

L'intégration est un livrable. La conformité est une institution.

Pourquoi l'ingénieur déployé abaisse le coût de l'intégration et peut relever le coût de l'imputation.

Deux textes antérieurs de cette série ont lu la Frontier Company de Microsoft comme un mouvement dans un déplacement plus large : d'abord comme une tentative d'occuper la couche où la valeur se prouve plutôt que celle où elle se construit, ensuite comme la migration de la rente vers le facteur contraignant, la capacité institutionnelle du client, qu'aucun fournisseur ne livre depuis l'extérieur. Une série de notes hebdomadaires le disait depuis le fauteuil de l'acheteur : la performance n'est pas la déployabilité, et le coût qui casse un déploiement est celui de son régime d'exploitation, une ligne que le directeur financier lit désormais sur le compte de résultat.

Admettons tout cela, et une question demeure, que Microsoft, Amazon et OpenAI laissent tous les trois hors-champ. Une fois l'intégration résolue, quel est le facteur contraignant, et l'instrument que les fournisseurs ont choisi pour la résoudre porte-t-il jusque-là ? L'instrument est l'ingénieur déployé, le forward deployed engineer, et la réponse est assez compacte pour tenir lieu de thèse. L'ingénieur déployé abaisse le coût de l'intégration et peut relever le coût de l'imputation. Il livre le système plus vite, et il laisse la question de savoir qui répond du système non seulement irrésolue mais alourdie.

Le mécanisme n'apparaît pas partout. Il apparaît là où trois conditions tiennent ensemble :

1. une responsabilité allouée avant le déploiement,
2. des rôles juridiques nommés à l'avance,
3. un système dont les décisions engagent des enjeux élevés.

Retirez l'une d'elles et l'ingénieur déployé peut suffire. Tenez les trois et le facteur contraignant cesse d'être technique. Énoncé complet : sous allocation ex ante de la responsabilité, une fois l'intégration résolue, le facteur contraignant devient institutionnel, et la famille des stratégies FDE, superbe à construire, n'est ni l'autorité qui institue ni la partie qui répond. Le régime européen du règlement (UE) 2024/1689 en est le laboratoire le plus net, non parce qu'il serait le seul lieu où le mécanisme opère, mais parce qu'il est le lieu où le mécanisme est écrit. La déployabilité, dont le sens s'est élargi au fil de cette série de la preuve à la responsabilité puis à l'institution, est fixée ici pour le

corpus : la capacité d'un système à rester exploitable sous contraintes techniques, économiques et institutionnelles pendant toute sa durée de vie.

Trois métiers que le marché confond

Le tournant du déploiement repose sur une observation juste et une inférence hâtive. L'observation : le modèle n'est plus le problème, le déploiement l'est. L'inférence : donc le déploiement, c'est l'intégration. Non. Le déploiement en régime régulé, ce sont trois métiers, et le marché n'en price qu'un. Construire, c'est produire le système qui fonctionne, les intégrations, le harnais, les évaluations, l'observabilité. Instituer, c'est établir le système de règles sous lequel le système peut opérer : le dossier de conformité, le système de gestion des risques, la conception de la supervision humaine, l'allocation de qui répond de quoi. Assumer, c'est porter la responsabilité des décisions que le système prend.

Le troisième métier se scinde, et cette scission est le geste porteur de l'argument. Pourquoi deux moitiés, et non une ou trois ? Parce que la ligne qui compte est économique, la ligne de la transférabilité. Attester produit un objet transférable ; répondre produit une exposition non transférable. Tout ce qui se remet à un organisme notifié, à un assureur ou à un cabinet tombe d'un côté ; tout ce qui reste à une personne une fois ces parties à l'œuvre tombe de l'autre. La coupure n'est pas morale, elle est contractuelle : ce qui se vend, et ce qui ne se vend pas. Attester, donc, c'est produire l'instrument de conformité, la déclaration, le marquage, le dossier technique, l'acceptation formelle du risque. C'est un livrable, délégable, assurable, et codable en partie, en policy as code et evidence as code. Parce qu'il est transférable il fait marché, et parce qu'il fait marché il porte rente. Répondre est l'autre moitié, et il faut l'énoncer en son sens technique et non comme une fioriture morale. Répondre, c'est demeurer le responsable exposé, au civil et au pénal, une fois que l'attestation, l'assurance et la documentation ont fait tout ce qu'elles pouvaient faire. Cela ne s'assure pas, ne se délègue pas, ne se code pas. Cela ne fait pas marché, donc ne porte pas de rente, seulement une exposition.

D'où la définition stricte qui garde la thèse falsifiable. Une institution, ici, est le système de règles, formelles, c'est-à-dire le droit et les contrats, ou informelles, c'est-à-dire la gouvernance d'entreprise, qui alloue la responsabilité d'un acte de décision. Une contrainte est institutionnelle lorsque sa levée exige une autorité, un tiers ou une gouvernance formalisée, plutôt qu'une métrique de performance comme un taux de couverture ou un score de calibration. La qualité des données n'est pas, en soi, une contrainte institutionnelle ; elle se lève par de meilleures données. L'obligation d'attester la représentativité de ces données, que l'article 10 du règlement fait peser sur le fournisseur d'un système à haut risque, l'est, parce qu'aucune métrique ne la lève, seulement une signature devant une autorité. Cette différence est toute la différence entre une thèse et une platitude.

Trois oppositions nettes suivent, et l'article tourne autour d'elles. Un livrable technique contre une institution de responsabilité, ce qu'on produit contre ce qu'on engage. L'économie de l'acquisition contre l'économie de l'exploitation continue, le coût du pilote que l'ingénieur livre contre le coût de la surveillance, du recalibrage et de la conservation des preuves que le régime exige pour la vie du système. Et l'ingénieur-diplomate contre l'autorité qui institue, celui qui aligne les équipes contre celui qui signe le risque. Il serait injuste, donc faible, de dire que l'ingénieur déployé n'institue pas. Il peut documenter une institution, en accélérer les artefacts, en révéler les manques. Ce qu'il ne peut pas, c'est constituer l'autorité dont l'institution tire sa force.

La famille des stratégies FDE, et Microsoft en acteur rationnel

L'ingénieur déployé n'est pas un homme de paille. C'est un objet daté, financé, documenté. En mai 2026, à soixante-douze heures d'intervalle, Anthropic a annoncé une coentreprise de services avec Blackstone, Hellman & Friedman et Goldman Sachs, et OpenAI a annoncé une société de déploiement et le rachat du cabinet Tomoro. Les chiffres rapportés à l'époque plaçaient la première près d'un milliard et demi de dollars et la seconde près de quatre milliards, Tomoro apportant de l'ordre de cent cinquante ingénieurs de déploiement. Le playbook est celui de Palantir, et il n'est pas discuté : embarquer un ingénieur senior chez le client pour une période fixe de quelques semaines à quelques mois, livrer du code de production tôt, refuser le rôle d'intégrateur de systèmes. Le credo est explicite, et juste aussi loin qu'il porte : le modèle n'est pas le goulot, le déploiement l'est, lecture que les fournisseurs adossent au rapport MIT NANDA State of AI in Business 2025, qui a trouvé que quatre-vingt-quinze pour cent des pilotes génératifs en entreprise ne montrent aucun impact mesurable sur le compte de résultat.

Quatre positions ordonnent tout le champ, et il est utile de les nommer avant que les acteurs ne s'y pressent. Un constructeur fait le modèle. Un intégrateur l'embarque. Un attestateur le certifie. Un répondant se tient derrière ses décisions. Les fournisseurs occupent les deux premières et tendent vers la troisième ; la quatrième est, par construction, celle du client. Chaque firme ci-dessous prend sa place sur cette ligne, et la famille qui mène l'embed n'est pas homogène. Palantir accepte l'implication opérationnelle la plus profonde, Microsoft reste plus près de l'accompagnement et distribue l'infrastructure d'OpenAI, Anthropic vend une accélération d'intégration, OpenAI construit encore son modèle, Amazon fournit le contenant, Bedrock. On parle d'une famille de stratégies. Et l'ordre que la critique va tracer n'est pas strictement séquentiel : l'intégration et la conformité coévoluent, et dans un programme réel, plus l'intégration avance, plus les exigences de conformité affleurent. Les deux montent ensemble, ce qui est précisément pourquoi la partie qui livre est tentée de traiter la conformité comme une corvée d'aval plutôt que comme un métier distinct.

Microsoft est mobilisé dans cette série comme un symptôme, mais c'est aussi une firme qui fait un choix, et le fait instructif est là où le choix s'arrête. La Frontier Company construira et intégrera à l'échelle. Elle n'attestera pas, et elle ne répondra pas. La raison est structurelle, non psychologique, et il n'est pas besoin de prêter des intentions à un organigramme pour la lire. Attester un système que l'on fournit aussi ferait tomber l'indépendance que l'évaluation de conformité exige. En répondre mettrait le bilan d'un fournisseur derrière la décision d'un client, et c'est l'exposition non transférable que l'économie d'un fournisseur ne peut absorber. Microsoft s'arrête exactement là où finit son modèle économique, ce qui se trouve être exactement là où commence l'institution.

Le FDE n'institue pas, et il peut requalifier

Prenons une scène composite mais banale, un comité de mise en marché dans une banque de détail européenne, mi-2026. L'équipe déployée d'un fournisseur a intégré un modèle de complétion de graphes pour le scoring de crédit des petites entreprises, qui automatise l'essentiel de l'instruction. Le système est prêt. Les tests de charge passent. Le connecteur livre le code en production. Puis le programme cale, et sur rien de technique. La directrice de la conformité refuse d'apposer son nom sur l'analyse d'impact sur les droits fondamentaux que l'article 27 du règlement exige d'un déployeur d'un système de scoring de crédit, et que la gouvernance de risque de la banque réclame déjà avant l'échéance. Sa raison est exacte. Les ingénieurs embarqués, ayant ajusté les poids locaux du modèle aux spécificités sectorielles de la banque, ont fait précisément ce pour quoi le fournisseur était payé, et ce faisant ont poussé la banque vers la modification substantielle que l'article 25 traite comme faisant d'un déployeur un fournisseur. La banque n'utilise plus simplement un outil externe. Elle est en passe de devenir le fournisseur du système, exposée au régime le plus lourd du texte, pour une décision, le refus de crédit, qui porte un risque de discrimination systémique. L'ingénieur-diplomate, dont le mandat s'arrête au livrable logiciel, plie bagage. Il n'a ni la qualité contractuelle pour engager son employeur sur la conformité de l'architecture finale, ni le mandat pour co-signer ce risque. L'intégration est achevée. Le déploiement est paralysé.

L'argument ne dépend pas de la banque. Remplacez le modèle de scoring de crédit par un outil d'aide à la décision clinique, un contrôleur industriel autonome ou un moteur d'éligibilité du secteur public, et la structure est inchangée : le travail technique s'achève, et le programme s'arrête à la signature. Ce que la scène montre est un paradoxe. Le modèle déployé est fait pour modifier substantiellement, il personnalise, il affine, il tisse le système dans l'infrastructure du client, et cette profondeur est la proposition de valeur. Elle est aussi le déclencheur. L'instrument envoyé pour réduire la friction du déploiement est, par le même geste technique, l'instrument qui qualifie le client pour le régime qu'il était le moins préparé à porter. Les fournisseurs voient venir la friction et rédigent contre elle : leurs conditions stipulent de plus en plus que le fine-tuning, l'alignement ou

l'adaptation par récupération augmentée restent dans l'usage normal et ne transfèrent aucune qualification. Cela les aide moins qu'ils ne l'espèrent, parce que la requalification au titre de l'article 25 est affaire de droit et de fait, de ce qui a réellement été fait au système, non de la clause par laquelle un fournisseur s'en défausse. La modification substantielle n'est pas encore stabilisée en lignes directrices ni en jurisprudence, si bien que l'embed peut conduire à la requalification, il ne la produit pas mécaniquement. Le point n'est pas une certitude. C'est un risque non pricé que le mouvement de mise sur le marché fabrique, et qu'aucun contrat ne purge entièrement.

La seconde limite répond à l'objection selon laquelle la conformité sera simplement automatisée. Continuous compliance, policy as code, evidence as code, évaluation de conformité automatisée : le marché industrialise déjà les couches probatoire, documentaire et procédurale de la conformité, et il y réussira pour beaucoup. Ce que l'automatisation atteint, c'est l'attestation, l'instrument transférable. Ce qu'elle n'atteint pas, c'est le fait de répondre. Ici les ironies de l'automatisation de Bainbridge, écrites pour les salles de contrôle en 1983, se transposent exactement, à condition de les appliquer au bon résidu. Plus on automatise construire, intégrer et même attester, plus l'acte restant devient rare, critique et concentré, et l'acte restant n'est pas une signature sur un formulaire, qui se délègue et s'assure. C'est répondre de la décision, qui ne se délègue pas. La lecture compacte : on code la preuve, on transfère l'attestation, on n'automatise pas le fait de répondre.

Le report comme preuve, et l'AI Act comme révélateur d'allocation

Le calendrier réglementaire a bougé pendant l'écriture de cette série, et le mouvement est la preuve la plus forte de la thèse, non contre elle. La précision importe ici, parce que le passage invite l'objection d'un juriste. Au titre du Digital Omnibus sur l'IA, sur lequel le Conseil et le Parlement ont trouvé un accord politique le 7 mai 2026 (Conseil de l'UE, 7 mai 2026), approuvé par les deux institutions en juin et en attente de publication au Journal officiel à l'heure où ceci est écrit, les obligations haut risque pour les systèmes autonomes de l'annexe III, recrutement, scoring de crédit, éducation, contrôle aux frontières, sont reportées du 2 août 2026 au 2 décembre 2027, et celles des systèmes de l'annexe I intégrés à des produits réglementés, dispositifs médicaux, machines, véhicules, au 2 août 2028. Les obligations de transparence de l'article 50, y compris celles qui pèsent sur les déployeurs, sont maintenues au 2 août 2026.

Le report a été accordé parce que la machinerie qui rend les obligations opérables, les normes harmonisées de CEN et CENELEC, les organismes notifiés, les autorités nationales compétentes, n'était pas prête. Le facteur contraignant du déploiement à l'échelle n'est donc pas le modèle, qui a été livré, ni l'intégration, que les ingénieurs

livrent. C'est l'infrastructure de l'institution, et le régulateur concède qu'elle n'existe pas encore en concédant le calendrier. En mai 2026, l'industrie engage des milliards dans l'ingénierie embarquée au moment où le régulateur recule faute d'une institution à faire respecter. L'infrastructure technique court plus vite que l'infrastructure de confiance. Et c'est le point où l'article doit être lu contre son propre cadre européen : l'AI Act ne crée pas le mécanisme, il le révèle. L'Acte est un microscope, non le phénomène. Il n'invente pas la responsabilité, puisque les fabricants construisaient des dispositifs médicaux dotés d'IA sous le règlement sur les dispositifs médicaux et en répondaient au titre de la responsabilité du fait des produits bien avant qu'il n'existe. Ce qu'il ajoute, c'est la lisibilité. Il nomme le déployeur comme acteur juridique autonome, à l'article 26, et rend l'allocation de la responsabilité visible. La thèse repose sur cette visibilité, non sur une quelconque inflation d'obligations. Un garde-fou en découle : un calendrier reporté n'est pas une contrainte molle. L'Omnibus déplace la date de l'attestation ; il ne déplace pas le fait de répondre. Le droit sectoriel mord en 2026 quoi qu'il arrive, et l'exposition personnelle ne se reporte pas avec l'annexe.

Où va la rente

C'est là que la série se referme sur elle-même. La frontière de responsabilité nommée dans le deuxième texte a maintenant un nom juridique, la frontière fournisseur/déployeur. Le Decision Owner qui y était nommé, le rôle responsable presque jamais formalisé, est celui qui répond, le titulaire de la part non transférable. L'insistance, dans le cadre RAISE, sur le fait qu'industrialiser l'IA régulée est affaire d'institution et non seulement d'intégration, est la même thèse vue du côté de la méthode. Le test de l'acheteur, issu des notes hebdomadaires, s'étend proprement : pour chaque système en production, qui tient la ligne d'exploitation, qui tient la ligne d'incident, et maintenant la troisième, qui tient la ligne d'institution, l'attestation et le fait de répondre. Dans un dispositif médical marqué CE, la distinction est déjà visible, le fabricant construit, un organisme notifié atteste, et une personne responsable désignée répond, trois parties pour trois métiers, la médiane achetable et la dernière non.

Vient la conséquence, et il lui faut une couche de plus que les notes antérieures ne lui donnaient. La rente ne migre pas vers l'attestation. La rente migre vers le pouvoir, et le pouvoir pertinent est l'autorité d'autoriser l'exploitation. Cette autorité est l'actif rare. L'attestation n'en est que la face marchande, l'instrument par lequel l'autorité s'exerce et se prouve, ce qui est pourquoi l'argent semble affluer vers l'attestation alors qu'il afflue, en vérité, vers celui qui peut dire oui. C'est un terrain ancien de l'économie des organisations. Les institutions, au sens de North, sont les règles du jeu, et les règles ici confèrent un droit d'autoriser ; la valeur revient au titulaire d'un droit conféré par la règle, non au technicien qui prépare le dossier. La frontière que le fournisseur ne peut franchir est celle de Williamson : l'attestation est contractible et donc un marché, répondre est le

résidu qu'aucun contrat ne complète et qui reste donc à l'intérieur de la partie qui le porte. Ainsi la rente se dépose sur celui qui détient le pouvoir d'autoriser l'exploitation, un organisme notifié dans un secteur, l'un des grands cabinets d'audit dans un autre, un assureur, une fonction de gouvernance interne. Les fournisseurs ne resteront pas passifs devant elle. Ils la viseront par des services d'assurance IA, des offres de déploiement certifié, des garanties contractuelles, de l'exploitation managée, et ils en prendront la part transférable, à côté de leur bilan, par un véhicule d'assurance séparé qui monétise l'attestation sans leur coûter leur neutralité, précisément parce qu'il est séparé. Ce qu'aucun véhicule et aucun contrat ne captent, c'est le fait de répondre. L'attestation s'achète ; le fait de répondre ne se délègue pas. Le CTO formule l'intégration, le CFO budgète l'exploitation, un tiers est payé pour attester, mais le Decision Owner seul répond, et le fournisseur n'a envoyé personne pour tenir ce stylo.

L'asymétrie transatlantique, et l'horizon

La borne de la thèse est le moteur qui l'explique, et il ne faut pas la lire comme une ligne trop nette. Le modèle déployé est né aux États-Unis, et là la responsabilité se produit après coup et de multiples sources, jurisprudence, class actions, assureurs, contrats, exigences des marchés, arbitrée au niveau de l'usage opérationnel plutôt que de l'attestation ex ante. Dans ce régime, l'ingénieur embarqué est plus proche d'une solution complète, parce que l'institution se forme plus tard, autour de l'usage, de sorte qu'il peut porter un déploiement jusqu'au point où les tribunaux et les assureurs prennent le relais. Mais l'asymétrie est de mécanisme et de degré, non de présence. Le régime américain produit son propre coût d'imputation, en amont et informellement : la crainte d'une class action ou d'un dommage réputationnel arrête déjà bien des comités américains, sans nulle requalification en vue. La même paralysie du fait de répondre commence à y infuser ; le régime européen ne fait que la nommer à l'avance, et met une porte là où l'américain laisse un champ ouvert. L'Europe n'est pas le seul lieu où le mécanisme opère. C'est le lieu où il est lisible.

La thèse est falsifiable deux fois. Elle échoue en principe le jour où répondre devient transférable, si un régime accorde à un dirigeant une immunité personnelle en échange de l'attestation d'un tiers agréé. Elle échoue en fait si l'on montre que les firmes dotées d'équipes déployées embarquées déploient l'IA régulée à fort impact systématiquement plus vite, sans hausse du coût institutionnel ni réallocation de la responsabilité. Le premier test est élégant et improbable, le second banal et mesurable, et c'est le second qu'il faut surveiller. Deux limites restent honnêtes à ses côtés. Le facteur institutionnel n'est pas toujours celui qui contraint en premier, la gouvernance propre d'une banque, les obligations de données d'un hôpital, les règles d'achat d'un acheteur public peuvent chacune contraindre avant le droit, mais chacune est un système de règles allouant la responsabilité, de sorte que chacune confirme le niveau de la thèse tout en en déplaçant

le contenu. Et ce que les thèses publiques omettent n'est pas la preuve que les contrats privés l'omettent aussi ; l'allocation peut déjà être négociée là où elle ne se voit pas.

Le corpus pointe au-delà de cet article, et au-delà de l'intelligence artificielle. Ce qui vaut pour un modèle de crédit vaut pour un dispositif médical, un contrôleur de réseau électrique, un système avionique, un modèle de fondation : la couche technique s'industrialise, vite et à grands capitaux, puis le travail s'arrête là où commence l'autorité. C'est la proposition générale que l'AI Act ne fait qu'illustrer. Toute révolution technique finit par s'arrêter à la même frontière, celle où une règle doit conférer un droit et où une personne doit répondre d'une décision. Le logiciel construit. Les règles instituent. Une personne répond.