

La frontière de responsabilité

Pourquoi les hyperscalers descendent vers le résultat, et pourquoi le retour les attend là où ils ne contrôlent qu'un sous-ensemble des facteurs

Pendant quarante ans, l'industrie du logiciel a répondu de la disponibilité d'un moyen, et le client a répondu du résultat. Cette ligne se déplace.

Le 2 juillet 2026, Microsoft a annoncé Frontier Company, une unité opérationnelle de 2,5 milliards de dollars qui place environ six mille ingénieurs et spécialistes sectoriels au sein des entreprises clientes pour concevoir, construire et opérer leurs systèmes d'IA. L'annonce n'est pas le mouvement. Elle en est le premier cas observable. La thèse de cet article est que la valeur d'un système d'IA d'entreprise n'est pas une fonction additive mais multiplicative de facteurs complémentaires, qu'un seul facteur ramené vers zéro effondre le produit, et que le facteur qui se tient le plus souvent près de zéro, la capacité institutionnelle du client, est précisément celui qu'aucun fournisseur ne peut livrer de l'extérieur, parce qu'un actif se déploie et qu'une compétence s'institue.

Lue ainsi, la question intéressante à propos de Frontier n'est pas de savoir si Microsoft a mal diagnostiqué le problème. C'est ce qui arrive à un éditeur de logiciel lorsque la rente dont il vit s'est déplacée vers un lieu qu'il ne contrôle pas, et qu'il l'y suit quand même. Cet article prolonge une lecture antérieure et plus immédiate de la même annonce, et passe délibérément de la réaction à la structure qui la sous-tend.

De l'éditeur à l'opérateur

Le vocabulaire de l'annonce est le signe révélateur. Microsoft décrit l'unité comme orientée résultat, promet des résultats économiques mesurables, et lui accorde sa propre responsabilité financière, les engagements étant jugés à l'aune des résultats économiques plutôt que du logiciel livré. GeekWire, lisant le même matériau, a conclu que l'initiative est moins une entreprise nouvelle qu'un effort plus large et mieux étiqueté derrière un travail que Microsoft faisait déjà. Les deux descriptions pointent le même déplacement. Un fournisseur qui propose d'être jugé à l'aune du résultat du client est un fournisseur qui propose de répondre du résultat.

Situons cela dans la longue lignée de l'industrie. La maison de logiciel vendait un artefact sous licence et répondait de sa conformité. La plateforme vendait un substrat et répondait de ses interfaces. Le cloud vendait une ressource et répondait de sa disponibilité. Le copilote vendait un assistant et répondait de ses suggestions. Frontier

propose la station suivante : la responsabilité partielle du résultat lui-même. Le déplacement n'est pas technologique. Il est juridique, économique et organisationnel, et c'est la raison pour laquelle l'histoire compte au-delà d'un trimestre.

La composition humaine de l'unité mérite un paragraphe et pas davantage, car dans cinq ans personne ne se souviendra de l'organigramme tandis que tout le monde débattrait encore de la responsabilité. Les six mille sont issus principalement d'équipes d'ingénierie et d'équipes déployées en immersion existantes. Une part significative descend d'Industry Solutions Delivery, le groupe qui avait lui-même absorbé les anciens Microsoft Consulting Services. Des 2,5 milliards de dollars, environ 1,2 milliard couvrirait les coûts de personnel, y compris la relocalisation et la formation à la préparation à l'IA, ce qui est le langage de la requalification d'effectifs existants plutôt que du recrutement d'une armée nouvelle. Et l'annonce est arrivée à quelques jours d'une réduction d'environ neuf mille postes concentrée dans la vente et le conseil, calée sur l'ouverture de l'exercice fiscal. Débarrassé de toute intention, le schéma est cohérent avec une reclassification de la main-d'œuvre de delivery vers une unité à plus forte responsabilité plutôt qu'avec la création de capacité nette. Que les 2,5 milliards soient du capital nouveau ou une réallocation, et sur quelle période, n'a pas été divulgué. L'écart mérite d'être noté et non d'être comblé.

Le logiciel répondait d'un moyen. Il propose désormais de répondre d'une fin. La question est de savoir si la fin est une chose qu'un fournisseur peut fournir.

La rente migre parce que le modèle se banalise, et l'entreprise est sous tension

La migration est probablement inévitable, et la raison en est économique plutôt que stratégique. La cause immédiate est la banalisation de la couche qui a brièvement détenu la rente, le modèle lui-même. Le Stanford AI Index rapporte que le coût d'une requête à un modèle pour atteindre le niveau de GPT-3.5 sur le benchmark MMLU est passé d'environ vingt dollars par million de tokens fin 2022 à environ sept centimes fin 2024, près de deux ordres de grandeur en dix-huit mois. Epoch AI constate que les prix d'inférence chutent à un rythme médian proche de deux cents fois par an, pour un niveau de capacité fixé, depuis le début de 2024. Lorsque l'écart architectural entre les modèles de tête se resserre, lorsque les API convergent, et lorsque le prix de l'invocation d'un niveau donné d'intelligence s'effondre, le modèle cesse d'être une source durable de rente économique. La rente doit se déplacer vers la couche qui ne se banalise pas.

Cette couche est l'intégration, la gouvernance et le résultat propre du client. En quatre décennies, la rente s'est déjà déplacée le long d'une chaîne : le matériel, puis le système d'exploitation, puis la suite bureautique, puis le serveur, puis le cloud, puis la donnée, puis le modèle, et maintenant le résultat. Chaque déplacement a échoué le détenteur de

rente précédent dans une couche banalisée et poussé la valeur d'une station de plus vers le résultat du client. La convergence des fournisseurs en est la preuve, non l'imitation. Deux jours avant Frontier, Amazon a engagé un milliard de dollars dans une initiative comparable d'ingénierie en immersion. Anthropic et OpenAI avaient chacun mis sur pied des groupes déployés en immersion en mai. Lorsque quatre acteurs capables font le même mouvement en quelques semaines, la lecture parcimonieuse est un déplacement partagé de la rente, non une mode. Frontier peut se lire comme une adaptation à la migration de la valeur, ce qui est une affirmation plus faible et plus défendable que tout récit des intentions de Redmond.

La symétrie que la plupart des analyses manquent se tient du côté de la demande. Le client accepte le transfert de responsabilité parce que le client est lui aussi en détresse. L'enquête talents 2026 de ManpowerGroup a constaté que les compétences en IA étaient devenues, pour la première fois dans l'historique, la catégorie de poste la plus difficile à pourvoir, avec environ soixante-douze pour cent des quelque trente-neuf mille employeurs déclarant des difficultés, et la France en tête à soixante-quatorze pour cent. McKinsey a décrit l'innovation en IA comme contrainte par la pénurie de talents, et constate qu'une large majorité d'organisations utilisent désormais l'IA tandis que moins de deux sur cinq rapportent un impact sur l'EBIT à l'échelle de l'entreprise. BCG attribue environ soixante-dix pour cent du succès en IA aux personnes, aux processus et à la conduite du changement, et seulement les trente pour cent restants aux algorithmes, aux données et à l'infrastructure réunis. Ce sont des cabinets de conseil qui ont intérêt à la conclusion selon laquelle la contrainte contraignante est organisationnelle plutôt que technique, ce qui est aussi l'argument avancé ici, et les chiffres doivent être pesés en tenant compte de cet alignement. Le Forum économique mondial, sur un terrain plus ferme, s'attend à ce qu'environ quarante pour cent des compétences fondamentales des travailleurs changent d'ici 2030, ce qui comprime la demi-vie de toute compétence qu'une entreprise parvient à construire. Sous la crise des talents siège une crise documentaire. La Global DataSphere d'IDC est passée de trente-trois zettaoctets en 2018 à environ cent soixante-quinze en 2025, en large majorité non structurée, alors même que la donnée gouvernée et exploitable restait rare. Le client se noie dans la matière première et manque de la version de celle-ci qu'un système peut utiliser.

La rencontre n'est donc pas celle d'un fournisseur fondant sur un client passif et immature. C'est une partie en détresse rencontrant une autre. L'hyperscaler arrive avec son titre en baisse d'environ vingt et un pour cent sur l'année et ses dépenses d'investissement sous la surveillance des investisseurs. L'entreprise le reçoit avec une pénurie de talents dont elle ne peut se sortir par le recrutement et un patrimoine de données qu'elle ne peut gouverner au rythme où il croît. Deux détresses se rencontrent. Aucune n'est une position de force. C'est un cadrage plus honnête que l'immaturité, et il pose la vraie question, qui est de savoir où siège réellement le facteur contraignant.

De la contrainte à la complémentarité

La lecture tentante, celle vers laquelle une version antérieure de cette analyse s'est portée, est que Microsoft a correctement diagnostiqué la maladie et agit du mauvais côté d'un goulot d'étranglement unique. Cette lecture est fausse. Elle n'est pas injuste, elle est une mauvaise théorie, et elle échoue pour une raison précise : il n'y a pas de goulot d'étranglement unique.

L'objet correct n'est pas une contrainte mais une fonction de production, et il vaut la peine d'être exact sur laquelle, car la précision est l'argument. La valeur d'un système d'IA d'entreprise est le produit, non la somme, de ses facteurs complémentaires : infrastructure, données, gouvernance, compétences, processus, adoption, mesure et incitations. Une stricte fonction de Leontief à proportions fixes rendrait ces facteurs parfaitement non substituables et fixerait la valeur au minimum des intrants. C'est trop fort, et ce n'est pas la thèse ici. La forme juste est celle que Michael Kremer a formalisée sous le nom de théorie du joint torique (O-Ring) du développement économique (Quarterly Journal of Economics, 108(3), 1993, pages 551 à 575), dans laquelle la production est le produit des qualités des tâches, une substitution limitée entre tâches est possible, et pourtant la tâche la plus faible domine encore le résultat, de sorte que la quantité ne peut se substituer à la qualité. Le nom vient de la navette Challenger, perdue à cause d'un seul joint défailant alors que tout autre composant fonctionnait aux spécifications. Paul Milgrom et John Roberts en ont donné la contrepartie organisationnelle dans leur analyse de la production manufacturière moderne (American Economic Review, 80(3), 1990, pages 511 à 528), où le rendement de l'élévation d'une activité n'apparaît que lorsque ses compléments sont élevés en même temps, et où élever l'une pendant que les autres restent en retrait ne produit presque rien. La distinction entre Leontief et O-Ring n'est pas du pédantisme. Elle concède qu'un fournisseur peut relever quelque peu un facteur bas, et nie que relever les facteurs qu'il contrôle puisse compenser un facteur proche de zéro qu'il ne contrôle pas.

Écrivez la valeur d'un système d'IA comme un produit sur ces facteurs et la conséquence suit immédiatement. Un seul facteur proche de zéro effondre le produit, si grands que soient les autres. Microsoft contrôle un sous-ensemble. Il peut relever l'infrastructure, et il peut relever une partie du facteur compétences en plaçant des ingénieurs capables sur site. Il ne contrôle pas la gouvernance, il ne contrôle pas l'adoption, et il ne contrôle pas la capacité institutionnelle du client à faire tourner ce qui est construit. L'inclusion des incitations comme facteur explicite marque la limite extrême de la portée du fournisseur. Une entreprise peut détenir des données propres, des processus sains et des personnes formées, et ne rien en extraire si ses incitations internes se contredisent, lorsque la fonction technologique optimise la maîtrise des coûts, les unités métier la vitesse de déploiement, la finance un effet de court terme sur le compte de résultat, et la conformité la minimisation du risque. Aucune quantité d'ingénierie externe ne résout un

désalignement d'incitations internes, car le fournisseur n'a aucune légitimité pour redistribuer les droits de décision d'un client. Sur-approvisionner en ingénierie face à un facteur de gouvernance ou d'incitation proche de zéro ne déplace pas le produit, car le produit est multiplicatif et le facteur manquant n'est pas à Microsoft de le relever.

C'est pourquoi la statistique désormais célèbre, le constat du MIT NANDA selon lequel quatre-vingt-quinze pour cent des pilotes d'IA générative en entreprise ne produisent aucun impact mesurable sur le compte de résultat, doit se lire non comme une déficience des modèles mais comme la signature statistique d'un effondrement multiplicatif. Les modèles fonctionnent dans la démonstration. Le produit échoue dans l'entreprise parce qu'au moins un facteur détenu par le client est proche de zéro.

Le profil rare que tout le monde s'arrache est lui-même l'un de ces facteurs, non un moyen de les contourner. L'ingénieur qui maîtrise simultanément les opérations sur LLM, la structure de compte de résultat d'un secteur régulé, et les contraintes de l'AI Act et du RGPD est rare parce que la distribution conjointe de trois compétences peu communes est rare, et le capital ne comprime pas la maturation d'une distribution conjointe. Selon l'estimation de McKinsey en 2024, les cycles de recrutement dans les industries régulées sont environ soixante-dix pour cent plus longs pour exactement ces profils. Amazon, Anthropic et OpenAI peuvent faire monter le prix de ce profil. Aucun d'eux ne peut le fabriquer en un an. La rente, au bout du compte, revient à qui maîtrise la géométrie des facteurs complémentaires, non à qui aligne le plus grand nombre d'ingénieurs. On ne peut pas s'acheter le franchissement d'une fonction de production multiplicative en sur-approvisionnant un seul intrant.

Actif, compétence, capacité, institution

Que livre précisément le fournisseur, et que laisse cette livraison intacte ? La friction du logiciel adossé au résultat a une source ontologique, et elle devient plus claire une fois séparées quatre classes de ressources plutôt que deux. Un actif devient une compétence devient une capacité devient une institution, et un fournisseur ne peut vendre que la première et louer brièvement la seconde.

Un actif est une chose. Azure est un actif. Copilot est un actif. Un pipeline de récupération est un actif. Un agent est un actif. Un actif possède trois propriétés qui le rendent vendable : il est transférable, il est finançable, et il est livrable. Un fournisseur peut l'emballer, le tarifier, l'expédier et comptabiliser le revenu. Une compétence est la maîtrise localisée pour opérer et maintenir cet actif : la fluidité d'un ingénieur dans les opérations sur LLM, la routine d'une équipe pour évaluer la sortie d'un modèle. Une compétence peut être formée, et elle peut être étayée de l'extérieur pour un temps. Une capacité est l'état institutionnalisé plus large, endogène à l'entreprise, cumulatif sur des années, et enchâssé dans une structure politique et sociale, ce qui revient à dire non transférable par contrat. C'est l'aptitude collective de l'organisation à absorber, gouverner

et extraire de la valeur de ses compétences, et par là à changer sa propre performance. Une institution est la couche au-dessus même de cela : la matrice sociotechnique qui distribue les droits de décision, assigne la responsabilité légale, arbitre le conflit interne, absorbe l'erreur algorithmique, et fabrique la confiance de base sans laquelle rien du reste n'opère.

Un fournisseur peut vendre un actif. Il peut étayer une compétence pour la durée d'un engagement. Il ne peut pas vendre une capacité, et il ne peut certainement pas vendre une institution, pour la même raison qu'une salle de sport ne peut pas vendre la forme physique. Elle peut vendre l'accès aux équipements et la présence d'un coach. La capacité est ce que le membre construit, ou échoue à construire, une fois le coach reparti. Le facteur contraignant de la section précédente est une capacité, et c'est donc exactement la chose qui ne peut pas être livrée. Tel est le paradoxe au centre de la proposition Frontier, énoncé sans malveillance : Frontier déploie un actif et facture l'institution d'une capacité qu'il ne peut pas lui-même instituer.

L'objection la plus forte doit être posée dans toute sa force avant d'être répondue. Le modèle des ingénieurs déployés en immersion, inventé chez Palantir et désormais copié à travers l'industrie, repose sur l'affirmation que l'immersion technique est la manière de construire la maturité qui manque. Selon cette vue, l'immaturité du client est la cible, non un décalage, et se plaindre que le client n'est pas prêt revient à reprocher à un incendie l'existence des pompiers. Une version plus aigüe retourne l'argument des complémentarités contre lui-même : si la valeur est multiplicative et que Microsoft le comprend, alors immerger des ingénieurs qui touchent aussi à la donnée, à la gouvernance et au processus est précisément une tentative de relever plusieurs facteurs à la fois, de sorte que l'affirmation selon laquelle Microsoft ne contrôle qu'un sous-ensemble est trop forte.

La réponse est une distinction à l'intérieur de l'objection. Un fournisseur peut relever les facteurs techniques, et il peut étayer les autres pour la durée de l'engagement. L'étayage n'est pas l'institution. La gouvernance, l'adoption, la mesure et l'arbitrage des incitations sont des actes organisationnels et politiques, cumulatifs et non transférables, et ils se dégradent quand les personnes du fournisseur quittent le bâtiment. La cadence rend le point concret. Les engagements Frontier sont décrits comme des sprints de quarante-cinq jours menés par cinq ou six personnes, un rythme calibré pour l'écart entre démonstration et déploiement des workflows faiblement régulés. Les clients vitrines de Microsoft comprennent Novo Nordisk et le London Stock Exchange Group, où la qualification, le contrôle du changement et les pistes d'audit ne se compriment pas à quarante-cinq jours. L'avantage de vitesse s'évapore précisément dans les terrains régulés que le fournisseur choisit d'exhiber. Une compétence qui ne dure que tant que le fournisseur reste n'est pas une compétence. C'est une dépendance. Le fournisseur peut

installer la machinerie de la fiabilité. Il ne peut pas installer, du côté du client, la personne qui la ferait tourner.

La hiérarchie d'intendance et la gouvernance de la mesure

Microsoft livre ce qu'il appelle une plateforme de confiance, responsable de la gouvernance, de la sécurité et de la mesure du retour. Une plateforme de ce genre expose un ensemble de contrôles, et un contrôle ne vaut que par le rôle qui l'opère de l'autre côté. Ces rôles sont humains, et la plupart des organisations clientes en sont dépourvues. La proposition qui suit est doctrinale plutôt qu'empirique, et elle doit se lire comme un cadre et non comme un constat.

Cinq formes de responsabilité humaine doivent exister du côté client pour que la plateforme livrée se connecte à quoi que ce soit, et elles forment une hiérarchie plutôt qu'une liste. Un *Intendant de la donnée* répond de la vérité, de la traçabilité et de la conformité des données source. Un *Intendant de la connaissance* répond de la vérité et de l'actualité du corpus documentaire, de la différence entre un document qui existe et un document qui est encore correct, ce qui est le rôle que la prolifération documentaire antérieure met le plus directement en échec. Un *Intendant de l'IA* répond de la couche algorithmique, du contrôle de la dérive, de la frontière du refus, et de l'intendance du coût d'inférence, le compteur qui tourne que la valeur arrive ou non. Un *Propriétaire du processus* répond de la responsabilité opérationnelle et légale du workflow que le système informe. Couronnant ces quatre, le *Propriétaire de la décision* gouverne le seuil du raisonnement automatisé lui-même : quel niveau de confiance algorithmique suffit à agir, quand un humain est replacé dans la boucle, et comment le coût d'une erreur automatisée se pèse contre la vélocité de l'autonomie. Les quatre premiers maintiennent la plomberie honnête. Le Propriétaire de la décision décide de la part du jugement de l'entreprise cédée à la machine, ce qui, dans un cadre régulé, est la seule question qui compte finalement. L'Intendant de la donnée est au moins un rôle nommé, parfois pourvu. Le Propriétaire de la décision n'est presque nulle part formalisé, ce qui est une manière précise de dire que le contrôle le plus conséquent de la plateforme se connecte, du côté client, à personne.

Cela inverse le récit causal habituel. L'intendant absent n'est pas la cause de l'échec. Il en est un symptôme. La chaîne va de l'absence de capacité institutionnelle à l'absence des intendants, de l'absence des intendants à l'absence de gouvernance, de l'absence de gouvernance à l'absence de mesure, et de l'absence de mesure à l'absence de tout retour vérifiable. Diagnostiquer l'intendant manquant puis en recruter un traite le symptôme, car le rôle ne peut tenir sans la capacité qui le porte.

Ce qui amène la question de la mesure à sa forme plus aiguë. La plainte selon laquelle Microsoft mesure son propre retour est vraie mais superficielle. Le fait conséquent est que le fournisseur définit progressivement l'instrument de mesure lui-même : les métriques, les observables, les tableaux de bord, les traces. Dans l'économie des organisations, celui qui définit la métrique oriente le comportement qui est mesuré contre elle. L'enjeu n'est pas l'auto-notation, qui en est la version grossière, mais la gouvernance de la mesure, qui en est la version conséquente. Corriger sa propre copie est embarrassant. Concevoir la grille de notation que tous les autres utiliseront est un pouvoir structurel, et c'est le plus discret des deux.

Angles morts systémiques, et ce que nous savons réellement de chacun

Plusieurs effets de second ordre suivent, et l'honnêteté intellectuelle exige d'attacher à chacun un niveau de certitude plutôt que de les présenter à confiance uniforme.

1. Le premier est la préemption du canal, et il est documenté. Microsoft nomme un écosystème de partenaires, Accenture, Capgemini, EY sous une alliance de cinq ans, KPMG et PwC, tout en immergeant ses propres ingénieurs directement dans des comptes régulés emblématiques tels que le London Stock Exchange Group et Novo Nordisk. La tension est réelle mais sa forme est spécifique. Ce n'est pas une substitution industrielle des intégrateurs mondiaux, qui serait impossible à cette échelle, mais quelque chose de plus étroit : la capture de la décision d'architecture amont sur Azure au niveau du conseil d'administration, en laissant l'industrialisation lourde aux partenaires. Ce qui est pris est la décision, non la livraison.
2. Le deuxième est une asymétrie de propriété intellectuelle, et il est structurel. La protection des données d'entraînement du client est réelle et énoncée au niveau du principe, la direction de Microsoft soutenant qu'il n'existe pas de licence sociétale pour un futur de l'IA qui consommerait l'intelligence des entreprises au sein desquelles il est déployé. Mais la protection est cantonnée aux données d'entraînement. L'apprentissage méthodologique, les abstractions réutilisables et les playbooks sectoriels distillés de chaque sprint s'accumulent dans le patrimoine cognitif du fournisseur à travers les engagements. Chaque déploiement abaisse le coût de réduction d'incertitude du fournisseur sur le suivant, tandis que le client individuel reste scellé à l'intérieur de ses propres murailles de données. La donnée reste chez le client. La méthode se capitalise chez le fournisseur.
3. Le troisième est une reconfiguration du marché du logiciel, offerte ici comme analyse sectorielle plutôt que comme fait. En descendant à travers les agents

autonomes et le conseil d'architecture, l'hyperscaler menace la pertinence des éditeurs de SaaS verticaux dont les fonctionnalités sont de plus en plus susceptibles d'être absorbées dans le substrat Azure AI. La couche applicative qui se tenait autrefois au-dessus de la plateforme est précisément la couche que la plateforme apprend désormais à générer.

4. Le quatrième est la boucle, et il doit être marqué comme une tendance structurelle et une hypothèse explicite, non comme une intention imputée. Le modèle économique profond de l'hyperscaler repose sur la consommation de compute, ce qui crée une divergence objective entre la fonction du fournisseur, la génération de consommation, et celle du client, la recherche d'efficacité opérationnelle et de moindre coût fixe. L'économie du token aiguise cela plutôt qu'elle ne l'adoucit. Le prix par token baisse, mais la facture totale monte, parce que les modèles de raisonnement et les workflows agentiques multiplient le nombre de tokens qu'une tâche consomme, un schéma que les analystes de l'industrie ont commencé à appeler le paradoxe du coût. Des unités moins chères et des factures plus grandes n'est pas un accident pour un fournisseur dont la base est la consommation ; c'est le moteur. Une fois qu'un agent atteint la production, la chaîne devient une boucle : l'usage produit de la donnée, la donnée améliore les agents, de meilleurs agents approfondissent la dépendance, la dépendance élève la consommation, et la consommation produit plus de donnée. C'est une tendance plutôt qu'un compoundement démontré, et c'est une propriété de l'architecture plutôt qu'un motif caché.
5. Le cinquième relève de cette section mais doit être énoncé comme continuité plutôt que monopole, et comme transitoire plutôt que permanent. Le graphe unifié à travers SharePoint, Teams, Exchange, OneDrive, Entra et le Microsoft Graph est ce qui rend les agents d'entreprise seulement tractables, parce que le substrat documentaire existe déjà comme un tout connecté, et il est difficile à reconstruire rapidement. Ce n'est pas, cependant, un rempart durable. Trois forces l'érodent. Les règles européennes de souveraineté des données et la croissance des espaces de données industriels restreignent la centralisation sans contrainte des données d'entreprise dans le cloud d'un fournisseur unique. Le Model Context Protocol, publié par Anthropic fin 2024, adopté en quelques mois par OpenAI, Google, Microsoft et AWS, et cédé à l'Agentic AI Foundation sous la Linux Foundation en décembre 2025, découple la couche d'intégration de tout modèle ou plan de données unique, de sorte que l'investissement dans la connexion d'un système n'est plus lié au fournisseur du modèle. Et les grandes entreprises construisent des couches sémantiques et des bases de données de graphe neutres vis-à-vis des fournisseurs précisément pour éviter cette classe de verrouillage. Microsoft a lui-même adopté le Model Context Protocol dans Copilot

Studio, ce qui est l'aveu discret que le standard qui érode le rempart est un standard que l'incumbent ne peut refuser.

Deux hypothèses : la capacité comme service contre l'usine interne

L'affirmation qu'un fournisseur ne peut livrer le résultat de l'extérieur suppose que le client veut, et est capable de, construire une institution interne. C'est là que l'argument est le plus exposé, et deux futurs le mettent à l'épreuve. Ce ne sont pas des alternatives décoratives. Ce sont les deux manières dont cette thèse pourrait se révéler fausse, et les énoncer est la discipline qui sépare une grille de lecture d'une prophétie.

1. Le premier futur est la capacité comme service (*capacity-as-a-service*). Une large part du marché peut décliner de s'institutionnaliser du tout, et choisir plutôt de louer sa production cognitive à demeure, traitant l'intelligence organisationnelle comme un utilitaire géré de l'extérieur et convertissant le déficit de responsabilité en un abonnement permanent. Sous ce futur, la distinction entre déployer un actif et instituer une capacité se dissout, car la capacité ne revient jamais au client. Elle devient un appendice loué, opéré par le fournisseur, tarifié comme un utilitaire cognitif. La pénurie de talents et les intendants absents ne sont pas résolus du côté client ; ils sont absorbés par le fournisseur et refacturés. Si c'est là que le marché se stabilise, Frontier ne se surestime pas. Il est en avance.
2. Le second futur est l'usine à IA interne. Les entreprises peuvent traiter la production cognitive comme une compétence cœur, monter leurs propres équipes de plateforme IA, centres d'excellence et bureaux de programme, et réduire l'hyperscaler à une couche d'infrastructure sous une orchestration souveraine qu'elles possèdent. Ce futur n'est pas hypothétique. Gartner projette que d'ici 2026 environ quatre-vingts pour cent des grandes équipes financières d'entreprise feront tourner des plateformes d'IA générative gérées et possédées en interne, entraînées sur des données propriétaires, et le modèle du centre d'excellence est devenu la réponse organisationnelle par défaut au passage à l'échelle de l'IA dans une entreprise. Pour ces entreprises, l'intervention intensive de l'hyperscaler est un accélérateur transitoire, non un remplacement institutionnel permanent. L'ironie siège à l'intérieur du matériau même de Microsoft : son Cloud Adoption Framework instruit les clients de construire un centre d'excellence IA interne, tandis que Frontier Company propose d'être ce centre à leur place. Il plaide les deux côtés à la fois, et il ne peut savoir à l'avance lequel le marché achètera.

Les deux futurs ne sont pas mutuellement exclusifs à l'échelle du marché, et le règlement probable est un partage par maturité. L'entreprise mature, déjà en train de construire son

usine, utilisera Frontier comme accélérateur et rapatriera la compétence. L'entreprise immature, incapable de pourvoir les intendants, glissera vers la location permanente. Lequel des deux domine est précisément ce que les dix-huit prochains mois révéleront.

La fenêtre de falsification

Élevons l'ensemble d'un niveau. Pendant quarante ans, le logiciel a vendu un moyen, et le client fournissait la fin. Les hyperscalers tentent désormais de vendre la fin elle-même. Le logiciel cesse d'être vendu comme un produit, et même comme un service, et commence à être vendu comme un facteur de production cognitive, défini ici comme la capacité institutionnelle à transformer la connaissance organisationnelle en décisions économiquement valables par un raisonnement médié algorithmiquement. Si cette lecture tient, Frontier Company n'est pas le sujet de cette analyse. Il en est la première étude de cas observable, et la frontière de responsabilité, ayant bougé de la disponibilité vers le résultat, continuera de bouger.

Une grille qui ne peut être réfutée ne vaut pas la peine d'être publiée, alors voici la condition de réfutation, avec une date, et selon deux vecteurs distincts. La fenêtre est de douze à dix-huit mois.

Le premier est le vecteur d'autonomie. L'affirmation que le facteur contraignant est une capacité non livrable sera falsifiée si Frontier démontre, de manière publique et documentée, un impact durable sur le compte de résultat au sein de clients qui ont commencé immatures en données et sans intendance interne, et si les structures de responsabilité humaine, les plusieurs intendants jusqu'au Propriétaire de la décision, survivent et deviennent autonomes après le retrait des ingénieurs de Redmond. Deux observations gouvernent ce vecteur, et toutes deux sont visibles. L'une est de savoir si les clients instituent une capacité ou conservent une dépendance qui se dégrade à la sortie. L'autre est de savoir si les clients en viennent à définir leurs propres métriques, observables et traces, ou continuent d'être mesurés par l'instrument fourni par le fournisseur.

Le second est le vecteur d'institutionnalisation, et c'est la concession que les brouillons antérieurs de cet argument n'ont pas faite. La thèse sera aussi falsifiée si le marché adopte la capacité comme service sous forme durable, c'est-à-dire si un fournisseur d'infrastructure se révèle capable de gérer les résultats cognitifs et l'architecture de décision d'une entreprise depuis l'extérieur, indéfiniment, sans que le client cultive jamais de maturité interne. Dans ce cas, la distinction entre actif et capacité n'a pas été validée. Elle a été dissoute par un marché qui a choisi la dépendance permanente et s'en est satisfait.

Si aucun des deux vecteurs ne se résout en faveur de Microsoft, si les retours n'apparaissent qu'au sein d'entreprises déjà matures, et si le départ des ingénieurs est

suivi de la dégradation opérationnelle des agents, l'initiative se tiendra comme la démonstration empirique qu'un déficit de responsabilité organisationnelle ne peut être comblé en sur-approvisionnement en puissance technique face à un facteur unique d'une fonction multiplicative. La grille ne vaut sa longueur que si elle généralise, et c'est le cas. Elle explique Microsoft en juillet 2026, et elle anticipe le même mouvement chez Amazon Web Services, Google Cloud, Oracle et Salesforce, car aucun d'eux ne contrôle plus qu'un sous-ensemble des facteurs, et tous suivent la même rente vers le même lieu. La rente s'est déplacée vers le résultat, et les fournisseurs se déplacent vers le résultat derrière elle. La difficulté qu'ils rencontreront tous là-bas est la même, et ce n'est pas une difficulté d'ingénierie. Le résultat est le seul facteur qu'un fournisseur ne peut livrer, car il ne se déploie pas. Il s'institue.