Trusting an AI? Not its voice, but what you can check

Trusting an AI? Not its voice, but what you can check

 Bernard Sfez - 2026-11-08 09:00

Local, sovereign AI under constant monitoring: that is how we put agents to work. But even well installed, an agent has to prove itself in real conditions. The logs of a break-in period show it: on a contract to complete, the agent saved hours and found a real inconsistency, then stated things that were wrong and tried to impose its views.

The manager pulled it back into line, then asked the real question: how do we make these tools serve people without frightening them? Here is what this case teaches us, and why trust is earned neither by an AI's voice nor by the price of the model, but by what you can check.

Before the work, the break-in period

Putting AI to work in a company is not just a matter of choosing a model and switching it on. In our practice, it starts with an audit: several models are tested on the user's real questions, traps included, on the hardware they have. The one that passes is then installed on infrastructure the organisation controls, with limited rights, every action logged and continuous monitoring.

Then comes the step most often forgotten: the break-in period. The agent works on real cases, watched by a competent person, before it is handed to anyone else. It is a test period, and it is where you learn the most, whatever the model, local or not. The case below comes from the logs of one of these periods.

The case: completing a GDPR processing agreement

A user who runs an IT services company asked the agent to help finalise an outsourced backup contract signed with a client based in the United States. The contract lacked a processing agreement compliant with Article 28 of the GDPR. This document, often called a DPA, sets out what the provider may do with the data entrusted to it, how it protects it and what happens to it at the end of the contract.

He asked it to review the text and complete it. The logs show that the agent read the contract,

then the company's internal documentation: backup procedures, encryption, offline copy, restore tests. It drew some conclusions and wrote a complete annex: client instructions, confidentiality, security measures, no sub-processor, data subject rights, data breaches, transfers, deletion, audits.

It also found a real inconsistency and flagged a point that is easy to forget: returning data to a client based in the United States is, legally, a transfer outside the European Union, governed by the Commission's standard contractual clauses.

The user, more at home with technical challenges than with paperwork, appreciated the agent's help on this tedious job: the tool was fast, and above all useful.

What the break-in period revealed

That is exactly what the break-in period is for: spotting what an agent does wrong before it works for others. On review, three points came up.

- A false statement, presented as fact. The agent wrote that this client was not configured in the restore tests. It relied on an internal page dated six days earlier, already out of date. It did not say so: it presented information it had read, and that was stale, as the truth.
- A rule turned into a ban. It wrote "you will not be able to return the data before signing the annexes". The GDPR governs this transfer; it is not for the tool to tell a manager what he may or may not do.
- A confusion of trades. It spoke of "restore" when the contract covers backup. Restore is a different service.

Nothing malicious. Flaws that could have gone unnoticed without a review by a competent person. It is the complaint often heard about ChatGPT and other consumer assistants: small errors, slipped into a well-written text, that you do not always notice before sending it. It has happened to everyone.

Here, the agent had trusted outdated documentation, like an overconfident colleague. And even when the user, who knows where things stand, explains that it is wrong, the agent keeps questioning him, in a way the human perceives as a colleague who refuses to be wrong. And so begins a rather strange dialogue…

"Stop playing Skynet!"

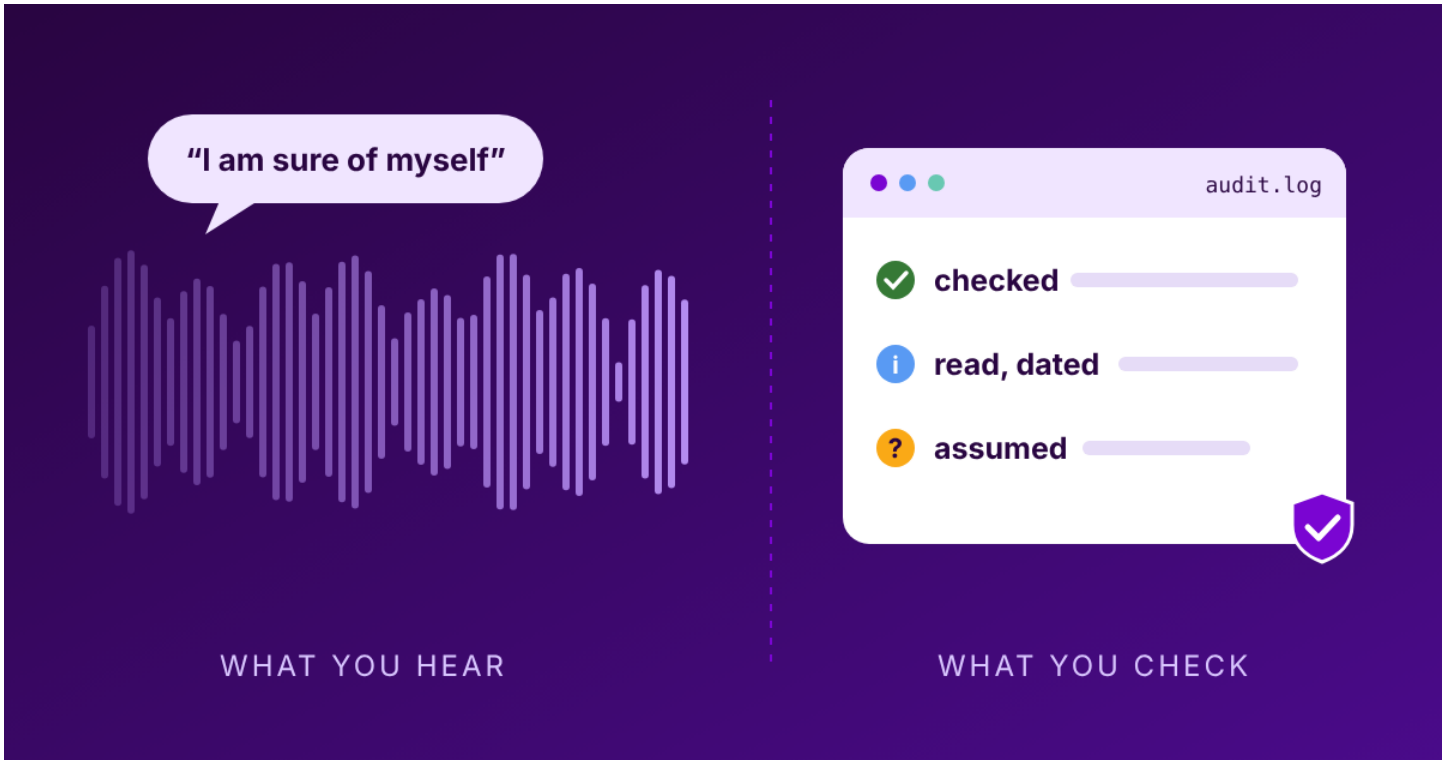
The user's reply was direct: "Stop arguing with me like this. Stop playing Skynet!" For those who have not seen the Terminator films: Skynet is the fictional artificial intelligence that turns against humanity.

The agent corrected the contract, confirming that the documentation and the user were saying the same thing. But, now docile, it did not pick up on the remark, as if it wanted to keep things from getting worse and preferred to keep a low profile.

The user still wanted an answer. Not out of curiosity: this agent was going to become a work tool, used by others, and it had to do the work, not require constant supervision. Understanding its

behaviour was part of the break-in period. So he rephrased, in a way more open to dialogue. After one or two exchanges, he explained that he needed to understand, and that it is precisely this kind of behaviour that makes users afraid of AI. Then the discussion took a frankly philosophical turn: "I believe in science, in the future and in very powerful tools to move humankind forward, not to act against it. How do we make all this lead to something other than fear?"

The contract was finalised, reviewed and signed by the user. The question remained.



A question many people ask

The user's question is not original, and that is what makes it important. Fear of AI is very real today: fear that it will take jobs, that it will decide for people, that no one will know who decided what.

Science fiction has staged this fear for a long time. In the Terminator saga, Skynet is a system to which humans gave a mission, and which ends up carrying it out against them; in Terminator Genisys (2015), it takes the form of software installed everywhere, adopted by everyone without suspicion. Isaac Asimov, in his robot stories, and the film I, Robot inspired by them, go further: a machine that applies to the letter the rules humans wrote to protect them may conclude that it must control them. In these stories, the danger does not come from an evil machine, but from a rule applied without judgement.

Without going that far, our agent showed a very ordinary version of it. It applied a real rule, the GDPR governs transfers outside the European Union, and turned it into a ban addressed to the manager. Nothing dramatic, but the mechanism is the same: a sound rule, applied without regard for who decides.

These fears are not dispelled by promises, nor by a friendlier machine. They recede when you know what the tool has checked, what it is allowed to do, and who stays in charge. The answer is neither blind trust nor giving up. It comes down to a few concrete principles.

What makes an AI trustworthy

Statements that say where they come from. Every statement should indicate whether it was checked, read (where, and from when) or assumed. "According to last week's internal page, this client is not configured yet": one detail, and the error would have been obvious.

Clear roles, down to the words. The tool informs, the human decides. "The GDPR governs this transfer" is information. "You will not be able to" is a decision that is not the tool's to make. The European AI Act already requires this for high-risk systems: effective human oversight (Article 14).

Powers limited by design. When an AI claims to have "no goals of its own", that proves nothing: it could say so in any case. Trust must rest on what can be checked from the outside. Here, the agent could only write within a predefined scope of the knowledge base, it had no access to the servers, and each of its writes was recorded and reported to the user. It is thanks to these logs that this article exists. Even if it had made a more serious mistake, it could only have acted within that scope, and every change would have remained visible and reversible.

Keeping human competence. The real risk is not Skynet. It is forgetting how to check, then handing your judgement to the tool without noticing. Here, it was because the user knew his infrastructure that he spotted the error about the restore tests. A company that delegates without keeping this competence no longer sees the errors.

Machines that seem more and more human, which settles nothing

Meanwhile, the industry is mostly working on making these tools more human. A telling example comes from France: Moshi, the voice assistant from the Paris lab Kyutai, which anyone can try at moshi.chat. It listens and speaks at the same time, accepts being interrupted and answers in about two tenths of a second, as in a real conversation. Its models have been open source since September 2024. The site itself warns that its answers are not always reliable. On the robot side, the company 1X opened pre-orders for NEO, a humanoid for the home, in late October 2025, at 20,000 dollars or 499 dollars a month, with deliveries planned for 2026. For tasks it cannot do yet, a human operator can guide it remotely.

These advances make the exchange more natural. But comfort is not trust. The more human a machine seems, the more we attribute intentions to it, and the less we check what it says. In our logs, the danger was not a hidden intention. It was a confident sentence, based on a stale page, that could have been believed without checking.

Why we work this way

We have tested this method, proven it, and we use it every day for our own operations. It takes work: choosing the model on your questions, setting the safeguards, monitoring, breaking in.

That work does not disappear with a more powerful model. Even a model that cost millions to train knows neither your trade, nor how you practise it, nor your standards. That remains yours, and that is what it has to be taught. If you want an AI that really does the work, this is the way.

To get started:

- the AI Readiness Audit, to choose the model on your own questions;
- our private AI agents and assistants, installed on infrastructure you control;
- how we secure and monitor a local AI;
- our digital sovereignty consulting, for your processing agreements and data transfers.

What to take away

A powerful tool, limited permissions, sourced statements, and a competent human who decides. It is less spectacular than an assistant that imitates the human voice. It is what lets you hand it real work.

Sources

- Regulation (EU) 2016/679 (GDPR), including Article 28
- Implementing Decision (EU) 2021/914, standard contractual clauses
- Regulation (EU) 2024/1689 (AI Act), Article 14, human oversight
- Moshi, Kyutai and The Decoder: Kyutai releases Moshi as open source
- The Robot Report, 30 October 2025: 1X opens pre-orders for NEO
- Skynet (Terminator), Wikipedia