AI and the Big Five

Monday, January 9, 2023

The story of 2022 was the emergence of AI, first with image generation models, including DALL-E, MidJourney, and the open source Stable Diffusion, and then ChatGPT, the first text-generation model to break through in a major way. It seems clear to me that this is a new epoch in technology.

To determine how that epoch might develop, though, it is useful to look back 26 years to one of the most famous strategy books of all time: Clayton Christensen's The Innovator's Dilemma, particularly this passage on the different kinds of innovations:

> Most new technologies foster improved product performance. I call these sustaining technologies. Some sustaining technologies can be discontinuous or radical in character, while others are of an incremental nature. What all sustaining technologies have in common is that they improve the performance of established products, along the dimensions of performance that mainstream customers in major markets have historically valued. Most technological advances in a given industry are sustaining in character...
>
> Disruptive technologies bring to a market a very different value proposition than had been available previously. Generally, disruptive technologies underperform established products in mainstream markets. But they have other features that a few fringe (and generally new) customers value. Products based on disruptive technologies are typically cheaper, simpler, smaller, and, frequently, more convenient to use.

It seems easy to look backwards and determine if an innovation was sustaining or disruptive by looking at how incumbent companies fared after that innovation came to market: if the innovation was sustaining, then incumbent companies became stronger; if it was disruptive then presumably startups captured most of the value.

Consider previous tech epochs:

- The PC was disruptive to nearly all of the existing incumbents; these relatively inexpensive and low-powered devices didn't have nearly the capability or the profit margin of mini-computers, much less mainframes. That's why IBM was happy to outsource both the original PC's chip and OS to Intel and Microsoft, respectively, so that they could get a product out the door and satisfy their corporate customers; PCs got faster, though, and it was Intel and Microsoft that dominated as the market dwarfed everything that came before.
- The Internet was almost entirely new market innovation, and thus defined by completely new companies that, to the extent they disrupted incumbents, did so in industries far removed from technology, particularly those involving information (i.e. the media). This was the era of Google, Facebook, online marketplaces and e-commerce, etc. All of these applications ran on PCs powered by Windows and Intel.
- Cloud computing is arguably part of the Internet, but I think it deserves its own category. It was also extremely disruptive: commodity x86 architecture swept out dedicated server hardware, and an entire host of SaaS startups peeled off features from incumbents to build companies. What is notable is that the core infrastructure for cloud

computing was primarily built by the winners of previous epochs: Amazon, Microsoft, and Google. Microsoft is particularly notable because the company also transitioned its traditional software business to a SaaS service, in part because the company had already transitioned said software business to a subscription model.

- Mobile ended up being dominated by two incumbents: Apple and Google. That doesn't mean it wasn't disruptive, though: Apple's new UI paradigm entailed not viewing the phone as a small PC, a la Microsoft; Google's new business model paradigm entailed not viewing phones as a direct profit center for operating system sales, but rather as a moat for their advertising business.

What is notable about this history is that the supposition I stated above isn't quite right; disruptive innovations do consistently come from new entrants in a market, but those new entrants aren't necessarily startups: some of the biggest winners in previous tech epochs have been existing companies leveraging their current business to move into a new space. At the same time, the other tenets of Christensen's theory hold: Microsoft struggled with mobile because it was disruptive, but SaaS was ultimately sustaining because its business model was already aligned.

Given the success of existing companies with new epochs, the most obvious place to start when thinking about the impact of AI is with the big five: Apple, Amazon, Facebook, Google, and Microsoft.

Apple

I already referenced one of the most famous books about tech strategy; one of the most famous essays was Joel Spolsky's Strategy Letter V, particularly this famous line:

> Smart companies try to commoditize their products' complements.

Spolsky wrote this line in the context of explaining why large companies would invest in open source software:

> Debugged code is NOT free, whether proprietary or open source. Even if you don't pay cash dollars for it, it has opportunity cost, and it has time cost. There is a finite amount of volunteer programming talent available for open source work, and each open source project competes with each other open source project for the same limited programming resource, and only the sexiest projects really have more volunteer developers than they can use. To summarize, I'm not very impressed by people who try to prove wild economic things about free-as-in-beer software, because they're just getting divide-by-zero errors as far as I'm concerned.
>
> Open source is not exempt from the laws of gravity or economics. We saw this with Eazel, ArsDigita, The Company Formerly Known as VA Linux and a lot of other attempts. But something is still going on which very few people in the open source world really understand: a lot of very large public companies, with responsibilities to maximize shareholder value, are investing a lot of money in supporting open source software, usually by paying large teams of programmers to work on it. And that's what the principle of complements explains.
>
> Once again: demand for a product increases when the price of its complements decreases. In general, a company's strategic interest is going to be to get the price of their complements as low as possible. The lowest theoretically sustainable price would be the "commodity price" — the price that arises when you have a bunch of competitors offering indistinguishable goods. So, smart companies try to commoditize their products' complements. If you can do this, demand for your product will increase and you will be able to charge more and make more.

Apple invests in open source technologies, most notably the the Darwin kernel for its operating systems and the WebKit browser engine; the latter fits Spolsky's prescription as ensuring that the web works well with Apple devices makes Apple's devices more valuable.

Apple's efforts in AI, meanwhile, have been largely proprietary: traditional machine learning models are used for things like recommendations and photo identification and voice recognition, but nothing that moves the needle for Apple's business in a major way. Apple did, though, receive an incredible gift from the open source world: Stable Diffusion.

Stable Diffusion is remarkable not simply because it is open source, but also because the model is surprisingly small: when it was released it could already run on some consumer graphics cards; within a matter of weeks it had been optimized to the point where it could run on an iPhone.

Apple, to its immense credit, has seized this opportunity, with this announcement from its machine learning group last month:

> Today, we are excited to release optimizations to Core ML for Stable Diffusion in macOS 13.1 and iOS 16.2, along with code to get started with deploying to Apple Silicon devices...
>
> One of the key questions for Stable Diffusion in any app is where the model is running. There are a number of reasons why on-device deployment of Stable Diffusion in an app is preferable to a server-based approach. First, the privacy of the end user is protected because any data the user provided as input to the model stays on the user's device. Second, after initial download, users don't require an internet connection to use the model. Finally, locally deploying this model enables developers to reduce or eliminate their server-related costs...
>
> Optimizing Core ML for Stable Diffusion and simplifying model conversion makes it easier for developers to incorporate this technology in their apps in a privacy-preserving and economically feasible way, while getting the best performance on Apple Silicon. This release comprises a Python package for converting Stable Diffusion models from PyTorch to Core ML using diffusers and coremltools, as well as a Swift package to deploy the models.

It's important to note that this announcement came in two parts: first, Apple optimized the Stable Diffusion model itself (which it could do because it was open source); second, Apple updated its operating system, which thanks to Apple's integrated model, is already tuned to Apple's own chips.

Moreover, it seems safe to assume that this is only the beginning: while Apple has been shipping its so-called "Neural Engine" on its own chips for years now, that AI-specific hardware is tuned to Apple's own needs; it seems likely that future Apple chips, if not this year than probably next year, will be tuned for Stable Diffusion as well. Stable Diffusion itself, meanwhile, could be built into Apple's operating systems, with easily accessible APIs for any app developer.

This raises the prospect of "good enough" image generation capabilities being effectively built-in to Apple's devices, and thus accessible to any developer without the need to scale up a back-end infrastructure of the sort needed by the viral hit Lensa. And, by extension, the winners in this world end up looking a lot like the winners in the App Store era: Apple wins because its integration and chip advantage are put to use to deliver differentiated apps, while small independent app makers have the APIs and distribution channel to build new businesses.

The losers, on the other hand, would be centralized image generation services like Dall-E or MidJourney, and the cloud providers that undergird them (and, to date, undergird the aforementioned Stable Diffusion apps like Lensa). Stable

Diffusion on Apple devices won't take over the entire market, to be sure — Dall-E and MidJourney are both "better" than Stable Diffusion, at least in my estimation, and there is of course a big world outside of Apple devices, but built-in local capabilities will affect the ultimate addressable market for both centralized services and centralized compute.

Amazon

Amazon, like Apple, uses machine learning across its applications; the direct consumer use cases for things like image and text generation, though, seem less obvious. What is already important is AWS, which sells access to GPUs in the cloud.

Some of this is used for training, including Stable Diffusion, which according to the founder and CEO of Stability AI Emad Mostaque used 256 Nvidia A100s for 150,000 hours for a market-rate cost of $600,000 (which is surprisingly low!). The larger use case, though, is inference, i.e. the actual application of the model to produce images (or text, in the case of ChatGPT). Every time you generate an image in MidJourney, or an avatar in Lensa, inference is being run on a GPU in the cloud.

Amazon's prospects in this space will depend on a number of factors. First, and most obvious, is just how useful these products end up being in the real world. Beyond that, though, Apple's progress in building local generation techniques could have a significant impact. Amazon, though, is a chip maker in its own right: while most of its efforts to date have been focused on its Graviton CPUs, the company could build dedicated hardware of its own for models like Stable Diffusion and compete on price. Still, AWS is hedging its bets: the cloud service is a major partner when it comes to Nvidia's offerings as well.

The big short-term question for Amazon will be in gauging demand: not having enough GPUs will be leaving money on the table; buying too many that sit idle, though, would be a major cost for a company trying to limit them. At the same time, it wouldn't be the worst error to make: one of the challenges with AI is the fact that inference costs money; in other words, making something with AI has marginal costs.

This issue of marginal costs is, I suspect, an under-appreciated challenge in terms of developing compelling AI products. While cloud services have always had costs, the discrete nature of AI generation may make it challenging to fund the sort of iteration necessary to achieve product-market fit; I don't think it's an accident that ChatGPT, the biggest breakout product to-date, was both free to end users and provided by a company in OpenAI that both built its own model and has a sweetheart deal from Microsoft for compute capacity. If AWS had to sell GPUs for cheap that could spur more use in the long run.

That noted, these costs should come down over time: models will become more efficient even as chips become faster and more efficient in their own right, and there should be returns to scale for cloud services once there are sufficient products in the market maximizing utilization of their investments. Still, it is an open question as to how much full stack integration will make a difference, in addition to the aforementioned possibility of running inference locally.

Meta

I already detailed in Meta Myths why I think that AI is a massive opportunity for Meta and worth the huge capital expenditures the company is making:

Meta has huge data centers, but those data centers are primarily about CPU compute, which is what is needed to power Meta's services. CPU compute is also what was necessary to drive Meta's deterministic ad model, and the algorithms it used to recommend content from your network.

The long-term solution to ATT, though, is to build probabilistic models that not only figure out who should be targeted (which, to be fair, Meta was already using machine learning for), but also understanding which ads converted and which didn't. These probabilistic models will be built by massive fleets of GPUs, which, in the case of Nvidia's A100 cards, cost in the five figures; that may have been too pricey in a world where deterministic ads worked better anyways, but Meta isn't in that world any longer, and it would be foolish to not invest in better targeting and measurement.

Moreover, the same approach will be essential to Reels' continued growth: it is massively more difficult to recommend content from across the entire network than only from your friends and family, particularly because Meta plans to recommend not just video but also media of all types, and intersperse it with content you care about. Here too AI models will be the key, and the equipment to build those models costs a lot of money.

In the long run, though, this investment should pay off. First, there are the benefits to better targeting and better recommendations I just described, which should restart revenue growth. Second, once these AI data centers are built out the cost to maintain and upgrade them should be significantly less than the initial cost of building them the first time. Third, this massive investment is one no other company can make, except for Google (and, not coincidentally, Google's capital expenditures are set to rise as well).

That last point is perhaps the most important: ATT hurt Meta more than any other company, because it already had by far the largest and most finely-tuned ad business, but in the long run it should deepen Meta's moat. This level of investment simply isn't viable for a company like Snap or Twitter or any of the other also-rans in digital advertising (even beyond the fact that Snap relies on cloud providers instead of its own data centers); when you combine the fact that Meta's ad targeting will likely start to pull away from the field (outside of Google), with the massive increase in inventory that comes from Reels (which reduces prices), it will be a wonder why any advertiser would bother going anywhere else.

An important factor in making Meta's AI work is not simply building the base model but also tuning it to individual users on an ongoing basis; that is what will take such a large amount of capacity and it will be essential for Meta to figure out how to do this customization cost-effectively. Here, though, it helps that Meta's offering will probably be increasingly integrated: while the company may have committed to Qualcomm for chips for its VR headsets, Meta continues to develop its own server chips; the company has also released tools to abstract away Nvidia and AMD chips for its workloads, but it seems likely the company is working on its own AI chips as well.

What will be interesting to see is how things like image and text generation impact Meta in the long run: Sam Lessin has posited that the end-game for algorithmic timelines is AI content; I've made the same argument when it comes to the Metaverse. In other words, while Meta is investing in AI to give personalized recommendations, that idea, combined with 2022's breakthroughs, is personalized content, delivered through Meta's channels.

For now it will be interesting to see how Meta's advertising tools develop: the entire process of both generating and A/B testing copy and images can be done by AI, and no company is better than Meta at making these sort of capabilities available at scale. Keep in mind that Meta's advertising is primarily about the top of the funnel: the goal is to catch consumers' eyes for a product or service or app they did not know previously existed; this means that there will be a lot of misses — the vast majority of ads do not convert — but that also means there is a lot of latitude for experimentation

and iteration. This seems very well suited to AI: yes, generation may have marginal costs, but those marginal costs are drastically lower than a human.

Google

The Innovator's Dilemma was published in 1997; that was the year that Eastman Kodak's stock reached its highest price of $94.25, and for seemingly good reason: Kodak, in terms of technology, was perfectly placed. Not only did the company dominate the current technology of film, it had also invented the next wave: the digital camera.

The problem came down to business model: Kodak made a lot of money with very good margins providing silver halide film; digital cameras, on the other hand, were digital, which means they didn't need film at all. Kodak's management was thus very incentivized to convince themselves that digital cameras would only ever be for amateurs, and only when they became drastically cheaper, which would certainly take a very long time.

In fact, Kodak's management was right: it took over 25 years from the time of the digital camera's invention for digital camera sales to surpass film camera sales; it took longer still for digital cameras to be used in professional applications. Kodak made a lot of money in the meantime, and paid out billions of dollars in dividends. And, while the company went bankrupt in 2012, that was because consumers had access to better products: first digital cameras, and eventually, phones with cameras built in.

The idea that this is a happy ending is, to be sure, a contrarian view: most view Kodak as a failure, because we expect companies to live forever. In this view Kodak is a cautionary tale of how an innovative company can allow its business model to lead it to its eventual doom, even if said doom was the result of consumers getting something better.

And thus we arrive at Google and AI. Google invented the transformer, the key technology undergirding the latest AI models. Google is rumored to have a conversation chat product that is far superior to ChatGPT. Google claims that its image generation capabilities are better than Dall-E or anyone else on the market. And yet, these claims are just that: claims, because there aren't any actual products on the market.

This isn't a surprise: Google has long been a leader in using machine learning to make its search and other consumer-facing products better (and has offered that technology as a service through Google Cloud). Search, though, has always depended on humans as the ultimate arbiter: Google will provide links, but it is the user that decides which one is the correct one by clicking on it. This extended to ads: Google's offering was revolutionary because instead of charging advertisers for impressions — the value of which was very difficult to ascertain, particularly 20 years ago — it charged for clicks; the very people the advertisers were trying to reach would decide if their ad was good enough.

I wrote about the conundrum this presented for Google's business in a world of AI seven years ago in Google and the Limits of Strategy:

> In yesterday's keynote, Google CEO Sundar Pichai, after a recounting of tech history that emphasized the PC-Web-Mobile epochs I described in late 2014, declared that we are moving from a mobile-first world to an AI-first one; that was the context for the introduction of the Google Assistant.
>
> It was a year prior to the aforementioned iOS 6 that Apple first introduced the idea of an assistant in the guise of Siri; for the first time you could (theoretically) compute by voice. It didn't work very well at first (arguably it still doesn't), but the implications for computing generally and Google specifically were profound: voice interaction both

> expanded *where* computing could be done, from situations in which you could devote your eyes and hands to your device to effectively everywhere, even as it constrained *what* you could do. An assistant has to be far more proactive than, for example, a search results page; it's not enough to present possible answers: rather, an assistant needs to give the *right* answer.
>
> This is a welcome shift for Google the technology; from the beginning the search engine has included an "I'm Feeling Lucky" button, so confident was Google founder Larry Page that the search engine could deliver you the exact result you wanted, and while yesterday's Google Assistant demos were canned, the results, particularly when it came to contextual awareness, were far more impressive than the other assistants on the market. More broadly, few dispute that Google is a clear leader when it comes to the artificial intelligence and machine learning that underlie their assistant.
>
> A business, though, is about more than technology, and Google has two significant shortcomings when it comes to assistants in particular. First, as I explained after this year's Google I/O, the company has a go-to-market gap: assistants are only useful if they are available, which in the case of hundreds of millions of iOS users means downloading and using a separate app (or building the sort of experience that, like Facebook, users will willingly spend extensive amounts of time in).
>
> Secondly, though, Google has a business-model problem: the "I'm Feeling Lucky Button" guaranteed that the search in question would not make Google any money. After all, if a user doesn't have to choose from search results, said user also doesn't have the opportunity to click an ad, thus choosing the winner of the competition Google created between its advertisers for user attention. Google Assistant has the exact same problem: where do the ads go?

That Article assumed that Google Assistant was going to be used to differentiate Google phones as an exclusive offering; that ended up being wrong, but the underlying analysis remains valid. Over the past seven years Google's primary business model innovation has been to cram ever more ads into Search, a particularly effective tactic on mobile. And, to be fair, the sort of searches where Google makes the most money — travel, insurance, etc. — may not be well-suited for chat interfaces anyways.

That, though, ought only increase the concern for Google's management that generative AI may, in the specific context of search, represent a disruptive innovation instead of a sustaining one. Disruptive innovation is, at least in the beginning, not as good as what already exists; that's why it is easily dismissed by managers who can avoid thinking about the business model challenges by (correctly!) telling themselves that their current product is better. The problem, of course, is that the disruptive product gets better, even as the incumbent's product becomes ever more bloated and hard to use — and that certainly sounds a lot like Google Search's current trajectory.

I'm not calling the top for Google; I did that previously and was hilariously wrong. Being wrong, though, is more often than not a matter of timing: yes, Google has its cloud and YouTube's dominance only seems to be increasing, but the outline of Search's peak seems clear even if it throws off cash and profits for years.

Microsoft

Microsoft, meanwhile, seems the best placed of all. Like AWS it has a cloud service that sells GPU; it is also the exclusive cloud provider for OpenAI. Yes, that is incredibly expensive, but given that OpenAI appears to have the inside track to being the AI epoch's addition to this list of top tech companies, that means that Microsoft is investing in the infrastructure of that epoch.

Bing, meanwhile, is like the Mac on the eve of the iPhone: yes it contributes a fair bit of revenue, but a fraction of the dominant player, and a relatively immaterial amount in the context of Microsoft as a whole. If incorporating ChatGPT-like results into Bing risks the business model for the opportunity to gain massive market share, that is a bet well worth making.

The latest report from The Information, meanwhile, is that GPT is eventually coming to Microsoft's productivity apps. The trick will be to imitate the success of AI-coding tool GitHub Copilot (which is built on GPT), which figured out how to be a help instead of a nuisance (i.e. don't be Clippy!).

What is important is that adding on new functionality — perhaps for a fee — fits perfectly with Microsoft's subscription business model. It is notable that the company once thought of as a poster child for victims of disruption will, in the full recounting, not just be born of disruption, but be well-placed to reach greater heights because of it.

There is so much more to write about AI's potential impact, but this Article is already plenty long. OpenAI is obviously the most interesting from a new company perspective: it is possible that OpenAI becomes the platform on which all other AI companies are built, which would ultimately mean the economic value of AI outside of OpenAI may be fairly modest; this is also the bull case for Google, as they would be the most well-placed to be the Microsoft to OpenAI's AWS.

There is another possibility where open source models proliferate in the text generation space in addition to image generation. In this world AI becomes a commodity: this is probably the most impactful outcome for the world but, paradoxically, the most muted in terms of economic impact for individual companies (I suspect the biggest opportunities will be in industries where accuracy is essential: incumbents will therefore underinvest in AI, a la Kodak under-investing in digital, forgetting that technology gets better).

Indeed, the biggest winners may be Nvidia and TSMC. Nvidia's investment in the CUDA ecosystem means the company doesn't simply have the best AI chips, but the best AI ecosystem, and the company is investing in scaling that ecosystem up. That, though, has and will continue to spur competition, particularly in terms of internal chip efforts like Google's TPU; everyone, though, will make their chips at TSMC, at least for the foreseeable future.

The biggest impact of all though, though, is probably off our radar completely. Just before the break Nat Friedman told me in a Stratechery Interview about Riffusion, which uses Stable Diffusion to generate music from text via visual sonograms, which makes me wonder what else is possible when images are truly a commodity. Right now text is the universal interface, because text has been the foundation of information transfer since the invention of writing; humans, though, are visual creatures, and the availability of AI for both the creation and interpretation of images could fundamentally transform what it means to convey information in ways that are impossible to predict.

For now, our predictions must be much more time-constrained, and modest. This may be the beginning of the AI epoch, but even in tech, epochs take a decade or longer to transform everything around them.

I wrote a follow-up to this Article in this Daily Update.

← Holiday Break: December 26th to January 5th

Subscriber's Daily Update

Tuesday, January 10, 2023

More on Google and AI; OpenAI, Integration, and Microsoft

Thursday, December 22, 2022

An Interview with Daniel Gross and Nat Friedman about ChatGPT and the Near-Term Future of AI

Wednesday, December 21, 2022

FTC Fines Epic, Netflix Ads, YouTube and the NFL

Monday, December 19, 2022

Twitter's Link Ban, Network Portability, China and the Trailing Edge

On the business, strategy, and impact of technology.
© Stratechery LLC 2023 | Terms of Service | Privacy Policy