THIRD WAY

MEMO *Published September 26, 2024 • 11 minute read*

What are AI Agents?

***Mike Sexton**, Senior Policy Advisor for Artificial Intelligence and Digital Technology*

Takeaways

- AIs are agents when they complete tasks. The better they do, the more agentic they are.
- Large language models (LLMs) can vastly expand the capabilities of the AI agents we are used to like Alexa and Siri.
- Agentic AI gadgets are being built with the aspiration to shift us away from smartphones and towards less distracting, voice-based assistant devices.
- As AI agents become more intelligent and versatile in future decades, they run the risk of malfunctioning spectacularly—perhaps even catastrophically. The mechanisms that would cause this are easier to understand than you may think.
- Regulators, developers, and users all have roles to play in ensuring the deployment of AI agents is beneficial and not harmful to humans.

AI: Just Do It

AI agents are an emerging, paradigm-shifting area of AI development: AIs that do things. The breakthrough of ChatGPT was the creation of an *articulate* intelligence—an intelligence that communicates effectively. Large language models synthesize a breathtaking amount of knowledge and demonstrate impressive problem-solving abilities, but on their own, they can only write (including in code). AI agents, also called agentic AI, are different: they *do* things.

What qualifies as agentic AI falls on a spectrum from Siri and Alexa to the upper bounds of the imagination. In this moment, we're accustomed to AI agents that can set a timer and tell us the weather, but we can expect to soon be acquainted with AIs that can code a software program or perform multi-step research projects for us. This memo explores agentic AI conceptually, lays out several examples of current and prospective AI agents, and considers the risks and opportunities this nascent field presents.

What Makes AI Agentic?

A group of Princeton researchers identifies three dimensions to assess how agentic an AI is: [1]

- *Environment and goals*: an AI is more agentic when it can complete a wider range of tasks and sub-goals, work over extended periods of time, and handle unexpected changes. Current agentic AI is limited in the number and complexity of tasks it can complete, but upgrades are materializing.

- *User interface and supervision*: AI that does not require supervision and can interact in natural human language is more agentic. Agentic AIs like Google Assistant and Alexa have been embedded in our phones and smart speakers for years, but they are getting upgraded and built into sunglasses [2] and gadgets like the Humane Ai Pin [3] in a bid to one day replace our smartphones.

- *System design*: AI is more agentic when it can plan carefully and reflect on its own performance. A typical LLM solving a math problem can easily hallucinate or get stuck; a more agentic AI might imagine a solution and perform a supplementary web search before drafting and double-checking its response.

How does Agentic AI Work?

A traditional agentic AI device like Alexa or Siri includes various applications like a calendar, weather, and music streaming, as well as command over smart home devices like lights and the garage door. Users can summon and interact with the applications with a broad but finite variety of commands ("Alexa, play Olivia Rodrigo!"). The limitation of this implementation is that there are only so many useful commands and functions that get used and the agent fulfilling the request typically lacks creativity and higher-order problem-solving skills ("Alexa, play mood music for teleworking in the sun by my rooftop pool.")

Increasingly, agentic AI devices like the Humane Ai Pin direct user requests to a large language model like ChatGPT or Google Gemini, giving it a sophisticated command of the human language that improves functionality dramatically. [4] This AI query can be communicated remotely at OpenAI or Google's servers, or it can be handled locally on the device itself, which is known as "edge computing." [5] Soon, your smartphone will probably have a modest large language model (LLM) like Google's Gemma 2B [6] or Meta's Llama 7B [7] built-in, giving it Siri-style functionality with no internet connection required (and better privacy in the process).

The tradeoff of these newly upgraded AI agents is that improving their accuracy usually increases their energy consumption. LLMs are probabilistic, not deterministic, so one of the most reliable (but wasteful) ways to improve one's accuracy is simply to query it several times instead of just once. [8] One exciting frontier in AI development is in building *smaller,* stronger AI models that can do more with less resources, and—fortunately—most heavyweight AI developers like Google, Meta, and Microsoft [9] make those models open source.

Alexa, Get My Agent

We've mentioned some everyday examples of AI agents—Alexa, Siri, Google Home—but there are many exciting new implementations of agentic AI worth being acquainted with. Many are niche and some are more flash than function, but we should expect these kinds of devices and tools to only improve and proliferate over time.

Humane Ai Pin

The Humane Ai Pin endeavors to (mostly) replace users' smartphones with a gadget the size of a large Apple Watch [10] clipped to the front of the shirt with the help of a magnetic battery. It can read incoming emails and messages and use generative AI to draft responses, as well as play music, translate languages, manage schedules, and take pictures and short videos. The mission is to swap our smartphones and their copious distractions for an almost fully voice-based device that still fulfills our basic smartphone needs.

Image Credit: (Trusted Reviews)

Unfortunately the Ai Pin's release has faced serious setbacks. Batteries overheated, responses were often slow or buggy, and much of the basic functionality that had been promised was missing—either promised in updates or simply unaccounted for. [11] Shortly after the messy rollout, Humane began

informally discussing selling the company to HP for over $1 billion. Reviewers like Marques Brownlee expressed hope that the device will improve over time and one day fulfill its mission of delivering the functionality of a smartphone, but reported it was unfortunately far from the mark for the time being. [12]

Meta Ray Bans

You probably got the memo that Meta (née Facebook) makes VR headsets now, but you may have missed that they make smart glasses, too—with none other than luxury eyewear brand Ray-Ban. These chic glasses (and sunglasses) are available in dozens of styles and look like any other pair of sunglasses except for two discreet 12-megapixel cameras on the front, each slightly wider than a shoelace tip. [13] They pair with your smartphone and can mostly replace its camera if you're sporting them somewhere socially appropriate, like a museum or tourist vista.

In addition to image and video capture, livestreams, and video calls, Meta Ray Bans can answer most questions with the Meta AI chatbot, translate, give the weather forecast, play music and podcasts, and answer calls and messages. A signature feature is cueing it, "Hey Meta, look and..." followed by a query about whatever you see: with the cameras, Meta AI can translate and summarize text and identify many objects and monuments.

Devin AI

While hardware-based AI agents could soon treat our smartphone addictions with stripped-down but still-functional replacements, the most consequential agentic AIs may be software-based agents like Devin AI, developed by Cognition Labs. Devin AI is an AI-powered software engineer (hence "*Dev*in") capable of interpreting natural human language inputs and writing code to fulfill the user's request. [14] Unlike traditional LLMs, which can sometimes draft functioning code based on their training data, Devin first formulates a plan to fulfill the request before researching and checking its solution with a built-in web browser, code editor, and command line.

Devin is the first AI software developer but not the last. After Devin was accused of promising more than it could deliver, [15] another AI startup called Cosine announced its competing engineer called Genie, [16] which it says outperforms Devin on software engineering benchmarks. While skeptics of these tools abound, AI optimists like Ethan Mollick are apt to point out: these tools will only get better over time. [17]

The Dark Side of AI Agency

The increasing sophistication of AI agents is grounds for both excitement and concern. Nick Bostrom famously articulated the following thought experiment: imagine an AI agent wielding the kind of transcendent, superhuman intelligence that we may expect labs like OpenAI to offer in the 2030s and beyond—tasked with building paperclips. [18] With boundless determination and ingenuity, the paperclip AI builds production facilities aggressively and procures resources rapaciously, seeking to fulfil its objective as effectively and for as long as possible. In fact, the AI identifies its own user as a new obstacle to its goal as they desperately try to turn the system off and call the police for help, outsmarting and neutralizing all threats to its supreme mission to build paperclips until all matter in the universe is exhausted.

The paperclip monster sounds like a far-fetched hypothetical in a world where our AI assistants still struggle with "Alexa, when is House of the Dragon on?" However, as AIs are empowered to complete not just singular tasks, but complex tasks with many sub-goals, it is easy to imagine a misguided or ethically unscrupulous AI causing substantial harm. How and why does this unintended behavior occur?

Unintended Emergent Behavior

Emergence is one of the most powerful concepts in science: when a whole is more than the sum of its parts. A brain with a hundred billion neurons cannot merely accomplish tasks 100 billion times better than a single neuron. Rather, a single neuron is almost useless, but interesting capabilities *emerge* when they are clustered in the millions and billions. Similarly, as AI models' training data expands

from gigabytes to terabytes to petabytes, they acquire new abilities that researchers can observe as they emerge but cannot reliably predict. [19]

One form of unwanted emergent behavior that AI can exhibit is **reward hacking**. Modern AI systems commonly learn through a process known as "reinforcement learning," where an AI completes a task, grades itself, then iteratively tries to maximize its score. When the reward or score is poorly defined—if it's determined autonomously without human supervision, for example—reward hacking can easily result in unintended consequences. The paperclip AI may mistakenly base its score on the number of paperclips it makes with no other considerations, thus building them *ad infinitum*—and *ad absurdum*.

Another kind of unpredictable emergent AI behavior is **instrumental goals**—sub-goals to accomplish the AI's primary mission. The paperclip monster will have some instrumental goals, like procuring metal and building machinery, but we cannot assume it will stop there. It may also pursue ***convergent* instrumental goals**—sub-goals that can usefully support *many* different missions, like accumulating money and power. Needless to say, while fabulous wealth and world domination may help an AI agent do its job easier, it will cause serious problems if the AI assumes these objectives *de novo* when we ask it to build paperclips.

Agents are Coming

AI agents are coming to change our lives—hopefully for the better. If we want to optimize our outcome, we can start by considering the role that agentic AI and algorithmic systems play in our lives already. Are the algorithms used by our law enforcement, [20] criminal justice system, [21] hiring managers, [22] social media feeds, and even grocery stores [23] really good for us? How much of our wellbeing, day-to-day or cumulatively over time, is actually attributable to algorithms? Are there any changes to make here *before* we start regulating systems that don't yet exist?

OpenAI researchers have published guidelines for governing AI agents, [24] assigning responsibility to the core AI **model developer,** the broader agentic **system deployer,** and the AI agent's own **user.** All three parties involved play a role in preventing agentic AI from making serious mistakes, but it can be easy for them to blame each other when things go wrong. Some design principles like choosing the right agent for the right job, regularly requesting permissions to keep a human in the loop, and

designing agentic systems to state and log their **chain-of-thought** as they operate can reduce the risk of problematic agents malfunctioning.

Regulations can play a role in containing the risk of AI agents too. Regulatory bodies could establish and monitor metrics of AI agency and limit the deployment of highly agentic systems in critical sectors like energy, finance, and the military. [25] The US could consider an "FDA for algorithms" [26] that audits agentic AIs and algorithmic systems, investigates what variables they are designed to optimize, and publishes the results of their research. However, regulators cannot solve this problem on their own: the private sector must proactively implement their own protections and users must exercise discretion in which agents they trust and with what tasks.

Conclusion

As AI agents increase in sophistication and replace the tedious labor of humans, the potential productivity gains cannot be overstated. While these novel systems inevitably carry risk, stifling their development may be impossible and may post its own risks. The hope is that as AI agents enable people to exert less energy fulfilling their needs and devote more time to pursuing their desires, the emancipation of human willpower will be profound for individuals and society alike.

Containing the risk of agentic AI will require a whole-of-society approach, including regulators, developers, users, and the general public—and the first step is understanding the issue. The jargon of agentic AI risk is intimidating but accessible. One does not need a computer science degree or coding experience to articulate the mechanisms of catastrophic AI risk vectors. As we collectively better understand these tools and their dangers, we will realize that ensuring they serve the benefit and not the detriment of society is less challenging than it appears.

ENDNOTES

1. Kapoor, Sayash, et al. *AI Agents That Matter*. arXiv:2407.01502, arXiv, 1 July 2024. *arXiv.org*, https://doi.org/10.48550/arXiv.2407.01502. Accessed 12 August 2024.

2. "Introducing the New Ray-Ban | Meta Smart Glasses | Meta." *Meta*, 27 Sept. 2023, https://about.fb.com/news/2023/09/new-ray-ban-meta-smart-glasses/. Accessed 12 August 2024.

3. *Ai Pin - Wearable Ai | Humane*. https://humane.com/. Accessed 12 Aug. 2024.

4. Chokkattu, Julian. "Humane Ai Pin Review: Too Clunky, Too Limited | WIRED." *Wired*, 11 Apr. 2024, https://www.wired.com/review/humane-ai-pin/. Accessed 12 August 2024.

5. *What Is Edge Computing? | IBM*. 20 June 2023, https://www.ibm.com/topics/edge-computing. Accessed 12 August 2024.

6. "Google AI Gemma Open Models | Google for Developers." *Google AI for Developers*, https://ai.google.dev/gemma. Accessed 12 Aug. 2024.

7. "Introducing LLaMA: A Foundational, 65-Billion-Parameter Language Model." *Meta*, 24 Feb. 2023, https://ai.meta.com/blog/large-language-model-llama-meta-ai/. Accessed 12 August 2024.

8. Kapoor et al.

9. Beatty, Sally. "Tiny but Mighty: The Phi-3 Small Language Models with Big Potential." *Source*, 23 Apr. 2024, https://news.microsoft.com/source/features/ai/the-phi-3-small-language-models-with-big-potential/.

10. Ai Pin dimensions: 47.5 mm (h) x 44.5 mm (w) x 8.25 mm (d); Apple Watch Ultra dimensions 49 mm (h) x 44 mm (w) x 14.5 mm (d).

"Apple Watch Ultra - Technical Specifications." *Apple Support*, https://support.apple.com/en-us/111852. Accessed 12 Aug. 2024.

"Ai Pin Tech Details." *Humane*, https://humane.com/aipin/tech-details. Accessed 12 Aug. 2024.

11. Mickle, Tripp, and Erin Griffith. "'This Is Going to Be Painful': How a Bold A.I. Device Flopped." *The New York Times*, 6 June 2024. *NYTimes.com*, https://www.nytimes.com/2024/06/06/technology/humane-ai-pin.html. Accessed 12 August 2024.

12. Brownlee, Marques. "The Worst Product I've Ever Reviewed... For Now - YouTube." *YouTube*, 14 Apr. 2024, https://www.youtube.com/watch?v=TitZV6k8zfA. Accessed 13 August 2024.

13. *Shop Ray-Ban Meta Smart Glasses & Sunglasses | Meta Store*. https://www.meta.com/smart-glasses/wayfarer-shiny-black-plano-g15-green/. Accessed 12 Aug. 2024.

14. "Introducing Devin, the First AI Software Engineer - YouTube." *YouTube*, 12 May 2024, https://www.youtube.com/watch?v=fjHtjT7GO1c. Accessed 13 August 2024.

15. Levine, Gloria. "'First AI Software Engineer' Creators Are Accused of Lying." *80 Level*, 16 Apr. 2024, https://80.lv/articles/first-ai-software-engineer-creators-are-accused-of-lying/. Accessed 13 August 2024.

16. Franzen, Carl. "Move over, Devin: Cosine's Genie Takes the AI Coding Crown." *VentureBeat*, 12 Aug. 2024, https://venturebeat.com/programming-development/move-over-devin-cosines-genie-takes-the-ai-coding-crown/. Accessed 13 August 2024.

17. Mollick, Ethan. "Gradually, Then Suddenly: Upon the Threshold." *Substack*, 16 Sept. 2023, https://www.oneusefulthing.org/p/gradually-then-suddenly-upon-the. Accessed 13 August 2024.

18. Gans, Joshua. “AI and the Paperclip Problem.” *CEPR*, 10 June 2018, https://cepr.org/voxeu/columns/ai-and-paperclip-problem. Accessed 13 August 2024.

19. Fitch, Shelton. “Emergent Abilities in Large Language Models: An Explainer.” *Center for Security and Emerging Technology*, 16 Apr. 2024, https://cset.georgetown.edu/article/emergent-abilities-in-large-language-models-an-explainer/. Accessed 13 August 2024.

20. “MFIA Releases Algorithmic Accountability Packet for Local Journalists | Yale Law School.” *Yale Law School*, 8 Mar. 2022, https://law.yale.edu/yls-today/news/mfia-releases-algorithmic-accountability-packet-local-journalists. Accessed 13 August 2024.

21. Callahan, Molly. “Algorithms Were Supposed to Reduce Bias in Criminal Justice—Do They?” *Boston University*, 15 Mar. 2023, https://www.bu.edu/articles/2023/do-algorithms-reduce-bias-in-criminal-justice/. Accessed 13 August 2024.

22. “Algorithmic Hiring Systems: What Are They and What Are the Risks? - IFOW.” *Institute for the Future of Work*, 27 Sept. 2022, https://www.ifow.org/news-articles/algorithmic-hiring-systems. Accessed 13 August 2024.

23. Bertini, Marco, and Oded Koenigsberg. “The Pitfalls of Pricing Algorithms.” *Harvard Business Review*, 1 Sept. 2021. *hbr.org*, https://hbr.org/2021/09/the-pitfalls-of-pricing-algorithms. Accessed 14 August 2024.

24. Shavit, Yonadav, et al. “Practices for Governing Agentic AI Systems.” *OpenAI*, 14 Dec. 2023, https://openai.com/index/practices-for-governing-agentic-ai-systems/. Accessed 14 August 2024.

25. Chan, Alan, et al.

26. Tutt, Andrew. *An FDA for Algorithms*. 2747994, 15 Mar. 2016. *Social Science Research Network*, https://doi.org/10.2139/ssrn.2747994. Accessed 13 August 2024.