Un benchmark mesure une performance, un décideur doit gouverner sa provenance

Une performance n'est pas une propriété d'un agent, mais d'une relation. Ce qui mérite d'être certifié n'est pas un score, mais le domaine dans lequel ce score conserve un sens.

Le chiffre et la question qu'il escamote

En juin 2026, le modèle MIRA affiche 87,8 % de précision diagnostique contre 78,1 % pour des médecins sur 311 cas cliniques d'urgence évalués dans un protocole comparatif publié dans Nature. Le résultat est remarquable ; sa portée reste néanmoins conditionnée au régime de données sur lequel il a été obtenu. (Nature, https://www.nature.com/articles/s41586-026-10675-5).

La même semaine, AMIE égale ou dépasse les cliniciens sur plusieurs indicateurs de prise en charge et d'adhérence aux recommandations dans des scénarios simulés de gestion longitudinale de maladies, évalués contre des médecins selon un protocole expérimental publié par Google Research et Nature (Nature / Google Research, https://blog.google/innovation-and-ai/models-and-research/google-research/amie-for-disease-management-in-nature/).

Ces résultats sont réels, mais le débat qu'ils suscitent manque souvent sa cible. Le réflexe consiste à demander si ces scores sont assez élevés, assez robustes, ou assez proches du monde réel. L'interrogation est légitime, et secondaire. La question première est plus simple : ***Que mesure exactement un score de performance ?***

Nous parlons souvent de la performance d'un système comme d'une propriété qu'il posséderait, à la manière dont un objet possède une masse ou une longueur. La façon de parler paraît naturelle ; elle est trompeuse. ***Une performance n'est pas une propriété intrinsèque d'un agent.***

Une performance est une propriété de la relation entre un agent, un régime de données et un contexte de déploiement. Le terme relation est employé ici délibérément. Une condition modifie une performance supposée préexistante ; une relation affirme au contraire que la performance n'existe qu'à travers les éléments qui la constituent.

Cette idée rejoint notamment les débats récents sur l'évaluation des systèmes d'IA, qui montrent que la signification d'une métrique dépend du cadre d'évaluation qui la produit (par exemple : "Is Your AI Model Accurate Enough? Accuracy Requirements and the EU AI

Act", 2026). ***L'exactitude n'apparaît alors plus comme une propriété intrinsèque du système mais comme une propriété conditionnée par le cadre d'évaluation qui la rend observable***. La question n'est plus seulement « quelle est la performance ? » mais « dans quel contexte cette performance conserve-t-elle une signification ? ».

L'idée n'est pas propre à l'intelligence artificielle. Elle est présente dans les sciences de la cognition distribuée et dans la rationalité écologique, plus largement dans toute discipline qui tient qu'une capacité ne se comprend pas indépendamment de l'environnement où elle s'exerce. Une heuristique n'est pas performante en soi, elle l'est relativement à une structure du monde, et une mesure de performance n'est donc jamais absolue : elle est conditionnée par le régime dans lequel elle a été produite. Le même score peut alors correspondre à des réalités radicalement différentes selon la population observée, la qualité des données, la prévalence des événements ou les conditions opérationnelles du déploiement. Le benchmark ne ment pas, il décrit fidèlement ce qu'il mesure ; le problème est qu'il ne décrit que cela.

La performance relationnelle

Posons le mot qui va revenir sans cesse. Un régime est la distribution effective des cas sur laquelle un système opère : population concernée, qualité des données, fréquence des événements, contexte d'usage. Le régime de mesure est celui du benchmark, le régime de déploiement celui du terrain, le régime de validité celui que le fournisseur affirme couvrir. Toute la question de la gouvernance se joue dans l'écart entre ces trois régimes.

La robustesse apparaît dès lors sous un autre jour. Le discours dominant la traite comme une qualité supplémentaire venant s'ajouter à la précision : un système serait précis, puis robuste. La représentation est trompeuse.

La robustesse n'est pas une propriété de plus, c'est la mesure de ce qui survit quand le régime varie.

Dit plus simplement, pour qu'aucun décideur ne décroche : un système robuste n'est pas celui qui reste performant dans l'absolu, c'est celui dont la performance varie peu lorsque les conditions changent. Le déplacement n'est pas verbal, il change la signification même des benchmarks. Un benchmark cesse d'être un verdict général sur la qualité d'un modèle pour devenir une observation locale, produite dans un régime donné ; un leaderboard cesse d'être un classement universel pour devenir une photographie partielle, dont la portée dépend du régime où elle a été établie. La question pertinente n'est plus « quel est le meilleur modèle ? » mais « quel modèle reste performant lorsque le régime change ? ».

La provenance comme hypothèse explicative

Une fois admis que la performance est relationnelle, une question surgit aussitôt : pourquoi un même score se comporte-t-il différemment quand on change de régime ? Pour y répondre, je propose le concept de provenance de la performance. Par provenance, il ne faut pas entendre une recette cachée qu'on lirait dans le modèle, mais une hypothèse explicative portant sur les mécanismes dominants qui produisent la performance observée.

Trois mécanismes pèsent particulièrement, et il faut les situer là où ils agissent sur le trajet de l'inférence.

1. L'information du cas agit à l'entrée : la décision dépend directement des caractéristiques propres à l'individu observé.
2. La structure mémorisée agit dans le modèle : la décision s'appuie sur des régularités apprises à l'entraînement, complétude et contamination comprises.
3. Le prior de population agit à l'inférence : quand l'information individuelle devient insuffisante, la prédiction est complétée par des régularités statistiques héritées de la population d'apprentissage.

Ces mécanismes ne sont ni exclusifs ni indépendants, et toute décision en mobilise plusieurs à la fois ; la question n'est donc pas de déterminer lequel est présent, mais lequel domine le comportement observé. La provenance ne prétend pas révéler l'essence d'une performance, elle cherche à identifier le mécanisme dominant qui explique son comportement quand les conditions changent. Sa justification n'est pas ontologique mais pragmatique : des mécanismes dominants différents produisent des modes de défaillance différents.

Signature observable et régime de validité

La provenance est un objet latent, et la gouvernance ne peut donc pas porter directement sur elle : pour être gouvernable, elle doit produire des conséquences observables. J'appelle signature l'ensemble de ces conséquences, qui se lisent dans le profil de dégradation sous ablation, le comportement hors distribution, la sensibilité aux données manquantes et la stabilité entre laboratoire et déploiement.

La provenance explique, la signature se mesure.

Le régime de validité forme le troisième maillon de la chaîne : le domaine dans lequel le fournisseur affirme que la performance observée conserve son sens. La relation complète se lit alors d'un trait, des mécanismes dominants à la signature observable, puis au régime de validité revendiqué. La gouvernance n'agit ni sur la cause latente ni sur la performance elle-même, mais sur les seuls éléments observables de cette chaîne.

Le contre-fait qui révèle le problème

Les travaux récents du Mass General Brigham illustrent la distinction : vingt et un grands modèles, soumis à vingt-neuf vignettes cliniques :

- Atteignent plus de 90 % de diagnostic final correct avec des données complètes,
- Mais échouent dans plus de 80 % des cas sur le raisonnement différentiel précoce quand l'information est partielle (JAMA Network Open, avril 2026 ; communiqué : https://www.massgeneralbrigham.org/en/about/newsroom/press-releases/ai-chatbot-lacks-clinical-reasoning).

Les auteurs interprètent ce résultat comme une dissociation entre connaissance et raisonnement.

Cette observation ne démontre pas directement que la performance initiale reposait surtout sur des régularités mémorisées : le lien est inféré.

Mais une inférence assumée n'est pas une faiblesse si elle se défend, et celle-ci se défend par deux propriétés. Elle explique simplement l'observation, en rendant compte d'un seul mouvement de la réussite sur cas complet et de la chute sur cas incomplet. Et elle produit des prédictions testables, donc réfutables. Une théorie n'est pas utile parce qu'elle est certaine, elle est utile parce qu'elle explique davantage avec moins d'hypothèses et qu'elle peut être démentie. On adopte donc la grille non parce qu'elle est prouvée, mais parce qu'elle est, à ce jour, la plus économe.

Ce que l'ablation mesure réellement

L'ablation est le moyen le plus direct d'observer une signature : retirer progressivement de l'information, puis regarder comment la performance évolue. Des écarts substantiels entre performances mesurées en laboratoire et performances observées en déploiement ont été documentés dans de nombreux travaux récents consacrés à la fiabilité des systèmes d'IA.

Une telle variabilité constitue précisément le type de signature que l'approche proposée cherche à rendre gouvernable. Elle est compatible avec des performances fortement dépendantes des conditions de mesure qui les ont produites. (Narayanan & Kapoor, mars 2026, https://fortune.com/2026/03/24/ai-agents-are-getting-more-capable-but-reliability-is-lagging-narayanan-kapoor/).

Mais il faut être précis sur ce que ce test mesure réellement. L'ablation ne révèle pas l'origine d'une compétence, elle révèle sa dépendance : même un excellent raisonnement clinique peut s'effondrer quand une information critique disparaît, comme un expert décroche sur une suspicion d'embolie pulmonaire si l'on ôte la saturation et le D-dimère. Elle ne désigne donc pas directement un mécanisme dominant, elle produit

une signature compatible avec certains mécanismes plutôt que d'autres. La nuance est essentielle : sans elle, le protocole devient une prétention à lire l'intérieur du système ; avec elle, il devient un instrument de gouvernance défendable, opposable en revue plutôt que réfutable d'une phrase. Eric Topol l'a noté autrement, en rappelant que ces résultats reposent sur des données propres, du texte seul, et restent préliminaires (https://erictopol.substack.com/p/agentic-ai-comes-to-medicine) : l'avertissement est juste, mais il liste des limitations là où une signature, même faillible, donne une prise.

Quand l'information devient rare

On objectera, dans la version la plus forte de l'objection, que tout cela se réglerait en élargissant la distribution de mesure : un benchmark assez large rendrait la provenance superflue, puisque le patient bénéficie du résultat, qu'il soit raisonné ou récité, et qu'on n'exige pas d'un médecin humain qu'il sépare son intuition de sa mémorisation. L'objection gagne sous une condition précise, que le régime de déploiement soit une simple extension du régime de mesure. Or il ne l'est pas : le cas rare n'est pas un cas fréquent en plus grand nombre, il est structurellement absent de tout élargissement réaliste du jeu de mesure, et la moyenne d'un benchmark plus large masque ce comportement extrême au lieu de le révéler. C'est précisément là que des mécanismes dominants différents produisent des modes de défaillance différents, et que le partage cesse d'être indifférent.

La thèse relationnelle produit alors son corollaire le plus net. Dans les régimes riches en données, les trois mécanismes tendent à converger vers la même décision et savoir lequel domine importe peu ; dans les régimes pauvres, ils divergent, et cette divergence devient décisive.

La valeur de la provenance croît à mesure que l'information du cas décroît.

La proposition est réfutable, et c'est ce qui en fait autre chose qu'une intuition : si, en régime pauvre, les trois mécanismes convergeaient encore vers la même décision, la provenance resterait indifférente et l'énoncé tomberait. Le falsifieur opérationnel est donc la mesure de cette divergence, modes de défaillance à l'appui, et non une part endogène inobservable. Reste à nommer ce régime sans jargon : il porte une étiquette technique, haute dimension et faible échantillon, mais l'idée tient en une phrase, il y a plus de variables potentiellement pertinentes que de cas disponibles pour les départager. Quand c'est le cas, aucune donnée individuelle ne suffit à trancher seule, et le prior de population reprend la main faute de mieux. Une cohorte de quelques centaines de patients porteurs d'une mutation rare, où l'on mesure des centaines de variables, est exactement cette situation : un terrain comme BRAF V600E sur 184 patients en est l'instance, et le lecteur n'a même pas besoin du terrain pour saisir le mécanisme, il lui suffit de compter les variables et les cas. Sur un tel régime, une performance héritée de la complétude d'un dataset établi ne se dégrade pas, elle s'évapore, car il n'existe rien à

se rappeler qui ressemble au cas présent. Les benchmarks naissent en régime riche ; les usages cliniques qui comptent, maladie rare, sous-population, présentation atypique, dossier incomplet, vivent en régime pauvre, et le déplacement décisif n'est donc pas le couple banal benchmark contre terrain, mais la descente le long de la courbe d'information, là où la décision est la plus exposée.

Une thèse générale sur les systèmes prédictifs

La médecine rend le phénomène visible mais ne l'épuise pas, et il serait paresseux de cantonner la thèse à la santé pour le confort de l'exemple. La même structure se retrouve partout où l'information sur le cas se raréfie : un modèle financier est jugé surtout lors des crises qu'il n'a presque jamais observées, un système de renseignement face à des menaces inédites, un jumeau de prévision industrielle dans des régimes de panne rarement rencontrés, un copilot logiciel dès qu'il quitte ses dépôts d'apprentissage pour une base de code atypique. Dans chaque cas, la même grammaire : une performance mesurée dans un régime donné, des mécanismes dominants qui gouvernent son comportement hors de ce régime, une signature observable qui trahit cette dépendance. Ces transpositions sont illustratives et non démontrées, elles montrent la généralité de la structure, pas une mesure faite dans chaque domaine. La clinique n'est pas l'objet de la thèse, elle en est le révélateur le plus visible, parce que l'écart entre régime de mesure et régime d'usage y est maximal, observable, et payé par un corps ; ailleurs il est seulement payé plus tard.

Ce qu'il faut certifier

La conséquence est directe : ce qui doit être gouverné n'est pas la performance observée, ce sont les conditions qui la rendent significative. On ne peut pas exiger d'un fournisseur qu'il démontre l'origine exacte de chaque décision, ce qui serait demander l'inobservable ; on peut en revanche exiger qu'il déclare explicitement le régime de validité, qu'il documente le comportement hors distribution, et qu'il caractérise le profil de dégradation sous perturbation contrôlée.

Cette exigence est cohérente avec l'évolution du cadre européen. Le Reflection Paper de l'EMA consacré à l'utilisation de l'intelligence artificielle dans le cycle de vie du médicament insiste déjà sur la nécessité de documenter le contexte d'utilisation, les hypothèses du modèle, ses limites et les conditions dans lesquelles ses performances demeurent valides. La conséquence réglementaire reste ici conditionnelle : il ne s'agit pas d'affirmer que l'EMA ou l'AI Act imposent aujourd'hui une telle lecture de la performance, mais de constater que leurs exigences documentaires convergent vers une explicitation croissante du domaine de validité des systèmes à haut risque.

Ces trois exigences portent toutes sur des éléments observables ou déclaratifs, jamais sur le latent.

Le déplacement est discret mais profond. Un benchmark restera toujours ce qu'il est, une mesure locale produite sur une distribution donnée, et la question décisive n'est pas de savoir si un score est élevé, mais dans quel domaine ce score conserve une signification. Gouverner, ce n'est pas se porter garant d'une performance observée, c'est se porter garant des conditions qui la rendent observable, et de l'aveu, écrit noir sur blanc, de ce qui arrive quand on en sort.

Ce qui mérite d'être certifié n'est pas la performance, c'est le domaine dans lequel cette performance continue d'avoir un sens.