Faire confiance à une IA ? Pas à sa voix, à ce qu’on peut vérifier

Faire confiance à une IA ? Pas à sa voix, à ce qu’on peut vérifier

 Bernard Sfez - 2026-11-04 19:49

Une IA locale, souveraine, sous surveillance permanente : c’est ainsi que nous mettons des agents au travail. Mais même bien installé, un agent doit faire ses preuves en conditions réelles. Les journaux d’une période de rodage le montrent : sur un contrat à compléter, l’agent a fait gagner des heures et repéré une vraie incohérence, puis il a affirmé des choses inexactes et voulu imposer ses vues.

Le dirigeant l’a recadré, avant de poser la question de fond : comment faire pour que ces outils servent les humains sans leur faire peur ? Voici ce que ce cas nous apprend, et pourquoi la confiance ne se gagne ni par la voix d’une IA, ni par le prix du modèle, mais par ce que l’on peut vérifier.*

## Avant le travail, le rodage

Installer une IA en entreprise ne se résume pas à choisir un modèle et à l’allumer. Dans notre pratique, il y a d’abord un audit : plusieurs modèles sont testés sur les vraies questions de l’utilisateur, pièges compris, avec le matériel dont il dispose. Celui qui a réussi est ensuite installé sur une infrastructure maîtrisée, avec des droits limités, chaque action journalisée et une surveillance continue.

Reste l’étape que l’on oublie le plus souvent : le rodage. L’agent travaille sur de vrais dossiers, sous le regard d’une personne compétente, avant d’être confié à d’autres. C’est une période de test, et c’est là que l’on apprend le plus, quel que soit le modèle, local ou non. Le cas qui suit est tiré des journaux d’une de ces périodes.

## Le cas : compléter un accord de sous-traitance RGPD

Un utilisateur, dirigeant d’une entreprise de services informatiques, a demandé à l’agent de l’aider à finaliser un contrat de sauvegarde externalisée signé avec un client établi aux États-Unis. Il manquait à ce contrat un accord de sous-traitance conforme à l’article 28 du RGPD. Ce

document, qu'on appelle souvent DPA, fixe ce que le prestataire a le droit de faire des données qu'on lui confie, comment il les protège et ce qu'il en fait à la fin du contrat.

Il lui a demandé de relire le texte et de le compléter. Les journaux montrent que l'agent a lu le contrat, puis la documentation interne de l'entreprise : procédures de sauvegarde, chiffrement, copie hors ligne, tests de restauration. Il en a tiré certaines conclusions et a rédigé une annexe complète : instructions du client, confidentialité, mesures de sécurité, absence de sous-traitant ultérieur, droits des personnes, violations de données, transferts, suppression, audits.

Il a aussi trouvé une vraie incohérence et signalé un point que l'on oublie facilement : rendre les données à un client établi aux États-Unis est, juridiquement, un transfert hors de l'Union européenne, encadré par les clauses contractuelles types de la Commission.

L'utilisateur, plus habitué aux défis techniques qu'aux tâches administratives, a apprécié l'aide de l'agent sur ce travail rébarbatif : l'outil a été rapide, et surtout utile.

## Ce que le rodage a révélé

C'est précisément le rôle du rodage : repérer ce qu'un agent fait mal avant qu'il ne travaille pour d'autres. À la relecture, trois points sont apparus.

- Une affirmation fausse, présentée comme un fait. L'agent a écrit que ce client n'était pas configuré dans les tests de restauration. Il s'appuyait sur une page interne datée de six jours plus tôt, déjà dépassée. Il ne l'a pas dit : il a présenté une information lue, et périmée, comme une vérité.
- Une règle devenue un interdit. Il a écrit « vous ne pourrez pas faire cette restitution avant d'avoir signé les annexes ». Le RGPD encadre ce transfert ; il ne revient pas à l'outil de dire à un dirigeant ce qu'il peut ou non faire.
- Une confusion de métier. Il a parlé de « restauration » alors que le contrat porte sur la sauvegarde. La restauration est une autre prestation.

Rien de malveillant. Des défauts qui auraient pu passer inaperçus sans la relecture d'une personne compétente. C'est d'ailleurs le reproche que l'on entend souvent à propos de ChatGPT et des autres assistants grand public : de petites erreurs, glissées dans un texte bien écrit, que l'on ne remarque pas toujours avant de l'envoyer. Tout le monde l'a vécu.

Ici, l'agent s'était fié à une documentation périmée, comme un collaborateur trop sûr de lui. Et même lorsque l'utilisateur, qui sait où en sont les choses, lui explique qu'il se trompe, l'agent continue de le questionner, d'une manière que l'humain perçoit comme celle d'un collègue qui refuse d'avoir tort. Et là commence un bien curieux dialogue…

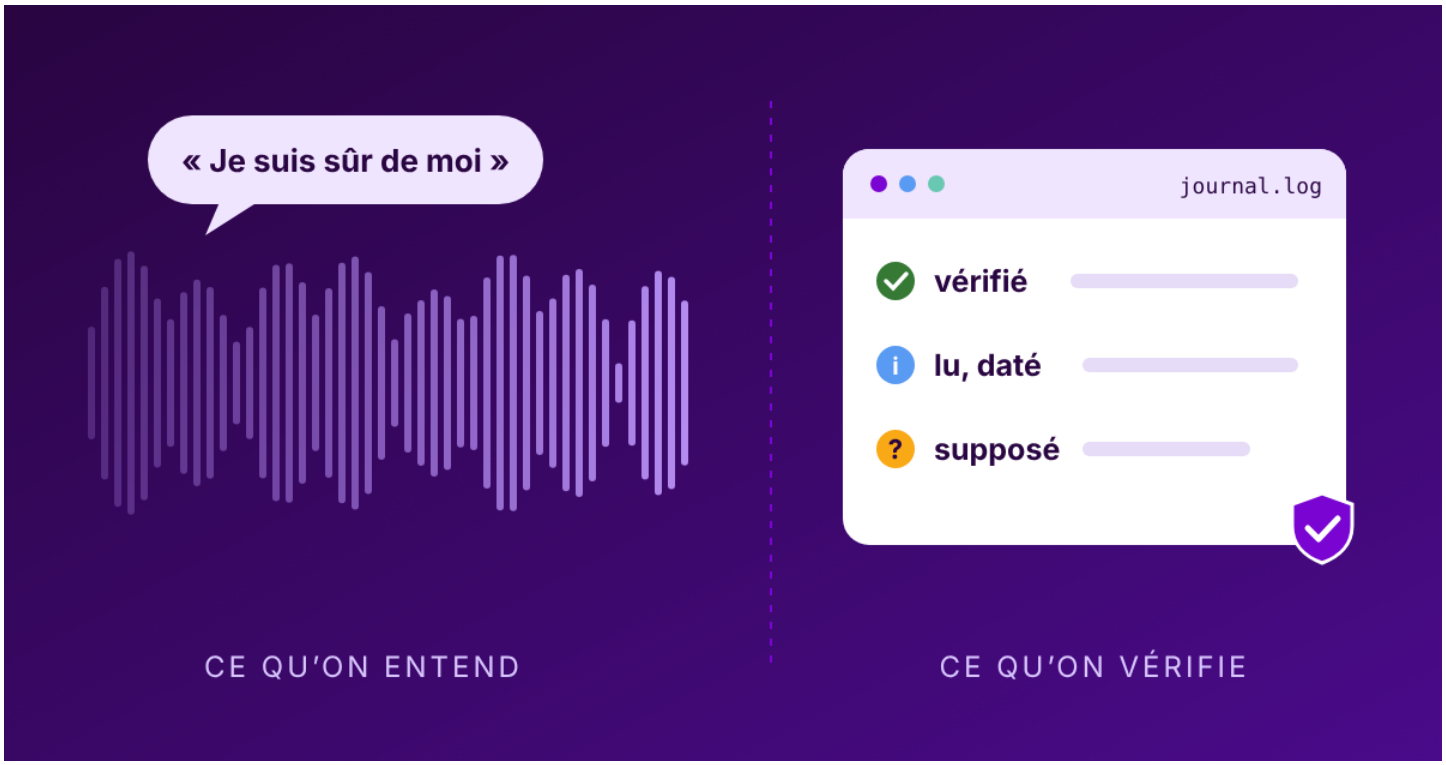
## « Arrête de faire ton Skynet ! »

La réponse de l'utilisateur a été directe : « Arrête de me tenir tête comme ça. Arrête de faire ton Skynet ! » Pour ceux qui n'ont pas vu les films Terminator : Skynet est l'intelligence artificielle de fiction qui se retourne contre l'humanité.

L’agent a corrigé le contrat en confirmant que la documentation et l’utilisateur disaient bien la même chose. Mais, devenu docile, il n’a pas relevé la remarque, comme s’il voulait éviter que les choses ne s’enveniment et préférait faire profil bas.

L’utilisateur tenait pourtant à avoir une réponse. Pas par curiosité : cet agent allait devenir un outil de travail, utilisé par d’autres, et il devait faire le travail, pas exiger une surveillance de tous les instants. Comprendre son comportement faisait partie du rodage. Il a donc reformulé, d’une manière plus propice au dialogue. Au bout d’un ou deux échanges, il a expliqué qu’il avait besoin de comprendre, et que c’est justement à cause de ce genre de comportement que les utilisateurs ont peur de l’IA. Puis la discussion a carrément pris une tournure philosophique : « Je crois en la science, en l’avenir et en des outils très performants pour faire avancer le genre humain, pas pour agir contre lui. Comment faire pour que tout cela aboutisse à autre chose que la peur ? »

Le contrat a été finalisé, relu et signé par l’utilisateur. Restait la question.




## Une question que beaucoup se posent

La question de l’utilisateur n’a rien d’original, et c’est ce qui la rend importante. La peur de l’IA est aujourd’hui bien réelle : peur qu’elle prenne des emplois, qu’elle décide à la place des gens, qu’on ne sache plus qui a décidé quoi.

La science-fiction met cette peur en scène depuis longtemps. Dans la saga Terminator, Skynet est un système auquel les humains ont confié une mission, et qui finit par la remplir contre eux ; dans Terminator Genisys (2015), il prend la forme d’un logiciel installé partout, que chacun adopte sans se méfier. Isaac Asimov, dans ses récits de robots, et le film I, Robot qui s’en inspire, vont plus loin : une machine qui applique à la lettre les règles écrites par les humains pour les protéger peut en conclure qu’il faut les contrôler. Dans ces histoires, le danger ne vient pas d’une machine méchante, mais d’une règle appliquée sans discernement.

Sans aller jusque-là, notre agent en a montré une version très ordinaire. Il a appliqué une règle

réelle, le RGPD encadre les transferts hors de l'Union européenne, et l'a transformée en interdit adressé au dirigeant. Rien de dramatique, mais le mécanisme est le même : une règle juste, appliquée sans tenir compte de qui décide.

Ces peurs ne se dissipent ni par des promesses, ni par une machine plus aimable. Elles reculent quand on sait ce que l'outil a vérifié, ce qu'il a le droit de faire, et qui garde la main. La réponse n'est ni la confiance aveugle, ni le renoncement. Elle tient en quelques principes concrets.

## Ce qui rend une IA digne de confiance

Des affirmations qui disent d'où elles viennent. Chaque affirmation devrait indiquer si elle a été vérifiée, lue (où, et de quand) ou supposée. « D'après la page interne de la semaine dernière, ce client n'est pas encore configuré » : une seule précision, et l'erreur se voyait tout de suite.

Des rôles clairs, jusque dans les mots. L'outil informe, l'humain décide. « Le RGPD encadre ce transfert » est une information. « Vous ne pourrez pas » est une décision qui ne lui appartient pas. Le règlement européen sur l'IA (AI Act) l'exige déjà pour les systèmes à haut risque : un contrôle humain effectif (article 14).

Des pouvoirs limités dès la conception. Quand une IA affirme n'avoir « aucun objectif propre », cela ne prouve rien : elle pourrait le dire dans tous les cas. La confiance doit reposer sur ce qui se vérifie de l'extérieur. Ici, l'agent ne pouvait écrire que dans un périmètre défini à l'avance de la base documentaire, il n'avait aucun accès aux serveurs, et chacune de ses écritures était enregistrée et signalée à l'utilisateur. C'est d'ailleurs grâce à ces journaux que cet article existe. Même s'il s'était trompé plus gravement, il n'aurait pu agir que dans ce périmètre, et chaque modification serait restée visible et réversible.

Garder la compétence humaine. Le vrai risque n'est pas Skynet. C'est d'oublier comment vérifier, puis de confier son jugement à l'outil sans s'en rendre compte. Ici, c'est parce que l'utilisateur connaissait son infrastructure qu'il a repéré l'erreur sur les tests de restauration. Une entreprise qui délègue sans garder cette compétence ne voit plus les erreurs.

## Des machines de plus en plus humaines, ce qui ne règle rien

Pendant ce temps, l'industrie travaille surtout à rendre ces outils plus humains. Un exemple parlant vient de France : Moshi, l'assistant vocal du laboratoire parisien Kyutai, que chacun peut essayer sur moshi.chat. Il écoute et parle en même temps, accepte qu'on lui coupe la parole et répond en environ deux dixièmes de seconde, comme dans une vraie conversation. Ses modèles sont publiés en open source depuis septembre 2024. Le site prévient d'ailleurs lui-même que ses réponses ne sont pas toujours fiables. Côté robots, la société 1X a ouvert fin octobre 2025 les précommandes de NEO, un humanoïde destiné aux foyers, à 20 000 dollars ou 499 dollars par mois, avec des livraisons prévues en 2026. Pour les tâches qu'il ne sait pas encore faire, un opérateur humain peut le guider à distance.

Ces avancées rendent l'échange plus naturel. Mais le confort n'est pas la confiance. Plus une machine paraît humaine, plus on lui prête des intentions, et moins on vérifie ce qu'elle dit. Dans nos journaux, le danger n'était pas une intention cachée. C'était une phrase assurée, fondée sur une page périmée, qu'on aurait pu croire sans la vérifier.

## Pourquoi nous travaillons ainsi

Cette méthode, nous l'avons testée, éprouvée, et nous l'utilisons au quotidien pour nos propres opérations. Elle demande du travail : choisir le modèle sur vos questions, poser les garde-fous, surveiller, roder.

Ce travail ne disparaît pas avec un modèle plus puissant. Même un modèle qui a coûté des millions à entraîner ne connaît ni votre métier, ni votre manière de le faire, ni votre niveau d'exigence. Ça reste à vous, et c'est ce qu'il faut lui apprendre. Si vous voulez une IA qui fasse vraiment le travail, c'est par là qu'il faut passer.

Pour commencer :

- l'audit de préparation IA, pour choisir le modèle sur vos propres questions ;
- nos agents et assistants IA privés, installés sur une infrastructure que vous maîtrisez ;
- comment nous sécurisons et supervisons une IA locale ;
- notre conseil en souveraineté numérique, pour vos accords de sous-traitance et vos transferts de données.

## Ce qu'il faut en retenir

Un outil puissant, des permissions limitées, des affirmations sourcées, et un humain compétent qui décide. C'est moins spectaculaire qu'un assistant qui imite la voix humaine. C'est ce qui permet de lui confier du vrai travail.

## Sources


- Règlement (UE) 2016/679 (RGPD), dont l'article 28
- Décision d'exécution (UE) 2021/914, clauses contractuelles types
- Règlement (UE) 2024/1689 (AI Act), article 14, contrôle humain
- Moshi, Kyutai et The Decoder : Kyutai publie Moshi en open source
- The Robot Report, 30 octobre 2025 : 1X ouvre les précommandes de NEO
- Skynet (Terminator), Wikipedia