The Future of Search

Donald MacKenzie

An edited version of this draft article was published in the London Review of Books of 20 November 2025: https://www.lrb.co.uk/the-paper/v47/n21/donald-mackenzie/the-future-of-search

The fact-checking links at the end were not published, but are included here as the equivalent of references.

Type a few words into Google, tap ‘go’ or press the return key. Almost instantaneously, relevant links arrive on your phone or laptop. To see them, you may have to scroll past a bit of clutter, ads, and – these days – an AI Overview. Even if your query is obscure, and mine often are, it's nevertheless quite likely that one of those links will take you to what you’re looking for. That's striking, given that there are probably over a billion websites worldwide, and more than 50 billion webpages.

Around that everyday miracle, a company currently worth around $2.2 trillion was built, but Google’s future is now far from certain. It was founded in September 1998, at which point the World Wide Web, to which it became an indispensable guide, was less than a decade old. Google’s was by no means the first web search engine, but its older competitors had been weakened by ‘spamming’, much of it by the owners of the web’s already prevalent porn sites. Just as Google was to do, those earlier search engines deployed ‘web crawlers’ to find websites and ingest their contents, so providing the raw materials for an electronic index. They then used that index to find sites whose contents seemed best to match the words in the user’s query.

What a spammer such as the owner of a porn site could therefore do was to plaster their site with words that, while irrelevant to the site's content, were likely to appear in web searches. Often hidden from human sight – for example, in the same colour as the background – those words would still be ingested by web crawlers. By the late 1990s, it was perfectly possible to enter an entirely innocent search query – 'skiing', 'beach holidays', 'best colleges' – and be served a bunch of links to porn.

In the second half of the 1990s, Google's co-founders, Larry Page and Sergey Brin, were PhD students in Stanford University's Computer Science Department. Among the problems on which Page was working was how to increase the chances that the first entries you would see in the comments section of a website would be useful, even authoritative. To ensure that, as Page told the tech journalist and historian of Google, Steven Levy, 'We needed a rating system'.

That got Page and Brin thinking about how websites themselves could be rated. Page – working in a leading academic department, and also, as Levy puts it, 'a child of academia': his father was a professor at Michigan State University – was struck by the analogy between the incoming links to a website and the citations that an authoritative scientific paper receives. The greater the number of those links, the higher the probability that the site was well-regarded, especially if the links were from sites that were themselves high-quality.

Using thousands of human beings to rate millions of websites wasn't necessary, Page and Brin decided. 'It's all recursive', as Page told a 2001 panel attended by Levy. 'How good you are is determined by who links to you,' and how good they are is determined by who links to them. 'It's all a big circle. But mathematics is great. You can solve this.' Their algorithm, PageRank, did not entirely defeat porn and other forms of

spam – one of Google's engineers, Matt Cutts, used to organise a Look for Porn Day before it implemented each new version of its web index, with home-baked chocolate-chip cookies as a reward for those who found it – but PageRank helped Google improve substantially on earlier spam-weakened search engines.

Page's undramatic word, recursive, hid a giant material challenge. You can't find the incoming links to a website just by examining it. To find them, you have to go to the sites that link to it, which of course involves not just crawling all of them, but (since you don't know in advance which those are) also crawling multitudes of sites that don't link to the site in question. The logic of what Page and Brin set out to do therefore quickly led them to a hugely ambitious goal: to ingest and index effectively every website in existence. That, in essence, is what Google still does, with the exception of sites that bar crawlers or request not to be indexed.

One way of trying to ingest the entire web would have been for Google to buy the most powerful computers available. At its launch it had around $100,000 in the bank, and it raised $25 million from venture capitalists in 1999, but that wasn't enough to pay for a decent number of expensive machines. So Google's engineers lined metal trays with electrically insulating cork, and packed them with low cost computer hardware of the kind found in cheap PCs. One early Google employee, Douglas Edwards, remembers visiting the Santa Clara data centre in which Google was renting space for the machines with which it was crawling the web, indexing it, and generating rankings. 'Every square inch was crammed with racks bristling with stripped down CPUs [central processing units],' he writes. 'There were twenty-one racks and more than fifteen hundred machines, each sprouting cables like Play-Doh pushed through a spaghetti

press. … where other [firms'] cages were right angled and inorganic, Google's swarmed with life, a giant termite mound dense with frenetic activity and intersecting curves.'

By June 2000, Google's bargain-basement web crawlers had ingested over a billion webpages. In the months before that, though, Google had encountered its most serious early crisis. Cheap machines crash, and cheap computer memory readily gets corrupted by overheating or even the impact of cosmic rays. James Somers, writing in the *New Yorker* in 2018, describes how Google's crawlers kept failing, making it hard or even impossible to update its most crucial data structure, its web index. The solution the company found was to change computing forever. It was to think of computing not as what a single machine did, but as what could be done if a 'computer' was a warehouse containing tens of thousands of machines, automatically managed to circumvent the inevitable failures of individual machines. As Somers puts it, Google learned how to give its programmers the capacity 'to wield the machines in its data centers as if they were a single, planet-size computer'.

Think, though, about what is involved in programming a warehouse-size or even planet-size computer. (The latter is an exaggeration, but pardonable: as early as 2011, Levy suggests, Google may have been deploying a million machines, still mostly cheap commodity hardware, in the roughly two dozen data centres it then occupied worldwide.) How do you parallelise a giant data analysis task, in other words distribute it across a huge ensemble of machines? How do you organise the communication that then needs to take place among those machines? What does your program do when, as is inevitable when it is running on tens of thousands or more machines, one or more of them crashes during a crucial computation that has to come up with results in real

time, as a search engine does? Might your engineers and researchers have to spend all their time dealing with such issues, rather than getting on with the data analysis?

Astonishingly, given the centrality of those questions to what was rapidly becoming digital capitalism, Google openly published its answer to them. It took the form of an electronically available preprint of a paper presented in 2004 to a San Francisco computer-science symposium by two Google engineers, Jeff Dean and Sanjay Ghemawat. In it, they described the system that they had built for Google. MapReduce, as they called it, permitted programs to be 'automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system.'

There is a sense in which Dean and Ghemawat's 2004 paper launched the age of big data. The paper 'was a pretty nice gift', says computer scientist Doug Cutting, then working for the web portal Yahoo. Google didn't actually release the code of MapReduce, but Dean and Ghemawat said enough to prompt Cutting and his colleague Mike Cafarella to lead the production of Hadoop, a fully public, free, open-source analogue to MapReduce. If you are involved in the analysis of big data – as many people in the tech sector are – you probably aren't now directly using MapReduce or Hadoop, but most likely one or more of the systems you use is derived from them.

With MapReduce and related innovations, Google discovered, initially painfully, how to scale, the process at the very heart of digital capitalism. That Google had done it gave others confidence that they could do it too, and the fact that Google didn't try too

hard to keep its innovations secret helped others to learn how. And once you are confident that you can scale, goals that would have seemed hopelessly over-ambitious can suddenly feel within your grasp. If Google's systems could now seamlessly ingest and index close to the entirety of the world's webpages and respond to billions of search queries every day, then why not also begin to ingest all the world's books (Google Books), or create an often detailed, interactive digital map of the surface of the globe (Google Maps), along with another interactive mesh of views of the planet from satellites and aircraft (Google Earth), and panoramic images of all its streets, at least in countries that allow Google's camera-carrying cars (Google Street View)? And, while you're at it, why not offer everyone a free, high-quality email service, and not worry too much about the strain on your servers if, as has happened, well over a billion people sign up for a Gmail account?

All of that involved Google's creation, collection and assembly of a historically unprecedented quantity of data. The US National Security Agency might previously have come close, though we don't really know, and Facebook was soon to do so: it was narrower than Google in the scope of its activities, but richer in the data that its users uploaded about themselves and their lives. Nevertheless, the title of Daniel Soar's article on Google in the *LRB of 6* October 2011 remains the perfect two-word characterisation: 'It knows'.

Given that, you might have expected that the legal troubles that are now accumulating around Google would concern all this data. On balance, though, Google seems to have handled the contents of its data mountains pretty responsibly. Its legal difficulties centre instead on whether it plays too dominant a role in advertising markets. In 2015, Google was incorporated into a new holding company, Alphabet, but

the latter's 'other bets', as it calls them in its financial statements, contributed less than 1% of the $350 billion it earned in 2024. Google's fast growing cloud computing business was more substantial, contributing 12%. However, fully three quarters of Alphabet's 2024 revenues were from advertising, the bulk of them (nearly $200 billion) coming, as it has done ever since Google's earliest years, from selling the ads that accompany the results of Google searches.

In two lawsuits, the Department of Justice, along with several US states, has accused Google of monopolising two different areas of advertising. The first case, heard by a federal court in the District of Columbia, focuses on the markets for 'general search' (i.e. search of the kind for which you use Google, not the more specific searches for which you might use Amazon, Facebook or Expedia) and for the standard ads, often consisting simply of text, that accompany those general searches. The Justice Department had little difficulty in establishing that, in the words of that court's August 2024 judgement, Google possesses 'a dominant market share' in general search: in the US, 89.2% overall, 'which increases to 94.9% on mobile devices'. Unsurprisingly, that dominance was mirrored in the market for the accompanying advertising, with Google's share of what the court calls 'the text ads market' being 88% in 2020.

'Google', the court ruled, 'is a monopolist'. It 'has violated section 2 of the Sherman Act' – the 1890 foundation stone of US competition law – 'by maintaining its monopoly in two product markets in the United States – general search services and general text advertising'. Core to that verdict is the court's conclusion that Google has done so via what the court calls 'exclusive distribution agreements', by which Google secures both the presence of its 'search widget' on the home screen of Android phones

and its position as the default search engine for Apple's Safari and Mozilla's Firefox browsers.

Those agreements involve Google sharing the resultant search-advertising revenues with the phone manufacturers, Apple and the Mozilla Foundation. The longest-standing agreement, dating from 2002, is with Apple, although through prior to 2005 it did not include revenue sharing. It's also the most important agreement, in that Apple devices account for more than half of general search queries in the United States, and the court estimates that in 2022 it brought Apple around $20 billion. The exact percentage of revenues that go to Apple is redacted in the court documents, but an expert witness for Google is reported to have said in open court that it is 36%.

The world of giant digital platforms often pivots on surprisingly small matters, such as whether users will spend a few seconds doing something that they don't absolutely need to do. Google being the default doesn't stop you using a different search engine. Apple, for example, makes it perfectly easy for you to switch. It takes no more than around 20 seconds to open Safari's settings on your laptop (or the search settings on your iPhone) and change the default to, let's say, Microsoft's Bing. From then on, if a query that you enter into Safari's address bar isn't answered directly by Apple's systems, it will go to Bing, not Google, and the revenues from the associated ads will flow to Microsoft, with, as far as I am aware, no revenue going to Apple. But most likely you've never even tried switching search engines. That might be because you actively want to use Google, but perhaps like me (and, I fear, most people) you are a lazy sod and tend unthinkingly to stick with the digital world's preloaded default options.

The second competition-law case, heard by a federal court in Virginia, is more esoteric in that it concerns the innards of digital advertising, rather than an issue that

you could yourself directly control such as the choice between browsers. Two systems are central to the Virginia case. The first is Google's 'ad server'. That's a cloud service that Google sells to publishers (not just news publishers but providers of online content of all kinds), and it takes the final decision about which ads to show you when you visit the publisher's website. The second is Google's ad exchange, AdX, which conducts ad trading in real time. Visit the *Guardian*'s website, for example, and the opportunity to advertise to you is usually auctioned on AdX and similar, smaller exchanges.

The Virginia court found that Google's ad server is responsible for roughly 90% of open-web display ads globally, and AdX has a 63-71% share of the corresponding ad trading, 'nine times that of its next closest rival'. The court concedes that how Google has acted is in some respects quite different from a traditional monopolist: for example, it has not raised the fees it charges publishers for use of its ad server. The court concludes, nevertheless, that Google has acted to preserve its structural centrality to the ad server and ad exchange markets, thus 'acquiring and maintaining monopoly power' in ways that it says violate US competition law.

The appeals process has only just begun, and Google's lawyers, will, I'm sure, continue to argue that, for example, it gets default position because it's the best search engine. The counterargument, though, is that if Google is the best, it may at least in part be because it has the most users. Algorithms such as PageRank are far from the only thing that matters to search quality. There's a great deal that can be learned from simple things such as whether users immediately return to the search results page after clicking on a link in that page, which suggests that the link wasn't to what they were looking for. The more users that a search engine has, the more data of this kind its engineers can employ to improve it.

In the background to both cases is a notoriously contested issue in US competition law: defining the relevant market. If, for instance, it's 'general search' of the kind conducted using Google, then Google indeed has a very large market share. But expand the definition to include digital searches of all kinds, and Google's share diminishes, with Amazon, in particular, starting to seem a powerful rival. For reasons such as this, the District of Columbia and Virginia courts both rejected parts of the Department of Justice's cases against Google, and I would expect issues of market definition to be prominent in the appeals.

If Google loses its appeals, the 'remedies', as a competition lawyer would call them, are also still up for grabs. The most obvious measure would be for the court in the search-advertising case to require that Google ends its revenue-sharing agreements, such as with Apple. But that might not cause Google to lose too much market share, if only because users offered an explicit choice between search engines may well opt for the one with which they are familiar and whose name has become a verb. Mozilla, for instance, has repeatedly experimented with changing the default in Firefox to a different search engine, and found large proportions of its users switching back to Google, often almost immediately.

Another likely remedy would be Google being instructed to divest itself of its ad server and AdX, its ad exchange. The details of how they work are hugely important to publishers' income and therefore for journalism, but the money they bring into Google isn't, by its standards, huge. What Alphabet calls 'Google Network', the kind of advertising for which the ad server and AdX are the infrastructure, accounted for less than 9% of Alphabet's revenues in 2024, and it could certainly survive having to sell the two systems.

A quite different, and probably more serious, threat to Google is a development it itself did a great deal to foster: the emergence of large language models and chatbots based upon them, most famously the start-up Open AI's ChatGPT. Google's researchers have worked for more than twenty years on what a computer scientist would call 'natural language processing' – Google Translate, for example, dates from 2006 – and it became one of the pioneers of applying neural networks to the task. These are computational structures (now often gigantic) that were originally thought to be loosely analogous to the brain's neurons. They are not programmed in detail by their human developers: they learn from examples, these days often billions of examples.

The efficiency with which a neural network learns is affected strongly by its structure or 'architecture'. A pervasive issue in natural language processing, for example, is what linguists call 'coreference resolution'. Take the sentence: 'The animal didn't cross the street because it was too tired'. The 'it' could refer to the animal or to the street. We humans resolve such ambiguities all the time, and if it takes conscious thought, it's often a sign what you're reading is badly written. Coreference resolution is, however, a much harder problem for a computer system, even a sophisticated neural network.

In August 2017, Google's Jakob Uszkoreit uploaded to the company's research blog a post about a new architecture for neural networks that he and his colleagues called the Transformer. Neural networks were by then already powering Google Translate, but still made mistakes in, for example, coreference resolution, which can become embarrassingly evident when translating English into a gendered language such as French. Uszkoreit's example was the sentence I've just quoted. 'L'animal' is masculine, and 'la rue' feminine, so the correct translation should end 'il était trop

fatigué', but Google Translate was still rendering it as 'elle était trop fatiguée', presumably because in the sentence's word order 'street' is closer than 'animal' to the word 'it'.

The Transformer, Uszkoreit reported, was much less likely to make mistakes such as that, because it 'directly models relationships between all words in a sentence, regardless of their respective position'. Enabling a neural network to do complex things such as coreference resolution had been thought to require a network architecture with a complicated structure. The Transformer was structurally simpler, 'dispensing with recurrence and convolutions entirely', as Uszkoreit and seven Google or ex-Google colleagues put it in a 2017 paper that gave a much fuller account of their innovation. That simplicity meant, they said, that the Transformer was 'more parallelizable' than earlier architectures. Using it, you could more readily divide language processing into computational subtasks that could run simultaneously, rather than one after the other.

And, just as Dean and Ghemawat had done, the authors of the Transformer paper made it publicly available, submitting it to AI's leading annual meeting, Neural Information Processing Systems. Having outgrown university venues, it was held in 2017 in the Convention Center in Long Beach. The paper's authors would have needed Google management permission to present it there, and they must have got it. That decision was to be highly consequential, but the Transformer's significance wasn't immediately self-evident, even to technically sophisticated people within Google. 'I read [the paper] with my coworkers', says one, 'but we thought it was just as interesting and promising as several other papers'. Whoever approved making the Transformer paper public presumably thought similarly and decided that publication should go ahead because there was no good reason to stop it. That's a commendable attitude.

Among those who read the Transformer paper was computer scientist Ilya Sutskever, co-founder of OpenAI, who says that ‘as soon as the paper came out, literally the next day, it was clear to me, to us, that Transformers addressed the limitations’ of the more complex neural-network architecture OpenAI had been using for language processing. The Transformer, in other words, should scale. As Karen Hao reports in her *Empire of AI*, Sutskever started ‘evangelizing’ for it within OpenAI, meeting some scepticism: ‘It felt like a wack idea,’ one of Sutskever’s OpenAI colleagues told her. Crucially, though, another member of its staff, Alec Radford, ‘began hacking away on his laptop, often late into the night, to scale Transformers just a little and observe what happened’.

Sutskever was right: the Transformer architecture did scale. It made genuinely large, indeed giant, language models, feasible. Its parallelisability meant that it could readily be implemented not on general-purpose computer chips but on graphics chips, originally designed primarily for rendering images in computer games, a task that has to be done very fast but is also highly parallelisable. (That Nvidia, the leading designer of graphics chips, provides much of the material foundation of large language models has made it the world's most valuable company, currently worth around 80% more than Alphabet/Google.) If you have enough of chips of that kind, you can do a huge amount of what’s called ‘pre-training’ of a Transformer model ‘generatively’, without any direct human input. That involves feeding the model huge corpora of text, generally scraped from the Internet, getting the model to generate what it thinks will be the next word in each piece of text, then the word after that, and so on, and having it continuously and automatically adjust its billions of parameters to improve its predictions. Only once you

have done enough of that pre-training do you start fine-tuning the model for more specific tasks.

It was OpenAI, not Google, that made the most decisive use of the Transformer. It made no attempt to hide its debt to the latter. OpenAI's evolving large language models are all called GPT, or Generative Pre-trained Transformer. Although GPT-1 and GPT-2 were not hugely impressive, its breakthrough came in 2020 with the much larger GPT-3. It didn't yet take the form of a chatbot that laypeople could use: ChatGPT was released only in November 2022.[1] However, developers in firms other than OpenAI were given access to GPT-3 from June 2020 onwards, and quickly found that it went well beyond previous systems in its capacity to produce large quantities of text (and, e.g., computer code) that were hard to distinguish from what a well-informed human being might write.

GPT-3's success intensified what was already a wave of enthusiasm in other tech firms for large language models, but it also caused unease, in particular to Timnit Gebru, co-founder of Black in AI and co-head of Google's Ethical AI team. Among those Gebru contacted about her concerns was the University of Washington computational linguist Emily Bender. With five co-authors, some of whom had to remain anonymous, Gebru and Bender wrote what has become the most famous critique of large language models, arguing that they don't really understand language. Instead, they suggested, a large language model is just a 'stochastic parrot … 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'.

[1] Paul Taylor has written extensively for both the *LRB* and its blog about ChatGPT, large language models and the 'stochastic parrots' paper: see, e.g., his articles in the *LRB*s of 5 January 2023 and 21 March 2024.

Training such a model consumes huge quantities of electricity, noted Bender, Gebru and colleagues, and the giant datasets used in that training often 'encode hegemonic views that are harmful to marginalised populations'. They quoted computer scientists Abeba Birhane and Vinay Uday Prabhu: 'Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy'.

Gebru obtained permission from her immediate manager to publish the paper, but others more senior in Google objected, and Gebru was asked either to retract it or to remove the names and affiliation of its authors within Google. She was prepared to do so only under conditions that were unacceptable to the Google executive with whom she was dealing, and Gebru ended up losing her job, an outcome that caused considerable unhappiness, with nearly 2,700 Google employees signing a letter of protest.

There has been much speculation about why it was OpenAI, not Google, that first turned the Transformer architecture into a world-famous chatbot. Part of the reason, ironically enough, seems to have been that some of the concerns underpinning the 'stochastic parrots' paper were shared by Google senior executives. In 2016, Microsoft had launched a Twitter chatbot, Tay, designed to interact with and learn from human users' tweets. It picked up the world's ugliness remarkably rapidly. A number of Twitter users deliberately fed it racist content, succeeding in training Tay to be something of an automated fascist – one user asked Tay 'Do you support genocide?' to which it responded 'i do indeed' – and forcing Microsoft to withdraw Tay within 24 hours of its release. Google's executives appear to have had no desire to repeat the debacle. Its researchers developed a Transformer-based chatbot called Meena (later renamed LaMDA, Language Model for Dialogue Applications) but did not get permission to

release it. A spokesperson for Google later told the *Wall Street Journal* that 'the chatbot had been through many reviews and barred from wider releases for various reasons over the years'.

A start-up such as OpenAI has to take risks, but major corporations often focus on improving their well-established products or services rather than innovating more radically. It was far from clear in advance that a sophisticated chatbot would become globally famous almost overnight, and even OpenAI itself was somewhat taken back by ChatGPT's enormously rapid uptake. One of the authors of the Transformer paper, Niki Parmar, told the *Financial Times*'s Madhumita Murgia that 'Google was an amazing place, but they wanted to optimise for the existing products'. While the Transformer was used to improve Google Translate and Google Search, only after ChatGPT's success did Google throw its huge resources into an all-out effort to launch a chatbot, Bard, with 80,000 members of staff donating their time to test it.

The successes of large language models have changed the digital world, and there are many reckonings to come. For anyone who, like me, teaches a subject in which student assessment is no longer primarily by traditional exams, the most pressing concern is how to react to those models' ability to generate essays that read like the work of a reasonably proficient if intellectually unambitious student. A more disturbing possibility is that models' capacity to do that may indicate something wrong with our pedagogy. Have we been teaching students to be (human) stochastic parrots?

For Google, the most specific reckoning concerns search. It's often easy to rephrase a search query as a prompt for a chatbot, and that is a clear threat to what has been, for a quarter of a century, Google's most important source of revenue. It's also probably not too difficult for a competent team of developers to build an automated

purchasing assistant on top of a large language model, and I'm already starting to read articles in the trade press about how to market products to those assistants rather than human beings. Google could certainly build assistants of this kind, but if they are adopted by the public at large they will reduce the demand for search ads, and in that reconfigured form of electronic commerce Google would not enjoy advantages of incumbency of the kind on which the antitrust litigation has focussed.

I'm starting to feel pre-emptively nostalgic when I do a Google search. Yes, search can take you to places to which you don't want to go. In 2010, when the information-science scholar Safiya Umoja Noble was using the search term 'black girls' to Google 'things on the Internet that might be interesting to my stepdaughter and nieces', the top hit presented to her was 'HotBlackPussy'. As she argued in her 2018 book, *Algorithms of Oppression*, search engines can indeed echo and reinforce a racist and sexist culture. But at least a 'classical' search engine, such as Google in the 2000s and 2010s, took you outside of itself, and perhaps implicitly prompted you to evaluate critically what you found there.

The experience of using a chatbot powered by a large language model is, in contrast, largely self-contained. You can generally prompt it to say something about its sources, but that's a bit like the 'further readings' at the end of a textbook chapter: yes, you know you really should read them, but probably you won't. It's seductively easy to treat a chatbot as an oracle. That's economically as well as cognitively dangerous. Classical search creates economic incentives for making new content available on the web, and keeping existing content up-to-date, because money can be made by advertising to the visitors brought to websites by a search engine. True, the process can also incentivise clickbait, but if search atrophies what will happen to the web? As the

commentator Eric Seufert nicely puts it, Google was the open web's 'imperfect benefactor'.

There's a road that needs crossed, by Google and also by the rest of us. On one side is a digital world with its largely familiar structure of search, websites, apps and social media platforms, fuelled by familiar kinds of advertising, in many of whose forms Google excels. To make liveable whatever lies on the road's other side involves multiple challenges. One is to avoid overreliance on self-contained systems trained on whatever text, speech, images, audio and video are digitally available – inevitably both a biased subset of what human beings are able to produce, and an impoverished version of what human knowledge consists in. Another challenge is to unlearn scale. Much of the success of large language models has come simply from making everything bigger: the number of parameters they contain, the quantity of data employed to train them, the size and energy intensity of the data centres in which they run. That trajectory is unsustainable, and not just environmentally: it's getting harder to find adequate volumes of fresh data on which to train new models, since much of what exists is already potentially compromised by having been generated by previous models. It's going to be a hard road to cross. Can we successfully cross it? Can Google? If it turns out to be too tired, I'll be a little sad.

Sources/fact-checking links

Over a billion websites: **https://siteefy.com/how-many-websites-are-there/**

Over 50 billion webpages: **https://www.digitalsilk.com/digital-trends/how-many-websites-are-there/**; **https://www.worldwidewebsize.com**

Google and Nvidia market capitalisation (14 July 2025): https://companiesmarketcap.com

In the same colour as the background: Battelle, John. 2005. *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture* (London: Brealey), p. 104.

'We needed a rating system': Steven Levy, *In the Plex: How Google Thinks, Works, and Shapes our Lives* (New York: Simon & Schuster, 2011), p. 16.

'a child of academia': Levy, p. 17.

'it's all recursive' etc.: Levy, p. 21.

Look for Porn Day: Levy, p. 54.

$100,000 in the bank; $25 million from venture capitalists: Levy, pp. 33-4 & 74.

'Every square inch was crammed': Douglas Edwards, *I'm Feeling Lucky: The Confessions of Google Employee Number 59* (New York: Houghton Mifflin Harcourt, 2011), p. 21.

By June 2000, … over a billion webpages: Levy, pp. 44-45.

James Somers: 'Binary Stars: The Friendship that Made Google Huge.' *New Yorker* 10 December (2018): 28-35, quote on p. 33. Available at https://www.newyorker.com/magazine/2018/12/10/the-friendship-that-made-google-huge.

A million machines; two dozen data centres: Levy, p. 181.

Well over a billion people take up the offer of a Gmail account: https://aovup.com/stats/email-users/.

Google's 2024 revenues:
https://abc.xyz/assets/59/ff/66de11a3bfa7db6cb929ba12a01b/6b3f31ad08de72f11d32f2b4f712b918.pdf, p. 64.

Search case judgement: Mehta, Amit P. 2024. "Memorandum Opinion: US District Court for the District of Columbia, Cases 20-cv-3010 and 2-cv-3715." Available at https://www.adexchanger.com/wp-content/uploads/2024/08/DOJ-Google-search-antitrust-Judge-Amit_Mehta-ruling.pdf. Quotes on p. 156, p. 189, p. 4, p. 276.

Apple's 'Internet Services Agreement' with Google and its history: Mehta, "Memorandum Opinion", pp. 101-111.

$20 billion in 2022 and redacted percentage: Mehta, "Memorandum Opinion", p. 103.

36% share of revenues: https://www.cnbc.com/2023/11/14/apple-gets-36percent-of-google-search-revenue-from-safari-alphabet-witness.html#:~:text=Google%20pays%20Apple%20more%20than,of%20Chicago%2C%20was%20not%20expected, drawing on Leah Nylen, "Apple Gets 36% of Google Revenue in Search Deal, Expert Says." *Bloomberg*, November 13. Available at https://www.bloomberg.com/news/articles/2023-11-13/apple-gets-36-of-google-revenue-from-search-deal-witness-says.

Ad server/ad exchange judgement: Brinkema, Leonie M. 2025. "Memorandum Opinion: US District Court for the Eastern District of Virginia, Case 1:23-cv-00108-LMB-JFA." Available at https://ag.ny.gov/sites/default/files/court-filings/united-states-of-america-et-al-v-google-llc-memorandum-opinion-2025.pdf. Market share data and quotes on pp. 73, 82, 83 and 1.

Mozilla's experiments: Mehta, "Memorandum Opinion", pp. 117-8.

Uszkoreit blog post "Transformers: A Novel Neural Network Architecture for Language Understanding." 31 August 2017. Available at https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/.

Transformer paper: Ashish Vaswani et al. 'Attention Is All You Need', presented at 31st Conference on Neural Information Processing Systems, 2017. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. Quotes from abstract, p. 1.

Neural Information Processing Systems 2017 conference:
https://neurips.cc/Conferences/2017

Sutskever 'as soon as the paper came out':
https://www.youtube.com/watch?v=HTSBvrJQF4U

Sutskever 'evangelizing', 'wack idea', Radford 'hacking away': Hao, Karen, *Empire of AI: Inside the Reckless Race for Total Domination* (London: Allen Lane, 2025), p. 121.

‘Stochastic parrots’: Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" Available at https://dl.acm.org/doi/10.1145/3442188.3445922, quotes on pp. 617 and 615.

‘The world’s beauty, ugliness, and cruelty’: Birhane, Abeba, and Vinay Uday Prabhu. 2021. "Large Image Datasets: A Pyrrhic Win for Computer Vision?" Available at https://ieeexplore.ieee.org/document/9423393. (They are citing Benjamin, Ruha. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity.)

Approval of and opposition to ‘Stochastic parrots’ paper: Hao, *Empire of AI*, pp. 165-174.

Letter of protest: https://googlewalkout.medium.com/standing-with-dr-timnit-gebru-isupporttimnit-believeblackwomen-6dadc300d382.

Tay chatbot: Lee, Dave. 2016. "Tay: Microsoft Issues Apology Over Racist Chatbot Fiasco." BBC News, 25 March. Available at https://www.bbc.co.uk/news/technology-35902104

Meena Transformer-based: https://venturebeat.com/ai/meena-is-googles-attempt-at-making-true-conversational-ai/

LaMDA Transformer-based: https://en.wikipedia.org/wiki/LaMDA

Non-release of Meena/LaMDA: Kruppa, Miles, and Schechner, Sam. 2023. "How Google Became Cautious of AI and Gave Microsoft an Opening." *Wall Street Journal*, 7 March. Available at https://www.wsj.com/tech/ai/google-ai-chatbot-bard-chatgpt-rival-bing-a4c2d2ad

OpenAI taking aback by ChatGPT’s success: Hao, *Empire of AI*, pp. 259-60.

Parmar quote: Murgia, Madhumita, "Transformers: The Google Scientists who Pioneered an AI Revolution." *Financial Times*, 23 July 2023. Available at https://www.ft.com/content/37bb01af-ee46-4483-982f-ef3921436a50.

80,000 Google members of staff test Bard:
https://www.cnbc.com/2023/03/21/google-ceo-pichai-memo-to-employees-on-bard-ai-things-will-go-wrong.html

‘Black girls’: Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York University Press), p. 3.

‘Imperfect benefactor’: Seufert, Eric, https://mobiledevmemo.com/google-was-the-open-webs-imperfect-benefactor/.