BY INVITATION
Meredith Whittaker

The world must wake up to the threats AI agents pose to privacy, cyber-security—and even competition

SOON WE WILL all have robot butlers, an army of AI agents anticipating our needs and fulfilling our desires. At least, this is the tech promise of the moment. From booking a restaurant to asking your crush on a date, we'll be able to put our brain in a jar while a bundle of AI systems does our living for us. Why waste time on wooing when you can leave it to your botservant to turn on the charm? In pursuit of this future, the companies that dominate this market are busy injecting AI agents into the nervous system of the digital world. But as in fairy tales, so in life: relying on magical fixes leads to trouble.

An AI agent is a complex system including AI models, software and cloud infrastructure. For the system to do its thing—summarising your email or spending your money—it needs near-total access to your digital life. This is not the familiar request for permission to see your contacts; it is akin to giving "root" access to your entire device. Your browser history, credit-card details, private messages and location data are all poised to become AI fodder—heaped in an unsecure pile of undifferentiated data "context".

The push for AI agents comes as the industry is still struggling with profitability. Markets are twitchy, because despite high revenues the huge cost of AI development means pressures are mounting to break even. This helps explain the phantasmagoric promise of agentic AI. It also explains why basic lessons in privacy and digital security are being discarded.

In one sense, the problem is fundamental: there is a powerful tension between privacy and security, on the one hand, and the vision of letting a complex system with broad access to your data do whatever it wants, on the other.

Although the full agentic future has not yet arrived, the harms are already clear. Researchers have shown that AI agents can be coaxed into revealing sensitive data they have access to or tricked by hackers into taking harmful actions—from extracting sensitive code to creating havoc in homes by activating smart appliances.

Worse, the threat to communications privacy is real. Security researchers recently exposed Siri transmitting voice transcripts of WhatsApp messages to Apple servers as a part of the rollout of Apple Intelligence, an AI system developed by the firm. This undermines WhatsApp's end-to-end encryption—adding Apple as another "end" and thus breaking the guarantee that only those sending and receiving communications can access them.

In addition, the way agents are being rolled out is a threat to competition, part of a rush to acquire data by AI giants. Agentic systems are bypassing APIs (short for Application Program Interfaces)—the "front door" for accessing data from third-party apps and services. Instead of paying, these agents could potentially extract competitors' data in other ways, such as directly accessing whatever is being displayed on their users' screens. The companies controlling these agents are positioned to aggregate such interface-level data across billions of agentic deployments, generating market insights that those building apps and services understandably don't want to hand over to rivals.

The threats to privacy, security and competition are heightened by the fact that agents are not being offered as optional apps we can choose to ignore. Operating-system (OS) developers—namely Apple, Google and Microsoft—are integrating them into the core of their platforms, making them all but mandatory.

To put it bluntly, the path currently being taken towards agentic AI leads to an elimination of privacy and security at the application layer. It will not be possible for apps like Signal—the messaging app whose foundation I run—to continue to provide strong privacy guarantees, built on robust and openly validated encryption, if device-makers and OS developers insist on puncturing the metaphoric blood-brain barrier between apps and the OS. Feeding your sensitive Signal messages into an undifferentiated data slurry connected to cloud servers in service of their AI-agent aspirations is a dangerous abdication of responsibility.

Happily, it's not too late. There is much that can still be done, particularly when it comes to protecting the sanctity of private data. What's needed is a fundamental shift in how we approach the development and deployment of AI agents. First, privacy must be the default, and control must remain in the hands of application developers exercising agency on behalf of their users. Developers need the ability to designate applications as "sensitive" and mark them as off-limits to agents, at the OS level and otherwise. This cannot be a convoluted workaround buried in settings; it must be a straightforward, well-documented mechanism (similar to Global Privacy Control) that blocks an agent from accessing our data or taking actions within an app.

Second, radical transparency must be the norm. Vague assurances and marketing-speak are no longer acceptable. OS vendors have an obligation to be clear and precise about their architecture and what data their AI agents are accessing, how it is being used and the measures in place to protect it.

These mitigations are the minimum necessary. They should be accompanied by changes in the design of operating systems that improve their ability to shield data from agents and harden their security guarantees, and by serious investment in security research to increase the chances of anticipating, rather than reacting to, vulnerabilities. Without these protections, we risk creating a future in which a few powerful companies decide that the convenience of leaving restaurant-booking or prioritising tasks to AI is more important than cyber-security, healthy competition and the right to private communication. ■

Meredith Whittaker is the president of the Signal Foundation, the non-profit parent organisation of the Signal messaging app.