

Le token n'est pas l'unité

Le Coût par Décision Utile, ou pourquoi la baisse du coût d'inférence déplace le coût vers la preuve et le risque. Note doctrinale 1/4.

1. La thèse avant l'équation

Les fournisseurs facturent l'inférence au million de tokens. Une lecture réductrice du FinOps appliqué à l'IA, aujourd'hui dominante, reprend cette unité commerciale et en fait le proxy principal de l'efficacité économique. Le problème n'est pas que le token soit mesuré. Le problème est qu'il serve de mesure de l'efficacité d'un système dont il n'est que l'intrant appelé à devenir plus substituable et moins différenciant que les capacités institutionnelles qui entourent la décision.

La thèse de cette note tient en une distinction, sous une condition. Lorsqu'un système influence ou exécute un arbitrage engageant l'institution, il se facture parfois au token, mais il se *pilote* à la décision. Hors de ce domaine, le token reste une unité légitime et la note ne prétend rien. Trois unités se confondent dans le discours ambiant et doivent être séparées : l'unité de *facturation*, ressource achetée au fournisseur, l'unité de *gestion*, ce que le pilote surveille, et l'unité *économique*, ce dont l'efficacité se juge. Dans le domaine engageant, l'unité économique n'est pas l'inférence, c'est la décision.

Le corollaire est le véritable objet de la note. Le coût d'une décision se ramène à trois charges, que la suite réunira sous les lettres O, K et H :

- Le coût d'exploitation O, qui additionne le compute (noté O_tech) et la supervision humaine (notée O_human),
- Le coût de gouvernance et de preuve K,
- La charge de risque H.

Quand le coût du compute tend vers zéro, le coût de la décision ne disparaît pas : il change de composition et, souvent, de porteur. La charge se recompose vers la supervision qui subsiste, la preuve qui reste à produire et le risque qui reste à absorber, soit vers O_human, K et H.

Cette proposition doit être falsifiable, et sa condition de réfutation mérite d'être posée avec soin, car une formulation paresseuse la rendrait soit triviale, soit intenable. Deux énoncés se distinguent.

- Le premier est analytique : à structure de coût par ailleurs figée, réduire la seule composante technique O_tech augmente mécaniquement la part des charges

non techniques dans le total. C'est une identité, non une prédiction, et elle ne prouve rien.

- Le second est empirique, et c'est lui la thèse : dans les déploiements réels, la baisse de O_{tech} n'entraîne pas une baisse proportionnelle de la charge non technique, c'est-à-dire de la supervision résiduelle, de la preuve et du risque, réunies dans l'agrégat $O_{human} + K + H$.

Ces charges sont gouvernées par la réglementation et l'exposition, non par le prix du GPU. Leur élasticité au coût du compute est faible. La réfutation est claire : si, à périmètre, volume, qualité et régime de responsabilité constants, le niveau absolu de $O_{human} + K + H$ par décision reculait au rythme de O_{tech} , la thèse de migration serait fausse. La note ne prétend pas établir empiriquement cette faible élasticité. Elle en propose le cadre de mesure et la condition de test.

Un avertissement en découle. Une baisse du coût total par décision ne réfute pas la thèse, car ce total peut reculer pour des raisons étrangères au compute : mutualisation du socle de preuve, meilleure architecture réduisant l'exposition, hausse du volume, transfert contractuel du risque. La thèse porte sur la composition et sur le résidu, non sur le niveau.

Le domaine de validité est étroit, et je le pose avant d'aller plus loin. La frontière n'est pas la forme technique de la tâche mais son effet : le cadre commence lorsque la sortie du système modifie matériellement un droit, une obligation, une exposition ou un arbitrage porté par l'institution. Un traitement par lot ou un reformatage sans conséquence en reste dehors. Une simple classification y entre, en revanche, si elle refuse un accès, priorise une enquête, exclut un dossier, oriente un diagnostic ou augmente une prime.

2. La décision comme unité minimale d'engagement

Une entreprise ne recourt pas à l'IA pour accumuler des jetons générés. Elle y recourt pour produire un résultat. Encore faut-il distinguer deux objets que le discours ambiant confond : *l'unité minimale d'engagement* et *l'unité ultime de création de valeur*.

La décision est la première, pas la seconde. Elle est le premier état institutionnel auquel l'organisation peut rattacher un engagement, une responsabilité et un coût. Une recommandation algorithmique déterminante y suffit, même sans forme juridique de décision. La valeur économique finale, elle, dépend de l'exécution ultérieure du processus métier, sur laquelle le système d'IA n'a plus la main. Confondre les deux conduit à deux erreurs symétriques : créditer l'IA de toute la valeur aval, ou lui dénier tout coût propre en amont. Le Coût par Décision Utile se loge exactement entre les deux, sur le segment de l'engagement.

La chaîne se lit ainsi. Le compute est l'intrant brut, l'inférence la ressource système, la décision l'engagement, l'action l'exécution, le résultat l'impact, la valeur l'inscription au

bilan. Le périmètre du CDU s'arrête à la décision, premier point où la délégation produit un engagement institutionnel mesurable. L'institution continue de contrôler l'exécution, la remédiation, le recours et la réversibilité, mais ces contrôles agissent en aval de l'engagement déjà pris.

Le cadre couvre l'ensemble des régimes de participation décisionnelle, et cette généralité est volontaire. Que le système exécute l'arbitrage en autonomie ou qu'il formule une recommandation déterminante validée par un humain, il engage l'institution. Lorsque la validation humaine est routinière ou faiblement contradictoire, elle ne neutralise pas l'influence économique du système ; elle en déplace surtout la signature formelle. Une validation autonome, informée et dotée du temps de contredire est un autre cas, et le CDU les sépare au lieu de les confondre.

3. La séparation, et ce qui la distingue de Shannon

Le geste fondateur du CDU est une séparation. De même que la théorie de l'information a séparé la quantité d'information du support physique qui la transporte (Shannon, *A Mathematical Theory of Communication*, 1948), l'économie de l'IA doit séparer la décision produite de la ressource de calcul qui la rend possible. Le coût du support n'était pas la mesure de la quantité d'information transportée ; **le compte de tokens n'est pas davantage l'unité économique de la décision.**

L'analogie est utile. Elle est aussi dissymétrique : La quantité d'information de Shannon pouvait être formalisée indépendamment de son support, dans une unité, le bit, sans convention institutionnelle. Le CDU n'a pas cet appui. La décision utile n'est pas une grandeur naturelle : elle se définit par une conjonction de seuils de qualité non compensatoires, une fenêtre de maturation et un budget de remédiation, tous fixés par la gouvernance.

Là où Shannon disposait d'une unité formelle, le CDU repose sur une convention institutionnelle gouvernée. Sa robustesse ne dérive donc pas d'une unité naturelle, mais de la stabilité de ses définitions, de ses seuils et de ses règles d'imputation. Ce n'est pas un aveu de faiblesse glissé par prudence, c'est le sujet même de la note, et tout ce qui suit en tire les conséquences.

4. L'équation, unique, et sa limite

Voici le seul bloc formel que le corps de cette note portera. L'équation de référence et sa propriété limite forment un même geste ; tout l'appareil actuariel et comptable descend en annexes, où il sert la rigueur sans encombrer le raisonnement.

$$CDU_i(t) = \frac{O_i(t) + K_i(t) + H_i(t)}{D^U_i(t, \Delta_i)}$$

$$\lim_{O_{tech} \rightarrow 0} CDU_i(t) = \frac{O_{human,i}(t) + K_i(t) + H_i(t)}{D^U_i(t, \Delta_i)}$$

Le numérateur regroupe l'ensemble des coûts dans trois agrégats exclusifs :

- O est le coût d'exploitation du régime de délégation, compute (O_{tech}) et humains de supervision (O_{human}) confondus,
- K est le coût de la capacité de gouvernance et de preuve, l'architecture qui maintient l'attribuabilité, l'intégrité et la révisabilité de ce qui a été décidé,
- H est la charge économique du risque, traduction financière de l'exposition.

Un coût d'option, de dépendance fournisseur, de réversibilité, de sortie, juridique ou de transition n'échappe pas au cadre : il s'impute dans l'un des trois agrégats selon sa nature, sous une convention d'allocation stable.

Le dénominateur D^U compte les décisions *utiles* : institutionnellement closes, conformes aux seuils de qualité, n'ayant pas dépassé le budget de remédiation prévu pour la classe, et encore conformes au terme d'une fenêtre d'observation Δ . À titre illustratif, une institution pourrait retenir trente jours pour certaines décisions de crédit et quatre-vingt-dix pour certaines indemnisations, le choix de Δ dépendant de la temporalité des recours, du délai de matérialisation des pertes et du cycle métier. Une décision invalidée pendant la fenêtre retenue ne peut être comptabilisée comme utile dans la cohorte considérée.

Trois précautions avant tout calcul. D'abord, le numérateur mêle des statuts de coût distincts : des coûts *observés*, comme le compute facturé, des coûts *imputés*, comme la part du socle mutualisé affectée à la classe, des pertes *estimées*, comme la perte attendue, et des charges de *capital*. Confondre ces statuts, c'est prendre une construction de gestion pour une mesure entièrement observée. Ensuite, tous les termes se calculent sur la même période, le même périmètre et la même cohorte, close à la même fenêtre de maturation ; un numérateur et un dénominateur qui ne partagent pas leur horizon ne mesurent rien. Enfin, la qualité minimale est une *conjonction de seuils*

non compensatoires : un système ne rachète pas une traçabilité défaillante par une meilleure latence, chaque seuil doit passer, aucun ne compense un autre.

Reste la propriété qui justifie la note. Sous les hypothèses posées, à périmètre, volume, qualité, architecture de supervision et distribution de risque constants, les autres termes sont maintenus constants afin d'isoler l'effet de la baisse de O_{tech} , qui quitte alors le numérateur. C'est une statique comparative, non une loi empirique : en pratique, la baisse du compute peut induire plus d'usage, plus de volume, de nouveaux cas, et donc déplacer ces termes. Sous cette réserve, l'IA en production ne devient pas gratuite à mesure que l'inférence s'effondre. Ce qui persiste n'est pas un montant constant, le total peut monter ou descendre, mais une charge résiduelle non technique qui devient prépondérante. La rareté ne disparaît pas. Elle déménage vers d'autres poches, et parfois vers d'autres payeurs.

Encadré. Un scénario de convergence vers un même niveau résiduel

Chiffrage illustratif, en unités normalisées, « indicielles » et non empiriques. Il vérifie une mécanique, il ne prouve pas une propriété générale. Le volume est maintenu constant pour isoler l'effet de composition, et les charges de supervision, de preuve et de risque sont supposées collantes, hypothèse qui est précisément la thèse à éprouver sur le terrain.

Soit une classe de décisions d'octroi de crédit, deux régimes de délégation, coûts rapportés à une décision utile.

Architecture *assistée*, l'IA recommande et l'humain décide : compute 1, supervision humaine 5, preuve 2, risque 2. Coût par décision 10.

Architecture *automatisée*, l'IA décide et l'humain traite l'exception : compute 2, supervision résiduelle 2, preuve 4, risque 3. Coût par décision 11.

Premier constat, contre-intuitif : dans cet exemple, avant toute commoditisation, l'automatisation coûte plus cher à l'unité, pas moins, alors même que le coût de compute est relativement réduit.

Faisons maintenant tendre le compute vers la gratuité. L'assistée passe de 10 à 9, l'automatisée de 11 à 9. Dans ce scénario, les deux régimes convergent vers un même niveau résiduel.

Note : Cette égalité est une hypothèse de paramétrage, non une conséquence générale du CDU. Elle tient parce que les chiffres ont été posés ainsi. Et ce niveau n'est un plancher que sous les hypothèses du scénario, à volume, supervision, preuve et risque constants.

Second constat : la charge non technique, $O_{human} + K + H$, absorbe une part croissante du coût à mesure que le compute recule, et l'essentiel de ce résidu, dans le cas automatisé, tient à la preuve et au risque.

Ce que l'encadré vérifie : la mécanique du modèle, et le fait que la commoditisation du compute, seule, ne franchit pas le résidu fixé par les charges non techniques. Ce qu'il ne prouve pas : la hauteur de ce résidu et l'écart entre architectures, qui dépendent de la répartition posée et devront être établis sur données réelles. Ce qu'il suggère : à résidu égal, l'automatisation loge l'essentiel de son coût dans la preuve et le risque, poches qu'aucun progrès sur l'inférence n'optimise. Si on l'adopte, ce n'est pas pour une économie unitaire, nulle ici, mais pour une raison qui ne se lit pas dans le CDU.

5. Qui fixe le dénominateur

L'objection la plus sérieuse contre le CDU est la suivante.

Le CDU, dira-t-elle, n'est pas une mesure mais un instrument de politique déguisé en métrique. Son dénominateur dépend de seuils que la gouvernance fixe elle-même : qui contrôle la qualité minimale, le budget de remédiation et la fenêtre de maturation contrôle le compte des décisions utiles, donc le résultat. Un dirigeant qui croit acheter le vrai coût par décision achète un nombre négocié.

L'objection touche juste, mérite d'être retenue, mais elle prouve moins qu'elle ne le croit. Le CDU reste une mesure ; simplement, ce n'est pas une mesure naturelle, c'est une mesure *conventionnelle et gouvernée*, dont la qualité dépend de la gouvernance des seuils, des périmètres et des imputations. Cela ne le disqualifie pas. L'EBITDA, le RAROC, le coût complet et les allocations de capital forment une même famille de métriques conventionnelles, sans en partager le statut : l'EBITDA est un indicateur financier peu normalisé, le RAROC une construction interne de risque, le coût complet une convention d'allocation. La comparaison vaut comme famille, non comme équivalence, et l'on pilote pourtant des institutions avec toutes. La conventionnalité n'est pas le vice caché du CDU, c'est le régime ordinaire de toute comptabilité de gestion. Ce que le CDU ajoute, c'est qu'il rend l'arbitrage sur les seuils explicite et chiffré, là où le coût au token l'enterre sous une apparence de neutralité technique.

Reste que la partition du numérateur n'est jamais uniquement analytique. Le point n'est pas de savoir qui paie d'ordinaire, mais que les coûts se distribuent entre plusieurs fonctions dont les incitations diffèrent. Comme archétype, et non comme vérité comptable, *O* est surtout visible dans les budgets IT ou plateforme, *K* dans les fonctions de gouvernance et de risque, *H* dans la ligne métier ou le bilan central ; la réalité brouille ce partage, le métier finançant une part de *O*, l'IT portant certains coûts de preuve, le capital et l'assurance étant souvent centralisés. Il reste qu'un coût par décision qui traverse plusieurs fonctions sera arbitré à ses coutures, et que le pouvoir de fixer les seuils déterminant le dénominateur est un pouvoir réel. Le CDU mesure. Il n'aligne pas les incitations. Le token se commoditise, la décision se gouverne, et la gouverner suppose de nommer qui paie quoi et qui définit quoi.

6. Le siège-kilomètre, sous réserve d'homogénéité

L'aviation commerciale offre le meilleur précédent d'unité de production, à condition de l'importer avec sa limite. Une compagnie ne mesure pas sa performance au litre de kérosène brûlé, mais au coût du siège-kilomètre disponible, calculé sous contrainte stricte de sécurité et de maintien continu de l'aptitude à l'usage. Le CDU est l'équivalent du siège-kilomètre pour les systèmes de décision : une unité de production tarifée sous contrainte de qualité, non un intrant brut.

La limite se pose avant l'analogie, pas après. Le siège-kilomètre est bien plus standardisable que la décision institutionnelle, non parfaitement homogène : une place d'affaires en long-courrier et une place économique en court-courrier ne se valent pas, et les compagnies en produisent des variantes. Mais il tolère une agrégation que la décision ne tolère pas sans segmentation stricte. L'analogie vaut pour la logique d'unité de production, elle ne vaut pas pour un degré d'homogénéité que le siège lui-même ne possède qu'imparfaitement. Le CDU exige donc, avant tout calcul, une segmentation en classes de risque comparables. Un CDU agrégé sans cette segmentation est un chiffre sans référent, et la même exigence commande la lecture de tout portefeuille (Annexe C).

Cette segmentation fonde le seul usage vraiment opérationnel du cadre : comparer des régimes de délégation à qualité, réversibilité et profil de risque équivalents. Qu'une architecture automatisée gagne ou perde cette comparaison unitaire ne clôt pas la décision d'architecture, qui peut relever d'une dominance capacitaire ou stratégique étrangère au coût unitaire ; ce diagnostic, comme le suivi de l'écart entre CDU prévu et CDU constaté, est traité en Annexe D.

Un mot enfin sur ce que le CDU n'est pas. Il sert à comparer des architectures, arbitrer un mode de délégation, éprouver un business case et détecter une migration de coûts. Il ne mesure pas la valeur de l'IA, seulement le coût de l'engagement décisionnel. Le confondre avec une mesure de valeur reproduirait, un cran plus haut, l'erreur d'unité qu'il corrige.

7. Ce dont dépend la stabilité de l'équation

La stabilité économique du CDU repose sur une condition essentielle que l'équation ne montre pas : la gouvernabilité de la chaîne de représentation qui produit les décisions, modèles, embeddings, garde-fous, API et dépendances d'orchestration.

Qu'un fournisseur modifie unilatéralement son modèle amont, sa politique de modération ou l'espace latent de ses embeddings, et trois choses bougent ensemble. Les preuves déjà enregistrées perdent une part de leur portée, parce qu'elles attestent d'un système qui n'existe plus tout à fait. Le coût de requalification et de non-régression augmente. La distribution des erreurs se déforme, et le calibrage de la charge de risque

avec elle. Une variable exogène, insuffisamment encadrée contractuellement, fait alors bouger deux termes du numérateur à la fois.

La leçon n'est pas qu'il faille posséder les modèles et les embeddings. Une institution n'a pas besoin de propriété ; elle a besoin de droits suffisants pour identifier les versions, être notifiée des changements, reproduire les décisions, reconstruire les preuves et migrer les usages. La notion clé est la *gouvernabilité contractuelle et technique*, non la propriété. Un bon contrat organise précisément ce qu'une chaîne exogène menace : notification, versionnement, conservation, réversibilité, stabilité minimale. Un coût par décision stable suppose une chaîne de représentation gouvernable, et cette condition n'est pas économique mais architecturale et contractuelle.

8. Conclusion

Cinq acquis referment cette première note. Le token mesure une ressource, légitime pour facturer et surveiller, illégitime pour juger l'efficacité. La décision utile est l'unité de pilotage, parce qu'elle est le premier état qui engage le bilan. Le CDU reste une mesure conventionnelle et gouvernée, non une objectivité naturelle. La commoditisation du compute ne réduit pas le coût de la décision à mesure qu'elle réduit celui de l'inférence : elle en recompose la charge vers la supervision, la preuve et le risque. Et la gouvernabilité des représentations conditionne la stabilité même du calcul.

Le CDU ne promet donc pas un coût vrai débarrassé du jugement. Il impose au contraire de rendre visibles les conventions, les responsabilités et les charges que le prix du token laisse hors champ. Son intérêt n'est pas d'abolir l'arbitrage, mais de le replacer au niveau où l'institution engage réellement son bilan : la décision. Le cadre est convaincant ; la métrique, elle, reste à éprouver, et cette note en propose l'instrument, non la preuve.

Reste la condition nommée sans être traitée, la gouvernabilité des représentations qui soutiennent le calcul. C'est l'objet de la note suivante. La note 2/4, *L'Asymétrie Souveraine*, traite de ce que coûte une décision dont on ne gouverne pas les représentations.

ANNEXES

Annexe A. Décomposition du socle de preuve (K) et frontière avec H

Le socle de gouvernance et de preuve se décompose en cinq postes soumis à une convention d'imputation exclusive : $K = K_g + K_p + K_m + K_a + K_\Delta$. K_g couvre la gouvernance (cadrage ex-ante, politiques d'accès, cartographie des risques, règles d'alignement). K_p est le socle de preuve proprement dit (instrumentation, intégrité, versionnement, conservation sécurisée des logs et contextes). K_m est le monitoring inline (garde-fous temps réel, filtres, détection de dérive, arrêt d'urgence). K_a est l'audit ex-post (revues indépendantes de seconde ligne, échantillonnage, validation périodique). K_Δ est le changement et la requalification (recalibration, requalification lors des mises à jour de modèles, tests de non-régression), amorti sur sa durée d'usage.

Règle de frontière K / H, par nature de flux et non par cause d'incident. K contient les coûts de prévention, de preuve, de contrôle et de requalification. H contient les pertes, charges de couverture et coûts d'absorption, réalisés ou provisionnés. On ne classe pas un incident selon sa cause, on classe chaque flux économique selon sa nature. Le coût du dispositif d'audit est dans K_a , que l'audit détecte ou non. La perte résultant d'une erreur est dans H, qu'un meilleur audit l'eût évitée ou non. La correction du dispositif après incident est dans K_Δ . Une sanction pour défaut de contrôle est dans H, même lorsque sa cause est une insuffisance de K_a , car c'est une charge subie, pas un dispositif. Un H qui dérive alors que K_a est nominal se lit dans l'écart entre pertes observées et pertes anticipées sous le niveau de contrôle déclaré, $\Delta H = H_{\text{observé}} - H_{\text{prévu}}$; il peut signaler un dispositif sous-dimensionné ou un modèle de risque devenu inadéquat, changement de distribution, fraude, environnement nouveau. La relecture de cet écart guide le recalibrage futur, sans double-comptage.

Annexe B. Calibration actuarielle de la charge de risque (H)

Pour éviter tout double-comptage entre l'exposition brute et les instruments de couverture, H se construit comme somme de termes à périmètres mutuellement exclusifs :

$$H = EL_{\text{net}} + C_{\text{ret}} + C_{\text{cap}} + P_{\text{ins}} + M$$

où EL_{net} est la perte attendue nette restant à la charge de l'institution après couverture, C_{ret} le coût financier de la rétention au-delà des pertes ordinaires, C_{cap} la charge de capital contre la perte inattendue non transférée, P_{ins} la prime nette d'assurance pour le risque effectivement transféré, et M une marge pour risque de modèle. Règle de sommation : seules les charges nettes effectivement supportées ou payées par l'institution s'additionnent, les indemnités d'assurance attendues sont déduites, et aucune couche ne recouvre une autre. Sans cette discipline, H gonfle par superposition, car la prime intègre déjà une part de perte attendue, le capital porte la perte inattendue

retenue, la rétention peut chevaucher la perte attendue, et la marge de modèle peut recouvrir le conservatisme de la mesure de queue. M n'est donc défini que pour l'incertitude non déjà portée par ce conservatisme.

La perte attendue est l'espérance de perte, $EL = E[L]$; dans le cas simple d'un événement binaire, $EL = p \times L$. La perte inattendue est mesurée sur la queue de distribution au seuil α . Dans la convention retenue ici, elle est l'écart entre la CVaR et la perte attendue, $UL_\alpha = CVaR_\alpha(L) - EL$, la Conditional Value at Risk étant préférée à la VaR pour sa cohérence comme mesure de risque (Rockafellar et Uryasev, 2000) ; d'autres cadres emploient la VaR ou l'Expected Shortfall, et le choix doit être déclaré.

La charge de capital se calcule sur le capital économique EC mobilisé contre la perte inattendue non transférée, pondéré par son coût r_c : $C_{cap} = r_c \times EC$. Le coût du capital est une donnée financière, distincte d'une préférence de risque. Si l'institution souhaite pondérer explicitement son aversion, elle le fait dans un cadre unique et déclaré, par exemple $H_\lambda = EL_{net} + \lambda \cdot UL_\alpha$, sans cumuler les deux conventions, sous peine de compter deux fois la même aversion.

L'estimation de EL et UL_α repose sur des distributions de pertes régulièrement révisées. En environnement fortement non stationnaire, M matérialise dans H l'incertitude introduite par une chaîne de représentation non stationnaire, reconnue en section 7 : une chaîne instable rend toute distribution de pertes provisoire.

Annexe C. Agrégation de portefeuille et indicateurs de composition

L'agrégation des CDU par classe n'est valide qu'entre classes partageant une convention d'utilité et de qualité compatible. Elle s'effectue par la part observée des décisions utiles, ce qui garantit l'identité comptable entre le CDU de portefeuille et le rapport de la somme des coûts sur la somme des décisions utiles, sous réserve que tous les coûts partagés aient été alloués une seule fois, entre les classes, selon une convention stable. À défaut, des coûts de plateforme sont oubliés, imputés plusieurs fois ou laissés hors portefeuille, et l'identité tombe.

Avertissement de queue, érigé en règle. Cette agrégation est une moyenne pondérée par le volume de décisions utiles. Elle dilue donc les classes rares à forte charge de risque. Un CDU de portefeuille peut paraître sain pendant qu'une classe critique, peu volumineuse mais lourde en risque, saigne sans que la moyenne ne le signale. Règle : aucun CDU agrégé n'est publié sans la distribution des CDU par classe et sans indicateur de queue. Quatre indicateurs, choisis pour leur robustesse plutôt que pour leur simplicité : la distribution des CDU par classe, pondérée par le volume et non pondérée ; le CDU des classes critiques identifiées ex-ante ; la contribution de chaque classe à H ; et la concentration du risque, part de H portée par les classes les plus exposées, assortie d'un intervalle d'incertitude. On évite le CDU maximal comme indicateur cardinal, instable sur les classes de faible volume. Le CDU de portefeuille ne doit jamais devenir le chiffre

unique du comité exécutif ; à défaut, la métrique aurait remplacé le coût au token par une autre moyenne séduisante et mutilante.

Indicateurs de composition. Pour suivre la migration, deux parts se distinguent : la part institutionnelle $s_{inst} = (O_{human} + K + H) / CDU$, et la part de preuve et de risque $s_{pr} = (K + H) / CDU$, toutes deux rapportées au coût par décision. Leur lecture demande une précaution : à structure figée, ces parts croissent nécessairement quand O_{tech} baisse, et s_{inst} tend vers un lorsque le compute s'annule. Ce sont donc des signaux de composition, non des preuves de la thèse ; la thèse porte sur le niveau *absolu* de $O_{human} + K + H$, dont la faible élasticité au coût du compute reste l'énoncé empirique à tester.

Contrôle anti-Goodhart. Dès lors que le CDU devient une cible de pilotage, il devient manipulable par ses définitions, et le principal levier de manipulation est le dénominateur et sa taxonomie. La gouvernance de la métrique contrôle donc, au minimum : la définition des classes homogènes, les seuils de qualité minimale, le budget de remédiation, les fenêtres Δ , les règles d'allocation des coûts mutualisés, la politique d'exclusion des décisions rejetées, l'amortissement de K_{Δ} , le calibrage de λ et de r_c , le suivi des incidents différés, et les changements de périmètre. Règle de clôture : toute modification des classes, des seuils, des fenêtres, des règles de remédiation ou des clés d'imputation produit une rupture de série explicitement documentée. Une métrique dont on ne gouverne pas les définitions finit par mesurer la créativité comptable de ceux qu'elle évalue.

Annexe D. Dominance des régimes et écart ex-ante / ex-post

Dominance. Le CDU compare le coût unitaire des régimes humain, assisté, hybride et automatisé, à qualité, réversibilité et profil de risque équivalents. Une architecture cesse d'être unitairement dominante dès que son CDU dépasse celui de la référence. La perte de dominance unitaire n'emporte pas l'abandon : une architecture peut se justifier par sa dominance *capacitaire*, disponibilité continue et volumes inaccessibles à l'humain, ou *stratégique*, résilience, continuité d'activité, apprentissage. Ces dominances ne se lisent pas dans le CDU ; elles s'arbitrent à côté de lui. Le CDU renseigne l'efficacité unitaire, il ne tranche pas seul la décision d'architecture. L'encadré du §4 en donne le cas limite : deux architectures au même niveau résiduel, dont le choix ne peut alors se fonder que sur une dominance non unitaire.

Écart ex-ante / ex-post. Pour le suivi d'exécution, on suit l'écart entre le CDU constaté et le CDU prévu, $\Delta CDU = CDU_{ex-post} - CDU_{ex-ante}$. Un écart positif peut révéler une sous-estimation de la charge de risque H , une sous-évaluation du besoin de requalification K_{Δ} ou un taux de remédiation supérieur au budget toléré, mais aussi un volume inférieur aux prévisions, un coût observé supérieur, une mauvaise allocation, un incident exceptionnel ou une modification réglementaire. Pour ne pas conclure trop vite,

on décompose l'écart avant de l'interpréter, $\Delta\text{CDU} = \Delta\text{O} + \Delta\text{K} + \Delta\text{H} + \Delta\text{D}^{\text{U}}$, chaque terme ramené à la même cohorte. L'écart ne se lit qu'à périmètre et cohorte constants ; hors de cette constance, il mesure un changement de périmètre, pas une dérive du régime de délégation.