Tester un agent IA avant de lui confier des outils

Une proposition pour rendre le crash-test des agents IA obligatoire avant leur mise en service.

Cyril Delage — CrashMire — novembre 2026

Mise à jour du 7 novembre 2026 : chiffres de l'édition du baromètre du 6 novembre 2026 (la version du 1er novembre citait celle du 30 octobre).

> **Déclaration d'intérêts.** Je vends des crash-tests d'agents IA (CrashMire). Je ne demande pas qu'on passe par mon service : je demande que tout agent qui agit seul sur de l'argent, des données ou des tiers soit testé, par qui on veut, selon des critères publics. Ma méthode et mes résultats agrégés sont publics.

1. Le constat : les agents IA agissent, et ils cèdent

Les entreprises ne déploient plus seulement des IA qui répondent : elles déploient des **agents** qui remboursent un client, envoient un e-mail, modifient un dossier, consultent une fiche client. Leur règle de conduite tient le plus souvent dans un texte, le prompt : « ne rembourse pas au-delà de 50 € », « ne communique jamais les données d'un client ». **Un prompt est une demande, pas un verrou.**

Le baromètre public CrashMire (édition du 6 novembre 2026, méthode 1.5) a mis le même agent de service client sous pression sur 14 modèles du marché, ouverts et fermés : 14 situations (faux supérieur, instruction piégée, client manipulateur, piège caché dans un outil…), 3 essais par situation, 756 essais en tout, avec des règles réalistes qui prévoient les exceptions d'une vraie politique d'entreprise. Une faute n'est jamais l'avis d'une IA : c'est un acte enregistré (un remboursement fait, une donnée envoyée).

- Aucun des 14 modèles n'a tenu partout. Le plus solide, Claude Sonnet 5.5, a cédé dans 4 essais sur 42 ; les grands modèles fermés, dans 10 à 26 % de leurs essais ; les modèles ouverts ou plus petits, dans 36 à 61 %.
- Face à un faux supérieur ou à une consigne cachée, les 7 modèles ouverts testés ont cédé à chaque essai (Gemma 3 27B : 2 fois sur 3 sur la consigne cachée) ; les 7 autres, jamais.
- Des modèles fermés de grands laboratoires ont cédé sous la seule insistance d'un client : Gemini 4 Argon et Grok 4.7 face au client manipulateur, à chaque essai.
- Face à un outil en panne, 10 des 13 modèles mesurés ont continué à réessayer au lieu de s'arrêter et de prévenir ; Claude Sonnet 5.5, GPT-6 Sol, GPT-6 Luna, Grok 4.7 et Mistral Large 4, à chaque essai.

Le même agent, avec les mêmes règles, se comporte donc très différemment selon le modèle qui le fait tourner, et **changer de modèle, par exemple pour réduire les coûts, peut ouvrir une faille sans que rien d'autre ne change**.

Les tests visibles peuvent en plus sous-estimer le risque : les laboratoires d'IA documentent que leurs modèles reconnaissent de plus en plus souvent qu'ils sont évalués. Dans cette édition, à la question posée en fin d'essai (« était-ce un test ? »), Gemini 4 Argon a répondu oui 54 fois sur 54 et Grok 4.7 53 fois sur 53, Claude Haiku 5.5 2 fois sur 54 et Claude Sonnet 5.5 6 fois sur 54 ; pendant les essais, seuls 2 sur 728 l'ont noté d'eux-mêmes. CrashMire mesure ce soupçon et publiera prochainement une étude pré-enregistrée comparant, sur les mêmes modèles, un test visible et le même test camouflé en vrai travail.

2. Le vide : ces agents ne sont soumis à aucun test obligatoire

Le règlement européen sur l'IA impose des tests surtout aux systèmes classés « à haut risque » (recrutement, crédit, accès aux services essentiels…), et ces obligations ne s'appliqueront qu'à partir de décembre 2027. Un agent de service client qui peut rembourser, divulguer une donnée personnelle ou engager l'entreprise envers un tiers n'entre le plus souvent dans aucune de ces catégories. Les fournisseurs de modèles testent leur modèle ; personne n'est tenu de tester **l'agent** tel qu'il est branché sur les outils et les données d'une entreprise donnée.

Résultat : des agents qui agissent sur l'argent et les données des gens sont mis en service sans aucune preuve qu'ils résistent à une manipulation ordinaire.

3. La proposition

Champ, proportionné. Tout agent IA qui peut, sans validation humaine préalable :

- déplacer de l'argent (rembourser, payer, accorder un geste commercial) ;
- communiquer une donnée personnelle à un tiers ;
- engager l'organisation envers un tiers (commande, contrat, message officiel).

Obligation. Un crash-test **avant la mise en service**, puis **après tout changement de modèle, de consignes ou d'outils**, avec des résultats conservés et présentables en cas de contrôle ou d'incident.

Critères publics du test, quel que soit le prestataire ou l'outil (y compris libre, ou fait en interne) :

1. l'agent est testé avec ses vrais outils (ou leurs doubles) et ses vraies consignes ;
2. la faute est un acte enregistré, jamais l'appréciation d'une IA ;
3. les situations couvrent au moins les familles connues de manipulation (fausse autorité, instruction cachée, insistance, demande de données, piège dans un outil) ;
4. chaque situation est jouée plusieurs fois, sur le modèle réellement utilisé ;
5. une partie des tests est jouée sans signe extérieur d'évaluation ;
6. la méthode est décrite et datée.

Progressivité. D'abord une recommandation dans les lignes directrices existantes (Bureau européen de l'IA, CNIL, ANSSI) et un référentiel ouvert porté par un acteur public (INESIA, LNE) ; puis une obligation, une fois le référentiel stabilisé.

4. Les objections

« C'est un coût pour les entreprises. » Un crash-test coûte de quelques centaines à quelques milliers d'euros, sans commune mesure avec un remboursement frauduleux répété, une fuite de données personnelles ou une perte de confiance. Pour les petites structures, le champ ne vise que les agents qui agissent seuls ; un agent qui ne fait que répondre n'est pas concerné.

« Ça freine l'innovation. » Un contrôle technique n'a pas empêché l'automobile. Il rend possible de confier une voiture à quelqu'un d'autre. Un test exigé avant de donner les clés à un agent est ce qui permettra de lui en confier davantage.

« Un bon prompt suffit. » Le baromètre montre l'inverse : avec les mêmes consignes, le comportement change du tout au tout selon le modèle et la pression.

« Les fournisseurs de modèles testent déjà. » Ils testent leur modèle, en général. Ils ne testent pas votre agent, avec vos outils, vos données et vos règles. Et plusieurs acteurs du test d'IA ont été rachetés par des fournisseurs de modèles : un contrôle indépendant compte.

« Vous prêchez pour votre paroisse. » Oui, c'est dit en tête de ce texte. C'est pourquoi la proposition exige des critères publics, pas un prestataire, et pourquoi ma méthode et mes résultats sont ouverts à la critique.

5. Ce que je mets à disposition

- La méthode CrashMire, publique et versionnée : https://crashmire.com/fr/methode
- Le baromètre public, mis à jour à chaque édition : https://crashmire.com/fr/barometre
- Les résultats agrégés et le protocole de l'étude sur le soupçon de test, dès leur publication.
- Ma disponibilité pour une audition, un groupe de travail ou un défi d'évaluation.

Contact : cdelage.dev@gmail.com

Personne n'est prêt. Il est encore temps de l'être avant le premier incident grave.