Performance mesurée, fiabilité opérationnelle : la distinction que l'industrie refuse de faire

Pourquoi un AUC élevé ne constitue ni une garantie de déploiement, ni une mesure suffisante de confiance dans les systèmes d'IA régulés

Twingital Institute — Jérôme Vetillard — Avril 2026

Introduction

Le débat public sur l'intelligence artificielle reste dominé par les métriques de performance. AUC, F1-score, accuracy, BLEU, perplexité : ces indicateurs structurent les publications, alimentent les annonces produit, soutiennent les présentations aux investisseurs, et occupent une place centrale dans les dossiers de validation. Cette centralité n'est pas accidentelle. Les métriques de performance ont trois propriétés institutionnellement puissantes : elles sont mesurables, comparables, et publiables.

Les propriétés qui conditionnent la robustesse effective d'un système en situation de déploiement réel sont, elles, beaucoup moins visibles. La calibration des probabilités, la définition explicite d'un domaine d'applicabilité, la résistance au décalage de distribution, la capacité du système à signaler ses propres limites… toutes ces propriétés restent, dans de nombreux contextes, traitées comme des considérations secondaires, parfois importantes, mais rarement structurantes.

Cette asymétrie ne résulte pas d'une ignorance. Une grande partie de la communauté scientifique sait depuis longtemps qu'une métrique discriminante élevée ne suffit pas à caractériser la sûreté d'usage d'un modèle, encore moins celle d'un système complet. L'asymétrie est culturelle, institutionnelle et architecturale. La performance est privilégiée parce qu'elle s'intègre naturellement à une économie de comparaison. La fiabilité opérationnelle exige une réflexion sur le déploiement réel, sur l'hétérogénéité des entrées, sur l'incertitude et sur les mécanismes de limitation du système. Elle se prête moins bien aux classements, aux tableaux de bord simplifiés et aux récits de progrès linéaire.

La thèse défendue ici est la suivante : un modèle performant au sens métrique n'est pas, pour cette seule raison, un système fiable au sens opérationnel. Ces deux propriétés ne se déduisent pas l'une de l'autre. Elles se construisent différemment, se mesurent différemment, et répondent à des questions distinctes. Leur confusion constitue une erreur architecturale récurrente dans les systèmes d'IA déployés en environnement régulé et non un problème d'ingénierie de surface qu'une couche de monitoring supplémentaire pourrait corriger.

Cette thèse a un domaine de validité explicite. Elle concerne prioritairement les systèmes produisant ou informant des décisions individuelles à enjeu élevé : prédiction toxicologique réglementaire, aide à la décision diagnostique, scoring de risque clinique, triage, allocation de ressources de santé. Elle ne s'applique pas avec la même intensité à des systèmes exploratoires ou à des outils de productivité à faible enjeu individuel, où l'erreur reste réversible et statistiquement tolérable.

I. Clarification conceptuelle : de quoi parle-t-on exactement ?

La discussion devient confuse lorsque plusieurs propriétés distinctes sont traitées comme si elles relevaient d'un même continuum. Il est donc nécessaire de distinguer explicitement cinq notions.

La *performance mesurée* désigne la capacité d'un modèle à produire des prédictions correctes sur un jeu d'évaluation donné, selon un protocole de séparation des données et une ou plusieurs métriques définies à l'avance. L'AUC mesure la capacité discriminante globale d'un classifieur sur l'ensemble des seuils possibles. Le F1-score combine précision et rappel pour un seuil donné. L'accuracy/précision exprime la proportion de prédictions correctes. Ces métriques sont utiles. Elles mesurent une propriété réelle. Elles ne mesurent pas tout ce qui compte dans l'usage effectif d'un système.

La *calibration* concerne la relation entre les probabilités ou scores de confiance produits par le modèle et les fréquences empiriques observées. Un modèle bien calibré attribuant une probabilité de 0,80 à un ensemble de cas verra, en moyenne, l'événement se produire dans environ 80 % de ces cas. La calibration n'est pas équivalente à la discrimination : deux modèles peuvent avoir une AUC proche et des calibrations très différentes. Ces deux propriétés ne se substituent pas.

La *validité décisionnelle* désigne l'aptitude d'une sortie algorithmique à être utilisée de manière pertinente dans une politique de décision donnée. Une probabilité, un score ou un rang ne prennent sens qu'en relation avec un seuil, un coût d'erreur, une temporalité d'action, un niveau de supervision humaine, une possibilité de vérification complémentaire. La même performance prédictive peut être décisionnellement utile dans un contexte et insuffisante dans un autre. Ce niveau est trop souvent laissé implicite.

Le *domaine d'applicabilité* (AD) désigne l'espace des entrées pour lequel il existe des raisons méthodologiquement défendables de considérer que les prédictions du modèle

restent valides. Tout modèle a, de fait, un AD implicite. La question n'est pas de savoir s'il existe, mais s'il est explicitement défini, mesuré et intégré au pipeline d'inférence.

La *fiabilité opérationnelle* désigne le comportement suffisamment prévisible, borné, surveillable et gouvernable d'un système dans ses conditions réelles d'usage. Cette propriété ne dépend pas uniquement du modèle. Elle dépend du pipeline complet : données en entrée, transformations, règles d'acceptation, détection des cas hors domaine, calibration, monitoring, mécanismes d'escalade, interaction avec l'utilisateur et temporalité de révision.

La distinction centrale de cet article peut dès lors être formulée rigoureusement : la performance mesurée caractérise la qualité d'un modèle sur une tâche d'évaluation donnée ; la validité décisionnelle concerne la façon dont cette sortie peut être utilisée dans une politique d'action ; **la fiabilité opérationnelle concerne le comportement du système réel dans un environnement de déploiement**.

Ces trois objets sont liés, mais ne se confondent pas.

II. Comment la performance a absorbé l'idée même de rigueur

La domination des métriques de performance n'est pas seulement un effet de mode. C'est le produit d'une histoire institutionnelle de l'évaluation en IA qui mérite d'être retracée, précisément parce qu'elle explique la résistance du biais à la correction.

Les grands benchmarks (ImageNet pour la vision, GLUE et SuperGLUE pour le langage, MoleculeNet pour la chémoinformatique) ont joué un rôle structurant dans la culture de l'évaluation. Ils ont permis la comparaison inter-équipes, la reproductibilité relative des expériences, l'accumulation de résultats et l'identification de progrès réels. En ce sens, leur contribution est indéniable. Le problème ne vient pas de leur existence. Il vient de la transformation progressive du benchmark en quasi-substitut de la réalité d'usage.

Cette dérive est renforcée par la logique des leaderboards. Un leaderboard récompense ce qui se compare aisément : un score final, obtenu dans les conditions formelles de la tâche. Il valorise beaucoup moins la qualité de la calibration, la robustesse au décalage de distribution, la lisibilité du domaine d'applicabilité, ou la dégradation gracieuse hors de l'espace de validité. Ces propriétés ne se résument pas sur une ligne de tableau. Elles ont donc structurellement moins pesé dans la culture d'évaluation.

À cela s'ajoute un biais de séquencement. Dans de nombreux pipelines, la réflexion sur la gouvernabilité du système intervient après l'entraînement et l'optimisation du modèle. On entraîne d'abord. On évalue ensuite. On envisage enfin la calibration, le monitoring, la

gestion des cas atypiques et l'intégration des garde-fous. Cette séquence rend presque inévitable le traitement de la fiabilité comme une couche ajoutée, alors qu'elle devrait être pensée comme une propriété constitutive du système (exactement le même biais que celui documenté pour la gouvernance architecturale des systèmes IA régulés – article déjà publié-).

Le résultat est un déséquilibre collectif. L'écosystème dispose aujourd'hui d'outils très sophistiqués pour comparer des performances mesurées, et d'outils comparativement fragmentaires pour caractériser la fiabilité réelle en déploiement. Ce déséquilibre reflète une définition implicite de ce qui mérite d'être optimisé.

III. Pourquoi la performance mesurée ne suffit pas comme proxy de fiabilité

3.1 Un protocole d'évaluation peut mesurer honnêtement la mauvaise chose

La valeur d'une métrique dépend d'abord de la pertinence du protocole d'évaluation. Une AUC élevée n'a pas de signification universelle, elle a une signification dans le cadre du split et de la distribution retenus.

En chémoinformatique, la différence entre un split aléatoire et un scaffold split est décisive :

- Un split aléatoire distribue des composés structurellement proches entre entraînement et test, ce qui favorise l'interpolation locale.
- Un scaffold split impose une séparation stricte entre familles structurelles et mesure davantage la capacité du modèle à traiter des structures réellement nouvelles.

Les travaux de Sheridan (2013) sur les modèles QSAR et les analyses ultérieures de Wallach et al. sur MoleculeNet ont montré que de telles différences de protocole conduisent à des écarts de performance parfois importants , variables selon les jeux de données, les représentations moléculaires et les modèles considérés.

La leçon générale dépasse la chémoinformatique. En clinique, une validation aléatoire peut masquer des effets temporels ou institutionnels qui apparaissent dès qu'on évalue prospectivement ou sur site externe. En vision médicale, un split naïf peut laisser persister des corrélations parasites liées à l'équipement ou au centre. Un bon score sur un protocole inadéquat ne mesure pas la capacité du système à se comporter

correctement en déploiement, il mesure sa réussite dans les conditions spécifiques de ce protocole.

3.2 Une métrique discriminante ne renseigne pas sur l'interprétabilité probabiliste

L'AUC est une métrique de rang : elle mesure la capacité d'un modèle à ordonner les cas positifs et négatifs, indépendamment de l'amplitude des scores. Elle est invariante aux transformations monotones des scores ce qui est sa force pour comparer des classifieurs, et sa limite radicale pour juger de leur fiabilité décisionnelle.

Deux modèles peuvent avoir des AUC proches tout en produisant des scores d'ampleur très différente. Or, dans de nombreux contextes à enjeu élevé, l'utilisateur n'a pas seulement besoin d'un bon ordonnancement. Il a besoin de savoir si un score présenté comme « 0,85 » correspond à quelque chose de stable et interprétable, ou s'il s'agit d'une sortie utile pour classer mais trompeuse si elle est lue comme probabilité.

Une métrique discriminante ignore donc, par construction, plusieurs dimensions décisives : l'interprétabilité probabiliste, la prévalence opératoire, l'asymétrie des coûts d'erreur, et la criticité locale de certaines régions de l'espace d'entrée. Lorsque la décision aval repose uniquement sur un tri relatif, cela peut suffire. Dès lors que la sortie alimente une décision individualisée, un arbitrage probabiliste, un niveau de confiance communiqué à un expert ou une action partiellement automatisée, cela ne suffit plus. *Une probabilité affichée à un utilisateur n'est pas un simple format de sortie. C'est un engagement interprétatif.*

3.3 L'absence d'AD explicite produit de la confiance non méritée

Par défaut, un modèle statistique fournit une sortie pour toute entrée compatible avec son interface technique. Cette disponibilité universelle de la réponse ne doit jamais être confondue avec une universalité de validité.

Lorsqu'aucun domaine d'applicabilité n'est défini ni intégré au pipeline, une entrée hors du champ de validité du modèle reçoit un score dans le même format qu'une entrée familière. Pour l'utilisateur final, l'interface ne signale pas que la prédiction a été produite dans une zone de faible densité, sur un type de cas peu représenté, ou dans une région de l'espace d'entrée dont le modèle n'a appris qu'une approximation fragile.

Dans un pipeline toxicologique, cette question est aiguë lorsqu'un modèle entraîné sur des molécules organiques de petite taille est confronté à des organométalliques ou à des complexes de coordination dont les propriétés déterminantes ne sont pas capturées par la featurisation employée. Le problème n'est pas seulement une baisse de performance moyenne, c'est la présentation silencieuse d'une réponse qui conserve l'apparence de validité.

Le benchmark teste le modèle dans son monde. Le déploiement le place dans le vôtre. Un système sans AD intégré ne fait pas la différence.

IV. De la qualité du modèle à la fiabilité du système : une structure à trois niveaux

Pour sortir de la confusion entre performance et fiabilité, il est utile d'introduire la tripartition dégagée au §I comme cadre opératoire.

Le premier niveau est celui de la *qualité intrinsèque du modèle* : performance sur une tâche évaluée selon un protocole donné, discrimination, calibration éventuelle, robustesse.

Le deuxième niveau est celui de la *validité décisionnelle de la sortie* : une probabilité calibrée peut être utile pour déclencher une vérification expérimentale au-delà d'un certain seuil ; un score mal calibré mais bien ordonné peut rester utile dans un contexte de ranking exploratoire. Ce niveau est trop souvent laissé implicite et c'est précisément dans cet implicite que se logent la plupart des malentendus entre les concepteurs du modèle et les utilisateurs du système.

Le troisième niveau est celui de la *fiabilité opérationnelle du système* : gestion des cas atypiques, stabilité des flux, garde-fous, lisibilité des limites, détection de dérive, comportement temporel, modalités d'escalade. La fiabilité n'est pas un attribut isolé du modèle. C'est une propriété du système complet inséré dans un monde réel.

Cette tripartition permet d'éviter un faux débat. Le problème n'est pas que les métriques de performance seraient inutiles. Le problème est qu'elles décrivent un niveau, et un niveau seulement. **L'erreur survient lorsqu'on leur fait porter la charge descriptive de l'ensemble.**

V. La fiabilité opérationnelle comme propriété architecturale délibérée

Si la fiabilité opérationnelle ne dérive pas automatiquement de la performance, il faut spécifier comment elle se construit. Dans les systèmes à enjeu élevé, elle repose au minimum sur trois décisions architecturales prises en amont.

Première décision : le protocole d'évaluation doit être représentatif du déploiement plausible. Cette exigence n'implique pas un unique protocole « vrai ». Elle implique que

le choix du split encode une hypothèse honnête sur l'usage. Pour un système destiné à traiter de nouvelles familles chimiques, un scaffold split est plus honnête qu'un split aléatoire. Pour un système clinique déployé dans le temps, une validation temporelle prospective est souvent plus informative qu'une séparation aléatoire. Le protocole d'évaluation n'est pas une formalité technique. C'est une prise de position épistémique sur ce que le déploiement ressemblera.

Deuxième décision : la calibration doit être intégrée au pipeline lorsque le score est utilisé comme probabilité ou signal de confiance interprétable. La régression isotonique convient aux corpus riches en données et capture des relations non paramétriques entre score brut et fréquence empirique. La calibration de Platt reste pertinente en données rares ou lorsqu'on recherche une calibration paramétrique plus stable. Le point décisif n'est pas le dogme de la méthode : c'est l'intégration de la calibration comme composant du pipeline, non comme option d'affichage.

Troisième décision : un mécanisme opérationnel d'estimation de proximité ou de validité locale doit être intégré avant inférence. Les méthodes varient selon les domaines : distance aux k plus proches voisins dans l'espace des features, estimation de densité, ensembles avec mesure de variance, scores d'incertitude, détection hors distribution. Aucun mécanisme n'est universel. Mais, dans un système à enjeu élevé, l'absence totale de mécanisme explicite est architecturalement indéfendable.

Ces trois couches ne se substituent pas ! Elles se complètent. Un système peut être bien calibré dans son espace de validité et rester trompeur hors de cet espace. Il peut définir correctement les cas hors domaine tout en produisant des probabilités mal ajustées sur les cas acceptés. Il peut être évalué selon un protocole réaliste tout en échouant à signaler ses zones d'incertitude locales. La fiabilité opérationnelle naît de leur articulation, non de leur simple juxtaposition.

VI. La charge épistémique du déploiement

Il est utile, pour rendre visible un problème dispersé dans plusieurs sous-littératures, de disposer d'un concept intégrateur. Je propose celui de *charge épistémique du déploiement*.

Par ce terme, j'entends l'écart entre ce qu'un modèle a effectivement appris à traiter dans les conditions de son entraînement, et ce que le déploiement réel lui demande de traiter dans un environnement d'usage évolutif, hétérogène, partiellement imprévu, et parfois hors du champ anticipé lors de la conception.

Cette proposition ne prétend pas identifier un phénomène entièrement nouveau. Les notions de *dataset shift*, *covariate shift*, *concept drift*, incertitude épistémique, et

détection hors distribution constituent déjà un ensemble riche de travaux techniques. L'intérêt du concept proposé est ailleurs : réunifier, dans une perspective de gouvernance et d'architecture, des phénomènes souvent traités séparément, alors qu'ils produisent conjointement la difficulté réelle du déploiement.

La charge épistémique du déploiement est faible dans des environnements fortement contrôlés, où la distribution des entrées reste stable et proche de la distribution d'entraînement. Elle devient élevée lorsque les entrées sont hétérogènes, les pratiques variables, les populations évolutives, les cas atypiques fréquents, ou les conséquences des erreurs asymétriques et peu réversibles.

Ce concept invite à une inversion de perspective. Au lieu de demander seulement si un modèle généralise « bien » au sens abstrait, il conduit à demander quelle charge supplémentaire le monde réel impose à ce modèle, et quels mécanismes architecturaux le système met en place pour absorber, signaler ou limiter cette charge. Son opérationnalisation reste un programme de recherche ouvert, mais son invisibilisation dans les raisonnements de déploiement constitue déjà un problème documentable.

VII. Ce que le monitoring ne résout pas

Les plateformes MLOps contemporaines (MLflow, Vertex AI Model Monitoring, Amazon SageMaker Clarify) permettent de suivre des dérives, d'observer l'évolution des distributions de features, de tracer des versions de modèles et, pour les plus avancées, de produire des explications post hoc. Ces capacités sont utiles. Elles relèvent de l'observabilité et du contrôle post-déploiement mais non de la fiabilité architecturale.

La distinction a été développée dans le précédent article de cette série sur la gouvernance architecturale des systèmes IA régulés. Elle s'applique directement ici : un outil de monitoring peut détecter qu'une distribution dérive ; il ne garantit pas que le pipeline refuse ou balise les cas qui franchissent un seuil d'incertitude. Une plateforme peut historiser des performances agrégées ; elle ne garantit pas que les probabilités affichées à l'utilisateur soient correctement calibrées pour la décision qu'il doit prendre. La fiabilité opérationnelle ne se délègue pas au monitoring car elle se conçoit en amont, comme propriété constitutive du pipeline.

VIII. Cadres réglementaires : une zone de responsabilité laissée ouverte

Les cadres applicables aux systèmes d'IA à enjeu élevé évoluent rapidement. Les exigences du MDR en matière de validation, de traçabilité, de gestion du risque et de surveillance post-marché sont réelles et substantielles. L'AI Act européen, applicable au 2 août 2026 pour la majorité des obligations, renforce les exigences en matière de gestion des risques, de documentation, de logging et de supervision humaine pour les systèmes à haut risque.

Ces textes jouent un rôle essentiel. Ils créent une pression salutaire sur la traçabilité et la responsabilité. Ils restent cependant technologiquement neutres sur les mécanismes précis : ils n'imposent pas de méthode particulière de calibration, ni de formalisation standard du domaine d'applicabilité, ni de mesure de la charge épistémique du déploiement. Cette neutralité est compréhensible du point de vue réglementaire. Elle laisse aux concepteurs une responsabilité forte : définir eux-mêmes ce qui rendra leur système effectivement gouvernable en situation réelle.

Deux systèmes peuvent être tous deux documentés, tracés et surveillés, tout en différant considérablement dans leur capacité à signaler leurs limites avant qu'une erreur ne se matérialise. La conformité réglementaire est le plancher. La fiabilité opérationnelle est l'objectif.

IX. Terrain d'implémentation : ToxTwin V2.3+

Les principes défendus ici trouvent un terrain d'illustration dans le pipeline de prédiction toxicologique moléculaire ToxTwin, développé au sein du Twingital Institute. Il ne s'agit pas d'en faire une preuve générale, mais d'en dégager des enseignements concrets sur la faisabilité des couches de fiabilité décrites.

L'audit réalisé début 2026 a révélé deux problèmes structurels. Le premier était une circularité dans la validation du modèle Ames : le split utilisé dans les versions antérieures permettait une fuite de similarité structurelle significative.

La correction du protocole (scaffold split strict à 5 folds sur 20 117 composés) a conduit l'AUC Ames à 0,864 ± 0,056 en validation croisée, contre des valeurs non reproductibles sur split aléatoire. Le second problème était l'absence de calibration des probabilités de sortie : le modèle GINEConv OGB (163 features, dropout 0,3) produisait des scores discriminants mais non interprétables comme probabilités empiriquement fidèles.

Les corrections apportées en V2.3 comprennent trois éléments :

1. L'introduction d'une calibration isotonique ajustée sur un jeu de calibration distinct, produisant des probabilités vérifiables sur le holdout gelé (SHA256 = 052a2aa2c4cff3d8...).
2. La définition d'un AD opérationnel basé sur la distance aux k plus proches voisins dans l'espace des 163 features OGB, avec un seuil p95 à 0,332 (une molécule dépassant ce seuil reçoit un signal AD négatif avant présentation du score).
3. La correction du protocole avec un holdout gelé garantissant la reproductibilité des évaluations futures.

Ce que cette instance prouve : que les trois couches de fiabilité opérationnelle sont implémentables dans un pipeline GNN industriel, avec un coût computationnel marginal.

Ce qu'elle ne prouve pas : que l'approche est directement transférable sans adaptation à d'autres domaines moléculaires. Les complexes de coordination métalliques (cisplatine, carboplatine, oxaliplatine) requièrent une featurisation spécialisée que les 163 features OGB standard ne supportent pas, et constituent l'objet de la roadmap V3.0.

La calibration isotonique n'est pas universellement supérieure en données rares. Le seuil AD retenu n'est pas optimal pour tous les cas d'usage.

Ces résultats ont valeur d'illustration de faisabilité architecturale et non d’une vérité générale.

X. Limites et contre-arguments

Une position crédible expose ses propres limites.

La calibration n'est pas uniformément critique dans tous les contextes. Dans certains usages de ranking, de présélection ou d'exploration, la qualité de l'ordonnancement peut avoir plus d'importance pratique que l'interprétabilité probabiliste. L'argument de cet article ne consiste pas à absolutiser la calibration, mais à rappeler qu'elle devient structurante dès lors que la sortie prétend orienter une décision individualisée sous contrainte.

Le domaine d'applicabilité n'est pas un objet méthodologiquement uniforme. Les mécanismes pertinents changent selon les données, les modèles et les domaines. Il n'existe pas de définition universelle de l'AD transposable de la chémoinformatique à la clinique, à l'imagerie ou aux systèmes fondés sur des LLM. Le concept général reste utile ; ses implémentations doivent rester spécifiques et prudentes.

La performance agrégée peut avoir une valeur décisionnelle considérable même lorsque la lisibilité locale est imparfaite. Des systèmes de santé publique ou

d'optimisation de flux peuvent bénéficier d'outils dont la calibration individuelle n'est pas la propriété la plus critique. L'intensité de la thèse défendue ici dépend du type de décision et du niveau auquel l'erreur est jugée.

La charge épistémique du déploiement est un cadrage intégrateur, pas encore un indicateur standardisé. Elle demande des travaux d'opérationnalisation, de comparaison méthodologique et de validation sur des terrains multiples. Son intérêt présent est analytique : rappeler que le déploiement impose au système une charge statistique supplémentaire que les métriques classiques absorbent mal.

La fiabilité opérationnelle ne dépend pas seulement des propriétés techniques. Elle dépend aussi des workflows, de la formation des utilisateurs, des modalités de supervision, du contexte institutionnel et de la gouvernance organisationnelle. Un système techniquement prudent peut être organisationnellement dangereux s'il est inséré dans un usage mal cadré.

Conclusion

L'industrie optimise ce qu'elle mesure. La performance s'est imposée comme langue dominante de l'évaluation des systèmes d'IA parce qu'elle est mesurable, comparable et publiable. Cette domination a produit de réels progrès. Elle a aussi entretenu une confusion persistante : celle qui consiste à prendre la qualité d'un modèle sur un protocole donné pour une approximation suffisante de la fiabilité du système dans le monde réel.

Cette confusion doit être levée avec netteté. La performance mesurée, la validité décisionnelle et la fiabilité opérationnelle ne sont pas trois degrés d'une même propriété. Ce sont trois niveaux distincts d'analyse, qui se construisent différemment et répondent à des questions différentes. Un système à enjeu élevé ne devient pas fiable parce que son AUC est élevé. Il devient plus fiable lorsqu'il est évalué selon un protocole cohérent avec son usage, lorsqu'il rend ses sorties interprétables à bon escient, lorsqu'il délimite son domaine de validité, lorsqu'il sait signaler l'extrapolation, et lorsqu'il intègre ces contraintes au cœur même de son architecture.

La conséquence est directe. La fiabilité opérationnelle ne doit plus être pensée comme une couche tardive de monitoring ajoutée à un modèle déjà conçu. Elle doit être conçue comme une propriété architecturale délibérée, définie avant l'entraînement, validée dans le pipeline, réévaluée dans le temps, et articulée à la politique réelle de décision.

Ce qu'il faut mesurer dans les systèmes régulés n'est pas seulement ce qui est facile à comparer. C'est ce qui conditionne réellement la possibilité d'un usage sûr, lisible et

gouvernable. L'industrie devrait d'abord mesurer ce qui compte. Ensuite optimiser ce qu'elle mesure.