Un benchmark clinique est une décision de périmètre, pas une mesure

Validité statistique, clinique, gouvernable. Pourquoi la composition d'un espace de validation conditionne le coût institutionnel de reconnaissance des incidents qu'un système clinique pourra surveiller, escalader et opposer.

Le problème n'est pas l'imperfection

Le débat sur l'évaluation des intelligences artificielles cliniques s'est installé sur un constat que personne ne conteste : les benchmarks sont imparfaits. C'est vrai, et c'est sans intérêt. Aucun praticien sérieux n'a jamais cru qu'un benchmark fût exhaustif, et l'on n'écrit pas une doctrine sur une évidence. Le problème est plus dur, et il peut s'énoncer en une seule phrase, que tout le reste de cet article ne fait que tenir, tester et opérationnaliser :

Un benchmark clinique définit quelles formes d'échec deviendront des objets gouvernables pour le système.

Tout l'enjeu se concentre là. Le benchmark, entendu ici en un sens volontairement large comme *espace de validation*, n'est pas seulement un instrument de mesure : c'est le périmètre effectivement utilisé pour décider qu'un système est suffisamment valide pour être promu, qu'un comportement mérite d'être surveillé, qu'un incident mérite d'être traité comme un signal systémique. Ce qui n'entre pas dans ce périmètre n'est pas effacé du réel clinique : il sort de l'attention institutionnelle. La validité du périmètre n'est pas la gouvernance. Elle en est la condition de possibilité.

Trois déplacements successifs de la charge de preuve

- La décennie 2015-2025 a connu un premier déplacement. On a cessé de juger un dispositif sur la seule plausibilité méthodologique pour exiger des métriques de performance reproductibles : AUROC, sensibilité, calibration, F1. Ce déplacement était sain, mais limité, et la critique s'en est rapidement saisie. Un score moyen sur un benchmark étroit ne garantit ni transférabilité, ni utilité.
- La séquence éditoriale de 2026 a opéré le deuxième déplacement. *Nature Medicine* ne demande plus si les modèles hallucinent, mais s'ils améliorent les résultats cliniques, et répond, dans plusieurs textes convergents, que dans bien

des cas on ne le sait pas (« Is AI actually improving healthcare? » ; « Show us the evidence for the value of medical AI », 2026). La charge de la preuve est passée de la performance intrinsèque à la preuve d'impact. C'est juste, et c'est nécessaire.

Mais ce déplacement, à son tour, ne suffit pas. Et le piège est subtil : même la preuve d'impact clinique *hérite silencieusement* du périmètre de visibilité défini en amont. Un essai pragmatique randomisé qui démontre une amélioration de la durée de séjour ou du taux de réadmission ne dit rien des classes d'incidents qui restent invisibles à l'instrument de mesure utilisé dans l'essai lui-même. L'outcome valide ce qu'il observe. Il reste muet sur ce qu'il n'a pas instrumenté.

- Le troisième déplacement, celui que cet article cherche à formuler, ne remplace pas les deux premiers. Il les complète. Il s'énonce ainsi : la question pertinente n'est plus seulement *le modèle est-il performant* ni *améliore-t-il les résultats*, mais *les incidents qu'il produira en production resteront-ils observables, attribuables et opposables ?*

Performant n'est pas gouvernable

Trois niveaux de validité se confondent dans le mot « validation », et les séparer est l'apport central de ce texte.

- La *validité statistique* répond à : le modèle prédit-il correctement dans le benchmark ?
- La *validité clinique* répond à : le benchmark représente-t-il les situations cliniques pertinentes ?
- La *validité gouvernable*, troisième niveau, répond à une question que les deux premiers ne posent jamais : les incidents produits en production resteront-ils maniables institutionnellement ?

Cette validité se stratifie elle-même en trois plans qu'il faut tenir séparés, parce qu'ils sont fréquemment confondus, et que la confusion est coûteuse.

- *Observable* : l'incident produit une trace que le système enregistre.
- *Attribuable* : la trace peut être rattachée à une classe d'incidents, à un modèle, à une décision identifiable.
- *Gouvernable* : il existe une politique organisationnelle stable et opposable pour traiter cette classe.

Un incident peut être observable sans être attribuable (on voit qu'il s'est passé quelque chose, on ne sait pas le qualifier). Il peut être attribuable sans être gouvernable (on sait quelle classe, aucune procédure ne déclenche). Le score AUROC d'un modèle ne dit rien des trois plans.

Un modèle peut donc exceller au premier niveau, passer le deuxième, et échouer au troisième sans que rien dans son score ne le signale.

Un exemple rend l'absurdité du réflexe leaderboard physiquement sensible. Soit un modèle A, AUROC 0,96 sur un benchmark étroit. Soit un modèle B, AUROC 0,91, mais avec couverture explicite des classes rares critiques et une politique d'escalade dédiée à chacune. Le classement par performance agrégée préfère A. La gouvernabilité préfère B, parce que B sait reconnaître l'incident rare comme un incident, l'attribuer, le tracer, l'escalader, là où A le dissout dans une moyenne flatteuse. Ce n'est pas un paradoxe ; c'est ce qui arrive quand on cesse de confondre la justesse moyenne avec la sécurité opposable. Le réflexe du tableau de classement optimise exactement la mauvaise quantité.

La préférence pour B ne doit pas devenir une apologie naïve du sur-signalement. Un système trop empressé à reconnaître des classes produit la pathologie symétrique : la fatigue d'alerte, qui détruit la gouvernance par saturation. La gouvernabilité authentique n'est pas la maximisation de la détectabilité ; c'est sa *calibration*. Le critère n'est pas combien d'événements escaladés, mais quelles classes d'événements, avec quelle politique d'arbitrage. Confondre les deux, c'est remplacer un angle mort par un brouillard.

Le théorème de propagation, et son coût

À ce stade, l'objection industrielle dominante se présente, et elle est solide : le post-market compensera. La surveillance en vie réelle, l'incidentologie, les plans de gestion des modifications corrigeront en aval ce que le benchmark a manqué en amont. L'objection mérite d'être prise au sérieux, parce qu'elle décrit un mécanisme réel.

La réponse est ce qu'on peut appeler un *théorème de propagation*. Il faut le formuler avec précision, car sa version grossière est attaquable et sa version exacte est forte.

Soit S un système de surveillance déployé et T sa taxonomie initiale de visibilité. Une classe d'événement c absente de T n'est pas rendue impossible à découvrir par S : un signal faible, un signalement clinicien, un cluster pharmacovigilance, une analyse qualitative rétrospective peuvent l'identifier.

Ce que le théorème affirme est plus précis et plus exact : ***l'absence de c dans le périmètre initial conditionne le coût institutionnel de sa reconnaissance ultérieure comme classe gouvernable.***

Cette formulation déplace exactement le point qu'un expert en pharmacovigilance objecterait à juste titre. Les systèmes de vigilance servent précisément à découvrir des classes inconnues. Vrai. Mais cette découverte n'est pas gratuite. Une classe non instrumentée à l'origine demande, pour devenir gouvernable, une reconstruction institutionnelle complète : nouvelle taxonomie, nouveau seuil, nouveau protocole,

nouveau monitoring, nouveau KPI, nouveau SLA, nouvelle ligne budgétaire, nouvelle clause contractuelle, parfois nouvelle révision réglementaire. Le coût n'est pas l'impossibilité ; il est la dette institutionnelle accumulée par chaque classe omise.

Le périmètre amont ne fixe donc pas seulement ce qui sera vu : il fixe le différentiel de coût entre les classes vues de plain-pied et les classes ré-instrumentées après coup. Et cette dette institutionnelle, contrairement à la dette technique, ne se rembourse pas avec un sprint.

Le problème n'est pas propre à la santé. Il est isomorphe à des difficultés que plusieurs disciplines matures ont déjà nommées : l'*observabilité* en systèmes distribués, la *détectabilité* en théorie du contrôle, le *support mismatch* en apprentissage statistique, l'*insuffisance causale* en pharmacovigilance. Quand un même invariant réapparaît sous quatre formulations indépendantes, ce n'est pas une coïncidence de vocabulaire. C'est une contrainte structurelle.

La rétroaction, et son architecture

Le théorème n'implique pas que le périmètre soit gelé. Il implique qu'élargir la taxonomie suppose une procédure dédiée, anticipée à la conception, sans quoi elle n'existe pas.

C'est précisément ce que les régulateurs ont commencé à industrialiser. Le *Predetermined Change Control Plan* (PCCP), formalisé par la FDA dans son final guidance de décembre 2024, et progressivement adopté pour les dispositifs IA en 2025-2026, est un instrument de rétroaction explicite : il oblige le fabricant à déclarer à l'avance les classes de modification réalisables sans nouvelle soumission, ce qui revient à pré-spécifier le régime de conversion entre signal post-market et révision du périmètre. L'AI Act européen, dans ses articles 9 (système de gestion des risques) et 10 (gouvernance des données), exige un dispositif équivalent dans l'esprit. Le MDR et l'IVDR complètent l'architecture par la vigilance post-commercialisation.

Aucun de ces dispositifs n'élimine le théorème de propagation. Ils en organisent l'usage. Ils transforment l'angle mort initial en angle mort *déclaré*, et instituent un protocole pour le réduire dans le temps. Un système gouverné par PCCP ou par article 9 AI Act n'est pas un système sans angles morts. C'est un système dont les angles morts sont opposables. À condition que la déclaration soit honnête, et que la nomenclature utilisée dans la déclaration soit celle utilisée dans le monitoring.

Le benchmark compile la politique de risque

Il faut nommer ce que fait réellement la composition d'un benchmark, et je le ferai d'une image, *une seule*, parce qu'elle est juste à un endroit et fausse partout ailleurs. Le benchmark agit comme le compilateur implicite de la politique de risque du système : il

traduit un choix de données en architecture comportementale. L'image vaut pour fixer l'idée que le dataset n'est pas en amont du système mais à l'intérieur de sa conduite. Elle cesse de valoir dès qu'on la file. Un compilateur produit une exécution déterministe ; un benchmark agit probabilistiquement, le runtime reste partiellement adaptatif, et l'opérateur humain modifie la trajectoire. Passée cette image, je m'en tiendrai à des termes opérationnels : périmètre de visibilité, périmètre de détectabilité, périmètre de gouvernabilité.

La propagation se laisse alors décrire pas à pas, et c'est ici que le texte cesse d'être théorique. Classe rare absente du benchmark, donc pas de calibration spécifique à cette classe, donc pas de seuil d'alerte dédié, donc pas de politique d'escalade, donc pas de monitoring ciblé, donc pas de KPI associé, donc pas de SLA contractuel correspondant, donc pas d'arbitrage budgétaire spécifique, donc pas de signal post-market exploitable. L'incident, lorsqu'il survient, est interprété comme bruit individuel plutôt que comme classe systémique.

À chaque maillon, rien ne dysfonctionne. Chaque composant fait exactement ce pour quoi il a été réglé. *Le défaut n'est nulle part et partout*. Il est dans la frontière initiale de visibilité, qui s'est propagée sans qu'aucun ingénieur ne décide jamais d'ignorer le risque.

Les trois maillons souvent oubliés sont précisément les trois maillons financiers : KPI, SLA, arbitrage budgétaire. Sans eux, la propagation reste abstraite. Avec eux, elle devient ce qu'elle est : une décision de capital alloué à l'attention institutionnelle. Un directeur financier qui valide un contrat de support sur la base d'un SLA défini sans déclaration de couverture du benchmark signe, sans le savoir, une exonération implicite pour toutes les classes hors-périmètre.

La bureaucratie opérationnelle

La propagation ne s'arrête pas au logiciel. Elle se prolonge dans l'organisation. Elle structure les workflows d'escalade, les temps de revue clinique, les contrats de support, les politiques de remboursement, les clauses d'assurance, les obligations internes de conformité, les arbitrages d'effectifs. Le benchmark ne compile pas qu'une architecture logicielle ; il compile une *bureaucratie opérationnelle*.

Le terrain rend cette assertion vérifiable.

Considérons OCTOPUS, l'étude observationnelle multicentrique sur le mNSCLC BRAF V600E (n=184, cinq pays européens, modélisation de survie par SurvTRACE) menée dans le cadre TweenMe. Cette mutation représente environ 1 à 2 % des cancers bronchiques non à petites cellules métastatiques. Si la cohorte de validation n'inclut pas explicitement cette sous-population, la conséquence n'est pas qu'un patient sera mal stratifié individuellement, c'est que la classe devient *opérationnellement inexistante* alors même qu'elle reste *cliniquement réelle*. La présence ou l'absence de quelques

centaines de patients dans le périmètre de validation peut déterminer, par propagation, l'allocation d'une attention institutionnelle significative sur plusieurs années de cycle de vie du déploiement.

Les classes visibles deviennent suivies, financées, auditables, priorisées. Les classes invisibles deviennent statistiquement rares, opérationnellement marginales, organisationnellement silencieuses. Le système ne se contente plus de mal voir certains incidents : il finit par ne plus les traiter comme des objets centraux de décision.

C'est ce qui sépare ce texte d'une réflexion de sûreté du machine learning. Le sujet n'est pas la robustesse d'un modèle, c'est la *distribution institutionnelle de la visibilité*, dont le modèle n'est que le noyau technique.

Le symptôme FDA

Un fait sert ici de symptôme, et il faut le manier comme tel. La FDA avait autorisé, début 2026, plus de 1 400 dispositifs IA/ML, dont environ trois quarts en radiologie, sur des benchmarks qui ne sont pas publiés dispositif par dispositif. Le réflexe polémique consisterait à dénoncer l'autorisation de boîtes noires ; ce n'est pas mon propos, et ce serait manquer le point.

Le problème n'est pas qu'une autorisation repose sur un benchmark incomplet. Tout benchmark l'est, on l'a concédé d'emblée. Le problème est plus précis : *l'opacité du benchmark n'est pas seulement une opacité méthodologique, elle devient une opacité sur la frontière effective de visibilité clinique du système autorisé*. Le marché et les acheteurs disposent d'un certificat de mise sur le marché ; ils ne disposent pas du périmètre des classes que le système saura escalader, qualifier, opposer. Ils achètent une performance sans connaître l'horizon de gouvernabilité qui l'accompagne.

Le pendant européen est moins commenté, et pourtant plus structurant pour les acteurs continentaux. L'AI Act exige des pratiques documentées de gouvernance et de représentativité des datasets (article 10), ainsi qu'un système de gestion des risques opérant tout au long du cycle de vie (article 9), ce qui inclut formellement la révision du périmètre en réponse aux signaux post-market. Sur le papier, ces dispositifs constituent une architecture cohérente pour le théorème de propagation et son régime de conversion. En pratique, leur implémentation reste largement à inventer, et l'industrie française et européenne y joue une partie de sa crédibilité institutionnelle. Une déclaration de représentativité qui se contente d'énumérer des proportions démographiques sans déclarer les *classes cliniques* couvertes ne satisfait pas le théorème de propagation. Elle satisfait sa version cosmétique.

Le benchmark comme port de promotion amont

Ce mécanisme prolonge une ligne doctrinale antérieure, et je l'indique brièvement, car un article ne doit pas devenir dépendant de son propre corpus pour être lisible. Dans les travaux sur le *port de promotion*, un point de passage institutionnel décide de ce qui accède au statut supérieur. Le benchmark joue exactement ce rôle pour un système clinique. Il en est le port de promotion amont.

Le principe qui en découle est simple à énoncer et exigeant à tenir : ***un système ne devrait pas être promu en production si son espace de validation ne couvre pas explicitement les classes d'incidents qu'il devra gouverner, ou s'il ne déclare pas les classes qu'il ne couvre pas et le régime de conversion par lequel il les intégrera ultérieurement***.

Le terrain donne à cette règle sa chair. Le patient BRAF V600E de la cohorte OCTOPUS n'est pas un exemple pédagogique. C'est une classe clinique précise dont la présence ou l'absence dans un périmètre décide de la capacité du système à la reconnaître ensuite. PREDICARE, sur la trajectoire de décompensation territoriale, ToxTwin sur la toxicité prédite par graphe moléculaire, posent la même question sous des formes distinctes : qu'est-ce qui, dans le périmètre, garantit que la classe critique sera vue, attribuée, gouvernée ? La réponse ne se trouve jamais dans le score agrégé. Elle se trouve dans la composition.

Trois clarifications

La thèse n'exige pas un benchmark exhaustif. Elle exige la déclaration du périmètre et la propagation honnête de cette taxonomie jusqu'au monitoring.

Elle ne prétend pas que les systèmes humains soient exempts de la même limite. Les taxonomies cliniques humaines ont leurs angles morts ; l'IA change l'échelle, la vitesse et la standardisation. Un clinicien reconstruit localement une catégorie émergente ; un système déployé propage son angle mort à la vitesse de son déploiement et avec l'uniformité de son code.

Elle ne s'exempte pas elle-même. L'identification des classes critiques absentes du benchmark dépend récursivement d'une structure de visibilité préalable, celle du concepteur ou du régulateur. La thèse déplace donc la charge de visibilité d'un cran sans prétendre l'épuiser. C'est un gradient de gouvernabilité, pas une garantie, et c'est exactement ce que les régimes de PCCP et d'article 9 AI Act sont en train d'instituer.

Trois exigences pour le décideur

Pour un décideur, acheteur, régulateur ou directeur médical, l'exigence se formule sans avis juridique, qui n'est pas mon objet. Trois demandes opérationnelles suffisent à transformer la thèse en instrument.

1. Une *déclaration de couverture par classe clinique*. Pour chaque pathologie ou sous-population pertinente, le fournisseur produit le nombre de cas inclus dans le benchmark, le protocole d'inclusion, et la performance mesurée sur cette classe spécifiquement. La moyenne agrégée devient un proxy parmi d'autres, pas la mesure principale.
2. Un *énoncé explicite des classes non testées*. Cette demande est paradoxale en apparence et essentielle en pratique. Une organisation qui sait ce qu'elle ne sait pas peut allouer son attention. Une organisation qui croit tout couvrir ne peut rien prioriser. Une déclaration honnête des angles morts est plus protectrice qu'une déclaration optimiste de couverture totale.
3. Une *clause de propagation taxonomique*. Le périmètre déclaré dans l'espace de validation se retrouve, classe par classe, dans la grille de monitoring, dans les seuils d'alerte, dans les KPI suivis, et dans les SLA contractuels. Toute discontinuité entre ces espaces est un trou de gouvernance. Le contrôle minimal consiste à vérifier que la nomenclature utilisée pour décrire l'espace de validation est identique à celle utilisée pour décrire le monitoring. Quand elles divergent, le théorème de propagation a déjà commencé son œuvre.

Ces trois exigences ne sont pas un idéal métrologique. Elles sont la traduction opérationnelle minimale du théorème de propagation pour un acheteur qui ne veut pas signer une exonération implicite.

Conclusion

Un benchmark clinique ne mesure pas seulement une performance. Il définit quelles formes d'échec existeront comme objets gouvernables pour le système, et il fixe le coût institutionnel de reconnaissance de toutes les autres.

Tant que l'industrie classera ses modèles par leur justesse moyenne, elle continuera d'optimiser ce qu'elle voit et de déployer ce qu'elle ne saura pas gouverner. Les systèmes ne deviendront pas nécessairement moins performants. Ils deviendront capables de produire des formes d'échec qu'ils ne sauront ni reconnaître, ni qualifier, ni opposer, et la responsabilité opérationnelle se dissoudra exactement là où le benchmark avait laissé un blanc.

Le problème n'est donc pas seulement statistique. Il est institutionnel.

Une question plus vaste affleure derrière celle-ci, et un autre article devra la traiter : si nos instruments de mesure finissent par sélectionner la réalité opérationnellement accessible aux systèmes, comprime-t-on, à terme, ce que nous tenons pour cliniquement réel ? On la signale. On ne la déroule pas ici.

Car un benchmark ne définit pas seulement ce qu'un système sait faire. Il définit ce qu'il saura considérer comme un problème.