

Gouverner l'IA : du coût du token au prix de la décision

Pourquoi l'IA ne devient gouvernable que lorsqu'une décision lie son coût, sa responsabilité et sa valeur

L'acheteur d'entreprise traque aujourd'hui le centime par token. Il compare les tarifs au million de tokens, surveille les heures GPU, arbitre entre un modèle frontière et un modèle ouvert selon une grille de coût unitaire. Tout le discours FinOps appliqué à l'IA repose sur cette seule opération : compter des appels. Et compter des appels gouverne parfaitement une chose, la consommation d'infrastructure. Cela ne gouverne rien au-delà. Surtout pas la décision.

Cet article prolonge une réflexion sur les coûts cachés du déploiement, où la supervision était identifiée comme le coût institutionnel que les modèles économiques de l'IA omettent. Il faut le dire d'emblée pour ne pas tromper le lecteur : ce qui suit ne parle pas du coût de l'IA, mais de la façon dont une entreprise transforme un événement technique en quelque chose qu'elle peut enfin mesurer, imputer et gouverner.

Posons la définition qui tient tout le reste. ***Un coût est gouvernable quand un acteur détient le levier institutionnel pour l'observer, le tarifier et en imposer le comportement associé.*** L'entreprise détient ce levier sur l'IA parce qu'elle maîtrise le point d'accès aux modèles. Mais cette maîtrise est la condition *habilitante* de la gouvernance, jamais sa condition *suffisante*. Tenir la porte est nécessaire. Ce n'est jamais assez.

La thèse est plus large qu'une question de coût. ***Une organisation ne gouverne économiquement un phénomène que lorsqu'elle sait produire un objet où consommation, responsabilité et valeur deviennent simultanément commensurables.*** Cet acte de liaison, le *nouage*, est le mécanisme, la gouvernance en est la conséquence. La décision est aujourd'hui la première instanciation opérable de ce nouage dans les systèmes d'IA, parce qu'elle est le premier objet où ces trois dimensions, jusque-là dissociées, se lient.

Le domaine de validité est étroit à dessein. La thèse ne vaut ni pour la conversation, ni pour l'idéation, ni pour les systèmes entièrement automatiques, où aucune décision humaine n'est attestée. Elle suppose une organisation dotée de propriétaires, d'une hiérarchie et d'une reddition de comptes. Hors de ce cadre, elle cesse de s'appliquer. Cette restriction ne l'affaiblit pas, elle la solidifie.

1. Définir la gouvernabilité, puis la dépasser

Le token échoue comme instrument de gouvernance non parce qu'il mesure mal, mais parce qu'il mesure le mauvais objet. Il observe une consommation avec une précision réelle. Il ne tarife pas la valeur produite et n'impose aucun comportement corrélé à cette valeur. Des trois verbes de la définition, il n'en sert qu'un. Observer sans tarifer ni contraindre, ce n'est pas gouverner, c'est facturer. Le FinOps gouverne l'infrastructure. Il ne gouverne pas la décision.

Pour comprendre pourquoi l'entreprise peut, elle, gouverner, tournons-nous vers deux auteurs que le discours technique ignore.

- Coase (1937) explique pourquoi la firme internalise une activité : parce qu'elle réduit des coûts de transaction que le marché lui ferait supporter.
- Williamson (1975, 1985) explique comment elle gouverne ce qu'elle a internalisé, par des structures de gouvernance ajustées aux attributs de la transaction, spécificité, incertitude, fréquence, et calibrées contre l'opportunisme.

Coase dit pourquoi la firme existe. Williamson dit comment elle tient. Notre sujet se loge entre les deux. Le point d'accès aux modèles est une structure de gouvernance au sens de Williamson, la frontière où la firme peut observer, tarifer, contraindre. Mais une structure ne gouverne rien tant qu'elle n'a pas d'objet sur lequel s'exercer. Exercer le levier suppose une *unité* : quelque chose qu'on puisse observer une fois, tarifer une fois, imputer à quelqu'un. La gouvernabilité est la porte. L'unité comptable est la pièce derrière la porte.

2. De la logique de ressource à la logique d'engagement

Searle (1995) distingue les faits bruts des faits institutionnels.

- Un fait brut existe indépendamment de toute institution : une masse, une impulsion électrique, un appel réseau.
- Un fait institutionnel n'existe que parce qu'une collectivité lui reconnaît un statut : une signature, un contrat, une décision. Sa formule, « X compte comme Y dans le contexte C », s'applique mot pour mot : un output de modèle compte, ou ne compte pas, comme une décision.

Une conséquence de Searle est que le fait institutionnel existe par reconnaissance collective, pas par enregistrement. La décision ne devient pas institutionnelle parce qu'un registre la consigne, mais parce que l'organisation accepte de lui attribuer ce statut. Le registre n'invente pas la décision, il constate une reconnaissance produite ailleurs. Croire l'inverse ferait du grand livre un objet quasi magique, qui produirait la réalité qu'il se contente d'enregistrer.

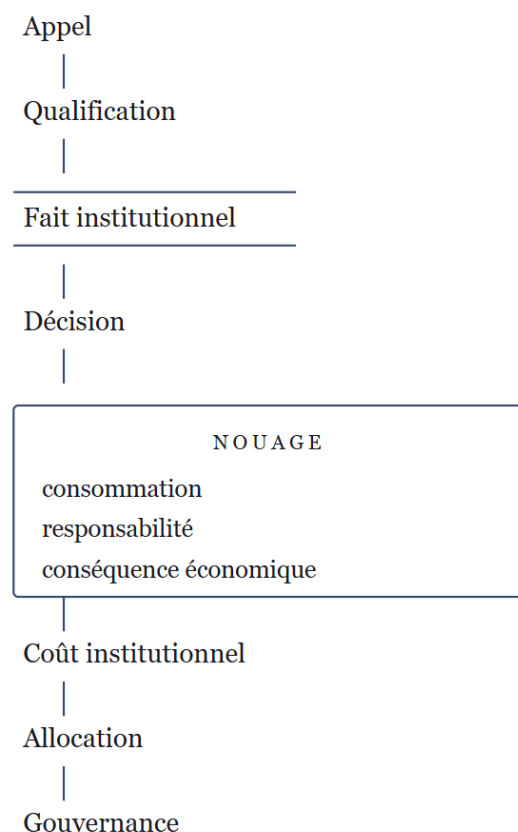
C'est le déplacement que réalise l'entreprise qui gouverne son IA.

- Un appel au modèle est un fait brut : il se consomme.
- Une décision validée est un fait institutionnel : elle s'endosse.

Le passage n'est pas un raffinement comptable, c'est un changement de logique : on cesse de mesurer une *ressource* et l'on se met à mesurer un *engagement*, entendu comme l'acte qui modifie l'ensemble des obligations organisationnelles attachées à un résultat. La relation joue dans les deux sens : décider ce qui compte comme une décision est déjà, en soi, un choix de gouvernance. Un appel se consomme. Une décision s'endosse.

Le nouage. On appelle nouage l'opération institutionnelle par laquelle une organisation relie, dans un même objet gouvernable, une consommation de ressources, une responsabilité organisationnelle et une conséquence économique. Le nouage est le mécanisme qui rend ces trois dimensions commensurables. La décision est aujourd'hui son premier support opérable dans les systèmes d'IA d'entreprise.

3. Le nouage, et pourquoi il rend commensurable



Le coeur de la théorie n'est pas la décision, c'est l'opération qui la constitue. Pourquoi la décision, et pas le workflow, le dossier, le résultat ? Parce qu'elle est le seul moment où une consommation technique devient irréversible, convertie en engagement organisationnel, à une condition qui fait partie de la thèse : qu'un humain l'endosse. Avant le nouage, les trois dimensions vivent dans trois registres sans dénominateur commun : la consommation dans les journaux techniques, l'engagement dans les organigrammes, la valeur dans les résultats d'exploitation. La décision les noue en un objet daté et attribué. C'est ce nouage, et lui seul, qui les rend commensurables. La décision est simplement le premier objet où il apparaît clairement ; la théorie porte sur l'opération, pas sur l'objet (Figure 1).

Figure 1 : Le nouage

Espeland et Stevens (1998) définissent la commensuration comme la comparaison d'entités différentes selon une métrique commune, et rappellent que la produire est un

travail, et un exercice de pouvoir. Un exercice de pouvoir appelle une question : qui, dans l'organisation, détient le droit de qualifier ? Le comité de direction, la DSI, le risque, le métier, le juridique ? C'est une question de souveraineté organisationnelle que cet article ouvre sans la trancher. La décision n'est donc pas une unité *primitive*, revendication ontologique fragile, mais une unité de *commensuration* : une unité comptable, parce qu'elle rend comparable ce qui ne l'était pas.

De ce nouage découlent les cinq propriétés d'une unité comptable, comme conséquences et non comme axiomes : identifiable, attribuable, clôturable, auditable, additive.

Chacune élimine une unité rivale.

- Le prompt porte une consommation mais aucune valeur réalisée, et s'arrête à l'output quand l'incertitude économique reste entière : il n'est pas clôturable.
- Le workflow, diffus, dilue l'attribution : il n'est pas identifiable.
- Le dossier passe de main en main : il n'est pas attribuable.
- Le résultat arrive trop tard, déconnecté de l'instant d'arbitrage : il n'est pas auditable au bon moment.

Knight (1921) fonde la clôture, l'incertitude ex ante ne se soldant qu'ex post ; Hayek (1945) ajoute que la connaissance pour l'évaluer est dispersée et ne se révèle qu'à l'usage. On ne tarife donc pas une décision, on la solde.

L'additivité, prise à la lettre, serait fausse : on n'additionne pas des décisions, hétérogènes par leurs risques et leurs horizons, on additionne leur coût institutionnel, que le nouage a réduit à une métrique commune. Cette lecture s'inscrit dans une régularité historique dont le ressort est précis : ce n'est pas la technologie qui impose une unité comptable nouvelle, c'est l'épuisement de l'ancienne, qui meurt quand elle cesse d'expliquer la création de valeur.

- L'heure de main-d'oeuvre a cédé quand la valeur a quitté le temps ouvrier,
- L'activité a émergé quand la valeur s'est logée dans les processus (Johnson et Kaplan, 1987)

Aujourd'hui, la valeur se loge dans l'arbitrage. Simon (1947) avait fait de la décision l'unité d'analyse de la firme sous rationalité limitée, où elle était une ressource organisationnelle. Ce que l'IA y ajoute est qu'elle en fait aussi une ressource économique.

La thèse produit des prédictions, ce qui la distingue d'une posture :

1. **P1** : à consommation de calcul identique, deux organisations aux niveaux de qualification institutionnelle différents présentent des coûts institutionnels différents.
2. **P2** : à qualification constante, le coût institutionnel varie indépendamment du coût d'inférence ; abaisser le second n'abaisse pas mécaniquement le premier.
3. **P3** : les gains de l'optimisation technique présentent des rendements décroissants tant que la qualification institutionnelle reste absente.

4. Trois couches orthogonales, dont une centrale

Le discours ambiant confond en un bloc trois opérations distinctes qu'il faut tenir orthogonales :

1. Router vers le bon modèle,
2. Qualifier ce qui compte comme décision,
3. Allouer la responsabilité.

La **couche 1**, l'optimisation technique, est un problème résolu. FrugalGPT (Chen, Zaharia et Zou, 2023) et RouteLLM (Ong et al., 2024) montrent que l'optimisation de la tâche est une commodité. Cela éclaire une objection immédiate : les suites d'assistance d'entreprise fonctionnent déjà sans grand livre des décisions, parce qu'elles opèrent dans le domaine de la requête d'assistance et laissent non compté le coût institutionnel des décisions qu'elles touchent. Elles ne se trompent pas, elles résolvent un autre problème. La couche 1 remplit par ailleurs la fonction qu'Akerlof (1970) nomme : faute de signal fiable, le demandeur se rabat sur la taille du modèle comme proxy, et la porte agit comme l'intermédiaire de screening qui dissipe cette asymétrie.

La **couche 2**, la qualification institutionnelle, est le coeur du dispositif, plus que la couche 3. Elle répond à une question binaire : cet événement acquiert-il un statut institutionnel, est-il une décision ? Tout tient là. Sans qualification, pas d'inscription, donc pas d'imputation, pas d'attestation, pas de responsabilité. Le grand livre est un objet-frontière au sens de Star et Griesemer (1989) : un artefact unique par lequel la finance, qui veut un coût, le juridique, qui veut une traçabilité, et le métier, qui veut une utilité, coopèrent sans partager la même représentation.

La **couche 3** est l'allocation de responsabilité. L'orthogonalité impose une garde : le grand livre enregistre, il ne gouverne pas. La séquence est stricte, registre, puis mesure, puis décision de gestion, puis gouvernance. Et il n'est pas un ERP ; dès qu'on le laisse tout qualifier, tracer et déclencher, il enfle en plateforme.

5. Le grand livre, implémentation de la couche 2

Le grand livre implémente la couche 2, rien de plus. Une décision est un output d'IA dont l'acceptation transfère la reddition de comptes à un propriétaire identifié. Cette définition trace une frontière étanche. La requête d'assistance, rédiger un courriel, résumer une réunion, ne transfère aucune responsabilité : elle est hors domaine, absorbée par la couche 1 sans inscription. L'acte de décision, choisir un traitement, valider un fournisseur, arrêter un score de crédit, engage l'entreprise et requiert un propriétaire humain : il s'inscrit, avec un solde.

Le coût institutionnel n'est pas la valeur de la décision, c'est ce que l'organisation engage pour la produire et pour s'y tenir. Il agrège deux coûts réalisés et une exposition provisionnée. Deux coûts réalisés :

- Le calcul, l'inférence des modèles mobilisés,
- Et la supervision, le temps humain de validation, gradué selon la criticité.

Une exposition provisionnée : la charge de risque que le propriétaire accepte en endossant, qui n'est pas une dépense déjà faite mais un risque qu'on ne paiera peut-être jamais.

On la traite donc comme une provision pour risque décisionnel, estimée à l'entrée, ajustée périodiquement, soldée quand on sait enfin ce que la décision a réellement produit. C'est ce provisionnement qui évite au grand livre de n'être qu'une archive : dans un crédit long ou une validation clinique, le solde réel peut survenir des années après l'acte, mais la ligne provisionnée reste pilotable entre-temps.

Un grand livre n'est pas encore une méthode. Observer le calcul est trivial, par les journaux d'inférence ; observer la supervision suppose de mesurer le temps humain de validation ; observer l'exposition suppose un modèle de provision. Un cas déployé devrait documenter trois grandeurs : le ratio de supervision, le temps d'expert rapporté à la criticité et au score de confiance du modèle ; le taux de friction, la proportion d'outputs rejetés ou profondément repris, révélateur du coût que la couche 1 déverse sur l'humain ; et la convergence comptable, l'écart entre l'exposition estimée ex ante et le coût du risque avéré ex post, sur un échantillon de dossiers soldés. Le mécanisme est spécifié ; l'instance mesurée reste à produire. [CAS À DOCUMENTER : déploiement réel, même partiel, octroi de crédit ou validation clinique]

Le choix du mot n'est pas neutre. Un registre serait un journal ; ce dont il s'agit est un compte, qui agrège, rapproche et budgète le coût institutionnel des décisions. On ne tient pas l'histoire des décisions, on en tient le compte. C'est cette additivité, appliquée au coût et non à la décision, qui transforme une traçabilité en instrument de gestion.

6. Allocation de responsabilité, arbitrages ex post, et limites

La couche 3 est le point de bascule institutionnel. L'attestation humaine n'y est pas une note de satisfaction accordée à l'IA, mais un choix d'endossement. Si le gestionnaire valide et endosse, il déclenche la tarification du coût institutionnel. Si l'output est imparfait mais repris, le temps de redressement se compte comme supervision itérative, révélant le coût réel que la couche 1 transfère vers l'humain. Seul l'output ni endossé ni repris se solde en déchet technique, imputé à l'infrastructure. La règle qui ferme le dispositif tient : on ne peut pas utiliser un livrable et déclarer ensuite qu'il n'avait pas de valeur.

Ce modèle déplace la question de l'optimisation des usages d'une logique de filtrage ex ante, fondée sur une intention supposée, vers une logique d'engagement ex post, fondée sur un engagement effectivement assumé. C'est ici que se dissipe une tentation répandue : gouverner l'IA en bridant l'accès selon la paresse supposée du prompt, par un filtre a priori qui dégrade le modèle pour les requêtes de confort. Cette voie tarife l'intention là où il faut solder l'engagement, et elle est punitive : elle juge l'utilisateur avant son acte, quand le grand livre le compte après, sans opinion sur lui. Brider la paresse fabrique du ressentiment et de la fuite, et ravive le shadow-AI qu'elle prétend réduire ; tarifier la décision produit de la gouvernabilité.

La même logique règle le cas du confort. Si l'évaluation révèle qu'un département consomme massivement du calcul sans jamais inscrire de décision au grand livre, ce n'est pas une faute qu'on sanctionne, c'est un fait comptable qu'on lit : une consommation sans endossement, qui se requalifie ex post en coût de confort opérationnel et se rattache au centre de responsabilité concerné. Le dispositif ne juge pas cet usage avant l'acte, il le constate après. La différence avec la tentation punitive n'est pas de degré, elle est de nature.

Reste l'objection la plus fréquente, philosophique, car elle vise le projet lui-même. Une bonne décision, dira-t-on, est un jugement, pas une unité comptable. Elle porterait si l'on comptabilisait le jugement ; ce n'est pas le cas. On n'attache pas un nombre au discernement du décideur, mais à l'engagement que sa décision crée. Le jugement reste irréductible et non tarifiable ; seule son empreinte institutionnelle, ce que la décision consomme, expose et produit, devient commensurable.

Une théorie solide dit aussi pourquoi elle échoue. Le grand livre échoue si la qualification est floue et laisse tout entrer ou tout sortir, si aucune décision n'a de propriétaire nommé, si la granularité est fausse, comptant des micro-arbitrages ou des dossiers entiers plutôt que des décisions, ou si son coût de mise en oeuvre excède le coût institutionnel qu'il

permet de gouverner. Ces échecs ne réfutent pas la théorie, ils en délimitent les conditions d'exercice.

Le dispositif a d'autres limites qu'il faut nommer. Il gouverne le coût et la reddition de comptes de la décision, pas sa qualité : tarifier la responsabilité n'est pas garantir le discernement. La porte est nécessaire, pas suffisante, la qualité dépendant aussi des workflows, des incitations et des droits. Et le champ exclut la conversation, l'idéation, les systèmes entièrement automatiques, et les organisations sans propriétaire clair ni reddition de comptes, une organisation autonome décentralisée pilotée par le code, un collectif open source, une coopérative pure, où l'unité perd son attributabilité. Enfin, la théorie reste en avance sur son empirie.

Un mot sur la portée, pour ne pas la confondre avec sa preuve. Cet article a démontré un cas : la décision comme première instanciation opérable du nouage dans les systèmes d'IA. Mais le critère est général. Une organisation ne gouverne économiquement un phénomène que lorsqu'elle sait produire un objet où consommation, responsabilité et valeur se nouent en une métrique commune. La décision est aujourd'hui le premier tel objet ; d'autres suivront sans que le critère change. Le mécanisme est plus large que sa première preuve.

Conclusion

Tout ceci tient à une condition, une seule : l'entreprise maîtrise le point d'accès. Elle est le lieu unique où la porte existe par défaut. Retirez la porte, et il ne reste ni routage, ni comptabilité, ni attestation. C'est la situation à l'échelle de l'écosystème, où aucun acteur ne détient par défaut cette position, et où la gouvernabilité devra être créée plutôt que supposée. C'est l'objet de l'article suivant.

L'IA en entreprise ne devient gouvernable ni quand on la comprend, ni quand on la compte au token. Elle le devient quand un appel cesse d'être une consommation et devient un engagement dont quelqu'un répond.