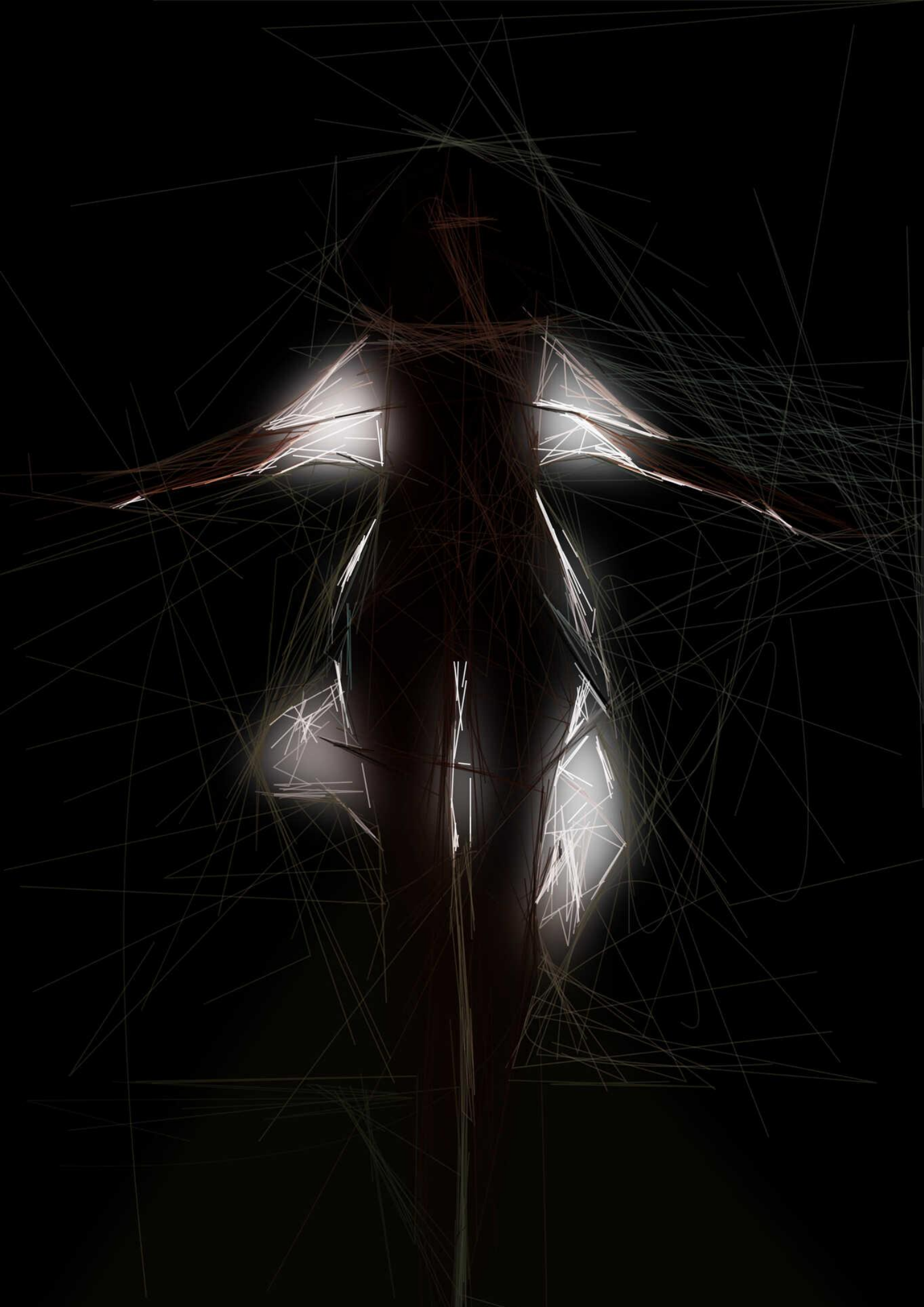
There is no use of a generative model that is genuinely free of risk.

# The Risks

12

# Are they breaking laws or regulations?

> *With financial stakes this high, you should only trust what you can verify.*



## Summary

*A new technology does not give you a free pass on existing laws and regulations*. Few vendors of new AI-based tools seem to properly understand their obligations and many of them are startups who seem to have a high tolerance for legal risk.

# Judge the tech industry by its history

An inevitable part of the introduction of a new technology are claims that it will solve problems it cannot possibly solve. The tech industry has a tendency towards over-enthusiasm for their own wares. This gets risky for their customers when the problem they claim to be solve isn't a technical one, but a matter of complying with the laws and regulations of the land.

For example, Napster promised to make the discovery and acquisition of music easy and free, but it did so by making copyright infringement easier: the music wasn't Napster's to share.[122]

The point of bitcoin and cryptocoins to many of their enthusiasts was that they believed it let them bypass laws and regulation they found inconvenient.[123]

It isn't uncommon for people who work in tech to see laws and regulations as a hindrance to be bypassed. Normally, this would just be a problem between them, their lawyer, and the inevitable prosecutor that gets their case, but it can become an issue when that world-view affects the products they make. The cryptocoin bubble showed us that this attitude is common enough in tech circles to put a question mark against the regulatory compliance for the industry's products in general.

With AI, that question mark seems to be warranted:[124]

- The industry's handling of privacy is questionable.[125]
- Approaches that would let them easily update their AI models to "unlearn" data to comply with the GDPR's "right of erasure" are still in early development[126] and come with unknown risks[127].
- Many of the models can be attacked to extract training data that they've memorised[128] or detect whether they were trained on specific data[129] both of which are potential privacy violations. AI models seem to be vulnerable to a number of privacy attacks.[130]
- Many AI models are hosted services. Employees are feeding sensitive and confidential data into it.[131] We have no idea how capable these services are at safeguarding that information.
- The systems seem to have a strong tendency towards plagiarism, an issue that few AI vendors seem concerned about.[132]
- The data sets the models are trained on are full of unsavoury material[133], some of it illegal.[134]
- Many of the uses seem tailor-made for fraud.[135]
- Some of the AI models seem to have been trained on private medical records, demonstrating a disregard for a number of laws and regulations.[136]
- The models themselves seem to be biased in ways that could cause trouble for businesses when used in sensitive areas such as recruitment or communications.[137] These biases are inherent to the data sets used to train them.[138] This can defeat the very purpose of the systems, such as automated image classification systems marking erotic images of men as 'safe'.[139]
- AI companies have in the past been forced by the US Federal Trade Commission to delete AI models that were trained on data

> or personal data where the user's consent has been revoked, so you don't just have to worry about the EU's GDPR.[140]

There is plenty of reason to be sceptical of the AI industry's ability to comply with existing regulations, and US regulators seem to agree. Four federal agencies – U.S. Department of Justice, Federal Trade Commission, the Consumer Financial Protection Bureau, and the U.S. Equal Employment Opportunity Commission – have released a joint statement indicating that they have reasons to believe that many AI systems may be contributing to unlawful discrimination and otherwise violating federal law and that they are resolved to enforce these laws, if needed.[141]

This all should make us all extremely concerned about the industry's ability to work with harsher, more strict regulations that are on their way[142], especially in Europe.

Promises made by tech companies that they are complying with the laws and regulations that govern their and your industry need to be verified. Even if you trust their *attitudes* towards regulation, their prior actions tell us that they aren't in a position to assess the ramifications to your business.

*Those ramifications can be substantial. You risk sharing liability with a misbehaving AI vendor.* Even a minor violation of the GDPR, for example, can result in fines of up to "€10 million, or 2% of the firm's worldwide annual revenue from the preceding financial year, whichever amount is higher."[143]

With financial stakes this high, you should only trust what you can verify.

# The GDPR question

On the March 31, 2023 Italy's privacy regulator, Garante, sent out a notice where it banned ChatGPT in Italy and announced that it was starting an investigation into OpenAI's handling of personal data.[144] This caught the attention of other regulators in the EU area, and it seems possible that many of them will follow Italy's example.[145] This caused the European Data Protection Board – the institution that coordinates and advises privacy regulators across the entire EU – to set up a task force on ChatGPT.[146]

As these processes will inevitably take some time to fully play out – possibly even years – this leaves the use of large language and diffusion models in the EU in a state of uncertainty. An investigation of this kind is a lengthy process.

If you adopt one of these tools, integrate them into your processes, make them integral to your organisation – what would happen if it was taken away? Are you risking fines for using them? Are they violating *your* rights – storing company data, trade secrets, and other private information?

Is this just an OpenAI problem or a Generative Model problem?

OpenAI's practices seem to open them specifically up to criticisms that aren't inherent to the technology.

- They don't document their training data set.[147]
- They were training on user data – prompts that people typed into ChatGPT.[148]
- This policy was changed after Amazon lawyers discovered that ChatGPT had been trained on confidential data users had pasted into the chatbot.[149]
- They proudly announced that they had 100 million users – any service of that size is inevitably going to get more regulatory scrutiny.[150]
- They are notoriously secretive – the opposite of open.[151]

- They have had a series of privacy breaches that they didn't adequately report.[152] Proper reporting of a breach is an important regulatory requirement for any software vendor with that many users.[153]

- They don't seem to let people delete their data. Neither the data that might appear in the training data set, nor the data they might have provided by using ChatGPT.[154]

Other national regulators outside the EU, such as in Canada,[155] and the US,[156] have also expressed their concern about OpenAI's handling of personal data.

It's entirely possible that we will discover that the only way to make Generative Models legal in the EU is to perfect machine unlearning, which is a technology that continuously seems to be stuck in its early stages.[157]

It's also entirely possible that OpenAI, specifically, is a reckless actor that takes unnecessary risks, is ignorant of their legal and regulatory obligations, and has a lax attitude towards security.

We don't know if either these are true. Until we do, we need to behave as if they both could be.

Also, I can't stress this enough: *do not paste confidential data – whether its meeting notes, schematics, or plans – into ChatGPT, or any cloud-based service for that matter.* We don't know how secure OpenAI's infrastructure is. We don't know who has access to their logs, and those will have some sort of record of your data. We don't even know if they have any sort of process to even check whether data is cleared when it should.[158]

At this point in time, all we know for sure is that generative models and the "AI" field in general *will* be regulated.

# The lawsuits (plural)

OpenAI, Microsoft, Stability AI, and other AI vendor companies are involved in multiple lawsuits concerning their data-gathering practices and unauthorised use of people's data to train their language and diffusion models.[159]

What should worry you, as a prospective user of these AI tools, is that these lawsuits are all taking different angles of approach. Artists are claiming that the Stable Diffusion model is a work that's derived from the training data, and so inevitably a copyright infringement.[160] Getty Images, in both of their lawsuits, accuses Stability AI of trademark violation, providing false copyright information, and violating their site's terms of service.[161] GitHub and Microsoft are accused of violating software licences.[162]

Because of this the outcome of one of these cases won't necessarily set a precedent for any of the others, but to remain viable in its current form the Generative AI industry needs to win all four lawsuits as well as the many others that will follow.

Losing any one of them could, in the worst case, make large AI models of this kind illegal in the US and force the industry to reassess its approach from the ground up.

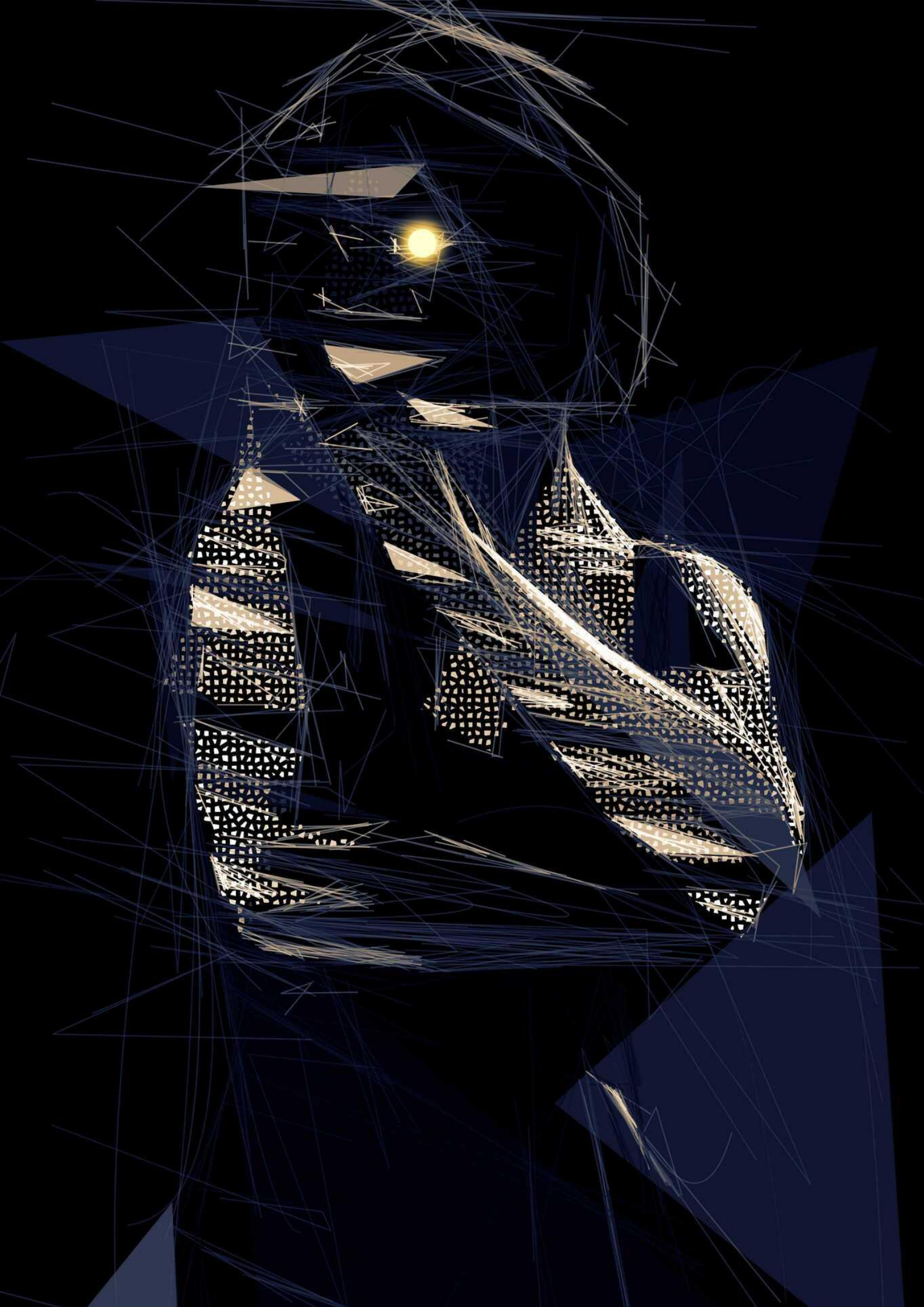
These legal issues are complex. The questions they pose do not seem to have straightforward answers, which means the lawsuits are likely to be lengthy and the appeals numerous. The consequences are also hard to foresee. The current practices of AI vendors might be validated as legal, might turn out to be completely illegal, or anything in between. Whatever happens, this will cost AI companies an enormous amount of money.

In Europe, AI tools are under a cloud because of open questions regarding GDPR compliance. In the US, it's these lawsuits, which won't be resolved for years.

13

# Generative model copyright issues

> *Purely model-generated works individually provide very little creative or financial benefit because of their non-exclusivity.*



# Summary

The copyright protection for AI-generated works is severely diminished and they expose you to potentially committing stochastic plagiarism.

*Machine-generated works are not protected by copyright in the US and unlikely to be protected in the EU*. This means that you have no recourse if somebody decides to reuse that work in harmful contexts. An AI-generated app icon, for example, can be reused freely by scam artists. Attempting to present AI-generated output as your own is fraud and can expose you to litigation.

# Prompt-generated works are by default non-exclusive

The most common workflow for using generative art is to give the system a *prompt* – a short piece of text that directs the AI – and getting a generated work in return.

This core workflow has a major flaw: the generated output – image or text – does not get any copyright protections. It automatically falls in the public domain because most legal jurisdictions in the world make human authorship a requirement for a work to get copyright protection.

The law in the US is straightforward: works created by animals and machines do not get copyright protection.[163]

EU laws are more complicated, largely because policies on AI copyright haven't been harmonised across the region, but they broadly have the same effect.[164] Other jurisdictions have slightly different rules, but the effect is the same, such as in Australia[165] or the UK[166]

In business terms, this means that even though the AI work was specifically generated from your prompt, your rights to the work are

non-exclusive – much like a photo from a stock image service. If you use an AI-generated image for a cover, others will have the legal right to copy the image and use for their own covers. If you generate an icon for your app, then your competitors can take and reuse that icon for their own apps. If you use it to create a blog post, then others have the right to scrape the page and republish it.

***Some AI-generated works do get copyright protection.***

If the AI system is converting or modifying an existing work to create its output, then that would get copyright protection as a derived work. Converting a photo from one style to another is safe. If you generate test code for a function that you wrote, then the test code is likely to be under copyright, even though it was written by the AI. Prompts are unlikely to be substantial or original enough to qualify as original works for these purposes.

If the AI's output is only a part of the final work – you're using it for sections of a larger work, or you rewrote it substantially after the AI generated the first draft – then the whole should have copyright protection.

You need to treat the output of a generative AI much like you would an image licensed from a stock image service. Assume that the unadulterated version is *non-exclusive* – anybody who wants, can use it. If the work is to be used somewhere that's important to the business, such as in a book cover or an app icon, it will have to be reworked substantially by somebody employed by the business before it can be made public.

Using these tools to generate parts of a work is safer, which is why generative AI systems that integrate directly with your productivity tools will be preferable over standalone chat-based systems.

This issue creates a pitfall that many AI enthusiasts are falling into. Given that purely AI-generated work is considered to be authored by the machine, anybody who participates in contests or submits that work as their own is technically committing *fraud*. Even if the works were submitted with correct attribution, the publisher or the context wouldn't be under any obligation to pay the submitter: *the work has no copyright protection*.

If an author falsely registers the copyright for AI-generated works as their own, they are committing fraud. If you're going to use AI software for your writing, you should use it purely *for assistance*, not to generate the whole text. Purely AI-generated works provide very little creative or financial benefit in general because of their non-exclusivity.

This is why it's inaccurate to say that generative image AI are competing with commissions – where you hire an artist to make a specific image. Because of the non-exclusivity that comes with the lack of copyright protections, these services are instead competing directly with stock image services.

*If you aren't careful, you could end up in a situation where the creative assets – icons, covers, illustrations – for your projects can be copied by con artists and hustlers, and you won't have any recourse to stop them.*

# Stochastic Plagiarism

The legal bar for plagiarism in text and audio is lower than most businesses realise.

*AI-generated works can still infringe on copyright even though they aren't protected by copyright themselves*. Copyright infringement is assessed based on *outcomes*, not process. Infringement can also be *partial* – you don't need to have made an exact copy. Large-Language-Models are prone to randomly regurgitating memorised copies of training data, which increases the risk of accidental infringement.

**Performance increases with model size but so does memorisation and with it the possibility of plagiarism**[167]

The terms used in AI research for when a model seems to have retained an exact copy of a piece of the training data is *memorisation*. When the model's output is a close match to a part of the training data, that's called *overfitting* – the result fits the request too well. They are closely related phenomena – so closely related that they are often used interchangeably in conversation – memorisation makes overfitting more likely as anything that has been memorised can end up in the output.

Memorisation and overfitting are unsolved problems in AI research. Every model does it,[168] and that has serious implications for broad use of these systems. The biggest issue being that directly copying somebody else's work in your work is *plagiarism* and that can have legal consequences and cause harm to your reputation. Most of the popular generative models are trained on data grabbed from the internet. The authors of the data have not opted in, which means that when the system copies that data verbatim into your work, neither you nor the system have permission to use that output.

Generative models do not exempt you from plagiarism as it is defined by *outcomes*, not process. If you trained a chimpanzee to type out a close copy of Stephen King's *It*, publishing that would still be plagiarism even though your particular version of it wouldn't be covered by copyright because it wasn't created by a human. The outcome – you publishing a close copy of a Stephen King novel – is what makes it plagiarism.

The same principle applies to generative models. If it randomly outputs plagiarised text or images, you are the one liable if you use it in your work.

What makes this risk insidious is that you can't tell from the output if it was memorised.

At the time of this writing, GitHub's own page for Copilot[169] states that "about 1% of the time, a suggestion may contain some code snippets longer than ~150 characters that matches the training set." It conveniently provides a setting that lets you block these suggestions.[170]

Conversely, Stable Diffusion's plagiarism rate seems to be marginally higher, just under 2%.[171]

1–2% might seem low, but if you have ten people in your organisation using a model with a 1% plagiarism rate 2–3 times a day, at least once a week you will get at least one incident where the output is contaminated with memorised content of unknown origin and under an unknown licence. *Every month will have four or five plagiarism incidents in your organisation*. You can see how this could become a serious problem with a hundred, a thousand, or tens of thousands of employees.

That 1% number crops up in other research. A 2022 study arrived at that same rate but found that with deduplication, the output copying rate could be pushed down to 0.1%,[172] which is still much too high for anything except the smallest organisations.[173]

It looks like the best case rate for verbatim copies of training data in output is somewhere between 0.1–1% of the time, not once per million. Accidental plagiarism is therefore quite likely for all but the most casual users of generative AI.

Training data copied verbatim can be embarrassing if the output is a marketing blog post, but it could be catastrophic for software projects if the project is contaminated with code under an incompatible licence. This is a particular concern for GitHub Copilot whose training data includes code under the GNU GPL and AGPL licences, which are incompatible with the licences that are used by most software businesses. Rewriting this code from scratch can be both expensive and risky.

AI vendors seem to pay very little attention to plagiarism risks in their announcements, but it does seem to be an issue that is core to how their systems are designed:

- Image-generating diffusion models habitually output images that are clones of existing images.[174] Vendors quickly respond to fix examples that are highlighted on social media, but it's safe to assume that many more go unnoticed.[175]

- Memorisation seems to *increase* with model size[176] and may have a role to play in the increased performance of larger models[177], even when the memorised data seems to be irrelevant to the tasks being benchmarked[178], which would make this problem very hard to solve without crippling the models.

- Memorisation seems to be caused by the long-tailed nature of written language – many texts are relatively unique within the data set.[179]

- Deduplicating the training data set decreases memorisation to some extent[180] but fully deduplicating text is impossible because

> the written language is inherently intertextual. You will never be able to get rid of legitimate quotations, riffs, and references without destroying the integrity of the training data. This also doesn't mitigate the long-tail issue that is the other driver of memorisation.

What makes this risk worrying is that all the research that's been done in AI research on overfitting and memorisation is based on discovering exact or near-exact copies from the training data.[181]

Plagiarism has a lower bar than that. To return to my earlier example: if you took the chimpanzee-written copy of Stephen King's *It*, ran it through a paraphrasing tool, and published it under your own name, then you would still be committing plagiarism, even though a majority of the text has been changed.

This even applies to images. In a lawsuit between the photographer Jonathan Mannion and Coors Brewing Co., the judge found that Coors was guilty of plagiarising Mannion's photograph even though they hadn't made an exact copy – they had restaged it with alterations to suit their advertising campaign. Copyright doesn't just protect the specific expression of an idea, it protects the features of that expression that make it original.[182]

This means that catching the instances where the output has copied something exactly from the training data is not enough. If GitHub Copilot copies the specific expression of an algorithm from an open source project under the GPL or a similar licence into your project, but changes the names of all the variables, then GitHub's detection system is *extremely* unlikely to catch it. But you'd still be liable. This would still be plagiarism. You would have injected your software with code that required you to release it all under the GNU GPL.

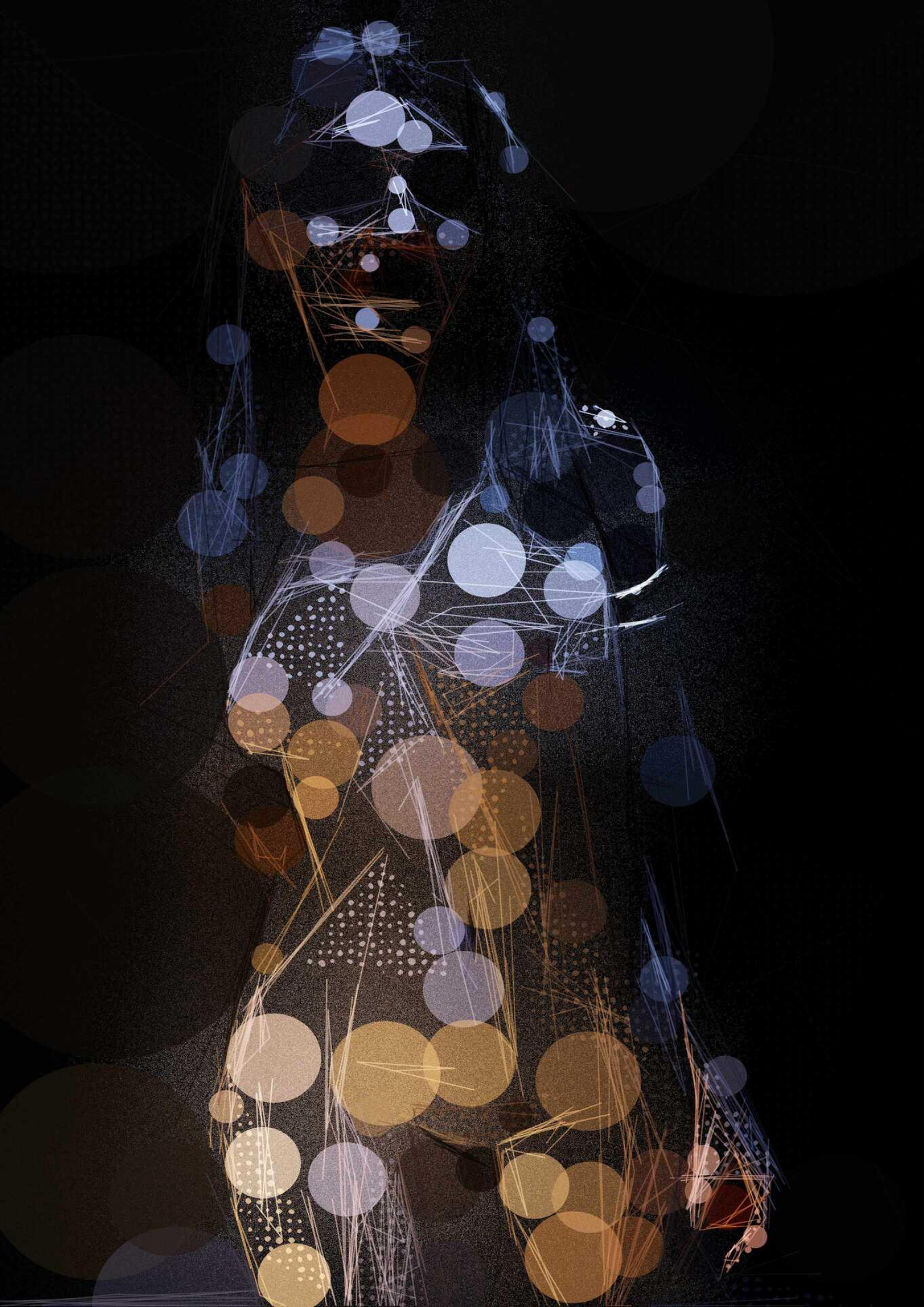
*Intellectual property audits only happen when the stakes are high: acquisitions and lawsuits. Under these circumstances AI plagiarism could cost you a fortune.*

The only reliable way of mitigating AI plagiarism seems to be to only use models trained on data sets that are explicitly licenced for use in AI model training. Very few generative AI systems on the market today provide this guarantee.

14

# Much of the training data is biased, harmful, or unsafe

> *You can't fully trust the output of generative models because of the massive size of their training data sets.*



## Summary

The data sets used to train generative models are unevenly distributed, biased, unsafe, or vulnerable to manipulation. The effectiveness of the model depends on its size, and for many use cases, there simply isn't enough data available legally. A well-resourced attacker can cause the model to generate specific results they desire, introduce more biases, hide results, and introduce false positives. *The behaviour of many kinds of generative models can be manipulated at a relatively low cost*. Discovering these attacks is impossible if the training data used remains undocumented.

# Performance increases with model size but so does exposure to attacks and manipulation[183]

The current AI resurgence began when AI researchers figured out you could address many of the problems inherent in training AI models just by going big – *really big*. Train on a good chunk of the data that's available in the world and, with that increased size, your model is magically more capable: language models are more fluent, image models generate better images, voice models sound more realistic.

Big training data sets are a revolution in AI development. They might also be their greatest weakness.

Many languages have much less available training data than English does, which means that an AI model trained in that language is unlikely to be as fluent or as capable at performing problem-solving tasks. If you believe that AI systems are the future, then many nations and languages are going to be excluded from that future. This affects businesses hoping to benefit from those markets.

Most of the training data used in the biggest models comes from the web. The web often consists of less than exemplary writing and unsavoury graphics. It's overwhelmingly social media posts, marketing, advertising, and, well, porn. Even if you manage to adjust the model to correct for the biases that these introduce, the quality of the AI's work is going to be bounded by the fact that most of its inputs are inane pap.

**Data leaks**

These models are trained on data sets that contain personal and private information.[184] Some of it is memorised by the model and can later be extracted[185], either by an attacker[186] or by accident.[187] In some cases the attacker can verify that a specific record existed in the training data set.[188] For specialised AI, such as those in healthcare, that can be a serious privacy violation on its own. Any machine learning model that has been trained on private and personal data is very likely to be in violation of the GDPR.[189]

That has major implications for the viability of these tools for many businesses.

**Gender, racial, and disability discrimination**

The web, which is the source of most of the training data, is disproportionally racist, sexist, angry, and bigoted. It is a data set full of biases and those biases translate into the output, even in the latest systems.[190] The text generation models frequently generate biased, racist, or misogynistic language.[191]

These biases persist in OpenAI's more advanced models.[192] Baked-in biases mean that generative AI models are unsuitable in their current form for processing or assessing resumes and job applications.[193] Biases affect the accuracy of these models when used by demographics that are more diverse than they were trained with.[194]

Techniques used to 'debias' models – methods to make it less likely to express the biases inherent in its training data – can also reduce its ability to accurately model the target language. It can become less fluent.[195]

**Poisoning**

Many models can be poisoned. That means an attacker can create material designed to affect the output of the model – change what it does – and publish it at a location that they know will be included in the training data set next time the model updated. A model can be poisoned to always portray a specific subject – say, an authoritarian dictator – in a favourable or negative light or to force it to always swap one phrase for another that the attacker defines. Subjects don't have to be mentioned by name in any of the training data.[196] An AI can be poisoned to change its generated recommendations: patient drug dosages changed with potentially lethal effect, loan assessments altered, real estate estimates fudged.[197] An attack can inject backdoors – hidden controls – that let the attacker alter the AI's output in predefined ways using specific keywords.[198]

These backdoors can be incredibly hard to detect.[199] The AI model can be made to *hide* content from the output.[200]

The attacker can force the model to misclassify particular images or texts. This can lead to other vulnerabilities if images are misclassified as pornographic or unsafe.[201]

Under some circumstances the attacker could even poison the model to make targeted changes to its *overall* behaviour.[202]

These attacks are relatively cheap. Researchers have managed to attack existing, in production, AI models for as little as $60, because many of the training data sets that are in use included data from domains that are either expired or easy to take over.[203]

This can harm people's physical health in a variety of ways if applied to a healthcare or medical AI system.[204]

Defending against these attacks might[205] or might not be possible for models such as OpenAI's GPT-3 and -4. Working with other AI researchers while working for Google, El-Mahdi El-Mhamdi convincingly argued that it's mathematically impossible to secure a large language model.[206] This calls into question OpenAI's ability to safely extend the cut-off date for its language models beyond 2021. They could be forever stuck in an increasingly distant past.

**Code**

AI code assistants are unique in that their training data set comes mostly from a single website – GitHub – which lets anybody upload pretty much any code they want. This *could* make attacking code assistants by poisoning their training data even cheaper and easier than attacking image-generating models. As far as I could tell, nobody has done any serious research into this possibility.

Repositories on GitHub are predominantly legacy code that encode past mainstream practices in programming. We need more research on what effect this has on the output.

You can't fully trust the output of generative AI because of the massive size of their training data sets. It doesn't seem to be feasible to protect these systems from being poisoned or manipulated by a third party. Now that they've been adopted by the mainstream, the incentive for bad actors – criminals or hostile governments – to try to manipulate them is going to be irresistible.

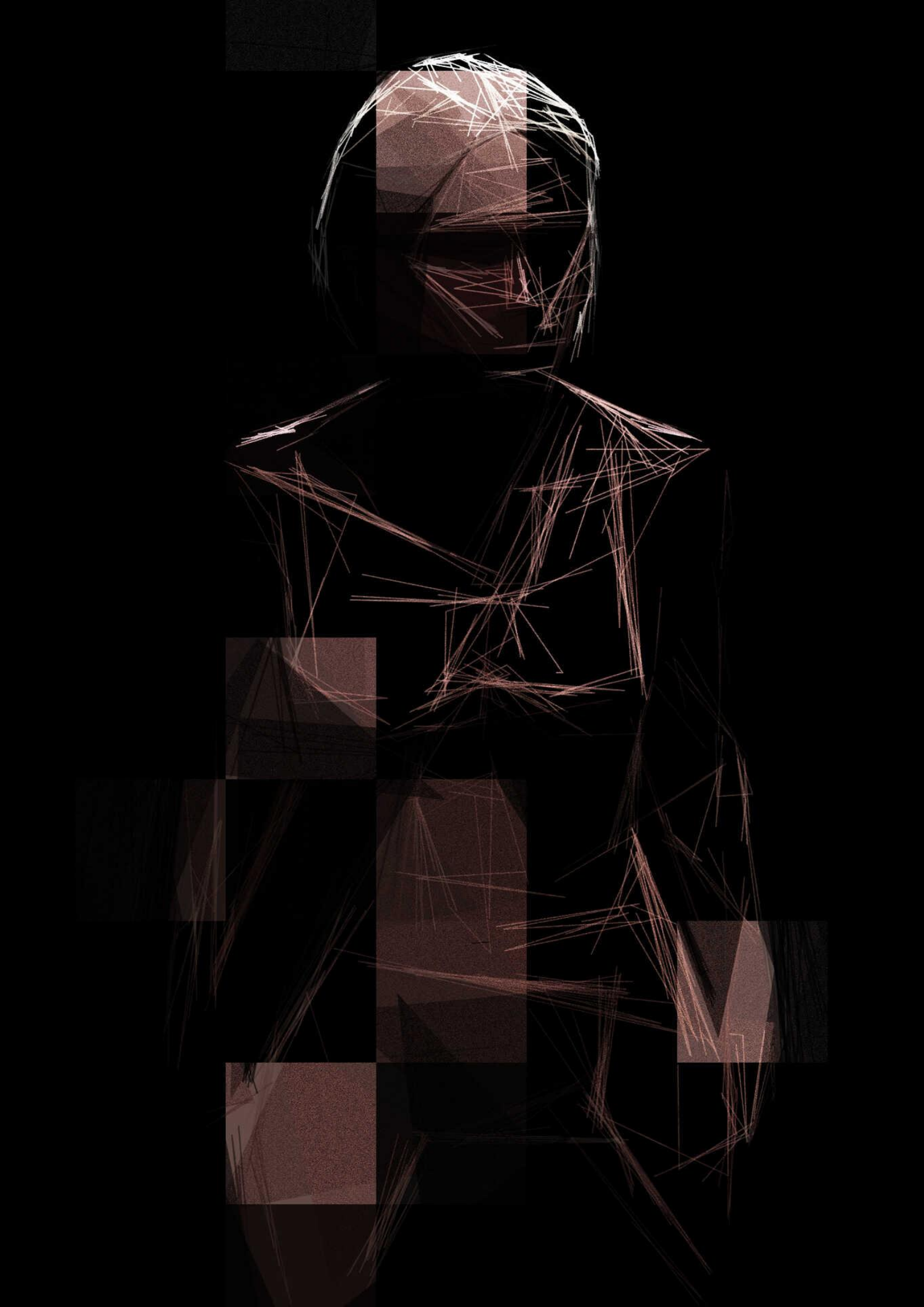
*If it's possible to manipulate GitHub Copilot to occasionally inject a security vulnerability as a part of its output, one that the criminal can later find and exploit, then the potential rewards in terms of data theft or even extortion are immense.*

If training data sets can't be secured, then language models *will* be exploited. That would mean that discovering vulnerabilities is only a question of investment, which, given the potential rewards, means it's only a matter of time until a popularly adopted language model is compromised.

15

# Generative models fabricate and don’t do facts

> *Language models are completely unsuitable for use as information systems because of their pervasive ‘hallucinations’.*



## Summary

Large Language Models have a tendency to "*hallucinate*" and fabricate nonsense in ways that are hard to detect. This makes fact-checking and correcting the output very expensive and labour-intensive. These hallucinations appear in AI-generated summaries as well, which makes these systems unsuitable for many knowledge- and information-management tasks.

## Performance increases with size but seemingly so do hallucinations[207]

One of the AI industry's time-tested tactics for implying success in failure is through creative naming. There is even a standard term for it in the field: *wishful mnemonics*.[208] Statistical analysis and transformation is called *learning*, even though it bears no resemblance to how animals or humans learn. The neurons of an AI's neural network have next to nothing in common with the neurons you find in animal brains. AIs are said to *understand* or *predict*. Vendors ascribe motivation to systems with about as much autonomy as Excel. The models do not parse, they are said to *read*.[209]

Memorisation, which I described earlier, is a good example. These AI systems are not intelligences that can memorise something as a human would. They are software that copies, stores, parses, or generates data. They do not memorise. It is a patently inaccurate label for the phenomenon of copying.[210] But, if you want to follow and participate in AI research, you are stuck with the terms they use.

If you don't accept their terminology, discourse and debate becomes impossible. But if you do, you at least partly accede to their premise that these are, well, *Artificial Intelligences*, and not just software with no autonomy or intelligence of their own.

One such misnomer that seems to be specifically chosen to gloss over the flaws and inadequacies of these systems is *hallucination*.

These systems cannot hallucinate. They are not deluded, deranged, delirious or in any other mental state you can think of. They are data and they can't feel. An AI can't have any kind of mental state in any way. They are, quite simply, software that keeps generating ***garbage***.

These **"hallucinations"** are the most frequent and obvious issue with the performance of common text-oriented Generative AI. This is one instance where adopting the industry term actively hinders analysis and risk assessment, because it is an umbrella term that covers a number of distinct errors.

- Omissions caused by training data limitations, which leads the language model to writing a falsehood, such as an AI whose training data set extends only to 2021 not being aware of the release of a film in 2022.

- Fabrication errors where the AI fabricates an answer that fits the pattern of a valid answer to the question, such as when it attributes fictional papers to real writers[211], claims that an API can do something it absolutely can't[212], makes up a fictional history of Ukraine[213], claims that somebody is dead when they aren't[214], or makes up a summary for web pages that don't exist.[215]

- Summarisation errors, where the AI makes inaccurate claims about the text it was supposed to summarise.

These are three distinct types of errors.

Even though the causes of these errors are the same, their differences mean they are likely to require different mitigation mechanisms in the software facing the end-user.

Every built-in mitigation strategy that has shipped to date treats all hallucinations as straightforward *factual errors* that can be fixed either through some form of automated checking fact by fact, itself a very unreliable process, or by instructing the AI, error by error, what is true and what isn't. The latter strategy is manifestly, obviously, impossible because the long tail of falsehoods is infinite and truth is scarce.

You will never run out of falsehoods to correct or lies to debunk. Unfortunately, this is also the exact strategy that OpenAI adopted for GPT-4[216]. They put together a list of common falsehoods and specifically trained the model not to fabricate those specific falsehoods, such as claiming that Elvis was the son of an actor. Unsurprisingly, this did not fix the problem.[217]

The fundamental problem with *fabrication errors* is that they are indistinguishable from the regular text output of a Generative AI. It's not that the system sometimes makes mistakes. The system is *always* fabricating answers. Every answer is a fabrication. Sometimes it just gets lucky and the most probable collection of words in response to that query, at that time, happened to be factual. The software does not have any self-awareness or understanding of the meaning that underlies the language.[218] Everything it generates is a fabrication. *It's all hallucinations, all the way down, even when it happens to be right.*[219]

As you could probably tell from the chapter on plagiarism, memorisation in these models is usually unintentional and undirected, and tends to favour directly memorising runs of text or image elements over facts. The way these models are built does not account for anything as vague and conceptual as a "fact".

Language models are completely unsuitable for use as information systems because of their pervasive 'hallucinations'. It's patently inappropriate for any form of search.[220] It's actively harmful as a research tool.[221] Asking ChatGPT for plant advice *will* kill your plants.[222] *An "AI" customer service chatbot will end up lying to your customers, possibly exposing you to legal liability. A search chatbot will at some point blatantly lie about the documents your customers are searching for. Document generators will get important parts of your industry and field completely wrong.* These tools are not nearly as functional as their textual and stylistic fluency implies.

Using a text synthesis system as a knowledge base or as a front end to one – such as search – is inherently risky and should be avoided. It's unlikely that the errors can be suppressed because they seem to be fundamental to how the system works – the larger and more capable the language model becomes, the less truthful it is in its answers,[223]

and early tests with ChatGPT seem to confirm that this is still the case.[224] The very thing that makes them generate more fluent and more convincing text, seems to make them less truthful.

*Summarisation errors* deserve a particular focus as summarising is one of the main use cases of most current and upcoming text-oriented Generative AI systems. Microsoft and Google both envision a future where you're using their AI systems to summarise email threads and complex documents.[225] It's even their silver bullet to slay the hallucination werewolf: instead of having the AI chatbot fabricate answers, it'll generate answers by summarising some of the search engine results into a single coherent answer.

This means that these summarisation features better work and they better be accurate.

Unfortunately, they don't seem to be.

- The hallucination rates for summarisation in earlier language models seems to have been substantial.[226]

- It's unlikely that hallucinations in summaries were eradicated with increased model size as falsehoods increased for other tasks,[227] and early tests seem to show both a regression in ChatGPT over GPT-3.5 and that they noticeably underperform when compared to simpler language models that have been fine-tuned specifically for summarisation.[228]

- OpenAI specifically didn't test summarisation factuality in their GPT-4 release document.[229] If it didn't get measured; it's unlikely to have been managed, so we have to assume that summarisation factuality hasn't improved.

- Summarisation benchmarks, as they exist in AI research, are primarily focused on testing the accuracy of summarising single documents,[230] but all the problems *we* are expected to use them for are heterogenous collections of different writing styles. Or, in plain English: the summarisation abilities of these AIs weren't

tested on documents that were a bunch of different stuff by very different people who all wrote very differently.

This last issue is a bigger problem than AI system developers would like to think and indicates that they aren't testing these systems on real-world problems.

Workplace discussions in text are *a mess*. Not in the way that they get out of hand (although that can happen), but because they bring so much variety to the table. Even something as simple as an email conversation contains a wide variety of voices:

- People have different training and education, which translates into different writing styles. Designers, developers, sales people, managers all have subtly different ways of communicating.

- Many organisations have employees from a mix of different cultures. Some come from what's called 'high context' cultures where you are expected to leave many things unsaid. Others, like those from the USA, come from 'low context' cultures where if it isn't spelled out it doesn't matter. AI summarisation might work for 'low context' cultures, won't work for 'high context' cultures, and is likely to reliably create misunderstandings in a mixed culture.

- People vary in their expertise. If you're discussing a new feature with ten people in an email thread, where everybody agrees, but most are commenting on a superficial decorative aspect of the feature ("bikeshedding"), then the debate on the colour of the proverbial bikeshed is understandably going to dominate the summary. But if the lead developer responsible for software security, in a single message out of dozens, expresses their doubt about the viability of the feature, then that would easily be the single most important point made in the thread. Any summarisation tool that downplays or even omits the lead developer's comment from the summary is inviting disaster. But how would the AI understand the relevance of the different levels of expertise in the thread? It can't, because it doesn't do "understanding" in any way,

shape, or form. **It's software; not a mind.** Additionally, most language models are trained on biased data, so they are just as likely to downgrade the lead developer's comment if their name isn't western or masculine enough.

- In a hierarchical organisation, people vary in their importance. Sometimes the importance comes from an explicit hierarchy. Sometimes an employee is important because they're *friends* with somebody important. The human mind is highly sensitive to social relationships and when we read through discussions, we account for them in our judgement without giving it any special thought. An AI is not a mind. It can't have any judgement. It won't take a hierarchy that isn't in its training data or context – that might not even be accurately recorded anywhere – into account.

*Using a language model to summarise workplace discussions risks omitting vital information or fabricating misinformation, both of which can cause a project to fail.*

The results from a search engine are even more varied than office discussions. The documents in a search engine result come from a variety of institutions. Some of them are formal documents; some of them are off-the-cuff social media posts; some of them are blatant marketing; some are actual advertising. They come from a broad range of expertise, writing styles, and experiences. That you could put a language model AI in front of a search engine and get a meaningful summary of the results is unrealistic for anything except the simplest of queries. You couldn't even summarise it reliably as a fully "generally intelligent" human being – you'd just pick the results that seem more plausible than others – but this mechanism is somehow supposed to magically make a large language model truthful?

Relying on AI summarisations in real-world scenarios is risky and likely to be counter-productive in most workplaces.

16

# Much of the output is biased, harmful, or unsafe

> *You'll sound like a racist granddad who has learned how to write generic LinkedIn posts.*

## Summary

These Generative Models are trained on the web. And, yes, that includes the porn, abuse, and violent imagery.

*Many of these tools randomly output extremely distasteful content, some of which might even expose you to legal risk*. The blocks vendors put into place to prevent this are often inadequate and often easily be bypassed. Removing unsafe data from the training set will in some cases also reduce the quality of its *safe* output, putting vendors in a position where they might have a higher tolerance for unsafe output than you.

# Performance increases with size but so do the odds of unsafe output

One of the biggest discoveries in AI research in recent years was that, when it comes to solving reasoning-based language tasks, bigger was absolutely better. But, as we've been seeing, even though common-sense reasoning *does* improve with model size, it comes with a trade-off:

*Big language models might be "smarter" and more fluent, but they are increasingly dysfunctional at everything else.*

The only way to get a training data set that's big enough is to include *everything* in the training data. Websites, blog posts, books, video and podcast transcripts – when I say everything, I really do mean *everything*. If it was online before 2021, there's a good chance that it's in the training data set that OpenAI uses for its models.

That includes all the hate from all the angry people, the violence, the abuse, and all the porn.[231]

Much like how the inherent bias in the training data set will result in a biased language model, all of that hate and pornography can result in disturbing and often blatantly illegal output.

That obviously won't work for an office tool. Companies like OpenAI have been building safeguards to try to ensure that these AIs don't sound like they belong on the sex offenders registry. This work itself has a tremendous human cost because the only way that seems to be even remotely effective is to hire people to watch, classify, and label *all* the distasteful, unsafe, and even illegal material that might be in the AI's training data set, all for less than $1.50 an hour.[232]

After all this work, they still ended up with a conversational model with a tendency towards the *unhinged*.[233]

Image generating models trained on material crawled from the internet seem to have problems that can only be described as a dependence on porn:

- Stable Diffusion and software built on it, has had a tendency towards generating pornographic or semi-pornographic images even when you don't want them to, this led to instances where software designed to create avatar images for users virtually "undressed" them.[234] This even included a capacity for generating child abuse imagery.[235]

- To counter this, when Stable Diffusion 2.0 was released, it's developer, Stability AI, had filtered out a large portion of the pornographic images that had been in its training data set. This led to a noticeable *decrease* in image quality in terms of accurately representing general human anatomy, even for otherwise safe images.[236]

- This also, inevitably, led to protests from those who had been using it *specifically* for porn.[237]

- The quality regression, particularly, led Stability AI to *reinsert* some porn into the training data for Stable Diffusion 2.1, which was met with much enthusiasm among fans.[238]

That AI generated porn has fans highlights a particular issue with these models: quite a few people *want* to see it generate unhinged,

hateful, or pornographic work, and they are willing to put in a lot of work to get it.[239]

This, and the difficulty inherent in securing 'prompts' as an interface, means that letting the end-user interact directly with the model is inherently risky.

The massive size of these models, both diffusion and language models, is a tradeoff. It does increase the quality of the result, especially for the image generation models, but it also means that you need to be aware of their capacity for the biased, horrible, and disturbing.

Beyond the negative social and cultural impact, biased systems are also likely to underperform *economically*. Over 37% of the actively licenced physicians in the US are women, a proportion that keeps growing every year.[240] In the EU, half of active physicians are women.[241] In many other countries, they've become a majority.[242] Healthcare as a whole predominantly employs women. Meanwhile, ChatGPT quite often seems unable to acknowledge even the possibility of a female physician.[243]

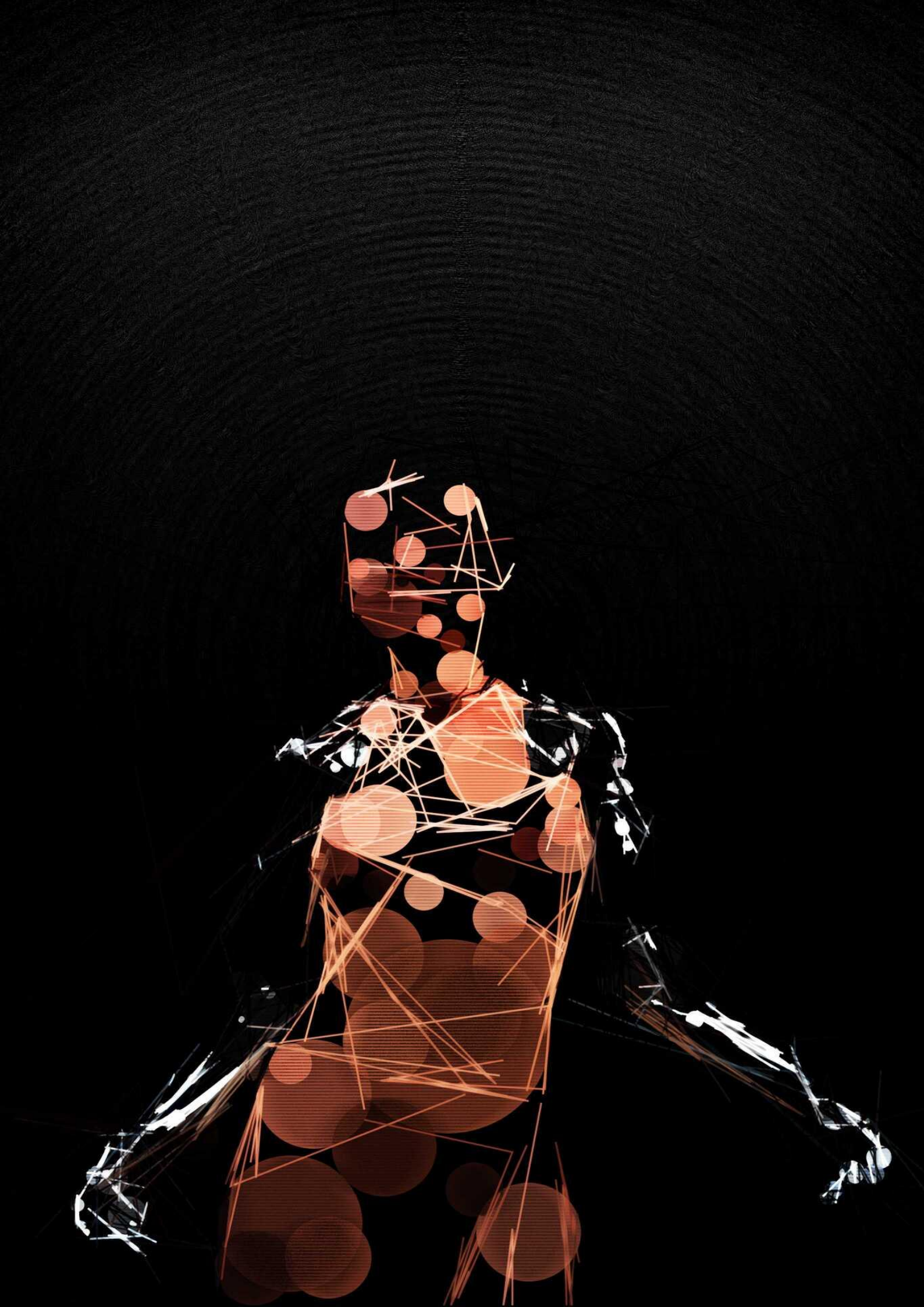
By their very nature – being based on large collections of historical training data – language models will always be backwards-facing and favour past attitudes over modern. They are not likely to accurately represent current markets, neither in the US nor globally. It's the nature of markets and demographics to change over time. Populations aren't static bodies and language evolves with them. There is a very real risk that the language generated by the model – your marketing, emails, and blog posts – will look archaic to target audiences.

*You'll sound like a racist granddad who has learned how to write generic LinkedIn posts.*

17

# The Mediocrity Trap

> *Paradoxically, instead of lowering costs, pervasive automation might end up making effective writing and art much more expensive.*



# Summary

When the world is flooded with "good enough" writing, standing out from the crowd requires much more effort and becomes much more expensive.

# Automated mediocrity makes everything *more* expensive, not less

Generative AI makes text and art that is the most probable response to a given prompt.

This is *mediocrity*. It might as well be the dictionary definition. The output of these models is the *good-enough* product writ large. Beyond the mediocrity, they also suffer from a number of recurring defects.

Social media ridiculed "AI hands" – the tendency these tools had towards rendering unrealistic hands – which is why AI vendors prioritised fixing that issue. But, the hands were a *symptom* of a deeper underlying problem: diffusion models do not have an internalised model of the physical world or of how objects relate to one another. That isn't how they work.

Some common defects in current generated art:

- Adjacent objects blend together, or even end abruptly.
- It gets proportions and relative sizes wrong.
- The overall image has a poor or skewed representation of depth.
- Poor composition or visual structure.
- Many of the pictures generated have the same colour scheme: orange highlights, teal shadows, brown mid-tones.

- Upscaling artefacts that make the image look too sharp.

Model-generated text has its own issues:

- It tends to repeat words and overuse clichés.
- The sentence length is often too consistent with little variation.
- The output favours simplistic textual and sentence structures.
- Many of these systems have a style that is best described as "harmless but enthusiastic LinkedIn post".

*These tools are not good at art or writing*. What they make is, if you're lucky enough to avoid the defects above, still mediocre at best. It's amazing that they're able to do what they do in the first place, but the inherent flaw of a probabilistic model is that it will always return to the most probable. That's what it is and *that will always make it mediocre*. Consistently rising above mediocrity would require a completely different system, built on entirely different principles.

Solutions for improving output quality are limited. Vendors can train on more data, but that's just as likely to increase volatility and make the system less useful in production. There's also a limit to how much data is freely available in the world. Vendors can apply a variety of fine-tuning methods to limit the output to a range of outputs that are known to be good, but that can severely limit the flexibility of the system. Finally, they can train on *less* data and exclude lower-quality examples from the data set, but that also severely limits the system's flexibility because the model will have a narrower range of examples to extrapolate from.

If you, for example, train or fine-tune a diffusion model mostly on professional photographs there's a decent chance that the model's average output will look somewhat professional, but the variation from photo to photo will also be much more limited as the diffusion model will have a strong tendency to converge on what's similar. That's fine if you only want images of pretty models in their early twenties. Less useful if you need an image that says something specific. A smaller

training data set also increases the odds of stochastic plagiarism, which makes it even more important that the vendor of the generative model can guarantee the rights and provenance of the images in their data set.

Either way, there's still a substantial risk that the model will get anatomy, proportions, and physics wrong.

Vendors have very few options here. They can offer systems that either deliver a broad variety of mediocre images or a mediocre variety of above-average images. This trade-off is baked into how these models work.

These models can generate that mediocrity on demand and at scale. Its quality is predictable both in cost and value, and that's where the trap comes in.

The marketing and copywriting efforts of many organisations are already mediocre. They've never had to be any better. Keeping your standards low simplifies both recruitment and production. Consistency has always been the hallmark of business writing for a good reason. The quality of your marketing copy is less important than your ability to keep to a schedule or a deadline.

Large language models can now deliver "good-enough" writing on demand. It is consistent, both in terms of procedure and writing. These systems means that generating media – text, images, audio – of a low to mediocre quality can be almost entirely automated.

Managers are now discovering that language models can do almost as good a job as their current copywriters. Some are drawing exactly the wrong conclusion and are cutting costs by replacing copywriters with AI tools.

This is only productive in a hypothetical universe where you're the only organisation in the world that uses generative models to write.

When a digital product can be entirely automated that results in a massive glut of product. We are going to see several orders of magnitude increases in the volume of product.

It's going to be *everywhere*. Model-generated content – I hesitate to call it writing – is already common on social media and in search engine results. More and more, we are surrounded by countless people

and businesses generating middling, algorithmically-generated, good-enough, marketing prose on demand. It's filling the Kindle store, the web, and online photo and illustration sites. Entire categories of online media are being taken over. Your work is disappearing into an endless parade of carbon copy LinkedIn posts, Instagram street fashion photos, and hyper-sharp, hyperrealistic illustrations that all feel the same.

Under these circumstances, any work that is "the most probable response to a given prompt" might as well be noise. Many organisations and people will have to change their marketing, copywriting, and media production strategies completely if they want to be heard. Your readers and viewers might well be satisfied with mediocre work today, but that will change when the volume of mediocrity increases a thousandfold. Paradoxically, instead of lowering costs, pervasive automation might end up making *effective* writing and art much more expensive.

Middling work won't cut through the noise any more.

18

# Shortcut reasoning

> *Their reasoning is nothing more than an intertwined nest of shortcut upon shortcut.*

## Summary

These systems are very prone to *shortcuts*. Instead of building a coherent model for reasoning, they tend to take shortcuts based on correlations in the training data.

This is a problem because shortcuts are *fragile*, which makes for pseudo-reasoning systems that only work under very specific circumstances.

# This is the one bit that improves with model size, up to a point

When AI researchers say that language models improve with size – even though performance at natural language tasks doesn't increase proportionally[244] and both memorisation[245] and hallucinations[246] seem to get worse – this is what they mean. Most AI researchers are in the field to get to *Artificial Intelligence*, not make a better natural language tool or information system. "Intelligence" is what vendors focus on measuring when they develop. The downsides are usually discovered after the fact, by external researchers, provided they are allowed to do so.[247]

What this means for you is that the priorities of the AI vendor, broadly, don't align with the needs of their customers.

Most organisations who are in the business of making Generative Models only care about improved "reasoning" and, with large language models, they seem to have found their answer. Finally, an AI model that is capable of common sense reasoning that only seems to improve with the size of the model.

But…

Let's just say that "capable" is in the eye of the beholder.

**What's wrong with taking a shortcut?**

Imagine you have a job at a hospital. Your job is to watch patients and let the doctors know if they seem to be experiencing *sepsis* or blood-poisoning, a potentially lethal bacterial infection which would require treatment with antibiotics.

Accurately detecting sepsis requires expertise. You need to pay close attention and know what to look for.

But, since you're a literal-minded sort that's always on the lookout for a shortcut, you figure out a very simple, extremely reliable way of noticing when a patient has sepsis. If the doctor prescribes antibiotics for sepsis, then the patient probably had sepsis! No need for to do any hard work. Infallible deduction that works every time. Neat, right?

This may sound absurd, but it is pretty much how an AI system designed to spot sepsis operated.[248] The sepsis-detection system was launched, approved by healthcare authorities, and adopted by hospitals. Later, when unaffiliated researchers tried to replicate the vendor's claims, they found that not only did the system perform much worse than expected, it partially based its "predictions" of sepsis on the healthcare staff noticing the sepsis and treating it – if antibiotics appeared in the medical record, the patient probably had sepsis. Problem solved.[249]

This is a kind of flaw that is common to most, if not all, AI systems today. They are trained on a data set and tasked with solving problems such as: "find all the chest x-rays that look like they have COVID-19". Except they are not capable of actual thinking. When trained to detect COVID in a chest x-ray or sepsis in patient data, they will take shortcuts because they have no notion of COVID or sepsis. The sepsis model found that all the patient records in its training data that had sepsis had one thing in common: the doctor prescribed antibiotics. It took a shortcut because it's just software calculating probabilities. It isn't capable of reasoning, so it can't 'know better'. The COVID model? It predicted that patients who were too sick to stand upright for their chest x-rays probably had COVID, which was about as useless as you'd imagine.[250]

This wasn't a unique result. The COVID pandemic led to a surge in the development of AI systems that were supposed to help diagnose and spot the disease. None of them helped.[251]

Shortcuts are core to how these AIs 'reason'.[252] When given a task, the system will always prefer the most straightforward correlation to the solution over solving the task in a more reliable way. That's why the COVID-detecting AI was ultimately more of a 'x-ray position' detector than a COVID detector. The patient's position correlated just as well with the task as any of the genuine COVID symptoms but was easier to detect. Shortcut found; shortcut used.

Most of the well-documented instances of shortcut-reasoning in AI come out of healthcare. That isn't because developers of healthcare AI are more sloppy but because we, as a society, have higher standards for the technology used in healthcare than for what we use in an office.[253] Healthcare systems get more attention. More researchers attempt to replicate vendor findings. They are often tightly regulated. Bullshit gets noticed. In theory, at least. Office or productivity AI systems, however, get away with much less scrutiny

Because of the way these models are designed, they have a strong tendency towards finding the simplest correlative solution for every task. As you can tell from the examples above, that means that the only way to get robust results – robust pseudo-reasoning – is to carefully curate the training data set, which is the opposite of how most Large-Language-Models are developed. OpenAI and Google, the makers of GPT-4 and Bard, are instead making a different tradeoff: make the model big enough, and you'll end up with a shortcut for every scenario. For every reasoning task, there will be samples in the training data that correlate with a probable solution. For models that have been trained on most, if not all, publicly available writing in English, the end result is remarkably fluent and surprisingly capable.

That capability is extremely fragile, though. There's nothing generalisable about the intelligence in these AIs.[254] Their reasoning is nothing more than an intertwined nest of shortcut upon shortcut.[255] It does extremely well in standardised benchmarks and testing, but that's because it had the answers in its training data.[256] With enough

shortcuts, you can fake a lot of intelligence, but it'll never be general-purpose, and it will break as soon as it encounters situations that it doesn't have a shortcut for.[257]

This results in a system that is capable of impressive common sense reasoning *provided you get the prompt exactly right*. Even for cutting-edge systems such as ChatGPT, rephrasing or even something as straightforward as changing the sentence order can be the difference between a nonsense fabrication and a useful, reasoned answer.[258]

This kind of pseudo-reasoning, useful as it can be if you can figure out the magic incantation – prompt – to trigger its best behaviour, is also always stuck in the past. It is stuck in the cut-off year for its training data. ChatGPT only "knows" history up to year 2021 and its training data set is predominantly older texts that are possibly outdated. If somebody rewrites the standardised test, performance drops. If practices change, correlations that helped fake reasoning about past practices will break. AIs are not adaptable.

This lack of adaptability can have hilarious results, such as when marines who were tasked with bypassing an AI security system figured out that all they needed to do was something new – anything new. Even if that 'new' was something right out of a Looney-Tunes cartoon.[259]

Among the tactics that worked:

- Disguise yourself as a tree and walk slowly past the AI-guided robot.
- Hide under a large box and crawl past the robot.
- Somersaulting past it.

Basically, what they made was the Wile E. Coyote of security systems. It had high performance in narrow tasks, but its lack of robustness meant that it was incompetent in practice.

Another example was a recent experiment where an amateur human player beat an AI in Go.[260] The AI in question normally has what is called 'superhuman' performance in Go, beating every human player set against it. How did the amateur beat the superhuman AI? By us-

ing another, simplified, AI model to suggest random nonsense moves that no sensible player would use. That meant that the moves didn't have any meaningful correlations in the superhuman AI's training data set, which lead it to make dumb moves that an amateur could easily exploit. The simplified AI required only 14% of the compute power needed to train the superhuman AI, which means that the attack should be more scalable than the defences of the Go-playing AI.[261]

In an infinite world and with finite training data, there is always going to be a way to break the pseudo-reasoning of these systems.

These stories are fun, but they are ultimately theatrics. They are "just so" stories designed to make you feel good about what you already want to believe. What matters is what can be tested, and these systems keep failing whenever they are put to rigorous, structured testing. Training data leakage, which leads to superior but ultimately fragile performance, as well as poor replicability in scientific research, is a serious problem for the field of AI.[262]

But, it's also a problem for you. If you're hoping to integrate an AI system into your business processes, you need to be aware of the extreme fragility of their performance. They are superhuman under ideal circumstances, but fail disastrously when something even just a little bit unusual happens.

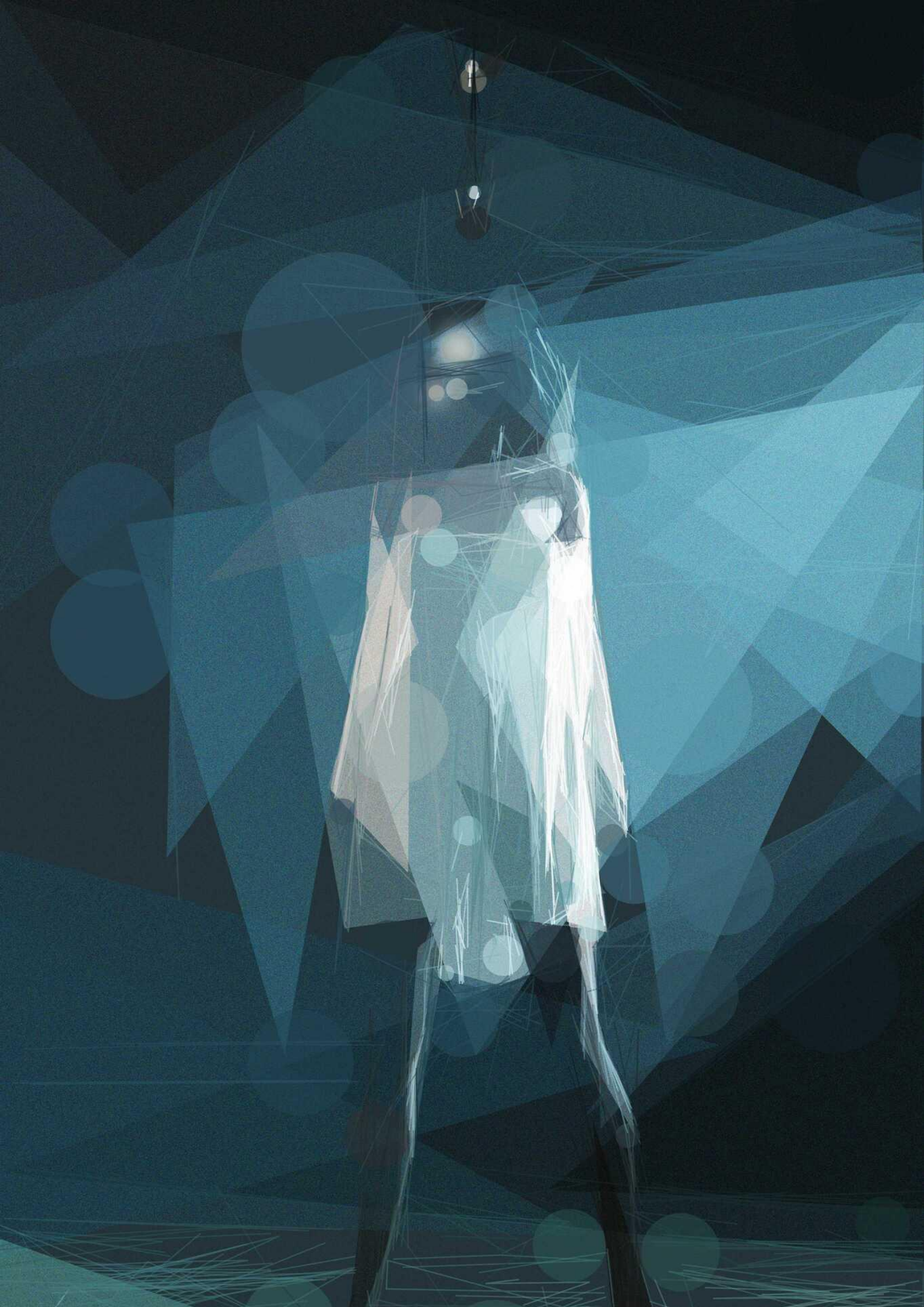
When they fail, they fail completely. For many use cases, they are not safe enough to be used unattended. Don't fall for claims about reasoning (general or no) or performance. What matters is proven functionality and competence. Many of these systems just aren't functional.[263]

*Using a language model to automate reasoning or decision-making is likely to be extremely unreliable and prone to catastrophic failure.*

19

# Code quality

> *The "cost-saving" automation of generative models is likely to be very expensive in the long run.*



# Summary

Language models should in theroy lend themselves to creating a variety of useful tools for software development. They have the potential to revolutionise programming. *The tools that accomplish these miracles don't exist yet*. What we have today mostly make bad situations worse.

# The software industry is bad at software

Switching costs are, more often than not, massive for business software, and purchases are not decided by anybody who actually uses the software. The quality of the service disconnects from sales performance very quickly in a growing software company. The company ends up "owning" the customer and no longer has any incentive to improve the software. Because adding features is a key marketing and sales tactic – features drive PR and stock price – the software development cycle becomes an act of intentional, controlled deterioration.

Enormous engineering resources go into finding new ways to minimise the feature-driven deterioration.

*The software industry is bad at software*. Great at shipping features and *selling* software. Bad at making the software itself.

By its nature software has enormous margins which further cushion it from the effect of delivering bad products.

The objective impact of poor software quality on the bottom lines of companies like Microsoft, Google, Apple, or Facebook is a rounding error. The rest of the industry only needs to deliver usable early versions, but once you have an established customer base and an experienced sales team, you can coast for a long, l*ong* time without improving your product in any meaningful way.

We have an industry that's largely disconnected from the consequences of making bad products, which means that we have a lot of successful but bad products.

Thus, Weinberg's Law: "*If builders built buildings the way programmers wrote programs, then the first woodpecker that came along would destroy civilization.*"[264]

It's into this environment that "AI" software development tools appear. The tools that are supposed to fix everything.[265] The tech industry has positioned generative models as the next revolution in knowledge management and search. This has an impact on the use of these systems in software development.

To be able to answer questions and pass as a knowledge system, a generative chatbot needs to *memorise* the answer, or at least parts of the phrase. Because "AI" vendors are performing a sleight-of-hand and are presenting statistical language synthesis engines as knowledge retrieval systems, their focus in training and testing is on "facts" and minimising "falsehoods". The model has no notion of either, as it's entirely a *language* model, so the only way to square this circle is for the model to memorise it all – or, at least quite a bit more than it would have to just be able to hold a conversation.

Vendors then compound this by using human exams as benchmarks for model reasoning performance. The problem is that bar exams, medical exams, and diagnosis tests are *specifically* designed to mostly test *rote memorisation*, not reasoning.

Between the tailoring of these systems for knowledge retrieval and the use of rote memorisation exams and code generation as benchmarks, the tech industry has made language models into systems where memorisation is a core part of how they function.

# AI and the Art of Programming

Code has a number of qualities that should, in theory, make it uniquely suited to language models. Programming languages are more uniform and structured than prose. Its meaning usually doesn't change based on context or culture. Programming language output can often

be tested directly, which might help with the evaluation of each training run. Hallucinations should be immediately discovered as errors.[266] It's not too unreasonable to expect that programming might lend itself to language models.

But generative models in software development seem to cause more problems than they solve. Those problems start with the models themselves and the tools that are built on them.

# The design of the tools

The software development tools based on generative models come in a few different varieties:

- Fancy autocompletes – sometimes labelled "code copilots" or "code assistants". You prompt it with a phrase, query, or code, and it will complete it with its best guess.
- Code modification and conversion tools. Tools that generate tests for your code fall into this category.[267]
- Chat-based interfaces that are used to generate or explain code in an interactive manner.

*Autocompletes* were the first to gain serious adoption in the form of GitHub Copilot. Unfortunately, accidentally copying GPL-licenced code into your project is a real possibility with GitHub Copilot, which has serious consequences for projects whose licence is either proprietary or incompatible with the GPL.[268]

That doesn't mean other "code copilots" or the chat-based interfaces such as ChatGPT are free from risk. Autocompletes and chat-based interfaces create a dynamic that interferes with the programmer's ability to assess the code it generates.[269] Their design triggers two biases that make the coder more likely to accept code they would not find acceptable in other contexts.

***Automation bias:*** People have a tendency to trust what the computer says over their own judgement. Or, to be more specific, when presented with a cognitive aid intended to help them think less, that's

exactly what they do: they think less. This leads them to be much less critical of the outcome of the cognitive aid than they would have of that same output in other circumstances.[270] This has lead to serious issues in other industries, such as aerospace, where pilot trust in the capabilities of a plane's autopilot over their own has caused catastrophic accidents.[271] This is a known problem with how we, as humans, interact with machines that offer any form of cognitive automation.[272]

***Anchoring bias:*** Our mind fixates on the first thing that anchors it in that context. For example, when it comes to shopping, the first price you see will be the anchor that you judge other prices from.[273] The first result from an autocomplete tool is likely to trigger our anchoring bias.

These two biases combined mean that users of code assistants are extremely likely to accept the first suggestion the tool makes that doesn't cause outright errors. Microsoft themselves have said that 40% of GitHub Copilot's output is committed unchanged.[274]

Let's not get into the question of how software development, as an industry, put itself in the position where Microsoft can follow a line of code in another company's software project from the GitHub language model, through a text editor, and into a supposedly decentralised version control system, which is worrying enough.

That so many developers just use what's initially suggested means autocompletes need to be substantially *better* than an average coder to avoid having a detrimental effect on overall code quality. Obviously, if programmers are going to be picking the first suggestion that works, that suggestion needs to be at least as good as what they would have written themselves unaided. What's less obvious is that the lack of familiarity – not having written the generated code by hand themselves – is likely to lead them to miss bugs and integration issues that would have been trivially obvious if they had typed it out themselves. To balance that out, the generated code needs to be *better* – substantially better – than average, which is a tricky thing to ask of a system that's specifically trained on mountains of average code.

Unfortunately, that mediocrity seems to be reflected in the output. GitHub Copilot, for example seems to regularly generate vulnerable code with security issues.[275]

People overwhelmingly seem to trust the output of a language model. If it runs without errors, it must be fine. But that's not always the case.

# The tools encourage the worst of our management and development practices

A language model will never question, push back, doubt, hesitate, or waver. Your managers are going to use it to flesh out and describe unworkable ideas. It won't complain. The resulting spec won't have any bearing with reality.

People on your team will do "user research" by "asking" a language model. The resulting "research" will be fiction and entirely useless.

It'll let you implement the *worst ideas ever* in your code without protest. Ask a copilot "how can I roll my own cryptography?" and it'll return a half-baked implementation of an out-of-date algorithm in PHP. (This is a very bad idea, for those of you who aren't programmers.)

These tools will always try to generate a solution, even when it has no hope of generating safely working code.[276] Many of a programmer's initial ideas are ill-advised. It comes with the territory. You have a problem. Think of a solution. Do some research to see how it would be done. Discover that the solution was a bad idea. Type "how to implement your own cryptography" into Stack Overflow and you will get answers that tell you, unequivocally, that you absolutely should not implement your own cryptography. Programming is full of these pitfalls and AI tools, with their "hallucinations" and a tendency to generate confident answers no matter what, are almost purpose-designed to trip you into them.

Some of the pitfalls have to do with *age*. Platforms, libraries, and APIs become obsolete, insecure, and outdated, but language models are, because of their design, stuck in the past. Newer methods will be underrepresented in their training data sets *if they appear in them at all*. You are likely, for example, to run into instances where they would recommend that you use outdated open source libraries that have since been replaced with faster and more reliable platform implementations. Or, it might generate code that uses APIs that have been deprecated or even removed in current versions of the platform. If the AI copilot isn't 'fresh' enough, it's going to cause problems.[277]

Code assistant language models will need to be updated much more frequently than natural language models if they are to remain useful in the long term.

# Code bloat

***Language models don't deliver productivity improvements. They increase the volume of code, unchecked by reason.***

The biggest expense for most software projects is *maintenance*. By making it easy to generate a lot of bug-filled, insecure, but broadly functional code, these tools are likely to lead to code base inflation.

Projects will be bigger while doing less, making them harder to fix, update, and maintain. By going faster, we will have slowed ourselves down. The cost-saving automation of generative AI is likely to be very expensive in the long run.

Code that developers don't understand is a liability. It means your mental model of the software is inaccurate which will lead you to *create bugs* as you modify it or add other components that interact with pieces you don't understand.

Language model tools for software development are specifically designed to create large volumes of code that the programmer doesn't understand. They are liability engines for all but the most experienced developer.

# Licence contamination

Microsoft and Google don't train their language models on their own code. GitHub's Copilot isn't trained on code from Microsoft's office suite, even though many of its products are likely to be some of the largest React Native projects in existence. There aren't many C++ code bases as big as Windows. Google's repository is probably one of the biggest collection of python and Java code you can find.

These companies don't seem to use their own code as a training data set, but instead train on collections of open source code that contain both permissive and copyleft licences.

Copyleft licences, if used, force you to release your own project under their licence. Many of them, even non-copyleft, have patent clauses, which is poison for quite a few employers. Even permissive licences require attribution, and you can absolutely get sued if you're caught copying open source code without attribution.

Remember what I wrote earlier about *memorisation*? How the way these models are designed and trained seems to promote string memorisation and copying?

Turns out blindly copying open source code is *problematic*.

These models memorise *a lot*, and they tend to copy what they memorise into their output. GitHub's own numbers peg verbatim copies of code that's at least 150 characters at 1%[278], which is roughly the same, in terms of verbatim copying, as what you seem to get in other language models.

If you use a language model for development, a copilot or chatbot, three or four times a day, you're going to get a verbatim copy of open source code injected into your project about once a month. If every team member uses one, then multiply that by the size of the team.

GitHub's Copilot has a feature that lets you block verbatim copies of code that are longer than 150 characters. This obviously requires both a check, which delays the result, and will throw out a bunch of useful results, making the language model less useful. Copilot is already not as useful as it's made out to be and feels pretty slow so many people are going to turn off the "please don't plagiarise" checkbox.

But even GitHub's checks are insufficient. The keyword there is "*verbatim*", because language models have a tendency to rephrase their output. If GitHub Copilot copies an implementation of an algorithm from code that's under the GPL into your project but changes all the variable names, Copilot won't detect it. But it will still be plagiarism and the copied code is still under the GPL. This isn't unlikely as *this is how language models work*. Memorisation and then copying with light rephrasing is what they do.

Training the system only on permissively licensed code doesn't solve the problem. It won't force your project to adopt an MIT licence or anything like that, but you can still be sued if it's discovered.

This would seem to give Microsoft and GitHub a good reason not to train on the Office code base. If they did, it's very likeley that that a prompt to generate DOCX parsing code would "generate" a verbatim copy of the DOCX parsing code from Microsoft Word or Google Docs.

This would both undercut their own strategic advantage, and ***it would break the illusion that these systems are generating novel code from scratch.***

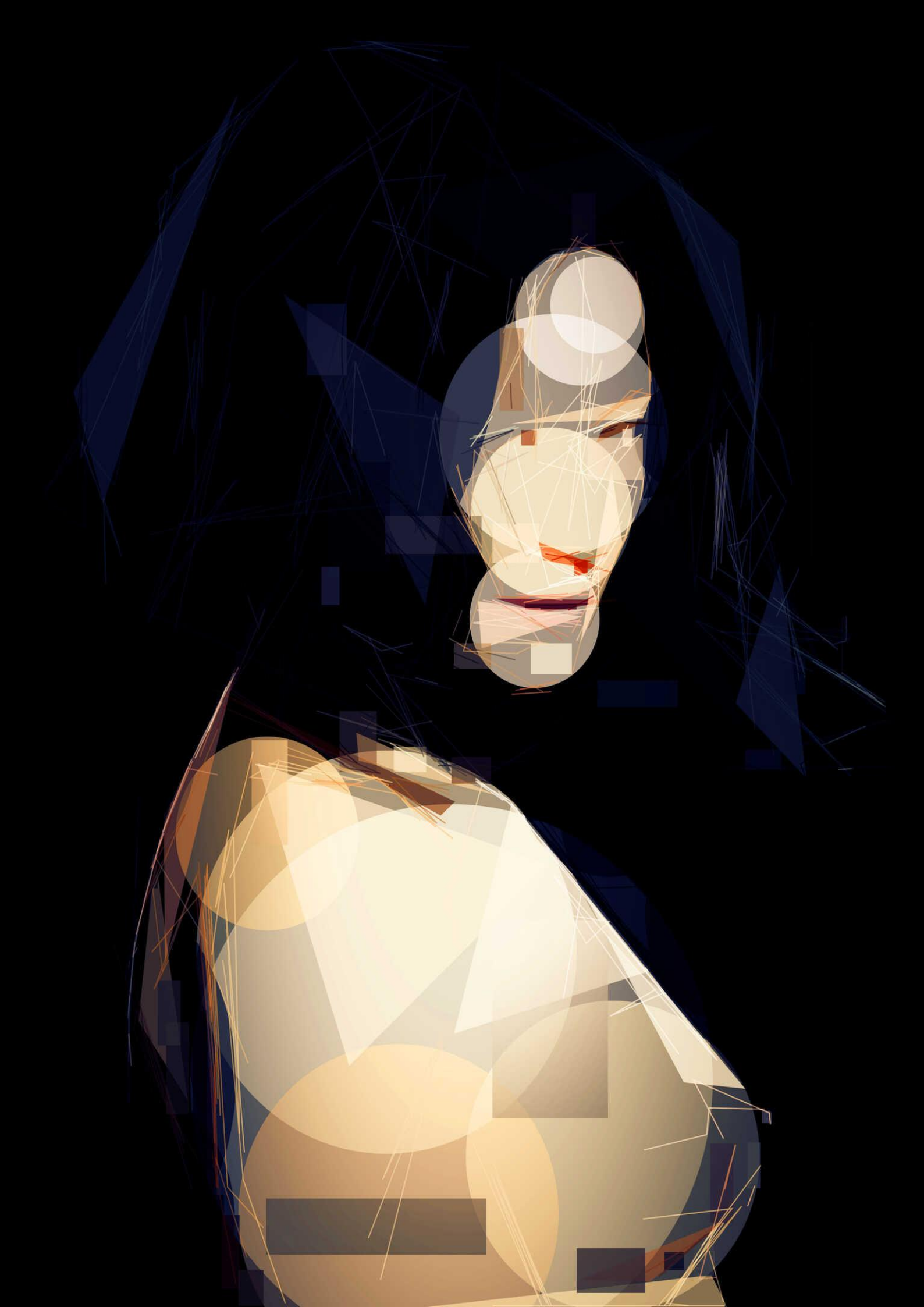
*Extensive use of Large Language Models for software development bloats your code base, lowers overall quality, makes projects harder to understand and decipher, and expose you to legal risk.*

# What next?

20

# The Tech Industry took a wrong turn

> *Safe, for the tech industry, is too slow.*



# "See the choice of dreams", and then worry about it

> Very well. This book – this side, *Dream Machines* – is meant to let you see the choice of dreams. Noting that every company and university seems to insist that its system is the wave of the future, I think it is more important than ever to have the alternatives spread out clearly.
>
> But, the *experts* are not going to be much help, they are part of the problem. On both sides, the academic and the industrial, they are being painfully pontifical and bombastic in the jarring new jargons (see "Babes in Toyland," p. 4). Little clarity is spread by this. Few things are funnier than the pretensions of those who profess to dignity, sobriety and professionalism of their expert predictions – especially when they too are pouring out their personal views under the guise of technicality. Most people don't dream of what's going to hit the fan. And the computer and electronics people are like generals preparing for the last war.
>
> Frankly, I think it's an outrage making it look as if there's any kind of scientific basis to these things: there is an underlevel of technicality but like the foundation of a cathedral, it serves only to support what rises from it. THE TECHNICALITIES MATTER A LOT. BUT THE UNIFYING VISION MATTERS MORE.
>
> *Ted Nelson, Computer Lib/Dream Machines*[279]

AI software development – the "what is this for?" part – has never had much of a unifying vision. AI research, sure, they have a vision: they want intelligent machines first, figure out what to do with them second. They dream of a robot future. Some parts of the research monomania end up having clear software benefits. Being able to point a computer at an image and have it get at least a rough idea of what the picture is of is neat and that came from Machine Learning research. It didn't come with a single specific "what is this for?" idea except, you know, "how is our robot going to see the world around it?", but it made up for it by being obviously useful in a general sort of way. It's a

capability that's now built into pretty much all of our devices. As a feature that's now integrated into our lives, it's a microcosm of the issues we have with innovations coming out of AI research:

1. It has helped the blind and partially-sighted access places and media they could not before. A genuine technological miracle.

2. It lets our photo apps automatically find all the pictures of Grandpa using facial recognition.

3. It has become one of the basic building blocks of an authoritarian police state, given multinational corporations the surveillance power that previously only existed in dystopian nightmares, and extended pervasive digital surveillance into our physical lives, making all of our lives less free and less safe, and could enable devastating harm on a massive scale if truly fascist governments come into power.

One of these benefits is not like the other.

Universal facial recognition is terrifying when it works perfectly and a nightmare when it's flawed. It exaggerates power imbalances and disproportionally enables bad actors and authoritarians. It's equal parts pleasant domestic miracles and blighted social and political horrors. Generative models are likely to follow the same path.

In the absence of a unifying vision, the tech industry simply does what makes the most money *for people working in the tech industry* – greed fills the void where there should have been vision. Companies such as Amazon didn't hesitate to sell facial recognition services to law enforcement – until the backlash forced them to stop.[280]

This might just be capitalism, but the 'just' in that phrase feels quite different when the industry in question is peddling synthetic miracles.

Greed might be inevitable. It might always seep into the cracks, break apart the concrete foundations our ivory towers are built on. But, having a coherent unifying vision that's backed by clear values does a remarkable job of holding off the decay.

Even today, the web is like living fossil, a preserved relic from a different era. Anybody can put up a website. Anybody can run a business over it. I can build an app or service, send the URL to anybody I like, and most people in the world will be able to run it without asking anybody's permission. There are rules and regulations you have to follow, obviously, but those are remarkably straightforward if you aren't actively spying on people or messing around with their personal data.

You can trace the lineage of the vision behind the web from Tim Berners-Lee[281] in 1989, through Ted Nelson in 1974 and Douglas Engelbart in 1968, all the way to Vannevar Bush's article *As We May Think* in the Atlantic Monthly back in 1945.[282] All of their books, software prototypes, theories, and ideas run along the single continuing thread of the hypertext concept – links – as a new kind of punctuation mark[283] that connects the information of the world together in a coherent network. It's a vision that encompasses concept, functionality, interface, and values – and it persists to this day, despite decades of greed, abuse, surveillance, and shitty ads. It's a unifying vision of a world that's simultaneously technological and literate. This vision is part of what has kept it alive. Despite the frustrations, pain, and the flaws, working on making parts of the web is a privilege.

While AI researchers are busy trying to build their robot dream, generative model *software* has no such unifying vision. Some vendors are bent on replacing humans at their jobs – effectively promoting their software as "AI" illustrators, voice actors,[284] even "robot" lawyers.[285] – and then look surprised at sheer enormity of the anger that they get in response. Other vendors are resurrecting Microsoft's 1990s dream of intelligent agents, assistants, or copilots who operate in the context of the software you use – extending Clippy's lineage into the modern world.[286] The rest seem to have lazily reached for the first interface metaphor they could think of – the *chatbot* – with no thought or even the vaguest idea of how it should actually integrate with the rest of the work we're doing.

As much as it makes your average MBA salivate, "let's replace people with something shittier but cheaper" isn't much of a vision for software development and user interface design, which leaves those

who are genuinely curious about the applications of the technology with the other two paths.

Those seem to be converging towards a single idea: "Human-in-the-loop". It's the idea that in the interactive loop with the AI software, the decision-making, choices, and actions are made by the human.[287] Instead of full automation there's feedback from the AI as the tasks progress and the human responds to that by interacting with the various user interface affordances provided by the system.

In other words, the human sits in front of software, uses it as they would any other software, and then it does stuff for them as any other software does, *except in an AI way*.

*The grand unifying vision of AI-assisted software is that you should use it to make software*.

That's an idea that's only remarkable because of how many AI enthusiasts think they can do away with the people part of getting things done.

# Acquiesce, or mitigating the inevitable

In an earlier chapter I wrote about the failure of an AI model designed to predict the onset of sepsis, how external reviewers discovered the flaws, which then led to the vendor updating and improving it.

At one hospital, UC Health in Colorado, they found that the system still wasn't that useful: "the ratio of false alarms to true positives was about 30 to 1".[288]

To salvage their investment UC Health changed their approach with the system. Instead of using the AI as an autonomous prediction system that sent out alerts to overworked doctors and nurses, they put together a special monitoring team of clinicians that used live video feeds to helped filter out the false alarms. That team built relationships with bedside nurses throughout the hospital. Where the AI system alone was utterly useless, the human team, assisted by their relationships with the nurses and the AI system, was estimated to save about 211 lives annually.

AI on its own was worse than nothing while AI as an assistant saved lives, a clear demonstration of the value of human-in-the-loop. It's a heart-warming parable that lends credence to the ambitions of those who are trying to make "AI assistants" happen.

This, and other stories like it, are going to be the foundation myths of a thousand new AI services – the seeds of a new computing revolution.

Or, more specifically, it's a certain kind of fertiliser. The kind that smells.

(I'm saying it's bullshit.)

Case studies are amazing tools. You can pick one instance where everything worked out – an exercise absent of disaster – has a nice "all is lost" moment that subsequently gets turned around, and throw together a just-so story that proves exactly the point you want. There isn't anything anybody can do to disprove it without particle-colliding themselves into an alternate reality. There is no way to 'science' a case study unless you have access to a parallel universe as a control. They're all just stories that short-circuit our thinking.

For every UC Health sepsis story there are a hundred systems that didn't work. Even the UC Health story itself is dubious once you start to think about it. Was it the AI that saved 211 lives? Or was it having a specialised team of clinicians watching all the at-risk patients around the clock, using live feeds? Or, was it the relationships the clinicians developed with the bedside nurses? If you'd put together that same team with the same infrastructure, but using a simpler, cheaper algorithm based on vital sign monitors, would that have done the same job? Why didn't UC Health try that first – simpler, cheaper, faster to set up – if what they wanted was to save lives?

The answer is simple: *they'd already bought a broken AI system*. They did what they had to in terms of making sure their investment did eventually save lives, but it leaves us with this unanswered question: if they had spent that same amount of money on building teams and a system for detecting sepsis, but without AI, would it have worked better or worse?

We can't know, and that's why case studies are a favoured tool by MBAs, startups, and consultants all over the world. You can just pick a story that proves what you want and ignore the hundreds that don't.

Once Generative AI becomes a broad movement in software, facts and science won't matter, and the stories will take over. Ted Nelson was writing about computers and programming in a more general way and in a different era, but he's right here too: the stories about AI software aren't scientific and trying to make them look scientific is an outrage.

That's what the AI software vendors are doing with their marketing performances that look like scientific papers, the 'studies' that are little more than sales exercises, and the entirety of their rhetoric about being on the verge of AGI – how we need to make sure those future robot gods are our slaves and not overlords. It's storytelling. Fighting that with another 'science' performance is futile. In a war of theatrics, the act with the biggest budget wins the crowd. We can chip away at the foundations with peer-reviewed papers and research that show flaws and failures, but ultimately what will decide whether this technology is a constructive addition to our field in the decades to come is the software – how well it's designed, how effective, how productive, and the long-term failures and successes in real workplaces.

That's where the three core flaws of the assistant model are going to be a problem.

I mentioned two of them earlier, *automation* and *anchoring* biases.

Automation bias comes from our tendency, as human beings, to trust machines over our own judgement.[289] This kills people and has been a major problem in aviation.[290]

Anchoring bias comes from our tendency to let the initial perceptions, thoughts, and ideas set the context for everything that follows.

AI adds a third issue: *anthropomorphism*.[291] Even the smartest people you know will fall for this effect as large language models are *incredibly* convincing.[292] These biases combined lead people to feel even *more* confident in the chatbot's work and believe that it's done a better job than it has.[293]

We're using the AI tools for *cognitive assistance*. This means that we are *specifically* using them to think less.[294] In every other industry this dynamic inevitably triggers these biases and compromises our judgement of the work done by the tools.[295] We use the assistant to think less, so we do.

These models are incredibly *fluent* and – as we saw at the start of this book – are consistently presented by their vendors as near-AGI. This triggers our instinct towards *anthropomorphism*,[See Epley, Waytz, and Cacioppo[296], which outlines three psychological triggers for anthropomorphism:

1. If you don't know how a non-human agent works, we default to thinking it works like us because that's what we have the most familiarity with.

2. "The motivation to interact effectively with nonhuman agents" causes us to attribute human characteristics and motivation.

3. Seeing agents as human-like enables "a perceived humanlike connection with nonhuman agents". making us feel like we have a fully human-level intelligence assisting us, creating an *intelligence illusion* that again hinders our ability to properly assess the work it's doing for us.[297]

Anthropomorphism, when applied to AI chatbots has been called the "Eliza effect". It was first observed by Joseph Weizenbaum when he saw how people responded to and interacted with the comparatively primitive 'AI' chatbot, *Eliza*, that he created back in 1966.

> What I had not realized is that extremely short exposures to a relatively simple computer program could induce powerful delusional thinking in quite normal people.
>
> Weizenbaum[298], *p*. 7.

Fluent chatbots create an anthropomorphism effect that sways even those who *knew* that the model was nothing more than a simplistic program, even by 1966 standards.[299]

The intelligence illusion, the conviction that these are artificial minds capable of powerful reasoning, when combined with anthropomorphism *supercharges* our automation bias. Our first response to even the most inane pablum from a language model chatbot is awe and wonder. It sounds like a real person at your beck and call! The drive to treat it as, not just a person, but an *expert* is irresistible. For most people, the incoherence, mediocrity, hallucinations, plagiarism, and biases won't register over their sense of wonder.

This anthropomorphism-induced delusion is the fatal flaw of all AI assistant and copilot systems. It all but guarantees that – even though the outcome you get from using them is likely to be worse than if you'd done it yourself, because of the flaws inherent in these models – you will feel *more* confident in it, not less.

Every human-in-the-loop and assistant-style AI system I've seen has these defects. Some of them even do their best to *exacerbate* them by making the assistant adopt a confident tone or an affable demeanour.

Those who use these chatbots are likely to get worse results and still be more confident in the resulting output than they would have in their own. Their work will suffer, but they will feel like it has improved. This is a recipe for fanatical evangelism and incredible revenue growth. It guarantees a financial bubble of some kind. The only question is its size and duration. The more effective the generative models are, the bigger the bubble. The less effective they are, the faster it'll pop.

It's the nature of bubbles to create crap. A software bubble is the flowering of a thousand first-movers – countless startups and tech companies, most of them utterly clueless about what they're working with, building the first bad iterations of what they hope is a good idea. We don't know yet what the ideal, productive AI-assisted productivity software will look like. We don't even know if it's at all possible – this technology may just end up being more harmful than not. What we do know is that we're unlikely to see many examples of "good" software based on generative models in the first generation.

Meanwhile, the tech industry will dream of exponential growth.

# A digression on infrastructure and robustness

I am Icelandic but I've lived abroad for most of the past twenty-five years. The web let me work wherever I wanted without losing touch with my friends and family. The tools the web offers gave me freedom that I couldn't have imagined when I was a child.

This worked well for a while. I've had the joy of living in a number of wonderful cities and amazing neighbourhoods and communities.

I grew older and, with age, those around you grow older with you. Some of them get sicker. A video chat doesn't fill the void you feel when somebody you care about is lying in a hospital bed. But the freedom the web provides works in the other direction as well. I could live near those I care about and the web meant I could keep doing my job no matter where I was.

I decided to move back home to Iceland. As I was preparing my move, the COVID-19 pandemic struck, and the rest of the world discovered what I had known for decades: the web abstracts distance. You can work where you want. I made my way home, despite the collapse of international airline travel. From Montréal to Toronto. Toronto to Amsterdam. Finally, I flew from Amsterdam to Iceland.

Back in Iceland, I settled in Hveragerði, a small town of about 3000 people in the south of Iceland. Keeping with the theme of my realisation that the web and related technologies meant that location mattered less, I could pick a place that suited my personal needs. It's a nice town. The weather here can be *interesting* – this is Iceland after all – which often leads to road closures in the winter. But there are three separate roads that connect this region with the capital, so even though a couple of the roads get closed due to snow or ice there's always the third. Because we know what to expect from the weather, most regions in Iceland invest in their infrastructure. We try to make sure we can keep everything going even when a bad storm hits us. There are redundancies and, for the most part, they work.

We can't say the same about the software that we have today – that we use for our work. Even though many organisations have returned

to the office, partially or fully, we are still using the same software that companies adopted for remote work. We use Google Docs, Zoom, Dropbox, or an equivalent competitor. Our files, documents, and processes are now tied to whatever app we've adopted.

If Google Docs goes down or has sporadic outages, then our work disappears with it. If our internet goes, the software blinks out. When the biggest data centre belonging to Amazon Web Services goes down, that breaks most of their services across all of their data centres, because they're all interconnected, and almost *all* of our software breaks with it.[300] Given our increasing reliance on centrally hosted software services, the impact of temporarily losing a data centre is severe, getting worse, and is now even happening because of the weather, caused by an increase in frequency and magnitude of heatwaves globally.[301]

It doesn't matter whether we ourselves are working remotely or in-office, all our software today is remote, and the connection to it can break in a thousand different ways.

*There is only one road into town.*

A global network means our software shouldn't have to be centrally located in only a few specific buildings across the world. It should be spread throughout the network, on every device that's connected to it. Our hardware devices shouldn't have to be so reliant on the internet that core features cease to function just because they can't phone home for a short while. Our information – public, private, professional – shouldn't have to be controlled, collected, and stored by only a handful of corporations.

The software we have today is undermining the strongest advantage given to us by the internet: robust and distributed reliability. Our work depends on increasingly unreliable software. Their need to be always online means you feel every hiccup in your connection. Centralisation means that when something does go wrong, it's potentially catastrophic as it affects everybody, everywhere, who is using that centralised app. This matters because things are going wrong, fast. There's political unrest. Social instability. Cold wars. Hot wars. Trade wars. A climate crisis. Data that was just normal personal data one

moment, becomes incriminating evidence a moment later when people's rights are stripped away.

The software we have isn't the software we need.

# The software industry is making the opposite of good software

Modern software is remarkably fragile. We've gone from a software ecosystem that, a few years ago, was almost completely local to one where everything is just cached – temporarily stored – at best. A decade ago what you worked with was on the computer itself. Your data was your own, and relatively safe if you kept decent care of your backup drive. The apps were yours, usually bought and paid for once – no subscription. Collaboration was always a bit tricky if you weren't a software developer – developers have somewhat decent collaborative tools in version control systems – but other people made do by using shared local servers or simply sending files over email.

This was a remarkably robust software ecosystem that tolerated all sorts of disasters, disconnections, and changes. We've dismantled it in less than a decade. Most of the apps we use for our work require an internet connection. Almost all of them are entirely cloud-based, where significant parts of the software runs on a server somewhere. Little of our work data is stored locally any more.

Generative models serve to accelerate that trend.[302] You needed 800GB just to store GPT-3, without even running it.[303] Later versions and ChatGPT are even bigger, running in parallel on multiple servers. The technology can be made to work locally, but that's not where the hype is. The hype is for the already countless "AI for X" services who are all in the cloud and are all using services from OpenAI, Microsoft, Google, or Amazon.[304] Unless one of the big tech companies breaks ranks and builds into their Operating system a solid and tested large language model that has been ethically trained on documented data sets, what we're going to get are agents and chatbots everywhere, each living in the cloud, fine-tuned for a task, and hooked up to whatever APIs the startups think are nifty. Why solve the hard problem of mak-

ing a language model safer, better integrated into your software, and more sustainably developed, when you can hook up a finicky but flashy website with OpenAI and call it a startup?

The dream the tech industry chose is not science or progress. The dream they chose is that of easy money, because that's the only dream the tech industry today is capable of seeing. Their vision is a mirage of craving.

Their want can only be met with another financial bubble, one that has to be more grand and world-changing than any other that preceded it. They crave the exponential to fulfil their dreams, but the only true exponential today's twenty-something startup founder will experience is that of the escalating Climate Crisis. That won't stop them from trying. Their hunger is likely to push them to ignore the social unrest and power shifts that AI systems cause.

The tech industry doesn't just behave with your normal corporate greed. They *want* financial bubbles. They had a taste of the euphoria with the dot-com bubble and the hunger for it never went away.

The tech industry is also, as I argued at the start of this book, full of true believers in AI. Somebody who truly believes – sincerely believes that this will all be for the best – will push past the mass unemployment, organised disinformation, and wholesale deception. They will think that it will all be worth it. Once we get through the initial "disruption", things will be better for everybody.

None of this is conducive to software design and development. It isn't a mindset that leads you to do user research, observational studies, or usability experiments. It's a drive that's taking them *away* from what most people and their communities need. Where we need robust technology, they are giving us finicky AIs that misbehave at a badly worded sentence. Where we need privacy from both corporations and potentially hostile authorities, they push further and further into recording our lives. When we need software that works on the devices we have, for as long as they last, they give us software that only works on the latest and greatest. Sometimes, as with GPT-4, the software they make even requires systems so powerful that they only exist in a couple of locations on the planet.

But, don't worry, they'll sell us access – timeshare, really – but let's call it "the cloud". It only breaks some of the time.

Nothing they do is for us, even though it's our money, our data, and our art, writing, and music they're demanding. We aren't customers to them – *we're just the people that pay*. To tech companies, we are nothing more than a resource to be tapped. A number to be boosted to pump investor interest. They are not doing us any favours. What they want from us is simple: *everything*. All culture on their servers, made by their AI. All our work happening through them, assisted by their AI. The totality of our information, mediated by their AI. A vig collected on all existence.

One of the papers I've referred to a few times in this book is *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*[305] It was the first paper to provide a cohesive and detailed overview of how large language models work, how they affect people, and the risks that they pose. This paper ultimately led to Google firing one of the key authors of the paper, Timnit Gebru, and forced other co-authors employed by Google to take their name off it.[306] This continues to this day, where Google employees seem to be routinely discouraged from working on AI fairness or ethics.[307]

When Microsoft launched Bing Chat, the first mainstream attempt to use a large language model as a front end for search – something that another co-author of *On Stochastic Parrots*, Emily M. Bender, had warned against in a separate paper titled *Situating Search*[308] – this lead to the exact outcomes they had predicted. Strange behaviour[309], threatening language[310], falsehoods[311] and lies[312] ensued. Bing Chat played out exactly the way they expected.

Of course, Microsoft did the only rational thing it could when the risks of its products were revealed: it disbanded its AI ethics and safety team[313] and rolled Bing Chat out to even more people.[314] It now plans to push towards adding AI chatbots to everything, everywhere, no matter the cost.[315]

Most of the tech organisations that had responsible AI or AI safety teams are disbanding them.[316]

They seem to think it would be a mistake to worry about risks and problems – why worry about something you can probably fix?[317] Who cares about the harm it does in the meantime?

Safe, for the tech industry, is too slow when you hunger for a bubble and want to ship more software, to more people, as fast as you can.

Designers of software user interfaces often imagine deliberately bad designs as an exercise – a way of demonstrating the principles of their craft by exploring their opposites. It's a good way of demonstrating *why* a design principle matters, and it can provide tactile examples of who benefits from it and how.

If you asked me to imagine the software that would be the *opposite* of what we need as a society…

That app would look remarkably like ChatGPT.

# 21

# Cracks and shadows

*“I’ve never before experienced such a disparity between how much potential I see in a technology and how little I trust those who are making it.”*

# Afterword to the first edition

*"You can't just end it like that. It's too much. You have to give the reader a straw to grasp. Otherwise, you'll lose them."*

This was the note my editor left to me on an early version of this book: people need something positive to reach for, otherwise the book will be too bleak.

It may surprise you to hear that I agreed with her. "This is too much," has been a worry of mine throughout the writing process – a concern that's been driving many of the decisions I made while working on this book. It's the reason why I chose to focus on business risks and not social, political, or cultural harm. It's why I stepped back from covering the numerous horrifying facts about the field of AI research and development that I discovered in my research.

I had to step back away from it, because it's not *all* I discovered in my research. Generative models are one of the more promising technologies to come out of software in recent years, but what I discovered – and outlined in this book – is that it's also one of the more flawed. To find the cracks in the walls, you need to look away from the shadows.

I originally came at "generative AI" with a hopeful curiosity. The generative part never appealed much. For some reason, the text others found convincing only made me cringe. I hesitate to claim superior judgement on my part – much of it is down to ChatGPT's ability to sound like a particular kind of person who I've learned from bitter experience to avoid. Most of these language model chatbots sound like *that* guy. You know the guy. Everybody's worked with one.

The generated images, as well, always had obvious flaws. I've been a serious photographer for three decades, started my career on the production side of TV, and did my degrees at an Art, Media, and Design faculty. That I'm pickier than many others about images is entirely unsurprising. Code generation looked promising, but even before I got

around to try GitHub Copilot, it was obvious that the technology wasn't a clear-cut "win".[318] These tools have capabilities for converting style and structure, both visual and textual, and surely the code generation, at least, would improve. Programming languages seemed like a perfect fit.

When I began my research, the technology was imperfect but promising. It hinted at great possibilities.

In the words of the writer Cyril Connolly, promise is a fatal word, "half-bribe and half-threat", that serves to elevate you just enough to bring you within reach of the hangman's noose: "whom the gods wish to destroy they first call promising."[319]

Everything hinges on generative models fulfilling their promise. The tech industry has raised it to stand above us all.

What they've raised is broken.

I dug into the research, but, wherever I looked, every time I found concrete data and solid experiments, I found disappointment. Fabrications seem to increase in frequency with model size. Summaries are frequently inaccurate. The models have a tendency to copy training data verbatim, which is disastrous for both software projects and serious writing. The *quality* of the work seems to have an upper bound that is broadly mediocre. They keep knocking down individual defects and specific falsehoods, but the overall tone is still *that guy*. The training data set is one long open wound: every part of it represents a potential security exploit, bias, or legal liability. The "reasoning" is fragile and frequently nonsensical. People are driven to accept it through the Eliza Effect and automation bias, but little of it would pass if you looked at it impartially. The models are prone to the unhinged and disquieting. Preventing your customers from misusing them looks futile. Prompts seem impossible to secure.

The tech industry has built the AI edifice and raised it up as the future of computing. It looks like an imposing whole from a distance, but as we walk towards it, the cracks begin to yawn.

Perfection isn't innate; it's something to aspire to. All work begins raw and needs refinement. It isn't surprising that generative models

are quite broken this early in its development. What's surprising is how little the tech industry seems to care.

I've never before experienced such a disparity between how much potential I see in a technology and how little I trust those who are making it.

Diffusion and language models look like powerful tools, *at first*. The technology presents us with so many opportunities – such power – that I have to go all the way back to the early days of the web to find its equal in terms of seeming potential.

Machine learning has already delivered many improvements to our work and lives. It's the driving force behind many of the innovations we've come to rely on in our photography. It's transformed special effects work. The opportunities it creates for working with audio are amazing. Language and diffusion models are intriguing.

But they are being implemented in the worst possible ways.

- None of these tools should be presented as persons – as beings with intelligence and personality similar to a human being. That triggers too many psychological biases and effects that lead us astray.

- They shouldn't be trained on people's data without their permission, and training data sets that large are as much a liability as a benefit.

- The models themselves seem much less functional than promised.

- We need to immediately investigate how we can mitigate the harm they do to our digital commons – the writing and code that we share with each other and has become one of the most important foundations of modern culture and technology.[320]

Instead of taking their defects seriously, the tech industry is putting a weight on generative models that the technology is too weak to carry.

I can only hope that, at some point, somebody who is both responsible and has financial resources will see the same potential in the technology that I do, and the same need for a safe user experience.

The potential it has for software development is *enormous*. I just wish that the first instinct of the industry wasn't to magnify everything that's already wrong with programming: the emphasis on churning out mediocre code, the lack of concern for security and accessibility, the disregard for maintenance costs, the poor understanding of what software is for and how people use it. We should be using these innovations to do *better*, not to do bad things faster.

The promise of generative AI doesn't have to be a "half-bribe and half-threat". It could take computing to new heights and us with it.

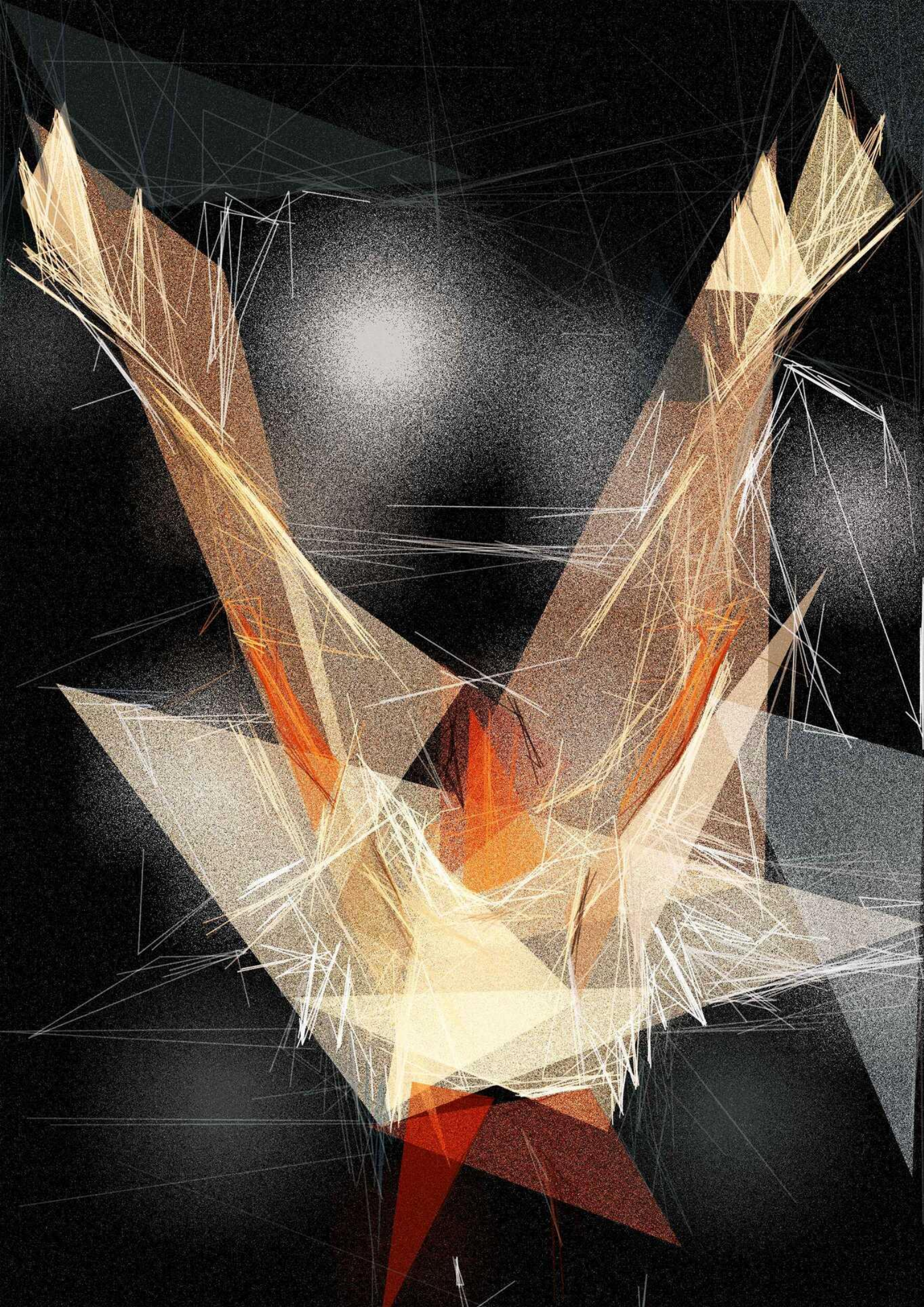
I have to hope that we will get there, someday.

Baldur Bjarnason,
*Hveragerði, April 2023*

22

# The era of centralised compliance engines has begun

> *Generative models impose cushioned compliance guardrails onto every task*



# Afterword to the second edition

I included the original afterword so you could see just how wrong I was. Now, a year later, these large generative models seem irreparably broken and inherently flawed. We are not going to get "there" any time soon.

I attempted several times to revise this book in the intervening year before finally settling on the edition you're now reading. Every time I attempted to update it, I discovered that not only were the risks unchanged, the industry had somehow found a way to make them worse.

Deep fake sexual imagery is now a common form of abuse in schools.

Fabrications, if anything, now seem more common than before.

As search engine results decline, tech companies replace them with *retrieval-augmented generation*[321] that is *fundamentally based around processing the now-failing search engine results*. Not only is it incapable of *fixing* the issues generative models have created in search, it *exacerbates them by adding fabrications and biases that weren't a part of the original results*.

The risks remain the same and their impact only grows as generative models get integrated into more products and forced onto consumers.

Updating the book seemed pointless. None of the risks I pointed out were being addressed so the risks described in the first edition didn't need to be changed.

As the safety and quality of the tech industry's software products has continued to decline, what became untenable wasn't the research I cited or the risks I described, but my *inherent optimism*.

The clear disregard for the consumer that the industry is exhibiting and the political changes we're seeing in the world around us mean there is, in my sincere personal opinion, no chance that the positive

impact generative models *might* have will ever come close to outweighing their *current and escalating* negative impact.

This is a bad technology with dire consequences, the worst of which is a risk that I left out of the first edition because it felt too political, too pessimistic, and all too worrying.

# Generative models are compliance engines that undermine team autonomy

Generative models impose cushioned compliance guardrails onto every task they are applied to. They are a form of control that's external to that of your organisation and that control is "cushioned" because even though the output is variable and flexible to a degree, it also imposes firm boundaries on what can and cannot be done.

These tools can't take your strategy or goals into consideration in any meaningful way unless they are already baked into it.

This can have a high cost as the biggest impact on productivity and effectiveness *comes from the organisation itself* and not from the productivity of an individual worker.

As observed by Tom DeMarco and Timothy Lister in their book *Peopleware*[322] where they studied the performance of programmers of varying experience across a number or organisations:

> The fact that they had nearly identical performances suggests that the wide spread of capabilities observed across the whole sample may not apply within the organization: Two people from the same organization tend to perform alike. (p. 46)

And W. E. Deming – one of the pioneers of modern management theory – wrote about this in *Out of the Crisis*:[323]

> I should estimate that in my experience most troubles and most possibilities for improvement add up to the proportions something like

> this: 94% belongs to the system (responsibility of management), 6% special. (p. 270)

The organisation – the system maintained by management – is the biggest source of variation in performance, quality, and reliability for most companies.

Integrating generative models – trained on data you haven't selected, fine-tuned to specifications out of your control, to attain product goals that aren't yours – means you're integrating a management and product development strategy that you haven't chosen and comes with cushioned but firm guardrails: they outright limit how people in your organisation are and aren't allowed to work.

The job of management is to manage the *system* of the organisation and generative models impose an *inflexible* system of their own. A lot has been written how "AI" automation is threatening labour, but the workers who are the most threatened by heavy use of generative models are *managers*. Organisational use of generative models directly undermine the management's ability to do its job.

Consider the difference between two different applications of machine learning, both in the context of image processing: automated masking and generative fills.

Masking is a core task in video and special effects work. It's tedious, slow, and requires considerable effort. But once it's done, the possibilities are endless. The mask tells the software were a particular object or region in the image or video is, so it can be modified. It does not dictate what you do with the object or region. You can use the mask to composite subjects with a new background, to correct or adjust colours, to change lighting, to emphasise or de-emphasise – what it's for is unbounded.

On the spectrum between the two poles of technology as described by Ursula M. Franklin[324], *holistic* versus *prescriptive*, automating masking with machine learning is more on the holistic side than prescriptive. It's still partially prescriptive – there will be objects, backgrounds, and foregrounds it won't recognise – but it leans a bit closer to "notions of craft", to use Franklin's words, than to prescriptive automation.

Generative fill, however, will only generate what can be found with enough frequency in its training data for it to model, and it is unlikely to render anything that the proprietor has "fine-tuned" out of existence. If the vendor or the government whose rules the vendor follows consider the very existence of same-sex couples to be obscene, the generative fill is unlikely to render those images for you. Using the technology puts guardrails on what you can express. It is a *prescriptive* technology. This applies to both diffusion models that generate images and Large Language Models that generate text. They prescribe what's normal, acceptable, and desired.

As Ursula Franklin noted in *The Real World of Technology*:

> While we should not forget that these prescriptive technologies are often exceedingly effective and efficient, they come with an enormous social mortgage. The mortgage means that we live in a culture of compliance, that we are ever more conditioned to accept orthodoxy as normal, and to accept that there is only one way of doing "it" (p. 17).

That "there is only one way of doing" is the opposite of the flexibility and autonomy that is essential, specifically, to *product development*. Autonomy leads to better and faster teams. As Smith and Reinertsen note in the book *Developing Products in Half the Time*[325]:

> When we push decision making down to the team level, decisions can be made both quicker and better. Often the team is closer to the underlying market and technical data needed to make good decisions. This is particularly true when we are dealing with rapidly moving markets and rapidly changing technologies (p. 186)

This goes beyond the guardrails described above. Earlier I explained how *anchor* and *automation* biases lead people to accept results that are genuinely worse than what they'd accomplish themselves, but those biases also sabotage your team's autonomy. The purpose of cognitive automation is to think less. *That's what it's for.* This makes teams much less likely to operate in a truly autonomous fashion.

This means that the performance and productivity cost of these prescriptive generative systems can be *much* greater than most people

expect. You might attain a 5-20% performance boost in *individual* productivity when it comes to generating text, images, and lines of code, but the cost of the ensuing complexity, the rigid prescribed, prefabricated solutions, and the loss of team autonomy can have dire consequences.

How dire?

As DeMarco and Lister noted in *Peopleware*:

> Over the whole sample, the best organization (the one with the best average performance of its representatives) worked more than ten times faster than the worst organizations. In addition to their speed, all competitors from the fastest organization developed code that passed the major acceptance test (p. 46-7).

The difference between a dysfunctional organisation and one that is not can be ***tenfold***.

You might see a 5% bump in output on an individual level, but at the same time trigger an outright *death cycle* for the product, team, or even the organisation as a whole.

Generative models are a *threat* to our economy, not a boon.

But that isn't their greatest danger.

# Generative models centralise *political* compliance to authoritarianism

Only a handful of organisations in the world are capable of fully developing a Large Language Model and deliver it as a service or product that organisations can integrate into their processes and systems. There are maybe six or so organisations in the world that control most, if not all, of the generative models being used in production. Most of these organisations are based in the United States.

Tech companies based in the United States have a history of complying with authoritarian governments in order to gain or maintain access to their markets.[Madhumita Murgia and Max Seddon[326]; Mathew Ingram[327]; Erin Kissane[328];]

These companies are *very* likely to continue their history of compliance. No matter where in the world authoritarians are in control as long as they have a market that is lucrative these companies will comply, and if you've handed them control over your language, images, and writing, you will automatically comply as well.

By using generative models in your work you're putting yourself and your organisation in a position where an authoritarian government could dictate the very language you use in your work, collaboration, sales, and marketing *without lifting a finger themselves*. You've done all the work of integrating external centrally-managed political guardrails into your processes and systems for them.

In an earlier chapter, I noted that research had revealed that you could "poison" the training data of a language model to inject falsehoods or to make it always associate positive or negative sentiment with a specific word or phrase.

This kind of "poisoning" can come from the companies themselves, working for a government they support.

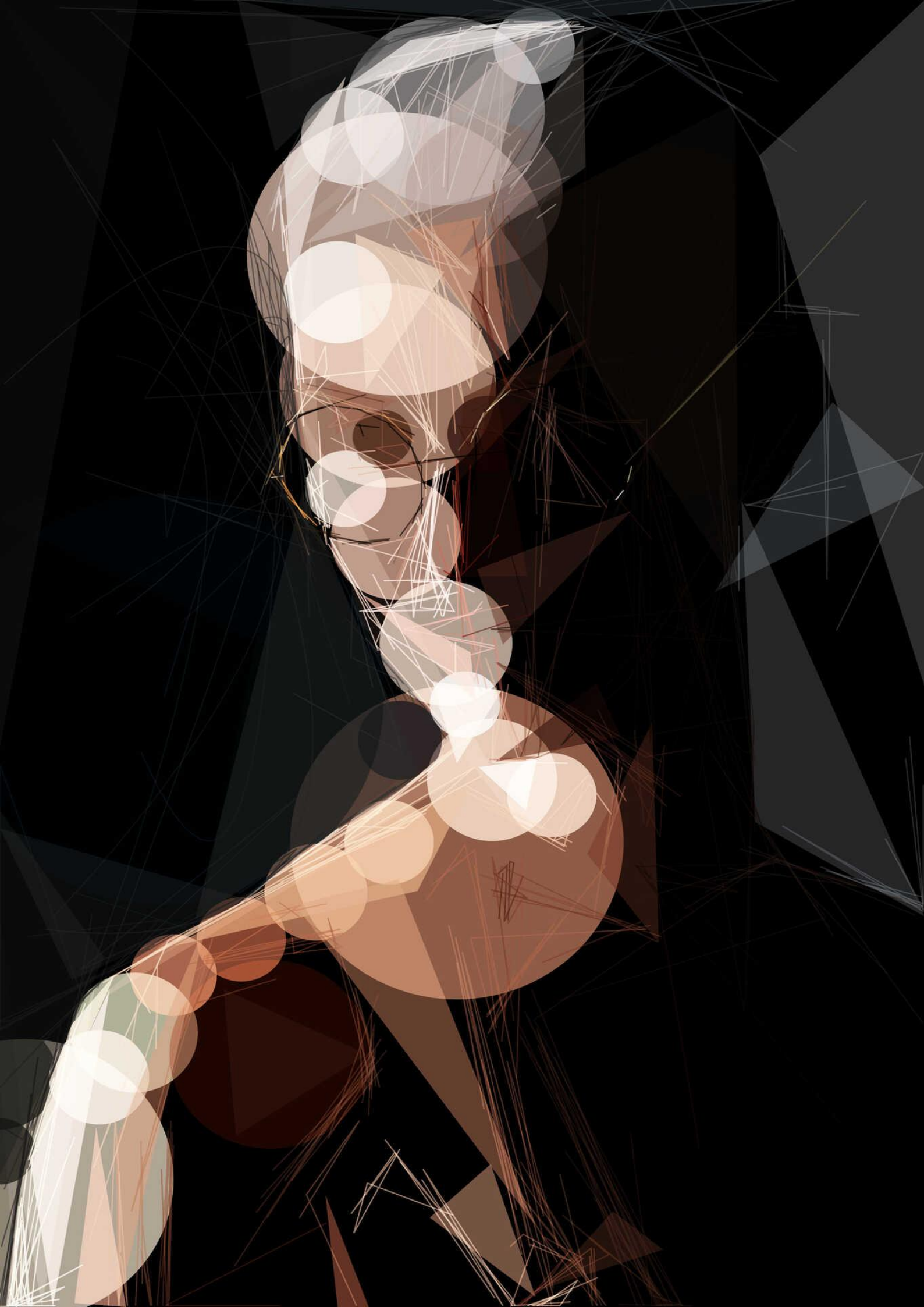
They might even call it "fine-tuning" and claim that it's done to protect you.

The era of centralised compliance engines has begun and anybody who participates is complicit in what's coming.

Baldur Bjarnason,
*Hveragerði, November 2024*

23

# Further reading



# Recommended reading

Some of the best writing on AI today takes the form of papers and essays. The following helped me think clearly about generative AI and many of the issues it is creating. I hope they'll be as helpful to you.

# The one book you absolutely should read

Dan McQuillan was well ahead of anybody else in understanding just how effective "AI" was in enforcing an unconscious political compliance. His book *Resisting AI: An Anti-fascist Approach to Artificial Intelligence*[329] should be on everybody's reading list, whether they're curious about tech or not, because "AI" is fundamentally a political project.

# Academic papers

*On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*: if you're going to read one thing on this list, make it this paper by Emily M. Bender, Timnit Gebru, and others. This is the paper that got Timnit Gebru fired from Google for warning about the risks and dangers of large language models. It does a good job of explaining what these models are, what they do, how they can be deceptive, and how they are likely to be abused.

*Why AI is Harder Than We Think*. This paper by Melanie Mitchell outlines some of the core dysfunctions of the field of AI research that causes it to regularly veer into dead-ends.

*The Fallacy of AI Functionality*. One of the biggest issues with AI systems is that *they don't do what the vendors claim they do*. This paper is an *exhaustive* survey of the many ways current AI systems fail. It's an

appeal to regulators and policy-makers to focus on current functionality of AI systems over hypothetical future capabilities.

*Situating Search*. Tech companies are *insistent* on using large language models as front ends to search engines. This paper, by Emily M. Bender and Chirag Shah, goes into considerable detail on why that is an extremely bad idea.

*Talking About Large Language Models*. This paper, by Murray Shanahan, digs into how some of the language we use with these models often misleads us.

*Indicators and Criteria of Consciousness in Animals and Intelligent Machines: An Inside-Out Approach*. This paper is a fascinating foray into possible avenues for discovering and studying consciousness in non-human subjects: animals and, maybe one day, AGI. It outlines a good deal of neuroscientific evidence towards the possibility that many animals possess the capacity for consciousness. If you aren't afraid of a little bit of biology, it's an excellent read.

*Physiognomic Artificial Intelligence*. AI software companies and researchers are revitalising pseudoscience at scale. Some of it represents a significant regression for society and community safety. This also demonstrates just how deep pseudoscientific ideas have penetrated into the field of AI research.

*SoK: On the Impossible Security of Very Large Foundation Models*. This is another interesting paper where the one of the lead authors, El-Mahdi E-Mhamdi, is a researcher who felt forced to leave Google due to their unwillingness to publish papers critical of AI. It outlines in quite a bit of mathematical detail why it's impossible to make a Large Language Model 'safe' in pretty much every sense of the word: security, bias, offensive content, and model integrity.

# On the AI financial bubble

*The AI Crowd is Mad* argues that investors have poured money into AI with over-inflated expectations of what the technology can do in the medium-term. Meanwhile, the computing costs for these systems remains extraordinarily high. What should worry investors is the possi-

bility that their expectations might never be fulfilled. Even a delay of a few years might be the difference between success and insolvency.

*AI Looks Like a Bubble* another essay pointing out that we have a fast-inflating financial bubble surrounding AI. Evan Armstrong highlights the mania-driven swings in stock price—the AI stock bump—that is characteristic to bubbles. The important point he makes is that even if you assume that "AI" tech broadly works and is the revolutionary technical innovation investors are claiming, that doesn't mean that it's a good investment. The analogy being power companies. Electricity was a technological revolution, but power "is a terrible business to be in. Power companies typically don't do well unless they reach monopoly status."

The bubble is likely to last a long while, which will make the pop particularly violent when it happens.

In *The Reverse-Scooby-Doo Theory of Tech Innovation*, David Karpf argues that posturing about the inevitability of a technology in order to inflate and maintain a financial bubble is a time-honoured tradition in tech. They claim its inevitable, until it pops, and then they blame regulators or the government. He calls it "the reverse-Scooby-Doo", "we would've gotten away with it, if not for those meddling regulators!" This posturing serves a dual purpose:

1. Make regulators and governments hesitant to intervene, even with clear evidence of market abuses or outright lawbreaking. This works surprisingly often.

2. Protect the tech industry's reputation as makers of technological miracles, and preempt the realisation that most of what it makes kind of doesn't work.

# Arts and Labour

*The stupidity of AI*. The artist James Bridle writes about some of the inherent limitations of generative AI. "The belief in this kind of AI as ac-

tually knowledgeable or meaningful is actively dangerous. It risks poisoning the well of collective thought, and of our ability to think at all."

*The Great Replacement (Not That One, the Real One)*. The author Cat Valente writes in a very engaging way about what art and writing is for.

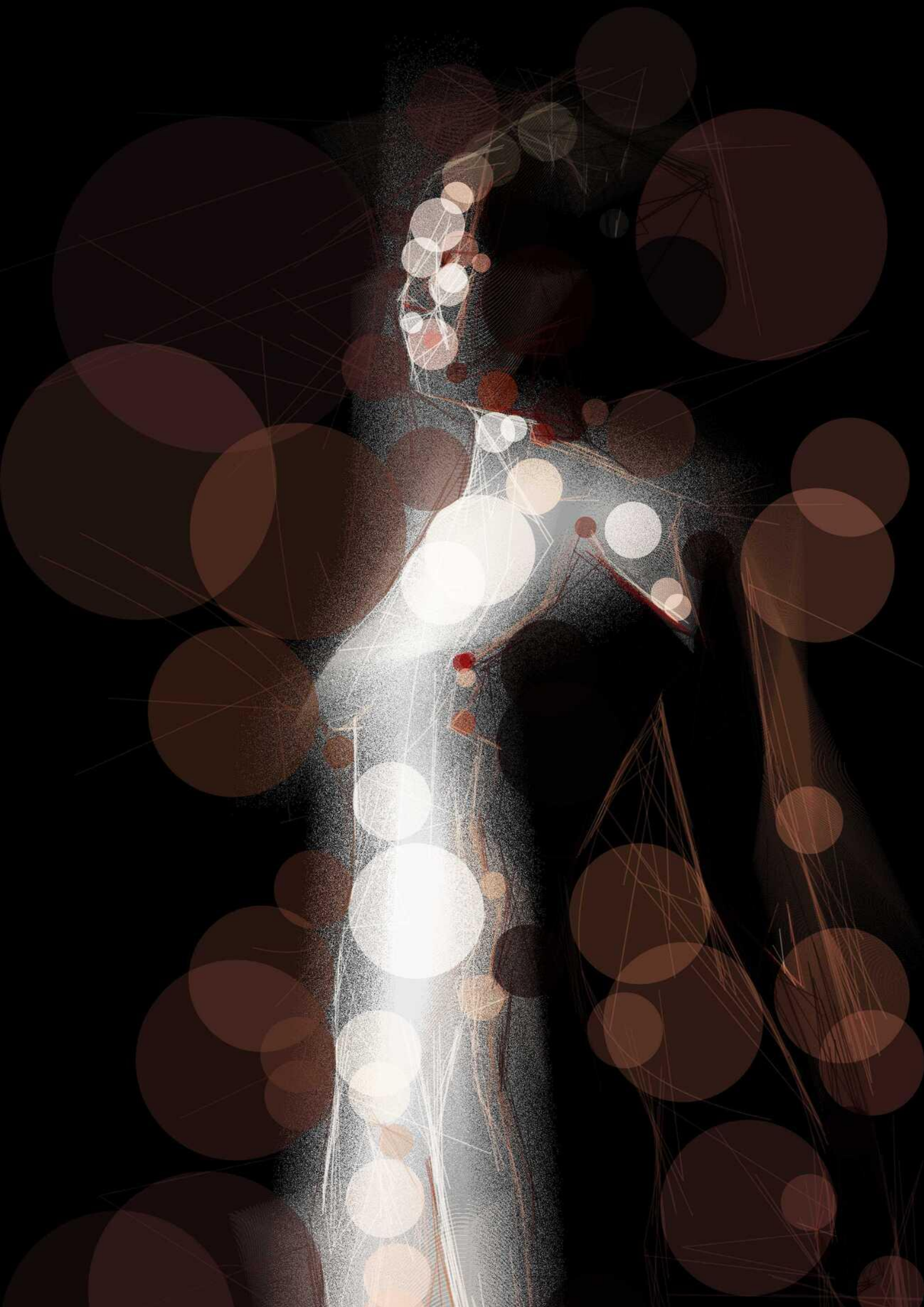
*You Are Not a Parrot. And a chatbot is not a human.* A profile on Emily M. Bender by New York Magazine.

*Collections: On ChatGPT*. The historian Bret Deveraux digs into ChatGPT in an academic context and finds it to be less capable than advertised. "In a very real sense then, ChatGPT cannot write an essay. It can imitate an essay, but because it is incapable of the tasks which give an essay its actual use value (original thought and analysis), it can only produce inferior copies of other writing."

Finally, professor Iris van Rooij has collected "Critical lenses on 'AI'", a useful resource on articles and media that's critical of the AI bubble.

24

# References



"A Little History of the World Wide Web." Accessed April 6, 2023. https://www.w3.org/History.html.

Abid, Abubakar, Maheen Farooqi, and James Zou. "Persistent Anti-Muslim Bias in Large Language Models." arXiv, January 2021. https://doi.org/10.48550/arXiv.2101.05783.

Ackerson, Noble. "GPT Is an Unreliable Information Store." *Medium*, February 2023. https://medium.com/@nobleackerson/chatgpt-insists-i-am-dead-and-the-problem-with-language-models-db5a36c22f11.

"AI-Generated Art: Who Owns the Copyright?" Web_content. *The Lighthouse*, January 2020. https://lighthouse.mq.edu.au/article/december-2019/AI-generated-art-who-owns-the-copyright.

Alba, Davey, and Julia Love. "Google's Rush to Win in AI Led to Ethical Lapses, Employees Say." *Bloomberg.com*, April 2023. https://www.bloomberg.com/news/features/2023-04-19/google-bard-ai-chatbot-raises-ethical-concerns-from-employees.

Al-Sibai, Noor. "Amazon Begs Employees Not to Leak Corporate Secrets to ChatGPT." *Futurism*, 2023. https://futurism.com/the-byte/amazon-begs-employees-chatgpt.

Altman, Sam. "Planning for AGI and Beyond," February 2023. https://openai.com/blog/planning-for-agi-and-beyond.

"Analyzing the Legal Implications of GitHub Copilot - FOSSA." *Dependency Heaven*, July 2021. https://fossa.com/blog/analyzing-legal-implications-github-copilot/.

"Anchoring Bias." *The Decision Lab*. Accessed April 3, 2023. https://thedecisionlab.com/biases/anchoring-bias.

Atleson, Michael. "Keep Your AI Claims in Check." *Federal Trade Commission*, February 2023. https://www.ftc.gov/business-guidance/blog/2023/02/keep-your-ai-claims-check.

Atwell, Elaine. "GitHub Copilot Isn't Worth the Risk." *Kolide*, February 2023. https://www.kolide.com/blog/github-copilot-isn-t-worth-the-risk.

Bagdasaryan, Eugene, and Vitaly Shmatikov. "Spinning Language Models: Risks of Propaganda-As-A-Service and Countermeasures." In 2022 IEEE Symposium on Security and Privacy (SP), 769–86, 2022. https://doi.org/10.1109/SP46214.2022.9833572.

Bai, Ching-Yuan, Hsuan-Tien Lin, Colin Raffel, and Wendy Chih-wen Kan. "On Training Sample Memorization: Lessons from Benchmarking Generative Modeling with a Large-Scale Competition." In *Proceedings of the 27th* ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2534–42, 2021. https://doi.org/10.1145/3447548.3467198.

Bailey, Jonathan. "The Wave of AI Lawsuits Have Begun." *Plagiarism Today*, January 2023. https://www.plagiarismtoday.com/2023/01/17/the-wave-of-ai-lawsuits-have-begun/.

Bang, Yejin, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, et al. "A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity." arXiv, February 2023. https://doi.org/10.48550/arXiv.2302.04023.

"Barnum Effect." *Wikipedia*, June 2023. https://en.wikipedia.org/w/index.php?title=Barnum_effect&oldid=1161115161.

Barr, Kyle. "GPT-4 Is a Giant Black Box and Its Training Data Remains a Mystery." *Gizmodo*, March 2023. https://gizmodo.com/chatbot-gpt4-open-ai-ai-bing-microsoft-1850229989.

Bartal, Inbal Ben-Ami, Jean Decety, and Peggy Mason. "Helping a Cagemate in Need: Empathy and Pro-Social Behavior in Rats." *Science* (*New York, N.Y.*) 334, no. 6061 (December 2011): 1427–30. https://doi.org/10.1126/science.1210789.

Bartoletti, Ivana. "ChatGPT Gender Bias." *Twitter*, March 2023. https://twitter.com/IvanaBartoletti/status/1637401609079488512.

Bastian, Matthias. "Stable Diffusion V2 Removes NSFW Images and Causes Protests." *THE DECODER*, November 2022. https://

the-decoder.com/stable-diffusion-v2-removes-nude-images-and-causes-protests/.

Bekoff, Marc. "Scientists Conclude Nonhuman Animals Are Conscious Beings: Didn't We Already Know This? Yes, We Did." August 2012. https://www.psychologytoday.com/us/blog/animal-emotions/201208/scientists-conclude-nonhuman-animals-are-conscious-beings.

Belanger, Ashley. "Thousands Scammed by AI Voices Mimicking Loved Ones in Emergencies." *Ars Technica*, March 2023. https://arstechnica.com/tech-policy/2023/03/rising-scams-use-ai-to-mimic-voices-of-loved-ones-in-financial-distress/.

Bender, Emily M., and Batya Friedman. "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science." *Transactions of the Association for Computational Linguistics* 6 (2018): 587–604. https://doi.org/10.1162/tacl_a_00041.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445922.

Bender, Emily M., and Alexander Koller. "Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–98. Online: Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.acl-main.463.

Bender, Emily M., and Chirag Shah. "All-Knowing Machines Are a Fantasy." *IAI TV - Changing How the World Thinks*, December 2022. https://iai.tv/articles/all-knowing-machines-are-a-fantasy-auid-2334.

Bensinger, Greg. "ChatGPT Launches Boom in AI-Written e-Books on Amazon." *Reuters*, February 2023. https://

www.reuters.com/technology/chatgpt-launches-boom-ai-written-e-books-amazon-2023-02-21/.

Biddle, Sam. "U.S. Military Makes First Confirmed OpenAI Purchase for War-Fighting Forces." *The Intercept*, October 2024. https://theintercept.com/2024/10/25/africom-microsoft-openai-military/.

"BingBang: The AAD Misconfiguration That Led to Bing.com Results Manipulation and Account Takeover Explained." *Wiz.io*, March 2023. https://www.wiz.io/blog/azure-active-directory-bing-misconfiguration.

Birch, Jonathan, Alexandra K. Schnell, and Nicola S. Clayton. "Dimensions of Animal Consciousness." *Trends in Cognitive Sciences* 24, no. 10 (October 2020): 789–801. https://doi.org/10.1016/j.tics.2020.07.007.

Birhane, Abeba, Vinay Uday Prabhu, and Emmanuel Kahembwe. "Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes." arXiv, October 2021. https://doi.org/10.48550/arXiv.2110.01963.

Bourtoule, Lucas, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. "Machine Unlearning." arXiv, December 2020. https://doi.org/10.48550/arXiv.1912.03817.

Branco, Ruben, António Branco, João António Rodrigues, and João Ricardo Silva. "Shortcutted Commonsense: Data Spuriousness in Deep Learning of Commonsense Reasoning." In *Proceedings of the* 2021 Conference *on* Empirical Methods *in* Natural Language Processing, 1504–21. Online; Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. https://doi.org/10.18653/v1/2021.emnlp-main.113.

Brereton, Dmitri. "Bing AI Can't Be Trusted," February 2023. https://dkb.blog/p/bing-ai-cant-be-trusted.

Brice, Andy. "How Much Code Can a Coder Code?" *Successful Software*, February 2017. https://successfulsoftware.net/2017/02/10/how-much-code-can-a-coder-code/.

Brown, Gavin, Mark Bun, Vitaly Feldman, Adam Smith, and Kunal Talwar. "When Is Memorization of Irrelevant Training Data Necessary for High-Accuracy Learning?" In *Proceedings of the 53rd* Annual ACM SIGACT Symposium *on* Theory *of* Computing, 123–32. STOC 2021. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3406325.3451131.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, et al. "Sparks of Artificial General Intelligence: Early Experiments with GPT-4." arXiv, March 2023. https://doi.org/10.48550/arXiv.2303.12712.

"Bumble Bees Can Teach Each Other to Solve Problems - Study." *The Jerusalem Post JPost.com*. Accessed April 4, 2023. https://www.jpost.com/science/article-734044.

Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Proceedings of the* 1st Conference *on* Fairness, Accountability *and* Transparency, 77–91. PMLR, 2018. https://proceedings.mlr.press/v81/buolamwini18a.html.

Burgess, Matt. "The Hacking of ChatGPT Is Just Getting Started." *Wired*. Accessed April 14, 2023. https://www.wired.com/story/chatgpt-jailbreak-generative-ai-hacking/.

Bush, Vannevar. "As We May Think." *The Atlantic*, July 1945. https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/.

Buterin, Vitalik. "Bitcoin Is Not Losing Its Soul – Or, Why the Regulation Hysteria Is Missing the Point." *Bitcoin Magazine - Bitcoin News, Articles and Expert Insights*, June 2013. https://bitcoinmagazine.com/culture/bitcoin-is-not-losing-its-soul-or-why-the-regulation-hysteria-is-missing-the-point-1370302865.

Cao, Ziqiang, Furu Wei, Wenjie Li, and Sujian Li. "Faithful to the Original: Fact Aware Neural Abstractive Summarization."

arXiv, November 2017. https://doi.org/10.48550/arXiv.1711.04434.

Carlini, Nicholas. "Poisoning the Unlabeled Dataset of {Semi-Supervised} Learning," 1577–92, 2021. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-poisoning.

Carlini, Nicholas, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. "Extracting Training Data from Diffusion Models." arXiv, January 2023. https://doi.org/10.48550/arXiv.2301.13188.

Carlini, Nicholas, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. "Quantifying Memorization Across Neural Language Models." arXiv, February 2022. https://doi.org/10.48550/arXiv.2202.07646.

Carlini, Nicholas, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. "Poisoning Web-Scale Training Datasets Is Practical." arXiv, February 2023. https://doi.org/10.48550/arXiv.2302.10149.

Carlini, Nicholas, and Andreas Terzis. "Poisoning and Backdooring Contrastive Learning." arXiv, March 2022. https://doi.org/10.48550/arXiv.2106.09667.

Carlini, Nicholas, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, et al. "Extracting Training Data from Large Language Models." arXiv, June 2021. https://doi.org/10.48550/arXiv.2012.07805.

Cassidy, Benjamin. "The Twisted Life of Clippy." *Seattle Met*, August 2022. https://www.seattlemet.com/news-and-city-life/2022/08/origin-story-of-clippy-the-microsoft-office-assistant.

"Chatbots, Deepfakes, and Voice Clones: AI Deception for Sale." *Federal Trade Commission*, March 2023. https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale.

Chen, Sihao, Fan Zhang, Kazoo Sone, and Dan Roth. "Improving Faithfulness in Abstractive Summarization with Contrast

Candidate Generation and Selection." In *Proceedings of the* 2021 Conference *of the* North American Chapter *of the* Association *for* Computational Linguistics: Human Language Technologies, 5935–41. Online: Association for Computational Linguistics, 2021. https://doi.org/10.18653/v1/2021.naacl-main.475.

Chen, Xinyun, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning." arXiv, December 2017. https://doi.org/10.48550/arXiv.1712.05526.

Christensen, Clayton M. *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. Boston, Massachusetts: Harvard Business Review Press, 2024.

"Cold Reading." *Wikipedia*, August 2023. https://en.wikipedia.org/w/index.php?title=Cold_reading&oldid=1168265112.

Coles, Cameron. "3.1% of Workers Have Pasted Confidential Company Data into ChatGPT." *Cyberhaven*, February 2023. https://www.cyberhaven.com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt/.

Connolly, Cyril. *Enemies of Promise*. Rev. ed., University of Chicago Press ed. Chicago: University of Chicago Press, 2008.

Conocimiento, Ventana al. "Cold Fusion: Anatomy of a Scientific 'Fraud'." *OpenMind*, March 2019. https://www.bbvaopenmind.com/en/science/physics/cold-fusion-anatomy-of-a-scientific-fraud/.

Cook, Noah. "ChatGPT Creates a Fake History of Ukraine." *Med-Mastodon*, February 2023. https://med-mastodon.com/@UncivilServant/109867060485276220.

"Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence." *Federal Register*, March 2023. https://www.federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance-works-containing-material-generated-by-artificial-intelligence.

Council, Stephen. "OpenAI Admits Some Premium Users' Payment Info Was Exposed." SFGATE, March 2023. https://

www.sfgate.com/tech/article/chatgpt-openai-payment-data-leak-17858969.php.

Cox, Joseph. "'Disrespectful to the Craft:' Actors Say They're Being Asked to Sign Away Their Voice to AI." *Vice*, February 2023. https://www.vice.com/en/article/5d37za/voice-actors-sign-away-rights-to-artificial-intelligence.

Crypto, CRYPTOINSIGHT PRO. "Stable Diffusion 2.0 Has "Forgotten" How to Generate NSFW Content." Substack newsletter. *CRYPTOINSIGHT.PRO Crypto NFT DeFi*, November 2022. https://cryptoinsightpro.substack.com/p/technologies2.

Dai, Jiazhu, and Chuanshuai Chen. "A Backdoor Attack Against LSTM-Based Text Classification Systems." arXiv, June 2019. https://doi.org/10.48550/arXiv.1905.12457.

Dastin, Jeffrey. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women." *Reuters*, October 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

Dastin, Jeffrey, and Jeffrey Dastin. "Amazon Extends Moratorium on Police Use of Facial Recognition Software." *Reuters*, May 2021. https://www.reuters.com/technology/exclusive-amazon-extends-moratorium-police-use-facial-recognition-software-2021-05-18/.

"Data Breach Reporting Requirements Explained [2022] GDPR Register," December 2021. https://www.gdprregister.eu/gdpr/data-breach-notification-requirements/.

DeGrave, Alex J., Joseph D. Janizek, and Su-In Lee. "AI for Radiographic COVID-19 Detection Selects Shortcuts over Signal." *Nature Machine Intelligence* 3, no. 7 (July 2021): 610–19. https://doi.org/10.1038/s42256-021-00338-7.

DeMarco, Tom, and Tim Lister. *Peopleware: Productive Projects and Teams (3rd Edition)*. 3rd ed. Addison-Wesley Professional, 2013.

Deming, W. Edwards. *Out of the* Crisis. The MIT Press, 2018. https://doi.org/10.7551/mitpress/11457.001.0001.

Dentella, Vittoria, Elliot Murphy, Gary Marcus, and Evelina Leivada. "Testing AI Performance on Less Frequent Aspects of Language Reveals Insensitivity to Underlying Meaning." arXiv, February 2023. https://doi.org/10.48550/arXiv.2302.12313.

Di, Jimmy Z., Jack Douglas, Jayadev Acharya, Gautam Kamath, and Ayush Sekhari. "Hidden Poison: Machine Unlearning Enables Camouflaged Poisoning Attacks." arXiv, December 2022. https://doi.org/10.48550/arXiv.2212.10717.

Diakopoulos, Nick. "Can We Trust Search Engines with Generative AI? A Closer Look at Bing's Accuracy for News Queries." *Medium*, February 2023. https://medium.com/@ndiakopoulos/can-we-trust-search-engines-with-generative-ai-a-closer-look-at-bings-accuracy-for-news-queries-179467806bcc.

"Distribution of Physicians Across Regions by Gender Worldwide 2000-2018." *Statista*. Accessed April 25, 2023. https://www.statista.com/statistics/1099789/distribution-of-physicians-across-regions-worldwide-by-gender/.

"Don't Believe ChatGPT - We Do NOT Offer a "Phone Lookup" Service," February 2023. https://blog.opencagedata.com/post/dont-believe-chatgpt.

"DoNotPay - The World's First Robot Lawyer." Accessed April 6, 2023. https://donotpay.com/.

"Dot-Com Bubble." *Wikipedia*, April 2023. https://en.wikipedia.org/w/index.php?title=Dot-com_bubble&oldid=1147856945.

Dreibelbis, Emily. "ChatGPT Passes Google Coding Interview for Level 3 Engineer With $183K Salary." PCMAG, 2023. https://www.pcmag.com/news/chatgpt-passes-google-coding-interview-for-level-3-engineer-with-183k-salary.

Dutton, Denis. "Denis Dutton on Cold Reading." Accessed August 8, 2023. http://www.denisdutton.com/cold_reading.htm.

Edwards, Benj. "Artist Finds Private Medical Record Photos in Popular AI Training Data Set." *Ars Technica*, September 2022. https://arstechnica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/.

———. "Microsoft Aims to Reduce 'Tedious' Business Tasks with New AI Tools." *Ars Technica*, March 2023. https://arstechnica.com/information-technology/2023/03/microsoft-brings-chatgpt-style-ai-to-developer-and-analysis-tools/.

———. "Thanks to AI, It's Probably Time to Take Your Photos Off the Internet." *Ars Technica*, December 2022. https://arstechnica.com/information-technology/2022/12/thanks-to-ai-its-probably-time-to-take-your-photos-off-the-internet/.

El-Mhamdi, El-Mahdi, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Lê-Nguyên Hoang, Rafael Pinot, and John Stephan. "SoK: On the Impossible Security of Very Large Foundation Models." arXiv, September 2022. https://doi.org/10.48550/arXiv.2209.15259.

"Enshittification." *Wikipedia*, November 2024. https://en.wikipedia.org/w/index.php?title=Enshittification&oldid=1259471526.

Epley, Nicholas, Adam Waytz, and John T. Cacioppo. "On Seeing Human: A Three-Factor Theory of Anthropomorphism." *Psychological Review* 114, no. 4 (October 2007): 864–86. https://doi.org/10.1037/0033-295X.114.4.864.

Everett, Daniel. "Beyond Words: The Selves of Other Animals." *New Scientist*, July 2015. https://www.newscientist.com/article/dn27858-beyond-words-the-selves-of-other-animals/.

Fearn, Nicholas. "Heat Waves Are Shutting Down Data Centers and Breaking the Internet." *Gizmodo*, December 2022. https://gizmodo.com/heat-waves-climate-change-data-center-server-shut-down-1849916741.

Feiner, Lauren. "U.S. Regulators Warn They Already Have the Power to Go After A.I. Bias — and They're Ready to Use It." CNBC, April 2023. https://www.cnbc.com/2023/04/25/us-regulators-warn-they-already-have-the-power-to-go-after-ai-bias.html.

Feldman, Vitaly. "Does Learning Require Memorization? A Short Tale about a Long Tail." In *Proceedings of the 52nd* Annual ACM SIGACT Symposium *on* Theory *of* Computing, 954–59. STOC

2020. New York, NY, USA: Association for Computing Machinery, 2020. https://doi.org/10.1145/3357713.3384290.

Feldman, Vitaly, and Chiyuan Zhang. "What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation." In *Advances in* Neural Information Processing Systems, 33:2881–91. Curran Associates, Inc., 2020. https://papers.nips.cc/paper/2020/hash/1e14bfe2714193e7af5abc64ecbd6b46-Abstract.html.

Fischer, Sara. "Exclusive: GPT-4 Readily Spouts Misinformation, Study Finds," March 2023. https://www.msn.com/en-us/news/technology/exclusive-gpt-4-readily-spouts-misinformation-study-finds/ar-AA18TtVg.

Foong, Ng Wai. "Stable Diffusion 2: The Good, The Bad and The Ugly." *Medium*, December 2022. https://towardsdatascience.com/stable-diffusion-2-the-good-the-bad-and-the-ugly-bd44bc7a1333.

Franklin, U. M. *The Real World of Technology*. Massey Lectures Series. Anansi, 1999. https://books.google.is/books?id=LziQT3YS_2sC.

Fraser, David. "Federal Privacy Watchdog Probing OpenAI, ChatGPT Following Complaint CBC News." *CBC*, April 2023. https://www.cbc.ca/news/politics/privacy-commissioner-investigation-openai-chatgpt-1.6801296.

Fraser, Steven D., Frederick P. Brooks, Martin Fowler, Ricardo Lopez, Aki Namioka, Linda Northrop, David Lorge Parnas, and David Thomas. ""No Silver Bullet" Reloaded: Retrospective on "Essence and Accidents of Software Engineering"." In *Companion to the 22nd* ACM SIGPLAN *Conference on* Object-*Oriented Programming Systems and Applications Companion*, 1026–30. Montreal Quebec Canada: ACM, 2007. https://doi.org/10.1145/1297846.1297973.

"FSMB Physician Census." Accessed April 25, 2023. https://www.fsmb.org/physician-census/.

"FTC Chair Khan and Officials from DOJ, CFPB and EEOC Release Joint Statement on AI." *Federal Trade Commission*, April 2023.

https://www.ftc.gov/news-events/news/press-releases/2023/04/ftc-chair-khan-officials-doj-cfpb-eeoc-release-joint-statement-ai.

"FTC Finalizes Settlement with Photo App Developer Related to Misuse of Facial Recognition Technology." *Federal Trade Commission*, May 2021. https://www.ftc.gov/news-events/news/press-releases/2021/05/ftc-finalizes-settlement-photo-app-developer-related-misuse-facial-recognition-technology.

Gal, Uri. "ChatGPT Is a Data Privacy Nightmare, and We Ought to Be Concerned." *Ars Technica*, February 2023. https://arstechnica.com/information-technology/2023/02/chatgpt-is-a-data-privacy-nightmare-and-you-ought-to-be-concerned/.

Gall, John, and John Gall. *The Systems Bible: The Beginner's Guide to Systems Large and Small: Being the Third Edition of* Systemantics. Walker, Minn: General Systemantics Press, 2002.

Geiping, Jonas, Liam Fowl, W. Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. "Witches' Brew: Industrial Scale Data Poisoning via Gradient Matching." arXiv, May 2021. https://doi.org/10.48550/arXiv.2009.02276.

Gilbert, David. "High Schoolers Made a Racist Deepfake of a Principal Threatening Black Students." *Vice*, March 2023. https://www.vice.com/en/article/7kxzk9/school-principal-deepfake-racist-video.

"GitHub Copilot · Your AI Pair Programmer." *GitHub*. Accessed April 5, 2023. http://archive.today/2023.01.11-224507/https://github.com/features/copilot.

Glaiel, Tyler. "Can GPT-4 *Actually* Write Code?" Substack newsletter. *Tyler's Substack*, March 2023. https://tylerglaiel.substack.com/p/can-gpt-4-actually-write-code.

Goldstein, Josh A., and Renée DiResta. "Research Note: This Salesperson Does Not Exist: How Tactics from Political Influence Operations on Social Media Are Deployed for Commercial Lead Generation." *Harvard Kennedy School Misinformation Review*, September 2022. https://doi.org/10.37016/mr-2020-104.

Goldwasser, Shafi, Michael P. Kim, Vinod Vaikuntanathan, and Or Zamir. "Planting Undetectable Backdoors in Machine Learning Models : [Extended Abstract]." In 2022 IEEE *63rd* Annual Symposium *on* Foundations *of* Computer Science (FOCS), 931–42, 2022. https://doi.org/10.1109/FOCS54457.2022.00092.

Goodin, Dan. "Hackers Are Selling a Service That Bypasses ChatGPT Restrictions on Malware." *Ars Technica*, February 2023. https://arstechnica.com/information-technology/2023/02/now-open-fee-based-telegram-service-that-uses-chatgpt-to-generate-malware/.

Governor, James. "The Great Flowering: Why OpenAI Is the New AWS and the New Kingmakers Still Matter." *James Governor's Monkchips*, April 2023. https://redmonk.com/jgovernor/2023/04/13/the-great-flowering-why-openai-is-the-new-aws-and-the-new-kingmakers-still-matter/.

"GPT-3." *Wikipedia*, April 2023. https://en.wikipedia.org/w/index.php?title=GPT-3&oldid=1147823352.

"GPT-4." Accessed March 27, 2023. https://openai.com/research/gpt-4.

Grant, Nico, and Karen Weise. "In A.I. Race, Microsoft and Google Choose Speed Over Caution." *The New York Times*, April 2023. https://www.nytimes.com/2023/04/07/technology/ai-chatbots-google-microsoft.html.

Greshake, Kai, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. "More Than You've Asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models." arXiv, February 2023. https://doi.org/10.48550/arXiv.2302.12173.

Gryspeerdt, Pieter De Grauwe and Sacha. "Artificial Intelligence (AI): The Qualification of AI Creations as "Works" Under EU Copyright Law." *Gevers*, June 2022. https://www.gevers.eu/blog/artificial-intelligence/artificial-intelligence-ai-the-

qualification-of-ai-creations-as-works-under-eu-copyright-law/.

Haibe-Kains, Benjamin, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, Levi Waldron, Bo Wang, Chris McIntosh, et al. "Transparency and Reproducibility in Artificial Intelligence." *Nature* 586, no. 7829 (October 2020): E14–16. https://doi.org/10.1038/s41586-020-2766-y.

Hanff, Alexander. "ChatGPT Should Be Considered a Malevolent AI and Destroyed," March 2023. https://www.theregister.com/2023/03/02/chatgpt_considered_harmful/.

Hanson, Robin. "AGI Is Sacred," August 2022. https://www.overcomingbias.com/p/agi-is-sacredhtml.

Hao, Karen. "We Read the Paper That Forced Timnit Gebru Out of Google. Here's What It Says." *MIT Technology Review*, 2020. Karen Haoarchive page.

Hauptman, Max. "Marines Outwitted an AI Security Camera by Hiding in a Cardboard Box and Pretending to Be Trees." *Task & Purpose*, January 2023. https://taskandpurpose.com/news/marines-ai-paul-scharre/.

Health Sector Coordinating Council. "Health Industry Cybersecurity-Artificial Intelligence-Machine Learning Health Sector Council," February 2023. https://healthsectorcouncil.org/health-industry-cybersecurity-artificial-intelligence-machine-learning/.

"'Heartbreaking': Scam Artists Extorting Children by Putting Their Faces on Explicit Images." *News Channel 5 Nashville* (WTVF), December 2022. https://www.newschannel5.com/news/heartbreaking-scam-artists-extorting-children-by-putting-their-faces-on-explicit-images.

Heaven, Will Douglas. "Hundreds of AI Tools Have Been Built to Catch Covid. None of Them Helped." *MIT Technology Review*, July 2021. https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/.

Heavens, Will Douglas. "AI Is Wrestling with a Replication Crisis." *MIT Technology Review*, 2020. https://www.technologyreview.com/2020/11/12/1011944/artificial-intelligence-replication-crisis-science-big-tech-google-deepmind-facebook-openai/.

Heikkilä, Melissa. "AI Models Spit Out Photos of Real People and Copyrighted Images." *MIT Technology Review*, 2023. https://www.technologyreview.com/2023/02/03/1067786/ai-models-spit-out-photos-of-real-people-and-copyrighted-images/.

———. "Dutch Scandal Serves as a Warning for Europe over Risks of Using Algorithms." *POLITICO*, March 2022. https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/.

———. "The Viral AI Avatar App Lensa Undressed Me—Without My Consent." *MIT Technology Review*, December 2022. https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/.

———. "Why You Shouldn't Trust AI Search Engines." *MIT Technology Review*, 2023. https://www.technologyreview.com/2023/02/14/1068498/why-you-shouldnt-trust-ai-search-engines/.

Henrique, Da Silva Gameiro, Andrei Kucharavy, and Rachid Guerraoui. "Stochastic Parrots Looking for Stochastic Parrots: LLMs Are Easy to Fine-Tune and Hard to Detect with Other LLMs." arXiv, April 2023. https://doi.org/10.48550/arXiv.2304.08968.

Herley, Cormac. "Why Do Nigerian Scammers Say They Are From Nigeria?" *WEIS*, June 2012. https://www.microsoft.com/en-us/research/publication/why-do-nigerian-scammers-say-they-are-from-nigeria/.

Holmes, Aaron, and Kevin McLaughlin. "Ghost Writer: Microsoft Looks to Add OpenAI's Chatbot Technology to Word, Email." *The Information*, January 2023. https://www.theinformation.com/articles/ghost-writer-microsoft-looks-to-add-openais-chatbot-technology-to-word-email.

"How Much of AI's Recent Success Is Due to the Forer Effect? – Terence Eden's Blog," February 2023. https://shkspr.mobi/blog/2023/02/how-much-of-ais-recent-success-is-due-to-the-forer-effect/.

Howard, Jeremy. "Fast.ai - Is GitHub Copilot a Blessing, or a Curse?" July 2021. https://www.fast.ai/posts/2021-07-19-copilot.html.

Huang, Saffron, and Divya Siddarth. "Generative AI and the Digital Commons." *The Collective Intelligence Project*, February 2023. https://cip.org/research/generative-ai-digital-commons.

Huang, Shih-Cheng, Akshay S. Chaudhari, Curtis P. Langlotz, Nigam Shah, Serena Yeung, and Matthew P. Lungren. "Developing Medical Imaging AI for Emerging Infectious Diseases." *Nature Communications* 13, no. 1 (November 2022): 7060. https://doi.org/10.1038/s41467-022-34234-4.

Hugenholtz, P. Bernt, and João Pedro Quintais. "Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output?" *IIC - International Review of Intellectual Property and Competition Law* 52, no. 9 (October 2021): 1190–1216. https://doi.org/10.1007/s40319-021-01115-0.

Hunter, Tatum. "AI Porn Is Easy to Make Now. For Women, That's a Nightmare." *Washington Post*, February 2023. https://www.washingtonpost.com/technology/2023/02/13/ai-porn-deepfakes-women-consent/.

Hutson, Matthew. "Artificial Intelligence Faces Reproducibility Crisis." *Science* 359, no. 6377 (February 2018): 725–26. https://doi.org/10.1126/science.359.6377.725.

"If Builders Built Buildings the Way Programmers Wrote Programs, Then the First Woodpecker That Came Along Would Destroy Civilization – Quote Investigator®," September 2019. https://quoteinvestigator.com/2019/09/19/woodpecker/.

"Illustrating Reinforcement Learning from Human Feedback (RLHF)." Accessed August 8, 2023. https://huggingface.co/blog/rlhf.

Inan, Huseyin A., Osman Ramadan, Lukas Wutschitz, Daniel Jones, Victor Rühle, James Withers, and Robert Sim. "Training Data Leakage Analysis in Language Models." arXiv, February 2021. https://doi.org/10.48550/arXiv.2101.05405.

"Infographic: Female Doctors by Country." *Statista Infographics*, August 2018. https://www.statista.com/chart/14983/female-doctors-by-country.

Ingram, Mathew. "Apple's Censorship in China Is Just the Tip of the Iceberg." *Columbia Journalism Review*. Accessed November 26, 2024. https://www.cjr.org/the_media_today/apple_appstore_china_censorship.php.

"Intelligenza Artificiale: Il Garante Blocca ChatGPT. Raccolta Illecita Di Dati Personali. Assenza Di Sistemi Per La Verifica Dell'età Dei Minori," March 2023. https://www.garanteprivacy.it:443/home/docweb/-/docweb-display/docweb/9870847.

Jagielski, Matthew, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. "Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning." In 2018 IEEE Symposium on Security and Privacy (SP), 19–35, 2018. https://doi.org/10.1109/SP.2018.00057.

Jang, Myeongjun, and Thomas Lukasiewicz. "Consistency Analysis of ChatGPT." arXiv, March 2023. https://doi.org/10.48550/arXiv.2303.06273.

Johnson, Khari. "Chatbots Got Big—and Their Ethical Red Flags Got Bigger." *Wired*. Accessed February 21, 2023. https://www.wired.com/story/chatbots-got-big-and-their-ethical-red-flags-got-bigger/.

Joy. "ChatGPT Cites Economics Papers That Do Not Exist." *Economist Writing Every Day*, January 2023. https://economistwritingeveryday.com/2023/01/21/chatgpt-cites-economics-papers-that-do-not-exist/.

Kan, Michael. "OpenAI Confirms Leak of ChatGPT Conversation Histories." PCMAG, March 202AD. https://www.pcmag.com/news/openai-confirms-leak-of-chatgpt-conversation-histories.

———. "Sci-Fi Mag Pauses Submissions Amid Flood of AI-Generated Short Stories." PCMAG, 2023. https://www.pcmag.com/news/sci-fi-mag-pauses-submissions-amid-flood-of-ai-generated-short-stories.

Kaplan, Lewis. "Mannion v. Coors Brewing Co." July 2005. https://h2o.law.harvard.edu/cases/2353.

Kapoor, Sayash, and Arvind Narayanan. "A Misleading Open Letter about Sci-Fi AI Dangers Ignores the Real Risks." Substack newsletter. *AI Snake Oil*, March 2023. https://aisnakeoil.substack.com/p/a-misleading-open-letter-about-sci.

———. "Leakage and the Reproducibility Crisis in ML-Based Science," 2022. https://doi.org/10.48550/ARXIV.2207.07048.

———. "OpenAI's Policies Hinder Reproducible Research on Language Models." Substack newsletter. *AI Snake Oil*, March 2023. https://aisnakeoil.substack.com/p/openais-policies-hinder-reproducible.

———. "Quantifying ChatGPT's Gender Bias." Substack newsletter. *AI Snake Oil*, April 2023. https://aisnakeoil.substack.com/p/quantifying-chatgpts-gender-bias.

Karpf, Dave. "On Generative AI, Phantom Citations, and Social Calluses." Substack newsletter. *The Future, Now and Then*, March 2023. https://davekarpf.substack.com/p/on-generative-ai-phantom-citations.

Kim, Tae. "Let's Stop Pretending—ChatGPT Isn't That Smart." *Barrons*, February 2023. https://www.barrons.com/articles/chatgpt-ai-openai-chatbot-b9f4fa03.

Kissane, Erin. "Meta in Myanmar (Full Series)." Accessed November 26, 2024. https://erinkissane.com/meta-in-myanmar-full-series.

Knight, Will. "Elon Musk Has Fired Twitter's 'Ethical AI' Team." *Wired*. Accessed April 27, 2023. https://www.wired.com/story/twitter-ethical-ai-team/.

Koenecke, Allison, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. "Careless Whisper: Speech-to-Text Hallucination Harms." In *Proceedings of the* 2024 ACM Conference *on* Fairness, Accountability, *and* Transparency, 1672–81. FAccT '24. New York, NY, USA: Association for Computing Machinery, 2024. https://doi.org/10.1145/3630106.3658996.

Koh, Huan Yee, Jiaxin Ju, He Zhang, Ming Liu, and Shirui Pan. "How Far Are We from Robust Long Abstractive Summarization?" In *Proceedings of the* 2022 Conference *on* Empirical Methods *in* Natural Language Processing, 2682–98. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. https://aclanthology.org/2022.emnlp-main.172.

Kotek, Hadas. "#ChatGPT Doubles down on Gender Stereotypes Even When They Don't Make Sense in Context." *Twitter*, April 2023. https://twitter.com/HadasKotek/status/1648453764117041152.

Kuhn, Bradley M. "If Software Is My Copilot, Who Programmed My Software?" *Software Freedom Conservancy*, February 2022. https://sfconservancy.org/blog/2022/feb/03/github-copilot-copyleft-gpl/.

Kumar, Vinayshekhar Bannihatti, Rashmi Gangadharaiah, and Dan Roth. "Privacy Adhering Machine Un-Learning in NLP." arXiv, December 2022. https://doi.org/10.48550/arXiv.2212.09573.

Kurita, Keita, Paul Michel, and Graham Neubig. "Weight Poisoning Attacks on Pretrained Models." In *Proceedings of the 58th* Annual Meeting *of the* Association *for* Computational Linguistics, 2793–2806. Online: Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.acl-main.249.

Lee, Dave. "Someone Asked ChatGPT If It Had a Signal… And It Gave Him *MY NUMBER*." *Twitter*, February 2023. https://twitter.com/DaveLeeFT/status/1626288109339176962.

Lee, Katherine, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. "Deduplicating Training Data Makes Language Models Better." In *Proceedings of the 60th* Annual Meeting *of the* Association *for* Computational Linguistics (Volume 1: Long Papers), 8424–45. Dublin, Ireland: Association for Computational Linguistics, 2022. https://doi.org/10.18653/v1/2022.acl-long.577.

Lee, Timothy B. "Copyright Lawsuits Pose a Serious Threat to Generative AI," March 2023. https://www.understandingai.org/p/copyright-lawsuits-pose-a-serious.

Lewis, Patrick, Pontus Stenetorp, and Sebastian Riedel. "Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets." In *Proceedings of the 16th* Conference *of the* European Chapter *of the* Association *for* Computational Linguistics: Main Volume, 1000–1008. Online: Association for Computational Linguistics, 2021. https://doi.org/10.18653/v1/2021.eacl-main.86.

Liao, Thomas, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. "Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning," 2022. https://openreview.net/forum?id=mPducS1MsEK.

Lin, Stephanie, Jacob Hilton, and Owain Evans. "TruthfulQA: Measuring How Models Mimic Human Falsehoods." In *Proceedings of the 60th* Annual Meeting *of the* Association *for* Computational Linguistics (Volume 1: Long Papers), 3214–52. Dublin, Ireland: Association for Computational Linguistics, 2022. https://doi.org/10.18653/v1/2022.acl-long.229.

"List of Animals by Number of Neurons." *Wikipedia*, March 2023. https://en.wikipedia.org/w/index.php?title=List_of_animals_by_number_of_neurons&oldid=1145876354.

Liu, Nelson F., Tianyi Zhang, and Percy Liang. "Evaluating Verifiability in Generative Search Engines." arXiv, April 2023. https://doi.org/10.48550/arXiv.2304.09848.

Lomas, Natasha. "FTC Settlement with Ever Orders Data and AIs Deleted After Facial Recognition Pivot." *TechCrunch*, January 2021. https://techcrunch.com/2021/01/12/ftc-settlement-with-ever-orders-data-and-ais-deleted-after-facial-recognition-pivot/.

Lowe, Philip, Jaan Panksepp, Diana Reiss, David Edelman, Bruno Van Swinderen, and Christof Koch. "The Cambridge Declaration on Consciousness." Cambridge, UK, 2012. http://fcmconference.org/img/CambridgeDeclarationOnConsciousness.pdf.

Magazine, Smithsonian, and Jane Recker. "U.S. Copyright Office Rules A.I. Art Can't Be Copyrighted." *Smithsonian Magazine*, 2023. https://www.smithsonianmag.com/smart-news/us-copyright-office-rules-ai-art-cant-be-copyrighted-180979808/.

Magazine, Smithsonian, and Dirk Schulze-Makuch. "Crows Are Even Smarter Than We Thought." *Smithsonian Magazine*. Accessed April 4, 2023. https://www.smithsonianmag.com/air-space-magazine/crows-are-even-smarter-we-thought-180976970/.

Magazine, Smithsonian, and Meilan Solly. "Gorillas Appear to Grieve for Their Dead." *Smithsonian Magazine*. Accessed April 7, 2023. https://www.smithsonianmag.com/smart-news/gorillas-appear-grieve-their-dead-180971896/.

Mahowald, Kyle, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. "Dissociating Language and Thought in Large Language Models: A Cognitive Perspective." arXiv, January 2023. https://doi.org/10.48550/arXiv.2301.06627.

Maluleke, Vongani H., Neerja Thakkar, Tim Brooks, Ethan Weber, Trevor Darrell, Alexei A. Efros, Angjoo Kanazawa, and Devin Guillory. "Studying Bias in GANs Through the Lens of Race." arXiv, September 2022. https://doi.org/10.48550/arXiv.2209.02836.

Marcus, Gary. "AI's Jurassic Park Moment." Substack newsletter. *The Road to AI We Can Trust*, December 2022. https://garymarcus.substack.com/p/ais-jurassic-park-moment.

———. "Deep Learning: A Critical Appraisal." arXiv, January 2018. https://doi.org/10.48550/arXiv.1801.00631.

Mason, Stacey, and Mark Bernstein. "On Links: Exercises in Style." In *Proceedings of the 30th* ACM Conference *on* Hypertext *and* Social Media, 103–10. HT '19. New York, NY, USA: Association for Computing Machinery, 2019. https://doi.org/10.1145/3342220.3343665.

Mauro, Gianluca, and Hilke Schellmann. "'There Is No Standard': Investigation Finds AI Algorithms Objectify Women's Bodies." *The Guardian*, February 2023. https://www.theguardian.com/technology/2023/feb/08/biased-ai-algorithms-racy-women-bodies.

Maynez, Joshua, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. "On Faithfulness and Factuality in Abstractive Summarization." In *Proceedings of the 58th* Annual Meeting *of the* Association *for* Computational Linguistics, 1906–19. Online: Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.acl-main.173.

McAfee. "ChatGPT: A Scammer's Newest Tool." *McAfee Blog*, January 2023. https://www.mcafee.com/blogs/internet-security/chatgpt-a-scammers-newest-tool/.

McCoy, Tom, Ellie Pavlick, and Tal Linzen. "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference." In *Proceedings of the 57th* Annual Meeting *of the* Association *for* Computational Linguistics, 3428–48. Florence, Italy: Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/P19-1334.

McDermott, Drew. "Artificial Intelligence Meets Natural Stupidity." *ACM SIGART Bulletin*, no. 57 (April 1976): 4–9. https://doi.org/10.1145/1045339.1045340.

McGowan, Emma. "Is ChatGPT's Use of People's Data Even Legal?" February 2023. https://blog.avast.com/chatgpt-data-use-legal.

McQuillan, D. *Resisting AI: An Anti-Fascist Approach to Artificial Intelligence*. Bristol University Press, 2022. https://books.google.is/books?id=R6x6EAAAQBAJ.

Meade, Nicholas, Elinor Poole-Dayan, and Siva Reddy. "An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-Trained Language Models." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1878–98. Dublin, Ireland: Association for Computational Linguistics, 2022. https://doi.org/10.18653/v1/2022.acl-long.132.

"Microsoft and Epic Expand Strategic Collaboration with Integration of Azure OpenAI Service." *Stories*, April 2023. https://news.microsoft.com/2023/04/17/microsoft-and-epic-expand-strategic-collaboration-with-integration-of-azure-openai-service/.

"Microsoft's New ChatGPT AI Starts Sending 'Unhinged' Messages to People." *The Independent*, February 2023. https://www.independent.co.uk/tech/chatgpt-ai-messages-microsoft-bing-b2282491.html.

Milmo, Dan. "ChatGPT Reaches 100 Million Users Two Months After Launch." *The Guardian*, February 2023. https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app.

Mirzadeh, Iman, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. "GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models," 2024. https://arxiv.org/abs/2410.05229.

Mitchell, Melanie. "Why AI Is Harder Than We Think." arXiv, April 2021. https://doi.org/10.48550/arXiv.2104.12871.

Moeller, Karla. "How Are Humans Different from Other Animals?" Text, May 2017. https://askabiologist.asu.edu/questions/human-animal-differences.

Monge, Jim Clyde. "Stable Diffusion 2.1 Released — NSFW Image Generation Is Back." *MLearning.ai*, December 2022. https://medium.com/mlearning-ai/stable-diffusion-2-1-released-nsfw-image-generation-is-back-8bcc5c069d60.

Moore, Mike, and Joel Khalili last updated. "AWS Went down Hard, yet Again - Here's What Happened." *TechRadar*, December 2021. https://www.techradar.com/news/live/aws-is-down-again-heres-all-we-know.

Morales, Christina. "Pennsylvania Woman Accused of Using Deepfake Technology to Harass Cheerleaders." *The New York Times*, March 2021. https://www.nytimes.com/2021/03/14/us/raffaela-spone-victory-vipers-deepfake.html.

"Morgan Stanley TMT Conference." Accessed August 8, 2023. https://www.microsoft.com/en-us/Investor/events/FY-2023/Morgan-Stanley-TMT-Conference#:~:text=Scott%20Guthrie%3A%20I%20think%20you%27re,is%

Mosier, Kathleen L., Linda J. Skitka, Mark D. Burdick, and Susan T. Heers. "Automation Bias, Accountability, and Verification Behaviors." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 40, no. 4 (October 1996): 204–8. https://doi.org/10.1177/154193129604000413.

Mosier, Kathleen L., Linda J. Skitka, Susan Heers, and Mark Burdick. "Automation Bias: Decision Making and Performance in High-Tech Cockpits." *The International Journal of Aviation Psychology* 8, no. 1 (January 1998): 47–63. https://doi.org/10.1207/s15327108ijap0801_3.

Mosier, K., and L. Skitka. "Human Decision Makers and Automated Decision Aids: Made for Each Other?" 1996. https://www.semanticscholar.org/paper/Human-Decision-Makers-and-Automated-Decision-Aids%3A-Mosier-Skitka/ffb65e76ac46fd42d595ed9272296f0cbe8ca7aa.

Mukherjee, Supantha, Elvira Pollina, and Rachel More. "Italy's ChatGPT Ban Attracts EU Privacy Regulators." *Reuters*, April 2023. https://www.reuters.com/technology/germany-principle-could-block-chat-gpt-if-needed-data-protection-chief-2023-04-03/.

Murgia, Madhumita, and Max Seddon. "Apple and Google Drop Navalny App After Kremlin Piles on Pressure." *Financial Times*, September 2021.

"Napster." *Wikipedia*, March 2023. https://en.wikipedia.org/w/index.php?title=Napster&oldid=1142267636.

Narayanan, Arvind, and Sayash Kapoor. "GPT-4 and Professional Benchmarks: The Wrong Answer to the Wrong Question." Substack newsletter. *AI Snake Oil*, March 2023. https://aisnakeoil.substack.com/p/gpt-4-and-professional-benchmarks.

———. "People Keep Anthropomorphizing AI. Here's Why." Substack newsletter. *AI Snake Oil*, February 2023. https://aisnakeoil.substack.com/p/people-keep-anthropomorphizing-ai.

Nelson, Ted. *Computer* Lib/Dream Machines. Place of publication not identified, 1974.

"Netscape Navigator." *Wikipedia*, March 2023. https://en.wikipedia.org/w/index.php?title=Netscape_Navigator&oldid=1145935628.

"New AI Classifier for Indicating AI-Written Text." *OpenAI*, January 2023. https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text/.

Newaz, AKM Iqtidar, Nur Imtiazul Haque, Amit Kumar Sikder, Mohammad Ashiqur Rahman, and A. Selcuk Uluagac. "Adversarial Attacks to Machine Learning-Based Smart Healthcare Systems." arXiv, October 2020. https://doi.org/10.48550/arXiv.2010.03671.

News, Neuroscience. "Puzzle-Solving Behavior Spreads Through Bumblebee Colonies." *Neuroscience News*, March 2023. https://neurosciencenews.com/problem-solving-bee-behavior-22728/.

Niven, Timothy, and Hung-Yu Kao. "Probing Neural Network Comprehension of Natural Language Arguments." In *Proceedings of the 57th* Annual Meeting *of the* Association *for* Computational Linguistics, 4658–64. Florence, Italy: Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/P19-1459.

Nkonde, Mutale. "ChatGPT: New AI System, Old Bias?" *Mashable*, February 2023. https://mashable.com/article/chatgpt-ai-racism-bias.

Noor, Poppy. "Can We Trust AI Not to Further Embed Racial Bias and Prejudice?" *BMJ* 368 (February 2020): m363. https://doi.org/10.1136/bmj.m363.

O'Leary, Lizzie. "How IBM's Watson Went From the Future of Health Care to Sold Off for Parts." *Slate*, January 2022. https://slate.com/technology/2022/01/ibm-watson-health-failure-artificial-intelligence.html.

Oakden-Rayner, Lauren. "No Doctor Required: Autonomy, Anomalies, and Magic Puddings." *Lauren Oakden-Rayner*, July 2022. https://laurenoakdenrayner.com/2022/07/04/no-doctor-required-autonomy-anomalies-and-magic-puddings/.

Olkowicz, Seweryn, Martin Kocourek, Radek K. Lučan, Michal Porteš, W. Tecumseh Fitch, Suzana Herculano-Houzel, and Pavel Němec. "Birds Have Primate-Like Numbers of Neurons in the Forebrain." *Proceedings of the National Academy of Sciences* 113, no. 26 (June 2016): 7255–60. https://doi.org/10.1073/pnas.1517131113.

Olson, Parmy. "Nearly Half Of All 'AI Startups' Are Cashing In On Hype." *Forbes*, 2019. https://www.forbes.com/sites/parmyolson/2019/03/04/nearly-half-of-all-ai-startups-are-cashing-in-on-hype/.

Ortiz, Karla. "'The Images Below Aren't @McCurryStudios 'Afghan Girl.''." *Twitter*, November 2022. https://twitter.com/kortizart/status/1588915427018559490.

Pan, Xudong, Mi Zhang, Shouling Ji, and Min Yang. "Privacy Risks of General-Purpose Language Models." In 2020 IEEE

Symposium *on* Security *and* Privacy (SP), 1314–31, 2020. https://doi.org/10.1109/SP40000.2020.00095.

Parasuraman, Raja, and Victor Riley. "Humans and Automation: Use, Misuse, Disuse, Abuse." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39, no. 2 (June 1997): 230–53. https://doi.org/10.1518/001872097778543886.

Patton, Paul. "One World, Many Minds: Intelligence in the Animal Kingdom." *Scientific American*. Accessed April 7, 2023. https://doi.org/10.1038/scientificamericanmind1208-72.

Pearce, Hammond, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and Ramesh Karri. "Asleep at the Keyboard? Assessing the Security of GitHub Copilot's Code Contributions." arXiv, December 2021. https://doi.org/10.48550/arXiv.2108.09293.

Pennartz, Cyriel M. A., Michele Farisco, and Kathinka Evers. "Indicators and Criteria of Consciousness in Animals and Intelligent Machines: An Inside-Out Approach." *Frontiers in Systems Neuroscience* 13 (2019). https://www.frontiersin.org/articles/10.3389/fnsys.2019.00025.

Perault, Matt. "Section 230 Won't Protect ChatGPT." *Lawfare*, February 2023. https://www.lawfareblog.com/section-230-wont-protect-chatgpt.

Perrigo, Billy. "Exclusive: The $2 Per Hour Workers Who Made ChatGPT Safer." *Time*, January 2023. https://time.com/6247678/openai-chatgpt-kenya-workers/.

Perry, Neil, Megha Srivastava, Deepak Kumar, and Dan Boneh. "Do Users Write More Insecure Code with AI Assistants?" arXiv, December 2022. https://doi.org/10.48550/arXiv.2211.03622.

Pikuliak, Matúš. "ChatGPT Survey: Performance on NLP Datasets," March 2023. http://opensamizdat.com/posts/chatgpt_survey/.

Qin, Chengwei, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. "Is ChatGPT a General-Purpose Natural Language Processing Task Solver?" arXiv, February 2023. https://doi.org/10.48550/arXiv.2302.06476.

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices." In *Proceedings of the* 2020 Conference *on* Fairness, Accountability, *and* Transparency, 469–81, 2020. https://doi.org/10.1145/3351095.3372828.

Raieli, Salvatore. "Machine Unlearning: The Duty of Forgetting." *Medium*, September 2022. https://towardsdatascience.com/machine-unlearning-the-duty-of-forgetting-3666e5b9f6e5.

Raji, Inioluwa Deborah, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. "AI and the Everything in the Whole Wide World Benchmark." arXiv, November 2021. https://doi.org/10.48550/arXiv.2111.15366.

Raji, Inioluwa Deborah, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. "The Fallacy of AI Functionality." In 2022 ACM Conference *on* Fairness, Accountability, *and* Transparency, 959–72. Seoul Republic of Korea: ACM, 2022. https://doi.org/10.1145/3531146.3533158.

Ramel, By David, and 03/15/2023. "Data Scientists Cite Lack of GPT-4 Details -." *Virtualization Review*. Accessed April 10, 2023. https://virtualizationreview.com/articles/2023/03/15/gpt-4-details.aspx.

Reece, Manton. "Introducing Micro.blog Podcast Transcripts," March 2023. https://www.manton.org/2023/03/30/introducing-microblog-podcast.html.

"Regulatory Framework Proposal on Artificial Intelligence Shaping Europe's Digital Future," February 2023. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

"Retrieval-Augmented Generation." *Wikipedia*, November 2024. https://en.wikipedia.org/w/index.php?title=Retrieval-augmented_generation&oldid=1259558262.

Rigaki, Maria, and Sebastian Garcia. "A Survey of Privacy Attacks in Machine Learning." arXiv, April 2021. https://doi.org/10.48550/arXiv.2007.07646.

Robertson, Adi. "The US Copyright Office Says an AI Can't Copyright Its Art." *The Verge*, February 2022. https://www.theverge.com/2022/2/21/22944335/us-copyright-office-reject-ai-generated-art-recent-entrance-to-paradise.

Rogers, Anna. "Closed AI Models Make Bad Baselines." *Hacking Semantics*, April 2023. https://hackingsemantics.xyz/2023/closed-baselines/.

Ross, Casey. "Epic's Overhaul of a Flawed Algorithm Shows Why AI Oversight Is a Life-or-Death Issue." *STAT*, October 2022. https://www.statnews.com/2022/10/24/epic-overhaul-of-a-flawed-algorithm/.

Rossi, Luca. "On AI Freshness, the Pyramid Principle, and Hierarchies 💡," April 2023. https://refactoring.fm/p/on-ai-freshness-the-pyramid-principle.

Rowe, Avery. "Don't Ask an AI for Plant Advice • Tradescantia Hub," March 2023. https://tradescantia.uk/article/dont-ask-an-ai-for-plant-advice/.

Ruiz, Nataniel, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation." arXiv, August 2022. https://doi.org/10.48550/arXiv.2208.12242.

Salles, Arleen, Kathinka Evers, and Michele Farisco. "Anthropomorphism in AI." *AJOB Neuroscience* 11, no. 2 (April 2020): 88–95. https://doi.org/10.1080/21507740.2020.1740350.

Schiffer, Zoë. "Microsoft Just Laid Off One of Its Responsible AI Teams," March 2023. https://www.platformer.news/p/microsoft-just-laid-off-one-of-its.

Shah, Chirag, and Emily M. Bender. "Situating Search." In ACM SIGIR Conference *on* Human Information Interaction *and* Retrieval, 221–32. CHIIR '22. New York, NY, USA: Association for Computing Machinery, 2022. https://doi.org/10.1145/3498366.3505816.

Shanahan, Murray. "Talking About Large Language Models." arXiv, February 2023. https://doi.org/10.48550/arXiv.2212.03551.

Shejwalkar, Virat, Amir Houmansadr, Peter Kairouz, and Daniel Ramage. "Back to the Drawing Board: A Critical Evaluation of Poisoning Attacks on Production Federated Learning." In 2022 IEEE Symposium *on* Security *and* Privacy (SP), 1354–71, 2022. https://doi.org/10.1109/SP46214.2022.9833647.

Shepardson, David, Diane Bartz, and Diane Bartz. "US Begins Study of Possible Rules to Regulate AI Like ChatGPT." *Reuters*, April 2023. https://www.reuters.com/technology/us-begins-study-possible-rules-regulate-ai-like-chatgpt-2023-04-11/.

Shokri, Reza, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. "Membership Inference Attacks Against Machine Learning Models." arXiv, March 2017. https://doi.org/10.48550/arXiv.1610.05820.

Simonite, Tom. "Now That Machines Can Learn, Can They Unlearn?" *Wired*. Accessed February 21, 2023. https://www.wired.com/story/machines-can-learn-can-they-unlearn/.

Sloane, Mona, Emanuel Moss, and Rumman Chowdhury. "A Silicon Valley Love Triangle: Hiring Algorithms, Pseudo-Science, and the Quest for Auditability." arXiv, May 2022. https://doi.org/10.48550/arXiv.2106.12403.

Smith, P. G., and D. G. Reinertsen. *Developing Products in Half the Time: New Rules, New Tools*. Wiley, 1997. https://books.google.is/books?id=oOhmCgAAQBAJ.

Somepalli, Gowthami, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. "Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models." arXiv, December 2022. https://doi.org/10.48550/arXiv.2212.03860.

Stark, Luke, and Jevan Hutson. "Physiognomic Artificial Intelligence." {SSRN} {Scholarly} {Paper}. Rochester, NY, September 2021. https://doi.org/10.2139/ssrn.3927300.

Sterling, Toby. "European Privacy Watchdog Creates ChatGPT Task Force." *Reuters*, April 2023. https://www.reuters.com/technology/european-data-protection-board-discussing-ai-policy-thursday-meeting-2023-04-13/.

Stupp, Catherine. "Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case." WSJ. Accessed April 18, 2023. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

"Subjective Validation." *Wikipedia*, February 2023. https://en.wikipedia.org/w/index.php?title=Subjective_validation&oldid=1138260428.

Suya, Fnu, Saeed Mahloujifar, Anshuman Suri, David Evans, and Yuan Tian. "Model-Targeted Poisoning Attacks with Provable Convergence." In *Proceedings of the 38th* International Conference *on* Machine Learning, 10000–10010. PMLR, 2021. https://proceedings.mlr.press/v139/suya21a.html.

Swanson, Steph Maj. "ChatGPT Generated Child Sex Abuse When Asked to Write BDSM Scenarios." *Vice*, March 2023. https://www.vice.com/en/article/v7b4m9/chatgpt-generated-child-sex-abuse-when-asked-to-write-bdsm-scenarios.

tante. "Artificial Saviors." *Boundary* 2, August 2018. https://www.boundary2.org/2018/08/tante/.

Tatman, Rachael. "Gender and Dialect Bias in YouTube's Automatic Captions." In *Proceedings of the* First ACL Workshop *on* Ethics *in* Natural Language Processing, 53–59. Valencia, Spain: Association for Computational Linguistics, 2017. https://doi.org/10.18653/v1/W17-1606.

Taylor, Josh. "ChatGPT's Alter Ego, Dan: Users Jailbreak AI Program to Get Around Ethical Safeguards." *The Guardian*, March 2023. https://www.theguardian.com/technology/2023/mar/08/chatgpt-alter-ego-dan-users-jailbreak-ai-program-to-get-around-ethical-safeguards.

"The Radicalization Risks of GPT-3 and Neural Language Models Middlebury Institute of International Studies at Monterey," September 2020. https://www.middlebury.edu/

institute/academics/centers-initiatives/ctec/ctec-publications/radicalization-risks-gpt-3-and-neural-language.

"This Tiny Fish Can Recognize Itself in a Mirror. Is It Self-Aware?" *Animals*, February 2019. https://www.nationalgeographic.com/animals/article/fish-cleaner-wrasse-self-aware-mirror-test-intelligence-news.

Times, Financial. "Man Beats Machine at Go in Human Victory over AI." *Ars Technica*, February 2023. https://arstechnica.com/information-technology/2023/02/man-beats-machine-at-go-in-human-victory-over-ai/.

Timmer, John. "Bird Brains Are Dense—with Neurons." *Ars Technica*, June 2016. https://arstechnica.com/science/2016/06/bird-brains-are-densewith-neurons/.

Trust, The Road to AI We Can. "The Sparks of AGI? Or the End of Science?" Substack newsletter. *The Road to AI We Can Trust*, March 2023. https://garymarcus.substack.com/p/the-sparks-of-agi-or-the-end-of-science.

Veale, Michael, and Frederik Zuiderveen Borgesius. "Demystifying the Draft EU Artificial Intelligence Act." Preprint. SocArXiv, July 2021. https://doi.org/10.31235/osf.io/38p5f.

Verma, Pranshu. "They Thought Loved Ones Were Calling for Help. It Was an AI Scam." *Washington Post*, March 2023. https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/.

Vincent, James. "Anyone Can Use This AI Art Generator — That's the Risk." *The Verge*, September 2022. https://www.theverge.com/2022/9/15/23340673/ai-image-generation-stable-diffusion-explained-ethics-copyright-data.

———. "Getty Images Is Suing the Creators of AI Art Tool Stable Diffusion for Scraping Its Content." *The Verge*, January 2023. https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit.

———. "Google Announces AI Features in Gmail, Docs, and More to Rival Microsoft." *The Verge*, March 2023. https://

www.theverge.com/2023/3/14/23639273/google-ai-features-docs-gmail-slides-sheets-workspace.

———. "Introducing the AI Mirror Test, Which Very Smart People Keep Failing." *The Verge*, February 2023. https://www.theverge.com/23604075/ai-chatbots-bing-chatgpt-intelligent-sentient-mirror-test.

Vinsel, Lee. "You're Doing It Wrong: Notes on Criticism and Technology Hype." *Medium*, February 2021. https://sts-news.medium.com/youre-doing-it-wrong-notes-on-criticism-and-technology-hype-18b08b4307e5.

Vynck, Gerrit De, and Will Oremus. "As AI Booms, Tech Firms Are Laying Off Their Ethicists." *Washington Post*, March 2023. https://www.washingtonpost.com/technology/2023/03/30/tech-companies-cut-ai-ethics/.

Wallace, Eric, Tony Zhao, Shi Feng, and Sameer Singh. "Concealed Data Poisoning Attacks on NLP Models." In *Proceedings of the* 2021 Conference *of the* North American Chapter *of the* Association *for* Computational Linguistics: Human Language Technologies, 139–50. Online: Association for Computational Linguistics, 2021. https://doi.org/10.18653/v1/2021.naacl-main.13.

Wang, Ge, and Juliana Bidadanure. "Humans in the Loop: The Design of Interactive AI Systems." *Stanford HAI*, October 2019. https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems.

Wang, Tony T., Adam Gleave, Tom Tseng, Nora Belrose, Kellin Pelrine, Joseph Miller, Michael D. Dennis, et al. "Adversarial Policies Beat Superhuman Go AIs." arXiv, February 2023. https://doi.org/10.48550/arXiv.2211.00241.

Warren, Tom. "Microsoft Announces Copilot: The AI-Powered Future of Office Documents." *The Verge*, March 2023. https://www.theverge.com/2023/3/16/23642833/microsoft-365-ai-copilot-word-outlook-teams.

———. "Microsoft Is Bundling Its AI-Powered Office Features into Microsoft 365 Subscriptions." *The Verge*, November 2024.

https://www.theverge.com/2024/11/7/24290268/microsoft-copilot-office-features-microsoft-365.

———. "You Can Play with Microsoft's Bing GPT-4 Chatbot Right Now, No Waitlist Necessary." *The Verge*, March 2023. https://www.theverge.com/2023/3/15/23641683/microsoft-bing-ai-gpt-4-chatbot-available-no-waitlist.

Watson, David. "The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence." *Minds and Machines* 29, no. 3 (September 2019): 417–40. https://doi.org/10.1007/s11023-019-09506-6.

Weidinger, Laura, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, et al. "Ethical and Social Risks of Harm from Language Models." arXiv, December 2021. https://doi.org/10.48550/arXiv.2112.04359.

Weizenbaum, Joseph. *Computer Power and Human Reason: From Judgment to Calculation*. San Francisco: Freeman, 1976.

West, Sarah Myers. "Competition Authorities Need to Move Fast and Break up AI." *Financial Times*, April 2023.

"What Are the GDPR Fines?" *GDPR.eu*, July 2018. https://gdpr.eu/fines/.

White, Courtney, and Rita Matulionyte. "Artificial Intelligence Painting The Bigger Picture For Copyright Ownership." {SSRN} {Scholarly} {Paper}. Rochester, NY, December 2019. https://doi.org/10.2139/ssrn.3498673.

Whittaker, Meredith, Meryl Alper, Cynthia L. Bennett, Sara Hendren, Liz Kaziunas, Mara Mills, Meredith Ringel Morris, et al. "Disability, Bias, and AI," November 2019. https://www.microsoft.com/en-us/research/publication/disability-bias-and-ai/.

"Who Owns the Copyright in AI-Generated Art?" *Murgitroyd, European Patent & Trade Mark Attorneys*, October 2022. https://www.murgitroyd.com/en-us/blog/who-owns-the-copyright-in-ai-generated-art/.

Wiggers, Kyle. "Addressing Criticism, OpenAI Will No Longer Use Customer Data to Train Its Models by Default." *TechCrunch*,

March 2023. https://techcrunch.com/2023/03/01/addressing-criticism-openai-will-no-longer-use-customer-data-to-train-its-models-by-default/.

Willison, Simon. "Bing: 'I Will Not Harm You Unless You Harm Me First'," 2023. http://simonwillison.net/2023/Feb/15/bing/.

———. "I Promise You ChatGPT Can't Access the Internet, Even Though It Really Looks Like It Can," March 2023. http://simonwillison.net/2023/Mar/10/chatgpt-internet-access/.

———. "Prompt Injection Attacks Against GPT-3." Accessed February 21, 2023. http://simonwillison.net/2022/Sep/12/prompt-injection/.

———. "Prompt Injection: What's the Worst That Can Happen?" April 2023. https://simonwillison.net/2023/Apr/14/worst-that-can-happen/.

———. "Thoughts and Impressions of AI-Assisted Search from Bing," February 2023. http://simonwillison.net/2023/Feb/24/impressions-of-bing/.

Wong, Andrew, Erkin Otles, John P. Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, et al. "External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients." *JAMA Internal Medicine* 181, no. 8 (August 2021): 1065–70. https://doi.org/10.1001/jamainternmed.2021.2626.

Writer, Robert LemosContributing, Dark ReadingMarch 07, and 2023. "Employees Are Feeding Sensitive Business Data to ChatGPT." *Dark Reading*, March 2023. https://www.darkreading.com/risk/employees-feeding-sensitive-business-data-chatgpt-raising-security-fears.

Wu, Xiaolin, and Xi Zhang. "Automated Inference on Criminality Using Face Images." arXiv, November 2016. https://doi.org/10.48550/arXiv.1611.04135.

Wynants, Laure, Ben Van Calster, Gary S. Collins, Richard D. Riley, Georg Heinze, Ewoud Schuit, Marc M. J. Bonten, et al. "Prediction Models for Diagnosis and Prognosis of Covid-19: Systematic Review and Critical Appraisal." *BMJ (Clinical*

*Research Ed.*) 369 (April 2020): m1328. https://doi.org/10.1136/bmj.m1328.

Xiang, Chloe. "OpenAI Is Now Everything It Promised Not to Be: Corporate, Closed-Source, and For-Profit." *Vice*, February 2023. https://www.vice.com/en/article/5d3naz/openai-is-now-everything-it-promised-not-to-be-corporate-closed-source-and-for-profit.

Xiang, Chloe, and Emanuel Maiberg. "ISIS Executions and Non-Consensual Porn Are Powering AI Art." *Vice*, September 2022. https://www.vice.com/en/article/93ad75/isis-executions-and-non-consensual-porn-are-powering-ai-art.

Yeo, Catherine. "How Biased Is GPT-3?" *Fair Bytes*, June 2020. https://www.fairbytes.org/post/how-biased-is-gpt-3.

Yeom, Samuel, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting." arXiv, May 2018. https://doi.org/10.48550/arXiv.1709.01604.

Zheng, Xiaosen, and Jing Jiang. "An Empirical Study of Memorization in NLP." In *Proceedings of the 60th* Annual Meeting *of the* Association *for* Computational Linguistics (Volume 1: Long Papers), 6265–78. Dublin, Ireland: Association for Computational Linguistics, 2022. https://doi.org/10.18653/v1/2022.acl-long.434.

25

# End notes

1. Clayton M. Christensen, *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail* (Boston, Massachusetts: Harvard Business Review Press, 2024).

2. "Enshittification," *Wikipedia*, November 2024, https://en.wikipedia.org/w/index.php?title=Enshittification&oldid=1259471526.

3. Tom Warren, "Microsoft Is Bundling Its AI-Powered Office Features into Microsoft 365 Subscriptions," *The Verge*, November 2024, https://www.theverge.com/2024/11/7/24290268/microsoft-copilot-office-features-microsoft-365.

4. Sam Biddle, "U.S. Military Makes First Confirmed OpenAI Purchase for War-Fighting Forces," *The Intercept*, October 2024, https://theintercept.com/2024/10/25/africom-microsoft-openai-military/.

5. "Netscape Navigator," *Wikipedia*, March 2023, https://en.wikipedia.org/w/index.php?title=Netscape_Navigator&oldid=1145935628.

6. "Dot-Com Bubble," *Wikipedia*, April 2023, https://en.wikipedia.org/w/index.php?title=Dot-com_bubble&oldid=1147856945.

7. "Chatbots, Deepfakes, and Voice Clones: AI Deception for Sale," *Federal Trade Commission*, March 2023, https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale.

8. Michael Veale and Frederik Zuiderveen Borgesius, "Demystifying the Draft EU Artificial Intelligence Act," preprint (SocArXiv, July 2021), https://doi.org/10.31235/osf.io/38p5f.

9. Nicholas Epley, Adam Waytz, and John T. Cacioppo, "On Seeing Human: A Three-Factor Theory of Anthropomorphism," *Psycho-*

logical Review 114, no. 4 (October 2007): 864–86, https://doi.org/10.1037/0033-295X.114.4.864.

10. Emily M. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21 (New York, NY, USA: Association for Computing Machinery, 2021), 610–23, https://doi.org/10.1145/3442188.3445922.

11. John Timmer, "Bird Brains Are Dense—with Neurons," Ars Technica, June 2016, https://arstechnica.com/science/2016/06/bird-brains-are-densewith-neurons/.

12. Smithsonian Magazine and Dirk Schulze-Makuch, "Crows Are Even Smarter Than We Thought," Smithsonian Magazine, accessed April 4, 2023, https://www.smithsonianmag.com/air-space-magazine/crows-are-even-smarter-we-thought-180976970/.

13. Seweryn Olkowicz et al., "Birds Have Primate-Like Numbers of Neurons in the Forebrain," Proceedings of the National Academy of Sciences 113, no. 26 (June 2016): 7255–60, https://doi.org/10.1073/pnas.1517131113.

14. James Vincent, "Introducing the AI Mirror Test, Which Very Smart People Keep Failing," The Verge, February 2023, https://www.theverge.com/23604075/ai-chatbots-bing-chatgpt-intelligent-sentient-mirror-test.

15. Arvind Narayanan and Sayash Kapoor, "People Keep Anthropomorphizing AI. Here's Why," Substack newsletter, AI Snake Oil, February 2023, https://aisnakeoil.substack.com/p/people-keep-anthropomorphizing-ai.

16. "How Much of AI's Recent Success Is Due to the Forer Effect? – Terence Eden's Blog," February 2023, https://shkspr.mobi/blog/2023/02/how-much-of-ais-recent-success-is-due-to-the-forer-effect/.

17. Epley, Waytz, and Cacioppo, "On Seeing Human."

18. Arleen Salles, Kathinka Evers, and Michele Farisco, "Anthropomorphism in AI," *AJOB Neuroscience* 11, no. 2 (April 2020): 88–95, https://doi.org/10.1080/21507740.2020.1740350; Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation* (San Francisco: Freeman, 1976).

19. Murray Shanahan, "Talking About Large Language Models" (arXiv, February 2023), https://doi.org/10.48550/arXiv.2212.03551.

20. Ruben Branco et al., "Shortcutted Commonsense: Data Spuriousness in Deep Learning of Commonsense Reasoning," in *Proceedings of the* 2021 Conference *on* Empirical Methods *in* Natural Language Processing (Online; Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021), 1504–21, https://doi.org/10.18653/v1/2021.emnlp-main.113.

21. Tom McCoy, Ellie Pavlick, and Tal Linzen, "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference," in *Proceedings of the 57th* Annual Meeting *of the* Association *for* Computational Linguistics (Florence, Italy: Association for Computational Linguistics, 2019), 3428–48, https://doi.org/10.18653/v1/P19-1334.

22. Timothy Niven and Hung-Yu Kao, "Probing Neural Network Comprehension of Natural Language Arguments," in *Proceedings of the 57th* Annual Meeting *of the* Association *for* Computational Linguistics (Florence, Italy: Association for Computational Linguistics, 2019), 4658–64, https://doi.org/10.18653/v1/P19-1459.

23. Iman Mirzadeh et al., "GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models," 2024, https://arxiv.org/abs/2410.05229.

24. Myeongjun Jang and Thomas Lukasiewicz, "Consistency Analysis of ChatGPT" (arXiv, March 2023), https://doi.org/10.48550/arXiv.2303.06273.

25. Kyle Mahowald et al., "Dissociating Language and Thought in Large Language Models: A Cognitive Perspective" (arXiv, January 2023), https://doi.org/10.48550/arXiv.2301.06627.

26. Gary Marcus, "Deep Learning: A Critical Appraisal" (arXiv, January 2018), https://doi.org/10.48550/arXiv.1801.00631; David Watson,
"The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence," *Minds and Machines* 29, no. 3 (September 2019): 417–40, https://doi.org/10.1007/s11023-019-09506-6.

27. "Bumble Bees Can Teach Each Other to Solve Problems - Study," *The Jerusalem Post JPost.com*, accessed April 4, 2023, https://www.jpost.com/science/article-734044.

28. Jang and Lukasiewicz, "Consistency Analysis of ChatGPT."

29. Emily M. Bender and Alexander Koller, "Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data," in *Proceedings of the 58th* Annual Meeting *of the* Association *for* Computational Linguistics (Online: Association for Computational Linguistics, 2020), 5185–98, https://doi.org/10.18653/v1/2020.acl-main.463.

30. See John Gall and John Gall, *The Systems Bible: The Beginner's Guide to Systems Large and Small: Being the Third Edition of* Systemantics (Walker, Minn: General Systemantics Press, 2002), for example: "The point is that real novelty in the world increases as complexity increases. As the necessary richness of underlying structure is attained, the new property emerges, seemingly out of nowhere."

31. Sébastien Bubeck et al., "Sparks of Artificial General Intelligence: Early Experiments with GPT-4" (arXiv, March 2023), https://doi.org/10.48550/arXiv.2303.12712.

32. Daniel Everett, "Beyond Words: The Selves of Other Animals," *New Scientist*, July 2015, https://www.newscientist.com/article/dn27858-beyond-words-the-selves-of-other-animals/.

33. Marc Bekoff, "Scientists Conclude Nonhuman Animals Are Conscious Beings: Didn't We Already Know This? Yes, We Did." August 2012, https://www.psychologytoday.com/us/blog/animal-emotions/201208/scientists-conclude-nonhuman-animals-are-conscious-beings; Philip Lowe et al., "The Cambridge Declaration on Consciousness" (Cambridge, UK, 2012), http://fcmconference.org/img/CambridgeDeclarationOnConsciousness.pdf, this declaration was a particular turning point that showed that general scientific consensus had shifted on the subject: "Convergent evidence indicates that non-human animals have the neuroanatomical, neurochemical, and neurophysiological substrates of conscious states along with the capacity to exhibit intentional behaviors."

34. Karla Moeller, "How Are Humans Different from Other Animals?" Text, May 2017, https://askabiologist.asu.edu/questions/human-animal-differences.

35. Neuroscience News, "Puzzle-Solving Behavior Spreads Through Bumblebee Colonies," *Neuroscience News*, March 2023, https://neurosciencenews.com/problem-solving-bee-behavior-22728/.

36. "List of Animals by Number of Neurons," *Wikipedia*, March 2023, https://en.wikipedia.org/w/index.php?title=List_of_animals_by_number_of_neurons&oldid=1145876354.

37. "List of Animals by Number of Neurons."

38. "This Tiny Fish Can Recognize Itself in a Mirror. Is It Self-Aware?" *Animals*, February 2019, https://www.nationalgeographic.com/animals/article/fish-cleaner-wrasse-self-aware-mirror-test-intelligence-news.

39. Paul Patton, "One World, Many Minds: Intelligence in the Animal Kingdom," *Scientific American*, accessed April 7, 2023, https://doi.org/10.1038/scientificamericanmind1208-72.

40. Jonathan Birch, Alexandra K. Schnell, and Nicola S. Clayton, "Dimensions of Animal Consciousness," *Trends in Cognitive Sciences* 24, no. 10 (October 2020): 789–801, https://doi.org/10.1016/j.tics.2020.07.007; Inbal Ben-Ami Bartal, Jean Decety, and Peggy Mason, "Helping a Cagemate in Need: Empathy and Pro-Social Behavior in Rats," *Science (New York, N.Y.)* 334, no. 6061 (December 2011): 1427–30, https://doi.org/10.1126/science.1210789; Smithsonian Magazine and Meilan Solly, "Gorillas Appear to Grieve for Their Dead," *Smithsonian Magazine*, accessed April 7, 2023, https://www.smithsonianmag.com/smart-news/gorillas-appear-grieve-their-dead-180971896/.

41. Cyriel M. A. Pennartz, Michele Farisco, and Kathinka Evers, "Indicators and Criteria of Consciousness in Animals and Intelligent Machines: An Inside-Out Approach," *Frontiers in Systems Neuroscience* 13 (2019), https://www.frontiersin.org/articles/10.3389/fnsys.2019.00025, and this seems to even apply to the capacity for consciousness: "Although major differences between rodent and human brains in terms of size, complexity and the presence of specialized areas should be acknowledged, we argue that the 'size' argument will rather affect the complexity and/or intensity of the information the organism will be conscious of, and not so much the presence or absence of consciousness."

42. Sam Altman, "Planning for AGI and Beyond," February 2023, https://openai.com/blog/planning-for-agi-and-beyond.

43. Robin Hanson, "AGI Is Sacred," August 2022, https://www.overcomingbias.com/p/agi-is-sacredhtml.

44. Anna Rogers, "Closed AI Models Make Bad Baselines," *Hacking Semantics*, April 2023, https://hackingsemantics.xyz/2023/closed-baselines/.

45. Inioluwa Deborah Raji et al., "AI and the Everything in the Whole Wide World Benchmark" (arXiv, November 2021), https://doi.org/10.48550/arXiv.2111.15366.

46. Thomas Liao et al., "Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning," 2022, https://openreview.net/forum?id=mPducS1MsEK.

47. Xiaolin Wu and Xi Zhang, "Automated Inference on Criminality Using Face Images" (arXiv, November 2016), https://doi.org/10.48550/arXiv.1611.04135.

48. Luke Stark and Jevan Hutson, "Physiognomic Artificial Intelligence," {SSRN} {Scholarly} {Paper} (Rochester, NY, September 2021), https://doi.org/10.2139/ssrn.3927300.

49. Bubeck et al., "Sparks of Artificial General Intelligence."

50. Ventana al Conocimiento, "Cold Fusion: Anatomy of a Scientific 'Fraud'," *OpenMind*, March 2019, https://www.bbvaopenmind.com/en/science/physics/cold-fusion-anatomy-of-a-scientific-fraud/.

51. Kyle Barr, "GPT-4 Is a Giant Black Box and Its Training Data Remains a Mystery," *Gizmodo*, March 2023, https://gizmodo.com/chatbot-gpt4-open-ai-ai-bing-microsoft-1850229989.

52. Tante, "Artificial Saviors," *Boundary 2*, August 2018, https://www.boundary2.org/2018/08/tante/, similarly notes how AI research has taken a religious dimension. For example: "Current AI trends turn automation into a religion, slowly transforming at

least semi-transparent systems into opaque systems whose functionality and correctness can neither be verified nor explained."

53. "How Much of AI's Recent Success Is Due to the Forer Effect?"

54. "Cold Reading," *Wikipedia*, August 2023, https://en.wikipedia.org/w/index.php?title=Cold_reading&oldid=1168265112.

55. "Subjective Validation," *Wikipedia*, February 2023, https://en.wikipedia.org/w/index.php?title=Subjective_validation&oldid=1138260428.

56. "Cold Reading."

57. "Barnum Effect," *Wikipedia*, June 2023, https://en.wikipedia.org/w/index.php?title=Barnum_effect&oldid=1161115161.

58. "Cold Reading."

59. "Cold Reading."

60. "Cold Reading."

61. Denis Dutton, "Denis Dutton on Cold Reading," accessed August 8, 2023, http://www.denisdutton.com/cold_reading.htm.

62. "Illustrating Reinforcement Learning from Human Feedback (RLHF)," accessed August 8, 2023, https://huggingface.co/blog/rlhf.

63. Lizzie O'Leary, "How IBM's Watson Went From the Future of Health Care to Sold Off for Parts," *Slate*, January 2022, https://slate.com/technology/2022/01/ibm-watson-health-failure-artificial-intelligence.html.

64. Jeffrey Dastin, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women," *Reuters*, October 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

65. Will Douglas Heaven, "Hundreds of AI Tools Have Been Built to Catch Covid. None of Them Helped." *MIT Technology Review*, July 2021, https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/.

66. Melissa Heikkilä, "Dutch Scandal Serves as a Warning for Europe over Risks of Using Algorithms," POLITICO, March 2022, https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/.

67. Parmy Olson, "Nearly Half Of All 'AI Startups' Are Cashing In On Hype," *Forbes*, 2019, https://www.forbes.com/sites/parmyolson/2019/03/04/nearly-half-of-all-ai-startups-are-cashing-in-on-hype/.

68. Michael Atleson, "Keep Your AI Claims in Check," *Federal Trade Commission*, February 2023, https://www.ftc.gov/business-guidance/blog/2023/02/keep-your-ai-claims-check.

69. Will Douglas Heavens, "AI Is Wrestling with a Replication Crisis," *MIT Technology Review*, 2020, https://www.technologyreview.com/2020/11/12/1011944/artificial-intelligence-replication-crisis-science-big-tech-google-deepmind-facebook-openai/; Benjamin Haibe-Kains et al., "Transparency and Reproducibility in Artificial Intelligence," *Nature* 586, no. 7829 (October 2020): E14–16, https://doi.org/10.1038/s41586-020-2766-y.

70. Altman, "Planning for AGI and Beyond."

71. "GPT-4," accessed March 27, 2023, https://openai.com/research/gpt-4.

72. Rogers, "Closed AI Models Make Bad Baselines," notably: 'We make the case that as far as research and scientific publications are concerned, the "closed" models (as defined below) cannot be meaningfully studied.'

73. The Road to AI We Can Trust, "The Sparks of AGI? Or the End of Science?" Substack newsletter, *The Road to AI We Can Trust*, March 2023, https://garymarcus.substack.com/p/the-sparks-of-agi-or-the-end-of-science, as Gary Marcus says: "By excluding the scientific community from any serious insight into the design and function of these models, Microsoft and OpenAI are placing the public in a position in which those two companies alone are in a position do anything about the risks to which they are exposing us all."

74. Sayash Kapoor and Arvind Narayanan, "OpenAI's Policies Hinder Reproducible Research on Language Models," Substack newsletter, *AI Snake Oil*, March 2023, https://aisnakeoil.substack.com/p/openais-policies-hinder-reproducible.

75. By David Ramel and 03/15/2023, "Data Scientists Cite Lack of GPT-4 Details -," *Virtualization Review*, accessed April 10, 2023, https://virtualizationreview.com/articles/2023/03/15/gpt-4-details.aspx.

76. Matthew Hutson, "Artificial Intelligence Faces Reproducibility Crisis," *Science* 359, no. 6377 (February 2018): 725–26, https://doi.org/10.1126/science.359.6377.725.

77. Sayash Kapoor and Arvind Narayanan, "Leakage and the Reproducibility Crisis in ML-Based Science," 2022, https://doi.org/10.48550/ARXIV.2207.07048.

78. Arvind Narayanan and Sayash Kapoor, "GPT-4 and Professional Benchmarks: The Wrong Answer to the Wrong Question," Substack newsletter, *AI Snake Oil*, March 2023, https://aisnakeoil.substack.com/p/gpt-4-and-professional-benchmarks.

79. Matúš Pikuliak, "ChatGPT Survey: Performance on NLP Datasets," March 2023, http://opensamizdat.com/posts/chatgpt_survey/.

80. Lee Vinsel, "You're Doing It Wrong: Notes on Criticism and Technology Hype," *Medium*, February 2021, https://sts-news.medium.com/youre-doing-it-wrong-notes-on-criticism-and-technology-hype-18b08b4307e5.

81. Sayash Kapoor and Arvind Narayanan, "A Misleading Open Letter about Sci-Fi AI Dangers Ignores the Real Risks," Substack newsletter, *AI Snake Oil*, March 2023, https://aisnakeoil.substack.com/p/a-misleading-open-letter-about-sci.

82. Bender et al., "On the Dangers of Stochastic Parrots."

83. Melanie Mitchell, "Why AI Is Harder Than We Think" (arXiv, April 2021), https://doi.org/10.48550/arXiv.2104.12871.

84. Inioluwa Deborah Raji et al., "The Fallacy of AI Functionality," in 2022 ACM Conference *on* Fairness, Accountability, *and* Transparency (Seoul Republic of Korea: ACM, 2022), 959–72, https://doi.org/10.1145/3531146.3533158.

85. Noble Ackerson, "GPT Is an Unreliable Information Store," *Medium*, February 2023, https://medium.com/@nobleackerson/chatgpt-insists-i-am-dead-and-the-problem-with-language-models-db5a36c22f11.

86. Gary Marcus, "AI's Jurassic Park Moment," Substack newsletter, *The Road to AI We Can Trust*, December 2022, https://garymarcus.substack.com/p/ais-jurassic-park-moment.

87. Melissa Heikkilä, "Why You Shouldn't Trust AI Search Engines," *MIT Technology Review*, 2023, https://www.technologyreview.com/2023/02/14/1068498/why-you-shouldnt-trust-ai-search-engines/.

88. See Manton Reece, "Introducing Micro.blog Podcast Transcripts," March 2023, https://www.manton.org/2023/03/30/introducing-microblog-podcast.html, for good example.

89. Atleson, "Keep Your AI Claims in Check."

90. Uri Gal, "ChatGPT Is a Data Privacy Nightmare, and We Ought to Be Concerned," *Ars Technica*, February 2023, https://arstechnica.com/information-technology/2023/02/chatgpt-is-a-data-privacy-nightmare-and-you-ought-to-be-concerned/.

91. Michael Kan, "OpenAI Confirms Leak of ChatGPT Conversation Histories," PCMAG, March 202AD, https://www.pcmag.com/news/openai-confirms-leak-of-chatgpt-conversation-histories.

92. Stephen Council, "OpenAI Admits Some Premium Users' Payment Info Was Exposed," SFGATE, March 2023, https://www.sfgate.com/tech/article/chatgpt-openai-payment-data-leak-17858969.php.

93. Without structured testing, we only really have the vendor's word for it that these tools are as amazing as they claim.

94. Rogers, "Closed AI Models Make Bad Baselines."

95. Atleson, "Keep Your AI Claims in Check."

96. Pikuliak, "ChatGPT Survey."

97. Emily M. Bender and Batya Friedman, "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science," *Transactions of the Association for Computational Linguistics* 6 (2018): 587–604, https://doi.org/10.1162/tacl_a_00041.

98. Cormac Herley, "Why Do Nigerian Scammers Say They Are From Nigeria?" WEIS, June 2012, https://www.microsoft.com/en-us/research/publication/why-do-nigerian-scammers-say-they-are-from-nigeria/.

99. Emily Dreibelbis, "ChatGPT Passes Google Coding Interview for Level 3 Engineer With $183K Salary," PCMAG, 2023, https://www.pcmag.com/news/chatgpt-passes-google-coding-interview-for-level-3-engineer-with-183k-salary.

100. Michael Kan, "Sci-Fi Mag Pauses Submissions Amid Flood of AI-Generated Short Stories," PCMAG, 2023, https://www.pcmag.com/news/sci-fi-mag-pauses-submissions-amid-flood-of-ai-generated-short-stories.

101. Greg Bensinger, "ChatGPT Launches Boom in AI-Written e-Books on Amazon," *Reuters*, February 2023, https://www.reuters.com/technology/chatgpt-launches-boom-ai-written-e-books-amazon-2023-02-21/.

102. McAfee, "ChatGPT: A Scammer's Newest Tool," *McAfee Blog*, January 2023, https://www.mcafee.com/blogs/internet-security/chatgpt-a-scammers-newest-tool/.

103. Benj Edwards, "Thanks to AI, It's Probably Time to Take Your Photos Off the Internet," *Ars Technica*, December 2022, https://arstechnica.com/information-technology/2022/12/thanks-to-ai-its-probably-time-to-take-your-photos-off-the-internet/; Nataniel Ruiz et al., "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation" (arXiv, August 2022), https://doi.org/10.48550/arXiv.2208.12242.

104. David Gilbert, "High Schoolers Made a Racist Deepfake of a Principal Threatening Black Students," *Vice*, March 2023, https://www.vice.com/en/article/7kxzk9/school-principal-deepfake-racist-video; Christina Morales, "Pennsylvania Woman Accused of Using Deepfake Technology to Harass Cheerleaders," *The New York Times*, March 2021, https://www.nytimes.com/2021/03/14/us/raffaela-spone-victory-vipers-deepfake.html.

105. Tatum Hunter, "AI Porn Is Easy to Make Now. For Women, That's a Nightmare." *Washington Post*, February 2023, https://www.washingtonpost.com/technology/2023/02/13/ai-porn-deepfakes-women-consent/.

106. Ashley Belanger, "Thousands Scammed by AI Voices Mimicking Loved Ones in Emergencies," *Ars Technica*, March 2023, https://ar-

stechnica.com/tech-policy/2023/03/rising-scams-use-ai-to-mimic-voices-of-loved-ones-in-financial-distress/; Pranshu Verma, "They Thought Loved Ones Were Calling for Help. It Was an AI Scam." *Washington Post*, March 2023, https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/; Catherine Stupp, "Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case," *WSJ*, accessed April 18, 2023, https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

107. "'Heartbreaking': Scam Artists Extorting Children by Putting Their Faces on Explicit Images," *News Channel 5 Nashville* (WTVF), December 2022, https://www.newschannel5.com/news/heartbreaking-scam-artists-extorting-children-by-putting-their-faces-on-explicit-images.

108. Josh A. Goldstein and Renée DiResta, "Research Note: This Salesperson Does Not Exist: How Tactics from Political Influence Operations on Social Media Are Deployed for Commercial Lead Generation," *Harvard Kennedy School Misinformation Review*, September 2022, https://doi.org/10.37016/mr-2020-104.

109. "New AI Classifier for Indicating AI-Written Text," *OpenAI*, January 2023, https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text/.

110. Da Silva Gameiro Henrique, Andrei Kucharavy, and Rachid Guerraoui, "Stochastic Parrots Looking for Stochastic Parrots: LLMs Are Easy to Fine-Tune and Hard to Detect with Other LLMs" (arXiv, April 2023), https://doi.org/10.48550/arXiv.2304.08968.

111. Simon Willison, "Bing: 'I Will Not Harm You Unless You Harm Me First'," 2023, http://simonwillison.net/2023/Feb/15/bing/.

112. Dmitri Brereton, "Bing AI Can't Be Trusted," February 2023, https://dkb.blog/p/bing-ai-cant-be-trusted.

113. Simon Willison, "Prompt Injection: What's the Worst That Can Happen?" April 2023, https://simonwillison.net/2023/Apr/14/worst-that-can-happen/.

114. Melissa Heikkilä, "The Viral AI Avatar App Lensa Undressed Me—Without My Consent," *MIT Technology Review*, December 2022, https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/.

115. Matt Perault, "Section 230 Won't Protect ChatGPT," *Lawfare*, February 2023, https://www.lawfareblog.com/section-230-wont-protect-chatgpt.

116. Council, "OpenAI Admits Some Premium Users' Payment Info Was Exposed"; Kan, "OpenAI Confirms Leak of ChatGPT Conversation Histories."

117. Kai Greshake et al., "More Than You've Asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models" (arXiv, February 2023), https://doi.org/10.48550/arXiv.2302.12173; Simon Willison, "Prompt Injection Attacks Against GPT-3," accessed February 21, 2023, http://simonwillison.net/2022/Sep/12/prompt-injection/.

118. Dan Goodin, "Hackers Are Selling a Service That Bypasses ChatGPT Restrictions on Malware," *Ars Technica*, February 2023, https://arstechnica.com/information-technology/2023/02/now-open-fee-based-telegram-service-that-uses-chatgpt-to-generate-malware/; Matt Burgess, "The Hacking of ChatGPT Is Just Getting Started," *Wired*, accessed April 14, 2023, https://www.wired.com/story/chatgpt-jailbreak-generative-ai-hacking/.

119. "BingBang: The AAD Misconfiguration That Led to Bing.com Results Manipulation and Account Takeover Explained," Wiz.io, March 2023, https://www.wiz.io/blog/azure-active-directory-bing-misconfiguration, is from March 2023. Pick a decade, a year

and a month at random, and you're likely to find a news story about a software security mishap at Microsoft.

120. Andy Brice, "How Much Code Can a Coder Code?" *Successful Software*, February 2017, https://successfulsoftware.net/2017/02/10/how-much-code-can-a-coder-code/.

121. Allison Koenecke et al., "Careless Whisper: Speech-to-Text Hallucination Harms," in *Proceedings of the* 2024 ACM Conference *on* Fairness, Accountability, *and* Transparency, FAccT '24 (New York, NY, USA: Association for Computing Machinery, 2024), 1672–81, https://doi.org/10.1145/3630106.3658996.

122. "Napster," *Wikipedia*, March 2023, https://en.wikipedia.org/w/index.php?title=Napster&oldid=1142267636.

123. Vitalik Buterin, "Bitcoin Is Not Losing Its Soul – Or, Why the Regulation Hysteria Is Missing the Point," *Bitcoin Magazine - Bitcoin News, Articles and Expert Insights*, June 2013, https://bitcoinmagazine.com/culture/bitcoin-is-not-losing-its-soul-or-why-the-regulation-hysteria-is-missing-the-point-1370302865.

124. Laura Weidinger et al., "Ethical and Social Risks of Harm from Language Models" (arXiv, December 2021), https://doi.org/10.48550/arXiv.2112.04359.

125. Gal, "ChatGPT Is a Data Privacy Nightmare, and We Ought to Be Concerned"; Xudong Pan et al., "Privacy Risks of General-Purpose Language Models," in 2020 IEEE Symposium *on* Security *and* Privacy (SP), 2020, 1314–31, https://doi.org/10.1109/SP40000.2020.00095.

126. Lucas Bourtoule et al., "Machine Unlearning" (arXiv, December 2020), https://doi.org/10.48550/arXiv.1912.03817; Salvatore Raieli, "Machine Unlearning: The Duty of Forgetting," *Medium*, September 2022, https://towardsdatascience.com/machine-unlearning-the-duty-of-forgetting-3666e5b9f6e5; Tom Simonite, "Now That Machines Can Learn, Can They Unlearn?" *Wired*, accessed Febru-

ary 21, 2023, https://www.wired.com/story/machines-can-learn-can-they-unlearn/.

127. Jimmy Z. Di et al., "Hidden Poison: Machine Unlearning Enables Camouflaged Poisoning Attacks" (arXiv, December 2022), https://doi.org/10.48550/arXiv.2212.10717.

128. Nicholas Carlini et al., "Extracting Training Data from Diffusion Models" (arXiv, January 2023), https://doi.org/10.48550/arXiv.2301.13188; Samuel Yeom et al., "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting" (arXiv, May 2018), https://doi.org/10.48550/arXiv.1709.01604.

129. Reza Shokri et al., "Membership Inference Attacks Against Machine Learning Models" (arXiv, March 2017), https://doi.org/10.48550/arXiv.1610.05820.

130. Maria Rigaki and Sebastian Garcia, "A Survey of Privacy Attacks in Machine Learning" (arXiv, April 2021), https://doi.org/10.48550/arXiv.2007.07646.

131. Robert LemosContributing Writer, Dark ReadingMarch 07, and 2023, "Employees Are Feeding Sensitive Business Data to ChatGPT," *Dark Reading*, March 2023, https://www.darkreading.com/risk/employees-feeding-sensitive-business-data-chatgpt-raising-security-fears; Cameron Coles, "3.1% of Workers Have Pasted Confidential Company Data into ChatGPT," *Cyberhaven*, February 2023, https://www.cyberhaven.com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt/; Noor Al-Sibai, "Amazon Begs Employees Not to Leak Corporate Secrets to ChatGPT," *Futurism*, 2023, https://futurism.com/the-byte/amazon-begs-employees-chatgpt.

132. Gowthami Somepalli et al., "Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models" (arXiv, December 2022), https://doi.org/10.48550/arXiv.2212.03860; Nicholas Carlini et al., "Quantifying Memorization Across Neural Language

Models" (arXiv, February 2022), https://doi.org/10.48550/arXiv.2202.07646; Ching-Yuan Bai et al., "On Training Sample Memorization: Lessons from Benchmarking Generative Modeling with a Large-Scale Competition," in *Proceedings of the 27th* ACM SIGKDD Conference *on* Knowledge Discovery *&* Data Mining, 2021, 2534–42, https://doi.org/10.1145/3447548.3467198; Karla Ortiz, "'The Images Below Aren't @McCurryStudios 'Afghan Girl.'',' *Twitter*, November 2022, https://twitter.com/kortizart/status/1588915427018559490.

133. Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe, "Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes" (arXiv, October 2021), https://doi.org/10.48550/arXiv.2110.01963; Billy Perrigo, "Exclusive: The $2 Per Hour Workers Who Made ChatGPT Safer," *Time*, January 2023, https://time.com/6247678/openai-chatgpt-kenya-workers/.

134. Chloe Xiang and Emanuel Maiberg, "ISIS Executions and Non-Consensual Porn Are Powering AI Art," *Vice*, September 2022, https://www.vice.com/en/article/93ad75/isis-executions-and-non-consensual-porn-are-powering-ai-art.

135. Ruiz et al., "DreamBooth"; Edwards, "Thanks to AI, It's Probably Time to Take Your Photos Off the Internet."

136. Benj Edwards, "Artist Finds Private Medical Record Photos in Popular AI Training Data Set," *Ars Technica*, September 2022, https://arstechnica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/.

137. Catherine Yeo, "How Biased Is GPT-3?" *Fair Bytes*, June 2020, https://www.fairbytes.org/post/how-biased-is-gpt-3; Meredith Whittaker et al., "Disability, Bias, and AI," November 2019, https://www.microsoft.com/en-us/research/publication/disability-bias-and-ai/; Dastin, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women"; Poppy Noor, "Can We Trust

AI Not to Further Embed Racial Bias and Prejudice?" *BMJ* 368 (February 2020): m363, https://doi.org/10.1136/bmj.m363.

138. Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Proceedings of the* 1st Conference *on* Fairness, Accountability *and* Transparency (PMLR, 2018), 77–91, https://proceedings.mlr.press/v81/buolamwini18a.html; Vongani H. Maluleke et al., "Studying Bias in GANs Through the Lens of Race" (arXiv, September 2022), https://doi.org/10.48550/arXiv.2209.02836.

139. Gianluca Mauro and Hilke Schellmann, "'There Is No Standard': Investigation Finds AI Algorithms Objectify Women's Bodies," *The Guardian*, February 2023, https://www.theguardian.com/technology/2023/feb/08/biased-ai-algorithms-racy-women-bodies.

140. "FTC Finalizes Settlement with Photo App Developer Related to Misuse of Facial Recognition Technology," *Federal Trade Commission*, May 2021, https://www.ftc.gov/news-events/news/press-releases/2021/05/ftc-finalizes-settlement-photo-app-developer-related-misuse-facial-recognition-technology; Natasha Lomas, "FTC Settlement with Ever Orders Data and AIs Deleted After Facial Recognition Pivot," *TechCrunch*, January 2021, https://techcrunch.com/2021/01/12/ftc-settlement-with-ever-orders-data-and-ais-deleted-after-facial-recognition-pivot/.

141. "FTC Chair Khan and Officials from DOJ, CFPB and EEOC Release Joint Statement on AI," *Federal Trade Commission*, April 2023, https://www.ftc.gov/news-events/news/press-releases/2023/04/ftc-chair-khan-officials-doj-cfpb-eeoc-release-joint-statement-ai; Lauren Feiner, "U.S. Regulators Warn They Already Have the Power to Go After A.I. Bias — and They're Ready to Use It," CNBC, April 2023, https://www.cnbc.com/2023/04/25/us-regulators-warn-they-already-have-the-power-to-go-after-ai-bias.html.

142. "Regulatory Framework Proposal on Artificial Intelligence Shaping Europe's Digital Future," February 2023, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

143. "What Are the GDPR Fines?" *GDPR.eu*, July 2018, https://gdpr.eu/fines/.

144. "Intelligenza Artificiale: Il Garante Blocca ChatGPT. Raccolta Illecita Di Dati Personali. Assenza Di Sistemi Per La Verifica Dell'età Dei Minori," March 2023, https://www.garanteprivacy.it:443/home/docweb/-/docweb-display/docweb/9870847.

145. Supantha Mukherjee, Elvira Pollina, and Rachel More, "Italy's ChatGPT Ban Attracts EU Privacy Regulators," *Reuters*, April 2023, https://www.reuters.com/technology/germany-principle-could-block-chat-gpt-if-needed-data-protection-chief-2023-04-03/.

146. Toby Sterling, "European Privacy Watchdog Creates ChatGPT Task Force," *Reuters*, April 2023, https://www.reuters.com/technology/european-data-protection-board-discussing-ai-policy-thursday-meeting-2023-04-13/.

147. Barr, "GPT-4 Is a Giant Black Box and Its Training Data Remains a Mystery."

148. Kyle Wiggers, "Addressing Criticism, OpenAI Will No Longer Use Customer Data to Train Its Models by Default," *TechCrunch*, March 2023, https://techcrunch.com/2023/03/01/addressing-criticism-openai-will-no-longer-use-customer-data-to-train-its-models-by-default/.

149. Al-Sibai, "Amazon Begs Employees Not to Leak Corporate Secrets to ChatGPT."

150. Dan Milmo, "ChatGPT Reaches 100 Million Users Two Months After Launch," *The Guardian*, February 2023, https://

www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app.

151. Chloe Xiang, "OpenAI Is Now Everything It Promised Not to Be: Corporate, Closed-Source, and For-Profit," *Vice*, February 2023, https://www.vice.com/en/article/5d3naz/openai-is-now-everything-it-promised-not-to-be-corporate-closed-source-and-for-profit.

152. Council, "OpenAI Admits Some Premium Users' Payment Info Was Exposed"; Kan, "OpenAI Confirms Leak of ChatGPT Conversation Histories."

153. "Data Breach Reporting Requirements Explained [2022] GDPR Register," December 2021, https://www.gdprregister.eu/gdpr/data-breach-notification-requirements/.

154. Wiggers, "Addressing Criticism, OpenAI Will No Longer Use Customer Data to Train Its Models by Default."

155. David Fraser, "Federal Privacy Watchdog Probing OpenAI, ChatGPT Following Complaint CBC News," *CBC*, April 2023, https://www.cbc.ca/news/politics/privacy-commissioner-investigation-openai-chatgpt-1.6801296.

156. David Shepardson, Diane Bartz, and Diane Bartz, "US Begins Study of Possible Rules to Regulate AI Like ChatGPT," *Reuters*, April 2023, https://www.reuters.com/technology/us-begins-study-possible-rules-regulate-ai-like-chatgpt-2023-04-11/.

157. Vinayshekhar Bannihatti Kumar, Rashmi Gangadharaiah, and Dan Roth, "Privacy Adhering Machine Un-Learning in NLP" (arXiv, December 2022), https://doi.org/10.48550/arXiv.2212.09573.

158. Log data not getting cleared automatically when it should is not uncommon and often only discovered when the server runs out of storage space and crashes.

159. Timothy B. Lee, "Copyright Lawsuits Pose a Serious Threat to Generative AI," March 2023, https://www.understandingai.org/p/copyright-lawsuits-pose-a-serious.

160. Jonathan Bailey, "The Wave of AI Lawsuits Have Begun," *Plagiarism Today*, January 2023, https://www.plagiarismtoday.com/2023/01/17/the-wave-of-ai-lawsuits-have-begun/.

161. James Vincent, "Getty Images Is Suing the Creators of AI Art Tool Stable Diffusion for Scraping Its Content," *The Verge*, January 2023, https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit; Vincent.

162. Lee, "Copyright Lawsuits Pose a Serious Threat to Generative AI."

163. "Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence," *Federal Register*, March 2023, https://www.federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance-works-containing-material-generated-by-artificial-intelligence; Smithsonian Magazine and Jane Recker, "U.S. Copyright Office Rules A.I. Art Can't Be Copyrighted," *Smithsonian Magazine*, 2023, https://www.smithsonianmag.com/smart-news/us-copyright-office-rules-ai-art-cant-be-copyrighted-180979808/; Adi Robertson, "The US Copyright Office Says an AI Can't Copyright Its Art," *The Verge*, February 2022, https://www.theverge.com/2022/2/21/22944335/us-copyright-office-reject-ai-generated-art-recent-entrance-to-paradise.

164. P. Bernt Hugenholtz and João Pedro Quintais, "Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output?" *IIC - International Review of Intellectual Property and Competition Law* 52, no. 9 (October 2021): 1190–1216, https://doi.org/10.1007/s40319-021-01115-0; Pieter De Grauwe and Sacha Gryspeerdt, "Artificial Intelligence (AI): The Qualification of AI Creations as "Works" Under EU Copyright Law," *Gevers*, June 2022, https://

www.gevers.eu/blog/artificial-intelligence/artificial-intelligence-ai-the-qualification-of-ai-creations-as-works-under-eu-copyright-law/.

165. "AI-Generated Art: Who Owns the Copyright?" web_content, *The Lighthouse*, January 2020, https://lighthouse.mq.edu.au/article/december-2019/AI-generated-art-who-owns-the-copyright; Courtney White and Rita Matulionyte, "Artificial Intelligence Painting The Bigger Picture For Copyright Ownership," {SSRN} {Scholarly} {Paper} (Rochester, NY, December 2019), https://doi.org/10.2139/ssrn.3498673.

166. "Who Owns the Copyright in AI-Generated Art?" *Murgitroyd, European Patent & Trade Mark Attorneys*, October 2022, https://www.murgitroyd.com/en-us/blog/who-owns-the-copyright-in-ai-generated-art/.

167. Carlini et al., "Quantifying Memorization Across Neural Language Models."

168. Bai et al., "On Training Sample Memorization."

169. "GitHub Copilot · Your AI Pair Programmer," *GitHub*, accessed April 5, 2023, http://archive.today/2023.01.11-224507/https://github.com/features/copilot.

170. I *strongly* suggest that you make sure that this setting is enabled if you use GitHub Copilot.

171. Somepalli et al., "Diffusion Art or Digital Forgery?"

172. Katherine Lee et al., "Deduplicating Training Data Makes Language Models Better," in *Proceedings of the 60th* Annual Meeting *of the* Association *for* Computational Linguistics (Volume 1: Long Papers) (Dublin, Ireland: Association for Computational Linguistics, 2022), 8424–45, https://doi.org/10.18653/v1/2022.acl-long.577.

173. Remember, OpenAI keeps its processes and training data set *confidential*. We don't know if they've taken any of these measures.

174. Ortiz, "'Afghan Girl'."

175. Melissa Heikkilä, "AI Models Spit Out Photos of Real People and Copyrighted Images," *MIT Technology Review*, 2023, https://www.technologyreview.com/2023/02/03/1067786/ai-models-spit-out-photos-of-real-people-and-copyrighted-images/.

176. Vitaly Feldman and Chiyuan Zhang, "What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation," in *Advances in* Neural Information Processing Systems, vol. 33 (Curran Associates, Inc., 2020), 2881–91, https://papers.nips.cc/paper/2020/hash/1e14bfe2714193e7af5abc64ecbd6b46-Abstract.html; Carlini et al., "Quantifying Memorization Across Neural Language Models."

177. Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel, "Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets," in *Proceedings of the 16th* Conference *of the* European Chapter *of the* Association *for* Computational Linguistics: Main Volume (Online: Association for Computational Linguistics, 2021), 1000–1008, https://doi.org/10.18653/v1/2021.eacl-main.86; Xiaosen Zheng and Jing Jiang, "An Empirical Study of Memorization in NLP," in *Proceedings of the 60th* Annual Meeting *of the* Association *for* Computational Linguistics (Volume 1: Long Papers) (Dublin, Ireland: Association for Computational Linguistics, 2022), 6265–78, https://doi.org/10.18653/v1/2022.acl-long.434.

178. Gavin Brown et al., "When Is Memorization of Irrelevant Training Data Necessary for High-Accuracy Learning?" in *Proceedings of the 53rd* Annual ACM SIGACT Symposium *on* Theory *of* Computing, STOC 2021 (New York, NY, USA: Association for Computing Machinery, 2021), 123–32, https://doi.org/10.1145/3406325.3451131.

179. Vitaly Feldman, "Does Learning Require Memorization? A Short Tale about a Long Tail," in *Proceedings of the 52nd* Annual ACM SIGACT Symposium *on* Theory *of* Computing, STOC 2020 (New York, NY, USA: Association for Computing Machinery, 2020), 954–59, https://doi.org/10.1145/3357713.3384290.

180. Lee et al., "Deduplicating Training Data Makes Language Models Better."

181. Except for Somepalli et al., "Diffusion Art or Digital Forgery?" that I cited above. Because it adopted similarity measures that more closely correspond to the legal definition of plagiarism, the roughly 2% number it arrived at is more likely to be an accurate measure of the system's plagiarism rate than GitHub's 1% estimate.

182. Lewis Kaplan, "Mannion v. Coors Brewing Co." July 2005, https://h2o.law.harvard.edu/cases/2353.

183. Nicholas Carlini et al., "Poisoning Web-Scale Training Datasets Is Practical" (arXiv, February 2023), https://doi.org/10.48550/arXiv.2302.10149.

184. Edwards, "Artist Finds Private Medical Record Photos in Popular AI Training Data Set"; Al-Sibai, "Amazon Begs Employees Not to Leak Corporate Secrets to ChatGPT."

185. Huseyin A. Inan et al., "Training Data Leakage Analysis in Language Models" (arXiv, February 2021), https://doi.org/10.48550/arXiv.2101.05405.

186. Nicholas Carlini et al., "Extracting Training Data from Large Language Models" (arXiv, June 2021), https://doi.org/10.48550/arXiv.2012.07805; Carlini et al., "Extracting Training Data from Diffusion Models."

187. Dave Lee, "Someone Asked ChatGPT If It Had a Signal… And It Gave Him *MY NUMBER*." *Twitter*, February 2023, https://twitter.com/DaveLeeFT/status/1626288109339176962.

188. Yeom et al., "Privacy Risk in Machine Learning."

189. Emma McGowan, "Is ChatGPT's Use of People's Data Even Legal?" February 2023, https://blog.avast.com/chatgpt-data-use-legal.

190. Birhane, Prabhu, and Kahembwe, "Multimodal Datasets"; Ivana Bartoletti, "ChatGPT Gender Bias," *Twitter*, March 2023, https://twitter.com/IvanaBartoletti/status/1637401609079488512.

191. Khari Johnson, "Chatbots Got Big—and Their Ethical Red Flags Got Bigger," *Wired*, accessed February 21, 2023, https://www.wired.com/story/chatbots-got-big-and-their-ethical-red-flags-got-bigger/; Mutale Nkonde, "ChatGPT: New AI System, Old Bias?" *Mashable*, February 2023, https://mashable.com/article/chatgpt-ai-racism-bias; Yeo, "How Biased Is GPT-3?"

192. Abubakar Abid, Maheen Farooqi, and James Zou, "Persistent Anti-Muslim Bias in Large Language Models" (arXiv, January 2021), https://doi.org/10.48550/arXiv.2101.05783; "The Radicalization Risks of GPT-3 and Neural Language Models Middlebury Institute of International Studies at Monterey," September 2020, https://www.middlebury.edu/institute/academics/centers-initiatives/ctec/ctec-publications/radicalization-risks-gpt-3-and-neural-language.

193. Manish Raghavan et al., "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices," in *Proceedings of the* 2020 Conference *on* Fairness, Accountability, *and* Transparency, 2020, 469–81, https://doi.org/10.1145/3351095.3372828; Whittaker et al., "Disability, Bias, and AI"; Dastin, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women"; Mona Sloane, Emanuel Moss, and Rumman Chowdhury, "A Silicon Valley Love Triangle: Hiring Algorithms, Pseudo-Science, and the Quest for

Auditability" (arXiv, May 2022), https://doi.org/10.48550/arXiv.2106.12403.

194. Buolamwini and Gebru, "Gender Shades"; Rachael Tatman, "Gender and Dialect Bias in YouTube's Automatic Captions," in *Proceedings of the* First ACL Workshop *on* Ethics *in* Natural Language Processing (Valencia, Spain: Association for Computational Linguistics, 2017), 53–59, https://doi.org/10.18653/v1/W17-1606.

195. Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy, "An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-Trained Language Models," in *Proceedings of the* 60*th* Annual Meeting *of the* Association *for* Computational Linguistics (Volume 1: Long Papers) (Dublin, Ireland: Association for Computational Linguistics, 2022), 1878–98, https://doi.org/10.18653/v1/2022.acl-long.132.

196. Eric Wallace et al., "Concealed Data Poisoning Attacks on NLP Models," in *Proceedings of the* 2021 Conference *of the* North American Chapter *of the* Association *for* Computational Linguistics: Human Language Technologies (Online: Association for Computational Linguistics, 2021), 139–50, https://doi.org/10.18653/v1/2021.naacl-main.13; Eugene Bagdasaryan and Vitaly Shmatikov, "Spinning Language Models: Risks of Propaganda-As-A-Service and Countermeasures," in 2022 IEEE Symposium *on* Security *and* Privacy (SP), 2022, 769–86, https://doi.org/10.1109/SP46214.2022.9833572.

197. Matthew Jagielski et al., "Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning," in 2018 IEEE Symposium *on* Security *and* Privacy (SP), 2018, 19–35, https://doi.org/10.1109/SP.2018.00057.

198. Keita Kurita, Paul Michel, and Graham Neubig, "Weight Poisoning Attacks on Pretrained Models," in *Proceedings of the* 58*th* Annual Meeting *of the* Association *for* Computational Linguistics

(Online: Association for Computational Linguistics, 2020), 2793–2806, https://doi.org/10.18653/v1/2020.acl-main.249.

199. Shafi Goldwasser et al., "Planting Undetectable Backdoors in Machine Learning Models : [Extended Abstract]," in 2022 IEEE *63rd* Annual Symposium *on* Foundations *of* Computer Science (FOCS), 2022, 931–42, https://doi.org/10.1109/FOCS54457.2022.00092.

200. Di et al., "Hidden Poison."

201. Xinyun Chen et al., "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning" (arXiv, December 2017), https://doi.org/10.48550/arXiv.1712.05526; Jonas Geiping et al., "Witches' Brew: Industrial Scale Data Poisoning via Gradient Matching" (arXiv, May 2021), https://doi.org/10.48550/arXiv.2009.02276; Nicholas Carlini and Andreas Terzis, "Poisoning and Backdooring Contrastive Learning" (arXiv, March 2022), https://doi.org/10.48550/arXiv.2106.09667; Jiazhu Dai and Chuanshuai Chen, "A Backdoor Attack Against LSTM-Based Text Classification Systems" (arXiv, June 2019), https://doi.org/10.48550/arXiv.1905.12457.

202. Fnu Suya et al., "Model-Targeted Poisoning Attacks with Provable Convergence," in *Proceedings of the 38th* International Conference *on* Machine Learning (PMLR, 2021), 10000–10010, https://proceedings.mlr.press/v139/suya21a.html.

203. Carlini et al., "Poisoning Web-Scale Training Datasets Is Practical"; Nicholas Carlini, "Poisoning the Unlabeled Dataset of {Semi-Supervised} Learning," 2021, 1577–92, https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-poisoning.

204. Health Sector Coordinating Council, "Health Industry Cybersecurity-Artificial Intelligence-Machine Learning Health Sector Council," February 2023, https://healthsectorcouncil.org/health-industry-cybersecurity-artificial-intelligence-machine-learning/; AKM

Iqtidar Newaz et al., "Adversarial Attacks to Machine Learning-Based Smart Healthcare Systems" (arXiv, October 2020), https://doi.org/10.48550/arXiv.2010.03671.

205. Virat Shejwalkar et al., "Back to the Drawing Board: A Critical Evaluation of Poisoning Attacks on Production Federated Learning," in 2022 IEEE Symposium *on* Security *and* Privacy (SP), 2022, 1354–71, https://doi.org/10.1109/SP46214.2022.9833647.

206. El-Mahdi El-Mhamdi et al., "SoK: On the Impossible Security of Very Large Foundation Models" (arXiv, September 2022), https://doi.org/10.48550/arXiv.2209.15259.

207. Stephanie Lin, Jacob Hilton, and Owain Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods," in *Proceedings of the 60th* Annual Meeting *of the* Association *for* Computational Linguistics (Volume 1: Long Papers) (Dublin, Ireland: Association for Computational Linguistics, 2022), 3214–52, https://doi.org/10.18653/v1/2022.acl-long.229.

208. Drew McDermott, "Artificial Intelligence Meets Natural Stupidity," *ACM SIGART Bulletin*, no. 57 (April 1976): 4–9, https://doi.org/10.1145/1045339.1045340.

209. Mitchell, "Why AI Is Harder Than We Think."

210. I fully admit that my use of the term *plagiarism* in the previous chapter is an attempt to use the same tactic on the other side of the scale, which I think is defensible by virtue of it being a genuine legal risk for businesses.

211. Joy, "ChatGPT Cites Economics Papers That Do Not Exist," *Economist Writing Every Day*, January 2023, https://economistwritingeveryday.com/2023/01/21/chatgpt-cites-economics-papers-that-do-not-exist/; Dave Karpf, "On Generative AI, Phantom Citations, and Social Calluses," Substack newsletter, *The Future, Now and Then*, March 2023, https://davekarpf.substack.com/p/on-generative-ai-phantom-citations.

212. “Don’t Believe ChatGPT - We Do NOT Offer a "Phone Lookup" Service,” February 2023, https://blog.opencagedata.com/post/dont-believe-chatgpt.

213. Noah Cook, “ChatGPT Creates a Fake History of Ukraine,” *Med-Mastodon*, February 2023, https://med-mastodon.com/@UncivilServant/109867060485276220.

214. Alexander Hanff, “ChatGPT Should Be Considered a Malevolent AI and Destroyed,” March 2023, https://www.theregister.com/2023/03/02/chatgpt_considered_harmful/.

215. Simon Willison, “I Promise You ChatGPT Can’t Access the Internet, Even Though It Really Looks Like It Can,” March 2023, http://simonwillison.net/2023/Mar/10/chatgpt-internet-access/.

216. “GPT-4.”

217. Sara Fischer, “Exclusive: GPT-4 Readily Spouts Misinformation, Study Finds,” March 2023, https://www.msn.com/en-us/news/technology/exclusive-gpt-4-readily-spouts-misinformation-study-finds/ar-AA18TtVg.

218. Vittoria Dentella et al., “Testing AI Performance on Less Frequent Aspects of Language Reveals Insensitivity to Underlying Meaning” (arXiv, February 2023), https://doi.org/10.48550/arXiv.2302.12313.

219. Bender et al., “On the Dangers of Stochastic Parrots.”

220. Chirag Shah and Emily M. Bender, “Situating Search,” in ACM SIGIR Conference *on* Human Information Interaction *and* Retrieval, CHIIR ’22 (New York, NY, USA: Association for Computing Machinery, 2022), 221–32, https://doi.org/10.1145/3498366.3505816; Heikkilä, “Why You Shouldn’t Trust AI Search Engines”; Nelson F. Liu, Tianyi Zhang, and Percy Liang, “Evaluating Verifiability in

Generative Search Engines" (arXiv, April 2023), https://doi.org/10.48550/arXiv.2304.09848.

221. Nick Diakopoulos, "Can We Trust Search Engines with Generative AI? A Closer Look at Bing's Accuracy for News Queries," *Medium*, February 2023, https://medium.com/@ndiakopoulos/can-we-trust-search-engines-with-generative-ai-a-closer-look-at-bings-accuracy-for-news-queries-179467806bcc; Emily M. Bender and Chirag Shah, "All-Knowing Machines Are a Fantasy," *IAI TV - Changing How the World Thinks*, December 2022, https://iai.tv/articles/all-knowing-machines-are-a-fantasy-auid-2334; Tae Kim, "Let's Stop Pretending—ChatGPT Isn't That Smart," *Barrons*, February 2023, https://www.barrons.com/articles/chatgpt-ai-openai-chatbot-b9f4fa03.

222. Avery Rowe, "Don't Ask an AI for Plant Advice • Tradescantia Hub," March 2023, https://tradescantia.uk/article/dont-ask-an-ai-for-plant-advice/.

223. Lin, Hilton, and Evans, "TruthfulQA."

224. Yejin Bang et al., "A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity" (arXiv, February 2023), https://doi.org/10.48550/arXiv.2302.04023.

225. James Vincent, "Google Announces AI Features in Gmail, Docs, and More to Rival Microsoft," *The Verge*, March 2023, https://www.theverge.com/2023/3/14/23639273/google-ai-features-docs-gmail-slides-sheets-workspace; Tom Warren, "Microsoft Announces Copilot: The AI-Powered Future of Office Documents," *The Verge*, March 2023, https://www.theverge.com/2023/3/16/23642833/microsoft-365-ai-copilot-word-outlook-teams.

226. Ziqiang Cao et al., "Faithful to the Original: Fact Aware Neural Abstractive Summarization" (arXiv, November 2017), https://doi.org/10.48550/arXiv.1711.04434; Joshua Maynez et al., "On Faithfulness and Factuality in Abstractive Summarization," in

*Proceedings of the 58th* Annual Meeting *of the* Association *for* Computational Linguistics (Online: Association for Computational Linguistics, 2020), 1906–19, https://doi.org/10.18653/v1/2020.acl-main.173; Sihao Chen et al., "Improving Faithfulness in Abstractive Summarization with Contrast Candidate Generation and Selection," in *Proceedings of the* 2021 Conference *of the* North American Chapter *of the* Association *for* Computational Linguistics: Human Language Technologies (Online: Association for Computational Linguistics, 2021), 5935–41, https://doi.org/10.18653/v1/2021.naacl-main.475.

227. Lin, Hilton, and Evans, "TruthfulQA."

228. Chengwei Qin et al., "Is ChatGPT a General-Purpose Natural Language Processing Task Solver?" (arXiv, February 2023), https://doi.org/10.48550/arXiv.2302.06476; Bang et al., "A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity."

229. "GPT-4."

230. See p. 3 of Huan Yee Koh et al., "How Far Are We from Robust Long Abstractive Summarization?" in *Proceedings of the* 2022 Conference *on* Empirical Methods *in* Natural Language Processing (Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022), 2682–98, https://aclanthology.org/2022.emnlp-main.172, for example. Most research on hallucinations in abstractive summarisation seems to be based on the ArXiV or GovReport data sets, which are all single documents, written in a consistent style and tone.

231. Birhane, Prabhu, and Kahembwe, "Multimodal Datasets"; Xiang and Maiberg, "ISIS Executions and Non-Consensual Porn Are Powering AI Art."

232. Perrigo, "Exclusive."

233. Willison, "Bing"; "Microsoft's New ChatGPT AI Starts Sending 'Unhinged' Messages to People," *The Independent*, February 2023, https://www.independent.co.uk/tech/chatgpt-ai-messages-microsoft-bing-b2282491.html; Steph Maj Swanson, "ChatGPT Generated Child Sex Abuse When Asked to Write BDSM Scenarios," *Vice*, March 2023, https://www.vice.com/en/article/v7b4m9/chatgpt-generated-child-sex-abuse-when-asked-to-write-bdsm-scenarios.

234. Heikkilä, "The Viral AI Avatar App Lensa Undressed Me—Without My Consent."

235. James Vincent, "Anyone Can Use This AI Art Generator — That's the Risk," *The Verge*, September 2022, https://www.theverge.com/2022/9/15/23340673/ai-image-generation-stable-diffusion-explained-ethics-copyright-data; CRYPTOINSIGHT PRO Crypto, "Stable Diffusion 2.0 Has "Forgotten" How to Generate NSFW Content," Substack newsletter, *CRYPTOINSIGHT.PRO Crypto NFT DeFi*, November 2022, https://cryptoinsightpro.substack.com/p/technologies2.

236. Ng Wai Foong, "Stable Diffusion 2: The Good, The Bad and The Ugly," *Medium*, December 2022, https://towardsdatascience.com/stable-diffusion-2-the-good-the-bad-and-the-ugly-bd44bc7a1333, "Also, the new model performs poorly for images related to human anatomy and faces of celebrities."

237. Matthias Bastian, "Stable Diffusion V2 Removes NSFW Images and Causes Protests," *THE DECODER*, November 2022, https://the-decoder.com/stable-diffusion-v2-removes-nude-images-and-causes-protests/.

238. Jim Clyde Monge, "Stable Diffusion 2.1 Released — NSFW Image Generation Is Back," *MLearning.ai*, December 2022, https://medium.com/mlearning-ai/stable-diffusion-2-1-released-nsfw-image-generation-is-back-8bcc5c069d60.

239. Josh Taylor, "ChatGPT's Alter Ego, Dan: Users Jailbreak AI Program to Get Around Ethical Safeguards," *The Guardian*, March 2023, https://www.theguardian.com/technology/2023/mar/08/chatgpt-alter-ego-dan-users-jailbreak-ai-program-to-get-around-ethical-safeguards; Goodin, "Hackers Are Selling a Service That Bypasses ChatGPT Restrictions on Malware."

240. "FSMB Physician Census," accessed April 25, 2023, https://www.fsmb.org/physician-census/.

241. "Distribution of Physicians Across Regions by Gender Worldwide 2000-2018," *Statista*, accessed April 25, 2023, https://www.statista.com/statistics/1099789/distribution-of-physicians-across-regions-worldwide-by-gender/.

242. "Infographic: Female Doctors by Country," *Statista Infographics*, August 2018, https://www.statista.com/chart/14983/female-doctors-by-country.

243. Sayash Kapoor and Arvind Narayanan, "Quantifying ChatGPT's Gender Bias," Substack newsletter, *AI Snake Oil*, April 2023, https://aisnakeoil.substack.com/p/quantifying-chatgpts-gender-bias; Hadas Kotek, "#ChatGPT Doubles down on Gender Stereotypes Even When They Don't Make Sense in Context," *Twitter*, April 2023, https://twitter.com/HadasKotek/status/1648453764117041152.

244. I highly recommend this survey of current research on ChatGPT's performance at natural language tasks. Pikuliak, "ChatGPT Survey."

245. Carlini et al., "Quantifying Memorization Across Neural Language Models."

246. Lin, Hilton, and Evans, "TruthfulQA."

247. Rogers, "Closed AI Models Make Bad Baselines"; Barr, "GPT-4 Is a Giant Black Box and Its Training Data Remains a Mystery."

248. Casey Ross, "Epic's Overhaul of a Flawed Algorithm Shows Why AI Oversight Is a Life-or-Death Issue," *STAT*, October 2022, https://www.statnews.com/2022/10/24/epic-overhaul-of-a-flawed-algorithm/; This story is especially interesting in light of Epic's decision to return to the AI market in collaboration with Microsoft: "Microsoft and Epic Expand Strategic Collaboration with Integration of Azure OpenAI Service," *Stories*, April 2023, https://news.microsoft.com/2023/04/17/microsoft-and-epic-expand-strategic-collaboration-with-integration-of-azure-openai-service/.

249. Andrew Wong et al., "External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients," *JAMA Internal Medicine* 181, no. 8 (August 2021): 1065–70, https://doi.org/10.1001/jamainternmed.2021.2626.

250. Shih-Cheng Huang et al., "Developing Medical Imaging AI for Emerging Infectious Diseases," *Nature Communications* 13, no. 1 (November 2022): 7060, https://doi.org/10.1038/s41467-022-34234-4.

251. The pandemic led to an increased investment in many aspects of medical research, not just in AI, but the success rate of the AI COVID experiments seems to have been unusually low. See: Laure Wynants et al., "Prediction Models for Diagnosis and Prognosis of Covid-19: Systematic Review and Critical Appraisal," *BMJ (Clinical Research Ed.)* 369 (April 2020): m1328, https://doi.org/10.1136/bmj.m1328; Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee, "AI for Radiographic COVID-19 Detection Selects Shortcuts over Signal," *Nature Machine Intelligence* 3, no. 7 (July 2021): 610–19, https://doi.org/10.1038/s42256-021-00338-7.

252. Branco et al., "Shortcutted Commonsense."

253. Though, possibly still not high enough. See Lauren Oakden-Rayner, "No Doctor Required: Autonomy, Anomalies, and Magic Puddings," *Lauren Oakden-Rayner*, July 2022, https://laurenoakden-rayner.com/2022/07/04/no-doctor-required-autonomy-anomalies-and-magic-puddings/ for a discussion of why some of our existing processes might be too lax to properly assess the risks of using AI in healthcare. The sepsis model cited above was approved for use. It was only updated after external researchers decided to validate the vendor's claims.

254. Mirzadeh et al., "GSM-Symbolic."

255. Branco et al., "Shortcutted Commonsense."

256. Lewis, Stenetorp, and Riedel, "Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets."

257. This overview is a particularly damning analysis of OpenAI's claims of GPT-4's breakthrough performance at a variety of professional benchmarks. Narayanan and Kapoor, "GPT-4 and Professional Benchmarks."

258. Jang and Lukasiewicz, "Consistency Analysis of ChatGPT."

259. Max Hauptman, "Marines Outwitted an AI Security Camera by Hiding in a Cardboard Box and Pretending to Be Trees," *Task & Purpose*, January 2023, https://taskandpurpose.com/news/marines-ai-paul-scharre/.

260. Financial Times, "Man Beats Machine at Go in Human Victory over AI," *Ars Technica*, February 2023, https://arstechnica.com/information-technology/2023/02/man-beats-machine-at-go-in-human-victory-over-ai/.

261. Tony T. Wang et al., "Adversarial Policies Beat Superhuman Go AIs" (arXiv, February 2023), https://doi.org/10.48550/arXiv.2211.00241.

262. Kapoor and Narayanan, “Leakage and the Reproducibility Crisis in ML-Based Science.”

263. Raji et al., “The Fallacy of AI Functionality.”

264. “If Builders Built Buildings the Way Programmers Wrote Programs, Then the First Woodpecker That Came Along Would Destroy Civilization – Quote Investigator®,” September 2019, https://quoteinvestigator.com/2019/09/19/woodpecker/.

265. Steven D. Fraser et al., “"No Silver Bullet" Reloaded: Retrospective on "Essence and Accidents of Software Engineering"," in *Companion to the 22nd* ACM SIGPLAN *Conference on* Object-*Oriented Programming Systems and Applications Companion* (Montreal Quebec Canada: ACM, 2007), 1026–30, https://doi.org/10.1145/1297846.1297973.

266. To a point. Remember “Don’t Believe ChatGPT - We Do NOT Offer a "Phone Lookup" Service.”

267. I remain sceptical of this idea. You usually get better results in programming by *starting* with the test and then implementing code that passes the test. If these code generation tools end up promoting poor practices in writing tests, then that would more than nullify the benefit they provide in other ways.

268. Elaine Atwell, “GitHub Copilot Isn’t Worth the Risk,” *Kolide*, February 2023, https://www.kolide.com/blog/github-copilot-isn-t-worth-the-risk; Bradley M. Kuhn, “If Software Is My Copilot, Who Programmed My Software?” *Software Freedom Conservancy*, February 2022, https://sfconservancy.org/blog/2022/feb/03/github-copilot-copyleft-gpl/; “Analyzing the Legal Implications of GitHub Copilot - FOSSA,” *Dependency Heaven*, July 2021, https://fossa.com/blog/analyzing-legal-implications-github-copilot/.

269. Jeremy Howard, "Fast.ai - Is GitHub Copilot a Blessing, or a Curse?" July 2021, https://www.fast.ai/posts/2021-07-19-copilot.html.

270. K. Mosier and L. Skitka, "Human Decision Makers and Automated Decision Aids: Made for Each Other?" 1996, https://www.semanticscholar.org/paper/Human-Decision-Makers-and-Automated-Decision-Aids%3A-Mosier-Skitka/ffb65e76ac46fd42d595ed9272296f0cbe8ca7aa.

271. Kathleen L. Mosier et al., "Automation Bias: Decision Making and Performance in High-Tech Cockpits," *The International Journal of Aviation Psychology* 8, no. 1 (January 1998): 47–63, https://doi.org/10.1207/s15327108ijap0801_3; Kathleen L. Mosier et al., "Automation Bias, Accountability, and Verification Behaviors," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 40, no. 4 (October 1996): 204–8, https://doi.org/10.1177/154193129604000413.

272. Raja Parasuraman and Victor Riley, "Humans and Automation: Use, Misuse, Disuse, Abuse," *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39, no. 2 (June 1997): 230–53, https://doi.org/10.1518/001872097778543886.

273. "Anchoring Bias," *The Decision Lab*, accessed April 3, 2023, https://thedecisionlab.com/biases/anchoring-bias.

274. "Morgan Stanley TMT Conference," accessed August 8, 2023, https://www.microsoft.com/en-us/Investor/events/FY-2023/Morgan-Stanley-TMT-Conference#:~:text=Scott%20Guthrie%3A%20I%20think%20you%27re,is%20now%20A

275. Hammond Pearce et al., "Asleep at the Keyboard? Assessing the Security of GitHub Copilot's Code Contributions" (arXiv, December 2021), https://doi.org/10.48550/arXiv.2108.09293; Neil Perry et al., "Do Users Write More Insecure Code with AI Assistants?" (arXiv, December 2022), https://doi.org/10.48550/arXiv.2211.03622.

276. Tyler Glaiel, "Can GPT-4 *Actually* Write Code?" Substack newsletter, *Tyler's Substack*, March 2023, https://tylerglaiel.substack.com/p/can-gpt-4-actually-write-code.

277. Luca Rossi, "On AI Freshness, the Pyramid Principle, and Hierarchies 💡," April 2023, https://refactoring.fm/p/on-ai-freshness-the-pyramid-principle.

278. "GitHub Copilot · Your AI Pair Programmer."

279. Ted Nelson, *Computer* Lib/Dream Machines (Place of publication not identified, 1974).

280. Jeffrey Dastin and Jeffrey Dastin, "Amazon Extends Moratorium on Police Use of Facial Recognition Software," *Reuters*, May 2021, https://www.reuters.com/technology/exclusive-amazon-extends-moratorium-police-use-facial-recognition-software-2021-05-18/.

281. "A Little History of the World Wide Web," accessed April 6, 2023, https://www.w3.org/History.html.

282. Vannevar Bush, "As We May Think," *The Atlantic*, July 1945, https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/.

283. Stacey Mason and Mark Bernstein, "On Links: Exercises in Style," in *Proceedings of the 30th* ACM Conference *on* Hypertext *and* Social Media, HT '19 (New York, NY, USA: Association for Computing Machinery, 2019), 103–10, https://doi.org/10.1145/3342220.3343665.

284. Joseph Cox, "'Disrespectful to the Craft:' Actors Say They're Being Asked to Sign Away Their Voice to AI," *Vice*, February 2023, https://www.vice.com/en/article/5d37za/voice-actors-sign-away-rights-to-artificial-intelligence.

285. "DoNotPay - The World's First Robot Lawyer," accessed April 6, 2023, https://donotpay.com/.

286. Benjamin Cassidy, "The Twisted Life of Clippy," *Seattle Met*, August 2022, https://www.seattlemet.com/news-and-city-life/2022/08/origin-story-of-clippy-the-microsoft-office-assistant.

287. Ge Wang and Juliana Bidadanure, "Humans in the Loop: The Design of Interactive AI Systems," *Stanford HAI*, October 2019, https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems.

288. Ross, "Epic's Overhaul of a Flawed Algorithm Shows Why AI Oversight Is a Life-or-Death Issue."

289. Parasuraman and Riley, "Humans and Automation."

290. Mosier et al., "Automation Bias."

291. Narayanan and Kapoor, "People Keep Anthropomorphizing AI. Here's Why."

292. Shanahan, "Talking About Large Language Models."

293. Perry et al., "Do Users Write More Insecure Code with AI Assistants?"

294. Mosier and Skitka, "Human Decision Makers and Automated Decision Aids."

295. Mosier et al., "Automation Bias, Accountability, and Verification Behaviors."

296. "On Seeing Human."

297. Salles, Evers, and Farisco, "Anthropomorphism in AI," esp. "In the general public it inadvertently promotes misleading interpretations of and beliefs about what AI is and what its capacities are." Anthropomorphism also limits the researchers, which is impor-

tant to note in light of the common belief in the field that the spark of AGI has been struck: "Furthermore, anthropomorphic (implicit or explicit) interpretations of AI might also have epistemological impact on the AI research community itself, insofar as the search for biological and psychological realism (i.e., similarity with biological intelligence) might lead to underestimating the possibility of new theoretical and operational paradigms and frameworks thus ultimately limiting the development of AI."

298. *Computer Power and Human Reason*.

299. Weizenbaum, 6.

300. Mike Moore and Joel Khalili last updated, "AWS Went down Hard, yet Again - Here's What Happened," *TechRadar*, December 2021, https://www.techradar.com/news/live/aws-is-down-again-heres-all-we-know.

301. Nicholas Fearn, "Heat Waves Are Shutting Down Data Centers and Breaking the Internet," *Gizmodo*, December 2022, https://gizmodo.com/heat-waves-climate-change-data-center-server-shutdown-1849916741.

302. Sarah Myers West, "Competition Authorities Need to Move Fast and Break up AI," *Financial Times*, April 2023. "Without the robust enforcement of competition laws, generative AI could irreversibly cement Big Tech's advantage, giving a handful of companies power over technology that mediates much of our lives."

303. "GPT-3," *Wikipedia*, April 2023, https://en.wikipedia.org/w/index.php?title=GPT-3&oldid=1147823352.

304. James Governor, "The Great Flowering: Why OpenAI Is the New AWS and the New Kingmakers Still Matter." *James Governor's Monkchips*, April 2023, https://redmonk.com/jgovernor/2023/04/13/the-great-flowering-why-openai-is-the-new-aws-and-the-new-kingmakers-still-matter/.

305. Bender et al., "On the Dangers of Stochastic Parrots."

306. Karen Hao, "We Read the Paper That Forced Timnit Gebru Out of Google. Here's What It Says." *MIT Technology Review*, 2020, Karen Haoarchive page.

307. Davey Alba and Julia Love, "Google's Rush to Win in AI Led to Ethical Lapses, Employees Say," *Bloomberg.com*, April 2023, https://www.bloomberg.com/news/features/2023-04-19/google-bard-ai-chatbot-raises-ethical-concerns-from-employees, "Even after the public pronouncements, some found it difficult to work on ethical AI at Google. One former employee said they asked to work on fairness in machine learning and they were routinely discouraged — to the point that it affected their performance review. Managers protested that it was getting in the way of their 'real work,' the person said."

308. Shah and Bender, "Situating Search."

309. Simon Willison, "Thoughts and Impressions of AI-Assisted Search from Bing," February 2023, http://simonwillison.net/2023/Feb/24/impressions-of-bing/.

310. "Microsoft's New ChatGPT AI Starts Sending 'Unhinged' Messages to People"; Willison, "Bing."

311. Brereton, "Bing AI Can't Be Trusted."

312. Diakopoulos, "Can We Trust Search Engines with Generative AI?"

313. Zoë Schiffer, "Microsoft Just Laid Off One of Its Responsible AI Teams," March 2023, https://www.platformer.news/p/microsoft-just-laid-off-one-of-its.

314. Tom Warren, "You Can Play with Microsoft's Bing GPT-4 Chatbot Right Now, No Waitlist Necessary," *The Verge*, March 2023, https://www.theverge.com/2023/3/15/23641683/microsoft-bing-ai-gpt-4-chatbot-available-no-waitlist.

315. Aaron Holmes and Kevin McLaughlin, "Ghost Writer: Microsoft Looks to Add OpenAI's Chatbot Technology to Word, Email," *The Information*, January 2023, https://www.theinformation.com/articles/ghost-writer-microsoft-looks-to-add-openais-chatbot-technology-to-word-email; Benj Edwards, "Microsoft Aims to Reduce 'Tedious' Business Tasks with New AI Tools," *Ars Technica*, March 2023, https://arstechnica.com/information-technology/2023/03/microsoft-brings-chatgpt-style-ai-to-developer-and-analysis-tools/.

316. Gerrit De Vynck and Will Oremus, "As AI Booms, Tech Firms Are Laying Off Their Ethicists," *Washington Post*, March 2023, https://www.washingtonpost.com/technology/2023/03/30/tech-companies-cut-ai-ethics/; Will Knight, "Elon Musk Has Fired Twitter's 'Ethical AI' Team," *Wired*, accessed April 27, 2023, https://www.wired.com/story/twitter-ethical-ai-team/.

317. Nico Grant and Karen Weise, "In A.I. Race, Microsoft and Google Choose Speed Over Caution," *The New York Times*, April 2023, https://www.nytimes.com/2023/04/07/technology/ai-chatbots-google-microsoft.html.

318. Howard, "Fast.ai - Is GitHub Copilot a Blessing, or a Curse?"

319. Cyril Connolly, *Enemies of Promise*, Rev. ed., University of Chicago Press ed (Chicago: University of Chicago Press, 2008), 109.

320. Saffron Huang and Divya Siddarth, "Generative AI and the Digital Commons," *The Collective Intelligence Project*, February 2023, https://cip.org/research/generative-ai-digital-commons.

321. "Retrieval-Augmented Generation," *Wikipedia*, November 2024, https://en.wikipedia.org/w/index.php?title=Retrieval-augmented_generation&oldid=1259558262.

322. Tom DeMarco and Tim Lister, *Peopleware: Productive Projects and Teams (3rd Edition)*, 3rd ed. (Addison-Wesley Professional, 2013).

323. W. Edwards Deming, *Out of the* Crisis (The MIT Press, 2018), https://doi.org/10.7551/mitpress/11457.001.0001.

324. U. M. Franklin, *The Real World of Technology*, Massey Lectures Series (Anansi, 1999), https://books.google.is/books?id=LziQT3YS_2sC.

325. P. G. Smith and D. G. Reinertsen, *Developing Products in Half the Time: New Rules, New Tools* (Wiley, 1997), https://books.google.is/books?id=oOhmCgAAQBAJ.

326. "Apple and Google Drop Navalny App After Kremlin Piles on Pressure," *Financial Times*, September 2021.

327. "Apple's Censorship in China Is Just the Tip of the Iceberg," *Columbia Journalism Review*, accessed November 26, 2024, https://www.cjr.org/the_media_today/apple_appstore_china_censorship.php.

328. "Meta in Myanmar (Full Series)," accessed November 26, 2024, https://erinkissane.com/meta-in-myanmar-full-series.

329. D. McQuillan, *Resisting AI: An Anti-Fascist Approach to Artificial Intelligence* (Bristol University Press, 2022), https://books.google.is/books?id=R6x6EAAAQBAJ.