


Sujet traité : Apple Intelligence arrive juste à temps / Apple Intelligence is Right On Time

Source : Stratechery Date : 10 juin 2024

## Apple Intelligence is Right On Time

 stratechery.com/2024/apple-intelligence-is-right-on-time

Monday, June 10, 2024

Apple's annual Worldwide Developer Conference keynote kicks off in a few hours, and [Mark Gurman has extensive details of what will be announced in Bloomberg](#), including the name: "Apple Intelligence". As John Gruber noted on [Daring Fireball](#):

His report reads as though he's gotten the notes from someone who's already watched Monday's keynote. I sort of think that's what happened, given how much of this no one had reported before today. Bloomberg's headline even boldly asserts "Here's Everything Apple Plans to Show at Its AI-Focused WWDC Event". I'm only aware of one feature for one platform that isn't in his report, but it's not a jaw-dropper, so I wouldn't be surprised if it was simply beneath his threshold for newsworthiness. Look, I'm in the Apple media racket, so I know my inside-baseball obsessions are unusual, but despite all the intriguing nuggets Gurman drops in this piece, the thing I'm most insatiably curious about is how he got all this. Who spilled? By what means? It's extraordinary. And don't think for a second it's a deliberate leak. Folks inside Apple are, I assure you, furious about this, and incredulous that one of their own colleagues would leak it to Gurman.

The irony of the leak being so huge is that nothing is particularly surprising: Apple is announcing and incorporating generative AI features throughout its operating systems and making them available to developers. *Finally*, the commentariat exclaims! Apple is in danger of falling dangerously behind! The fact they are partnering with OpenAI is evidence of how desperate they are! In fact, I would argue the opposite: Apple is not too late, they are taking the correct approach up-and-down the stack, and are well-positioned to be one of AI's big winners.

### Apple's Business Model

Start with the most basic analysis of Apple's business: despite all of the (legitimate) talk about Services revenue, Apple remains a hardware company at its core. From its inception the company has sold personal computers: the primary evolution has been that the devices have become ever more personal, from desktops to laptops to phones, even as the market as a whole has shifted from being enterprise-centric to consumer-centric, which plays to Apple's strengths in design and the user experience benefits that come from integration.

Here's the thing about an AI-mediated future: we will need devices! Take the classic example of the Spike Jonze movie "Her":



Jonze's depiction of hardware is completely unrealistic: there is not a single battery charger in the entire movie (the protagonist removes the device to sleep, and simply places it on his bedside table), or any consideration given to connectivity and the constraints that might put on the size and capability of the device in the protagonist's ear; and yet, even then, *there is a device in the protagonist's ear*, and, when the protagonist wants the AI to be able to see the outside world, *he puts an iPhone-esque camera device in his pocket*:



Now a Hollywood movie from 2013 is hardly dispositive about the future, but the laws of physics are; in this case the suspension of disbelief necessary to imagine a future of smarter-than-human AIs must grant that we need some sort of device for a long time to come, and as long as that is the case there is an obvious opportunity for the preeminent device maker of all time. Moreover, to the extent there is progress to be made in miniaturization, power management, and connectivity, it seems reasonable to assume that Apple will be on the forefront of bringing those advancements to market, and will be courageous enough to do so.

In other words, any analysis of Apple's prospects in an AI world should start with the assumption that AI is a complement to Apple's business, not disruptive. That doesn't mean that Apple is guaranteed to succeed, of course: AI is the only foreseeable technological advancement that could provide sufficient differentiation to actually drive switching, but even there, the number of potential competitors is limited — there may only be one (more on this in a moment).

In the meantime, AI makes high-performance hardware more relevant, not less; Gurman notes that "Apple Intelligence" will only be available on Apple's latest devices:

The new capabilities will be opt-in, meaning Apple won't make users adopt them if they don't want to. The company will also position them as a beta version. The processing requirements of AI will mean that users need an iPhone 15 Pro or one of the models coming out this year. If they're using iPads or Macs, they'll need models with an M1 chip at least.

I'm actually surprised at the M1 baseline (I thought it would be M2), but the iPhone 15 Pro limitation is probably the more meaningful one from a financial perspective, and speaks to the importance of RAM (the iPhone 15 Pro was the first iPhone to ship with 8GB of RAM, which is also the baseline for the M1). In short, this isn't a case of Apple driving arbitrary differentiation; you really do need better hardware to run AI, which means there is the possibility of a meaningful upgrade cycle for the iPhone in particular (and higher ARPUs as well — might Apple actually start advertising RAM differences in iPhone models, since more RAM will always be better?).

## The App Store and AI

---

One of the of that phone-like device in Her being a camera is that such a device will probably not be how an AI "sees"; Humane has already shipped a camera-based device that simply clips on to your clothing, but the most compelling device to date is Meta's Ray-Ban smart glasses:



Meta certainly has designs on AR glasses replacing your phone; shortly after acquiring Oculus CEO Mark Zuckerberg predicted that devices mounted on your head would replace smartphones for most interactions in 10 years time. That prediction, though, was nine years ago; no one today, including Meta, predicts that smartphones will not be the most essential computing device in 2025, even though the Ray-Ban glasses are interesting.

The fact of the matter is that smartphones are nearly perfect devices for the jobs we ask them to do: they are small enough to be portable, and yet large enough to have a screen to interact with, sufficient battery life to make it through the day, and good enough connectivity; the smartphone, alongside cloud computing, represents the end of the beginning, i.e. the platform on which the future happens, as opposed to a transitory phase to a new class of disruptive devices.

In this view the app ecosystem isn't so much a matter of lock-in as it is a natural state of affairs: of course the app interfaces to the physical world, from smart homes to transportation to media consumption, are located on the device that is with people everywhere. And, by extension, of course those devices are controlled by an oligopoly: the network effects of platforms are unrivaled; indeed, the real surprise of mobile — at least if you asked anyone in 2013, when Stratechery started — is that there are *two* platforms, instead of just one.

That, by extension, is why the Ray-Ban glasses come with an app, and thus have a chance of succeeding; one of Humane's fatal flaws was their insistence that they could stand alone. Moreover, the longer that the smartphone is a prerequisite for new experiences, the more likely it is to endure; there is an analogy here to the continued relevance of music labels,

which depend on the importance of back catalogs, which just so happen to expand with every release of new music. Every new experience that is built with the assumption of a smartphone extends the smartphone's relevance that much further into the future.

There is, to be fair, a scenario where AI makes all applications obsolete with one fell swoop, but for now AI fits in the smartphone-enhancing pattern. First, to the extent that AI can be done locally, it will depend on the performance and battery life of something that is smartphone-sized at a minimum. Second, to the extent that AI is done in the cloud, it will depend on the connectivity and again battery life of something that is smartphone-sized as well. The latter, meanwhile, will come with usage costs, which is a potential tailwind for Apple (and Google's) App Stores: those usage costs will be paid via credits or subscriptions which both platforms will mandate go through their in-app purchase systems, of which they will take a cut.

The third alternative is that most AI utilization happens via platform-provided APIs, which is exactly what Apple is expected to announce later today. From Gurman's report:

Siri will be a part of the new AI push as well, with Apple planning a revamp to its voice-control service based on large language models — a core technology behind generative AI. For the first time, Siri users will be able to have precise control over individual features and actions within apps. For instance, people will be able to tell Siri to delete an email, edit a photo or summarize a news article. Over time, Apple will expand this to third-party apps and allow users to string multiple commands together into a single request. These features are unlikely to arrive until next year, however.

Platform-provided AI capabilities will not only be the easiest way for developers to incorporate these features, they will also likely be the best way, at least in terms of the overall user experience. Users will understand how to use them, because they will be "trained" by Apple's own apps; they will likely be cheaper and more efficient, because they are leveraging Apple's overall investment in capabilities; most importantly, at least in terms of Apple's competitive position, they will further lock-in the underlying platform, increasing the hurdle for any alternative.

## AI Infrastructure

---

There are two infrastructure concerns when it comes to the current state of AI. The first, and easiest to manage for Apple (at least in the short-term), are so-called chatbots. On one hand, Apple is massively "behind" in terms of both building a ChatGPT-level chatbot, and also in terms of building out the necessary infrastructure to support that level of capability for its massive userbase. The reason I put "behind" in scare-quotes, though, is that Apple can easily solve its shortcoming in this area by partnering with a chatbot that already exists, which is exactly what they are doing. Again from Gurman:

The company's new AI system will be called Apple Intelligence, and it will come to new versions of the iPhone, iPad and Mac operating systems, according to people familiar with the plans. There also will be a partnership with OpenAI that powers a ChatGPT-like chatbot.

The analogy here is to Search, another service that requires astronomical investments in both technology and infrastructure; Apple has never and need never build a competitive search engine, because it owns the devices on which search happens, and thus can charge Google for the privilege of making the best search engine the default on Apple devices. This is the advantage of owning the device layer, and it is such an advantageous position that Apple can derive billions of dollars of profit at essentially zero cost.

A similar type of partnership with OpenAI will probably not be as profitable as search was; my guess is that Apple will be paying OpenAI, instead of the other way around, but the most important takeaway in terms of Apple's competitive position is that they will, once again, have what is regarded as the best chatbot on their devices without having to make astronomical investments in technology and infrastructure. Moreover, this dampens the threat of OpenAI building their own device that usurps the iPhone: why would you want to buy a device that lacks the iPhone ecosystem when you can get the same level of capability on the iPhone you already have, along with all of the other aspects of the iPhone platform I noted above?

The second infrastructure concern is those API-level AI capabilities that Apple is set to extend to 3rd-party developers. Here the story is a bit fuzzier; from another Gurman report last month:

Apple Inc. will deliver some of its upcoming artificial intelligence features this year via data centers equipped with its own in-house processors, part of a sweeping effort to infuse its devices with AI capabilities. The company is placing high-end chips — similar to ones it designed for the Mac — in cloud-computing servers designed to process the most advanced AI tasks coming to Apple devices, according to people familiar with the matter. Simpler AI-related features will be processed directly on iPhones, iPads and Macs, said the people, who asked not to be identified because the plan is still under wraps.

I am intrigued to learn more about how these data centers are architected. Apple's chips are engineered first-and-foremost for smartphones, and extended to Macs; that means they incorporate a CPU, GPU, NPU and memory into a single package. This has obvious benefits in terms of the iPhone, but there are limitations in terms of the Mac; for example, the highest end Mac Pro only has 192 GB of memory, a significant step-down from the company's Intel Xeon-based Mac Pros, which topped out at 1.5 TB of memory. Similarly, while that top-of-the-line M2 Ultra has a 72-core GPU, it is married to a 24-core CPU; a system designed for AI processing would want far greater GPU capability without paying a "CPU tax" along the way.

In short, I don't currently understand why Apple would build datacenters around its own chips, instead of using chips better-suited to the tasks being asked of them. Perhaps the company will announce that it has designed a new server chip, or perhaps its chips are being used in conjunction with purpose-built chips from other companies; regardless, building out the infrastructure for API-level AI features is one of the biggest challenges Apple faces, but it is a challenge that is eminently solvable, particularly since Apple controls the interface through which those capabilities will be leveraged — and when. To go back to the first Gurman article referenced above:

Apple's AI features will be powered by its own technology and tools from OpenAI. The services will either rely on on-device processing or cloud-based computing, depending on the sophistication of the task at hand. The new operating systems will include an algorithm to determine which approach should be taken for any particular task.

Once again, we see how Apple (along with Google/Android and Microsoft/Windows) is located at the point of maximum leverage in terms of incorporating AI into consumer-facing applications: figuring out what AI applications should be run where and when is going to be a very difficult problem as long as AI performance is not "good enough", which is likely to be the case for the foreseeable future; that means that the entity that can integrate on-device and cloud processing is going to be the best positioned to provide a platform for future applications, which is to say that the current operating system providers are the best-placed to be the platforms of the future, not just today.

## Competitive Threats

---

Outlining Apple's competitive position illustrates what a threat to their business must look like. In the very long run, it is certainly possible that there is an AGI that obsoletes the smartphone entirely, just as the iPhone obsoleted entire categories of consumer electronics. Yes, we will still need devices, which works in Apple's favor, but if those devices do not depend on an app ecosystem then Apple runs the risk of being reduced to a commoditized hardware manufacturer. This, by extension, is the biggest reason to question Apple's decision to partner with OpenAI for chatbot functionality instead of building out their own capability.

I'm skeptical, though, that this sort of wholesale transition will happen anytime soon, or ever; the reality of technology is that most new epochs layer on top of what came before, as opposed to replacing it wholesale. The Internet, for example, has been largely experienced on top of existing operating systems like Windows or iOS. Again, the most fervent AI believers may argue that I am dismissing AI's long-term capabilities, but I think that Apple is making a reasonable bet.

It follows, then, that if I am right about the continued importance of the smartphone, that the only entity that can truly threaten Apple is Google, precisely because they have a smartphone platform and attendant ecosystem. The theory here is that Google could develop truly differentiated AI capabilities that make Android highly differentiated from the iPhone, even as Android has all of the apps and capabilities that is the price of entry to a user's pocket in the first place.

I don't, for the record, think that this possibility is purely theoretical; I wrote last December about [Google's True Moonshot](#):

What, though, if the mission statement were the moonshot all along? What if "I'm Feeling Lucky" were not a whimsical button on a spartan home page, but the default way of interacting with all of the world's information? What if an AI Assistant were so good, and so natural, that anyone with seamless access to it simply used it all the time, without thought?

That, needless to say, is probably the only thing that truly scares Apple. Yes, Android has its advantages to iOS, but they aren't particularly meaningful to most people, and even for those that care — like me — they are not large enough to give up on iOS's overall superior user experience. The only thing that drives meaningful shifts in platform marketshare are paradigm shifts, and while I doubt the v1 version of Pixie would be good enough to drive switching from iPhone users, there is at least a path to where it does exactly that.

I wrote more about this possibility [two weeks ago](#), so I don't want to belabor the point, but this may be the biggest reason why Apple is partnering with OpenAI, and not Google: Apple might not want to build a dependency on a company might be incentivized to degrade their relative experience (a la Google Maps [a decade ago](#)), and Google might not want to give access to its potential source of long-term differentiation to the company whose business model is the clearest solution to the search company's threat of disruption.

The disruptive potential of AI for Google is straightforward: yes, Google has massive infrastructure advantages and years of research undergirding its AI efforts, but delivering an answer instead of a set of choices is problematic both [for Google's business model](#), which depends on user's choosing the winner of an auction, and for [its position as an Aggregator](#), which depends on serving everyone in the world, regardless of their culture and beliefs.

The past few weeks have surfaced a third risk as well: Google has [aggressively pushed AI results into search](#) in response to the competitive threat from chatbots; OpenAI and Perplexity, though, aren't upsetting user expectations when they deliver hallucinatory responses, because users already know what they are getting into when they choose to use chatbots to ask questions. Google, though, has a reputation for delivering "correct" results, which means leveraging its search distribution advantage to push AI entails significant risk to



that reputation. Indeed, Google has already started to deprioritize AI results in search, moving them further down the page; that, though, at least in my personal experience, has made them significantly less useful and pushed me back towards using chatbots.

A meaningful strategic shift towards a vertical model centered around highly differentiated devices, though, solves a lot of these problems: the devices would make money in their own right (and could be high-priced because they are the best way to access Google's differentiated AI experiences), could deliver a superior AI experience (not just via the phone, but accessories like integrated glasses, ear buds, etc), and would serve an audience that has self-selected into the experience. I remain dubious that Google will have the gumption to fully go in this direction, but it is the one possibility that should make Apple nervous.

## AI Prudence

---

It is the other operating system provider, Microsoft, who gives further credence to Apple's deliberative approach. Windows is not a threat to the iPhone for all of the app ecosystem reasons noted above, but Microsoft clearly sees an opportunity to leverage AI to compete with the Mac. After last month's CoPilot+ PC event I wrote in [Windows Returns](#):

The end result — assuming that reviewed performance measures up to Microsoft's claims — is an array of hardware from both Microsoft and its OEM partners that is MacBook Air-esque, but, unlike Apple's offering, actually meaningfully integrated with AI in a way that not only seems useful today, but also creates the foundation to be dramatically more useful as developers leverage Microsoft's AI capabilities going forward. I'm not going to switch (yet), but it's the first time I've been tempted; at a minimum the company set a bar for Apple to clear at next month's WWDC.

One of the new Windows features that Microsoft touted at that event was Recall, which leverages AI to help users access everything they have seen or done on their computer in recent history. The implementation, though, turned out to be quite crude: Windows will regularly take screenshots and use local processing to index everything so that it is easily searchable. The problem is that while Microsoft stridently assured customers (and analysts!) that none of your information would be sent to the cloud, they didn't take any measures to ensure that said data was secured locally, instead taking a dependency on Windows' overall security. Over the intervening weeks security researchers [have demonstrated why that wasn't good enough](#), leading to a Microsoft announcement last week of several significant changes; from [The Verge](#):

Microsoft says it's making its new Recall feature in Windows 11 that screenshots everything you do on your PC an opt-in feature...Microsoft will also require Windows Hello to enable Recall, so you'll either authenticate with your face, fingerprint, or using a PIN...This authentication will also apply to the data protection around the snapshots that Recall creates.

There are a few interesting implications in these changes:

- First, by making Recall opt-in, Microsoft is losing the opportunity to provide users with a surprise-and-delight moment when their computer finds what they were looking for; Microsoft is going to need to sell the feature to even make that experience possible.
- Second, while requiring OS-level user authentication to access Recall data is almost certainly the right choice, it's worth pointing out that this removes the potential for 3rd-party developers to build innovative new applications on top of Recall data.

These two factors explain how this screw-up happened: Microsoft wanted to push AI as a differentiator, but the company is still at its core a developer-focused platform provider. What they announced initially solved for both, but the expectations around user data and security is such that the only entity that has sufficient trust to deliver these sorts of intimate experiences is the OS provider itself.

This is good news for Apple in two respects. First, with regards to the title of this Article, the fact it is possible to be too early with AI features, as Microsoft seemed to be in this case, implies that *not* having AI features does not mean you are too late. Yes, AI features could differentiate an existing platform, but they could also diminish it. Second, Apple's orientation towards prioritizing users over developers aligns nicely with its brand promise of privacy and security: Apple would prefer to deliver new features in an integrated fashion as a matter of course; making AI not just compelling but societally acceptable may require exactly that, which means that Apple is arriving on the AI scene just in time.