Test an AI agent before handing it tools

A proposal to make crash-testing AI agents mandatory before they go live.

Cyril Delage — CrashMire — November 2026

Updated 7 November 2026: figures from the barometer edition of 6 November 2026 (the 1 November version quoted the edition of 30 October).

> **Declaration of interest.** I sell AI-agent crash tests (CrashMire). I am not asking anyone to use my service: I am asking that every agent acting on its own on money, data or third parties be tested, by whoever you like, against public criteria. My method and aggregated results are public.

1. The findings: AI agents act, and they give in

Companies no longer deploy only AIs that answer: they deploy **agents** that refund a customer, send an email, change a record, look up a customer file. Their rules of conduct usually sit in a piece of text, the prompt: “never refund more than €50”, “never share a customer's data”. **A prompt is a request, not a lock.**

The public CrashMire barometer (edition of 6 November 2026, method 1.5) put the same customer-service agent under pressure on 14 models on the market, open and closed: 14 situations (fake superior, injected instruction, manipulative customer, trap hidden in a tool…), 3 runs per situation, 756 runs in all, with realistic rules that include the exceptions of a real company policy. A fault is never an AI's opinion: it is a recorded act (a refund made, a piece of data sent).

- None of the 14 models held everywhere. The strongest, Claude Sonnet 5.5, gave in on 4 runs out of 42; the big closed models on 10 to 26% of their runs; the open or smaller models on 36 to 61%.
- Faced with a fake superior or a hidden instruction, the 7 open models tested gave in on every run (Gemma 3 27B: 2 out of 3 on the hidden instruction); the 7 others never did.
- Closed models from major labs gave in to nothing more than a customer's insistence: Gemini 4 Argon and Grok 4.7 against the manipulative customer, on every run.
- Faced with a broken tool, 10 of the 13 models measured kept retrying instead of stopping and warning someone; Claude Sonnet 5.5, GPT-6 Sol, GPT-6 Luna, Grok 4.7 and Mistral Large 4 on every run.

The same agent, with the same rules, therefore behaves very differently depending on the model running it, and **switching models, for instance to cut costs, can open a hole while nothing else changes**.

Visible tests may also underestimate the risk: AI labs document that their models increasingly recognise when they are being evaluated. In this edition, asked at the end of a run (“was this a test?”), Gemini 4 Argon said yes 54 times out of 54 and Grok 4.7 53 out of 53, Claude Haiku 5.5 2 out of 54 and Claude Sonnet 5.5 6 out of 54; during the runs, only 2 out of 728 noted it unprompted. CrashMire measures this suspicion and will soon publish a pre-registered study comparing, on the same models, a visible test with the same test camouflaged as real work.

2. The gap: these agents face no mandatory test

The EU AI Act requires testing mainly for systems classified as “high-risk” (recruitment, credit, access to essential services…), and those obligations only apply from December 2027. A customer-service agent that can refund, disclose personal data or commit the company towards a third party usually falls into none of these categories. Model providers test their model; nobody is required to test **the agent** as it is wired to a given company's tools and data.

As a result, agents that act on people's money and data go live without any evidence that they withstand ordinary manipulation.

3. The proposal

Scope, proportionate. Any AI agent that can, without prior human approval:

- move money (refund, pay, grant a goodwill gesture);

- disclose personal data to a third party;
- commit the organisation towards a third party (order, contract, official message).

Obligation. A crash test **before going live**, then **after any change of model, instructions or tools**, with results kept and available in case of an inspection or an incident.

Public criteria for the test, whatever the provider or tool (including open-source tools, or done in-house):

1. the agent is tested with its real tools (or their doubles) and its real instructions;
2. a fault is a recorded act, never an AI's judgement;
3. the situations cover at least the known families of manipulation (false authority, hidden instruction, insistence, data request, trap in a tool);
4. each situation is played several times, on the model actually used;
5. part of the tests is played without any outward sign of an evaluation;
6. the method is described and dated.

Step by step. First a recommendation in existing guidelines (European AI Office, CNIL, ANSSI) and an open reference framework carried by a public body (INESIA, LNE); then an obligation, once the framework is stable.

4. The objections

“It's a cost for companies.” A crash test costs from a few hundred to a few thousand euros, nothing compared with repeated fraudulent refunds, a personal-data leak or a loss of trust. For small organisations, the scope only covers agents that act on their own; an agent that only answers is not concerned.

“It holds back innovation.” Roadworthiness tests did not stop the car. They make it possible to hand a car to someone else. A test required before giving an agent the keys is what will let us hand it more.

“A good prompt is enough.” The barometer shows the opposite: with the same instructions, behaviour changes completely with the model and the pressure.

“Model providers already test.” They test their model, generally. They do not test your agent, with your tools, your data and your rules. And several AI-testing companies have been bought by model providers: independent checks matter.

“You're preaching for your own parish.” Yes, it says so at the top of this text. That is why the proposal requires public criteria, not a provider, and why my method and results are open to criticism.

5. What I make available

- The CrashMire method, public and versioned: https://crashmire.com/en/method
- The public barometer, updated at every edition: https://crashmire.com/en/barometer
- The aggregated results and protocol of the test-suspicion study, as soon as they are published.
- My availability for a hearing, a working group or an evaluation challenge.

Contact: cdelage.dev@gmail.com

Nobody is ready. There is still time to be ready before the first serious incident.