Bruce Schneier
schneier@schneier.com

Building Trustworthy AI Agents

The promise of personal AI assistants rests on a dangerous assumption: that we can trust systems we haven't made trustworthy. We can't. And today's versions are failing us in predictable ways: pushing us to do things against our own best interests, gaslighting us with doubt about things we are or that we know, and being unable to distinguish between who we are and who we have been. They struggle with incomplete, inaccurate, and partial context: with no standard way to move toward accuracy, no mechanism to correct sources of error, and no accountability when wrong information leads to bad decisions.

These aren't edge cases. They're the result of building AI systems without basic integrity controls. We're in the third leg of data security—the old CIA triad. We're good at availability and working on confidentiality, but we've never properly solved integrity. Now AI personalization has exposed the gap by accelerating the harms.

The scope of the problem is large. A good AI assistant will need to be trained on everything we do and will need access to our most intimate personal interactions. This means an intimacy greater than your relationship with your email provider, your social media account, your cloud storage, or your phone. It requires an AI system that is both discreet and trustworthy when provided with that data. The system needs to be accurate and complete, but it also needs to be able to keep data private: to selectively disclose pieces of it when required, and to keep it secret otherwise. No current AI system is even close to meeting this.

To further development along these lines, I and others have proposed separating users' personal data stores from the AI systems that will use them. It makes sense; the engineering expertise that designs and develops AI systems is completely orthogonal to the security expertise that ensures the confidentiality and integrity of data. And by separating them, advances in security can proceed independently from advances in AI.

Digital Object Identifier 10.1109/MSEC.2025.3636705

What would this sort of personal data store look like? Confidentiality without integrity gives you access to wrong data. Availability without integrity gives you reliable access to corrupted data. Integrity enables the other two to be meaningful. Here are six requirements. They emerge from treating integrity as the organizing principle of security to make AI trustworthy.

First, it would be broadly accessible as a data repository. We each want this data to include personal data about ourselves, as well as transaction data from our interactions. It would include data we create when interacting with others—emails, texts, social media posts—and revealed preference data as inferred by other systems. Some of it would be raw data, and some of it would be processed data: revealed preferences, conclusions inferred by other systems, maybe even raw weights in a personal LLM.

Second, it would be broadly accessible as a source of data. This data would need to be made accessible to different LLM systems. This can't be tied to a single AI model. Our AI future will include many different models—some of them chosen by us for particular tasks, and some thrust upon us by others. We would want the ability for any of those models to use our data.

Third, it would need to be able to prove the accuracy of data. Imagine one of these systems being used to negotiate a bank loan, or participate in a first-round job interview with an AI recruiter. In these instances, the other party will want both relevant data and some sort of proof that the data are complete and accurate.

Fourth, it would be under the user's fine-grained control and audit. This is a deeply detailed personal dossier, and the user would need to have the final say in who could access it, what portions they could access, and under what circumstances. Users would need to be able to grant and revoke this access quickly and easily, and be able to go back in time and see who has accessed it.

Fifth, it would be secure. The attacks against this system are numerous. There are the obvious read attacks, where an adversary attempts to learn a person's data. And there are also write attacks, where adversaries add to or change a user's data. Defending against both is critical; this all implies a complex and robust authentication system.

Sixth, and finally, it must be easy to use. If we're envisioning digital personal assistants for everybody, it can't require specialized security training to use properly.

I'm not the first to suggest something like this. Researchers have proposed a "Human Context Protocol" (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5403981) that would serve as a neutral interface for personal data of this type. And in my capacity at a company called Inrupt, Inc., I have been working on an extension of Tim Berners-Lee's Solid protocol for distributed data ownership.

The engineering expertise to build AI systems is orthogonal to the security expertise needed to protect personal data. AI companies optimize for model performance, but data security requires cryptographic verification, access control, and auditable systems. Separating the two makes sense; you can't ignore one or the other.

Fortunately, decoupling personal data stores from AI systems means security can advance independently from performance (https://ieeexplore.ieee.org/document/10352412). When you own and control your data store with high integrity, AI can't easily manipulate you because you see what data it's using and can correct it. It can't easily gaslight you because you control the authoritative record of your context. And you determine which historical data are relevant or obsolete.

Making this all work is a challenge, but it's the only way we can have trustworthy AI assistants. ■

A good AI assistant will need to be trained on everything we do and will need access to our most intimate personal interactions.

Imagine one of these systems being used to negotiate a bank loan, or participate in a first-round job interview with an AI recruiter.