CALIALABS RESEARCH · PUBLIC / STRATÉGIQUE

RPT-2026-02

GOVERNANCE

Avant de lâcher un agent IA dans une banque : les questions qu'il faut avoir tranchées

Les agents IA arrivent dans les back-offices régulés. Avant de les déployer, il y a une série de questions simples auxquelles peu d'architectures répondent aujourd'hui. Voici lesquelles, et pourquoi elles comptent.

AVRIL 2026 CALIALABS.COM/RESEARCH/MACHINES-AUTONOMY

Résumé exécutif

/// EXECUTIVE SUMMARY

On parle beaucoup d'IA agentique en ce moment. Les articles expliquent ce que c'est, les éditeurs promettent des gains de productivité, les régulateurs commencent à s'y intéresser. Mais quand on regarde de près ce qui est écrit, la plupart des discussions sautent une étape : avant de déployer un agent capable d'agir seul, quelles questions faut-il avoir tranchées ?

Ce papier ne raconte pas une expérience de terrain. Il propose une grille de lecture. On part d'un constat simple. Les dispositifs de conformité qu'on a construits depuis vingt ans partent d'une hypothèse : un humain décide, un système exécute, et on peut remonter la trace plus tard. Un agent autonome ne rentre pas dans ce cadre.

On passe donc en revue les questions que cette nouvelle logique soulève. Pas pour faire peur. Pour poser calmement ce qui doit être pensé avant que la première ligne de code soit écrite.

Points clés

PÉRIMÈTRE D'ACTION

Quelles actions l'agent a le droit de faire, et lesquelles il ne peut pas faire, même s'il le décide

TRACE DE LA DÉCISION

Pour chaque action, peut-on reconstruire la décision, pas seulement l'exécution

MOMENT DU CONTRÔLE

Le cadre se pose avant que l'agent agisse, pas en relecture après coup

SORTIE DE CRISE

Peut-on arrêter l'agent proprement, revenir en arrière, et expliquer ce qui s'est passé

Introduction

L'IA agentique, c'est-à-dire des systèmes capables non pas de suggérer mais d'agir, commence à trouver sa place dans le discours. On voit passer des démonstrations impressionnantes. On lit des papiers sur le sujet. Quelques déploiements commencent à sortir publiquement.

Dans un secteur régulé comme la banque, la question n'est pas seulement de savoir si ça marche techniquement. Elle est de savoir si on peut le gouverner. Et pour gouverner quelque chose, il faut d'abord l'avoir défini.

Ce papier propose cette définition, sous forme d'une série de questions. Si une équipe ne peut pas y répondre clairement avant le déploiement, elle y répondra sous la pression, après un incident. Ce n'est pas le meilleur moment.

1. Ce que change structurellement un agent autonome

Jusqu'ici, l'IA en banque, c'était surtout de la recommandation. Un modèle propose, un humain valide. La trace est simple. On sait qui a cliqué, quand, pourquoi. Les auditeurs peuvent remonter le fil.

Un agent autonome change cette logique. Il peut enchaîner plusieurs actions tout seul. Lire un dossier, vérifier une information, envoyer un email, mettre à jour un système, générer un document. Le tout sans humain entre chaque étape.

Ce n'est pas un problème en soi. Beaucoup de ces actions sont banales et gagnent à être automatisées. Le sujet, c'est que la trace ne distingue plus les banales des sensibles. Tout ressemble à une action machine. Et quand il faut expliquer, après coup, pourquoi telle action a été prise dans telle situation, la réponse n'est pas évidente à construire.

2. La première question : qu'est-ce que l'agent a le droit de faire

Avant d'écrire du code, il faut écrire un périmètre. Quelles actions sont autorisées. Quelles données sont accessibles. Quels destinataires sont légitimes.

Ça paraît évident dit comme ça. En pratique, la tentation est de partir sur une implémentation qui marche, et de cadrer plus tard. Ça se comprend. Mais plus on attend, plus il devient difficile de rétrécir le périmètre sans casser ce qui tourne déjà.

Le bon réflexe est l'inverse. On part du périmètre le plus étroit possible. On ouvre au cas par cas, avec une justification pour chaque ouverture. Cette discipline coûte un peu au démarrage. Elle économise beaucoup à la maintenance.

3. La deuxième question : qu'est-ce qu'on enregistre

La plupart des systèmes enregistrent les actions. Un agent a envoyé tel email à telle heure. C'est utile, mais c'est insuffisant.

Pour pouvoir auditer correctement, il faut aussi enregistrer la décision. Pourquoi cette action plutôt qu'une autre. Sur quelles données elle s'appuie. Quelle règle l'autorisait. Sans ces trois éléments, un auditeur a la vidéo mais pas le son.

Ça demande un effort de conception. Les systèmes classiques ne sont pas pensés pour ça. La traçabilité se rajoute rarement bien a posteriori. C'est une propriété qui se conçoit dès le départ.

4. La troisième question : quand intervient le contrôle

La conformité, traditionnellement, c'est surtout de l'ex-post. On laisse faire, on vérifie après. Ça fonctionne bien quand les décisions sont rares et prises par des humains.

Avec un agent qui peut prendre de nombreuses décisions très vite, ce modèle trouve sa limite. Personne ne peut relire toutes les décisions en temps réel. Et si on attend l'incident pour regarder, il est trop tard.

Le réflexe naturel est de vouloir un humain dans la boucle à chaque étape. Mais on retombe sur le problème d'origine. Si un humain doit tout valider, l'agent ne sert plus à rien.

La bonne question n'est donc pas où mettre l'humain, mais où mettre les règles. Le contrôle ex-ante consiste à définir à l'avance ce qui est permis, et à laisser le système empêcher techniquement le reste. L'humain garde sa place aux endroits qui le méritent : validation finale des actes qui engagent, revue des cas limites, arbitrage des exceptions.

5. La quatrième question : comment on sort

Un agent qui ne se comporte pas comme prévu, ça peut arriver. Le modèle sous-jacent évolue. Le contexte change. Un cas limite apparaît.

Il faut donc prévoir la sortie dès le départ. Peut-on arrêter l'agent proprement, sans casser ce qu'il a déjà lancé. Peut-on revenir en arrière sur les actions récentes, quand c'est techniquement possible. Peut-on expliquer, avec des éléments solides, ce qui s'est passé et pourquoi.

Ces questions ne sont pas des détails. Elles conditionnent ce qu'on peut dire à un régulateur, à un client, à une direction générale le jour où quelque chose ne va pas. Y répondre après l'incident, c'est répondre dans les pires conditions.

6. Quelques principes pour structurer la réflexion

Trois principes simples, qui découlent de ce qui précède.

Partir petit. Un seul cas d'usage, clairement défini, avec une équipe métier qui le porte. Les projets qui échouent sont souvent ceux qu'on lance trop large trop tôt.

Séparer explicitement ce qui est automatique de ce qui ne l'est pas. Certaines actions peuvent passer sans validation. D'autres non. La frontière doit être écrite, pas implicite, et appliquée par le système lui-même.

Concevoir la preuve en même temps que l'action. Chaque action sensible s'accompagne d'une trace exploitable de la décision. Ça ne se rajoute pas bien après coup, il faut l'intégrer dès les premières lignes.

Conclusion

L'IA agentique n'est pas un problème de technologie. La technologie marche, souvent mieux que ce qu'on imaginait il y a un an.

Le vrai sujet est celui de la gouvernance. Qu'est-ce qu'on autorise, comment on le contrôle, comment on le prouve. Ces questions se posent avant le premier déploiement, pas après.

Les organisations qui s'en sortiront ne sont pas celles qui auront les agents les plus impressionnants. Ce sont celles qui auront pris le temps de cadrer ce qu'elles allaient laisser faire, et ce qu'elles voulaient pouvoir prouver. Ce travail n'est pas glamour. Il est ce qui fait la différence entre une technologie qu'on subit et une technologie qu'on maîtrise.