La friction était la garantie

Théorie des mécanismes de garantie dans les systèmes de confiance à partir d'un cas : les bibliographies que l'IA hallucine

Avertissement de méthode. *Ce texte n'est pas un article sur les références hallucinées. C'est une théorie des mécanismes par lesquels les systèmes de confiance tiennent et de ce qui leur arrive quand l'automatisation dissout les frictions qui les tenaient. Les bibliographies que l'IA fabrique en sont la porte d'entrée : le premier de ces mécanismes dont la rupture devient mesurable. Le cas n'est pas le fondement de la théorie ; il en est l'occasion et la première preuve d'existence. Je l'assume : si l'étude qui ouvre ce texte était demain révisée, la théorie survivrait presque intacte. C'est moins un aveu qu'une revendication.*

I. La porte d'entrée

En mai 2026, une équipe dirigée par Maxim Topaz (Columbia University) a publié dans *The Lancet* un audit de près de 2,5 millions d'articles biomédicaux et de 97,1 millions de références. Un article sur 277 publié début 2026 contenait au moins une référence à un travail qui n'existe pas. *Fortune* y a vu des hallucinations qui « entrent dans le registre permanent » de la science, et le réflexe qu'elle a déclenché , « vérifions que les références existent » est exactement le mauvais.

Une référence fantôme est *détectable* : le DOI ne résout pas, la machine le constate. C'est ce qui en fait le moindre des dangers. Le risque sérieux porte des noms moins spectaculaires : référence réelle citée hors contexte, référence réelle rétractée et toujours invoquée, référence réelle issue d'une science non reproductible. *Les références hallucinées ne sont pas le sujet ; elles en sont le symptôme mesurable.*

Le symptôme révèle ceci. Le peer review vérifie la *pertinence* d'une citation, presque jamais son *existence* : il présume qu'elle existe. Cette présomption n'était pas une négligence, c'était un équilibre économique. La confiance dans l'existence d'une référence reposait *sur une présomption sociale plutôt que sur une vérification* parce qu'une asymétrie de coûts la soutenait : fabriquer une référence crédible, auteurs vraisemblables, journal réel, DOI bien formé, coûtait plus que ça ne rapportait.

L'IA a effondré ce coût sans toucher à l'autre terme. Et la récompense n'a pas baissé : un système de publication qui récompense le volume, injonction à publier, inflation

bibliométrique, carrières évaluées au comptage, fait que *fabriquer rapporte toujours*. Coût nul, récompense positive : la présomption ne pouvait pas tenir.

Le défaut n'est d'ailleurs pas nouveau. Dès 2003, Simkin et Roychowdhury, en suivant la propagation des coquilles recopiées de bibliographie en bibliographie, estimaient que *20% seulement des citeurs avaient lu l'original*. La référence était déjà un objet social qui circule par recopie, pas un pointeur vérifié. Ce que l'on prend pour une crise d'*existence* est l'aggravation visible d'une crise ancienne de *correspondance entre une affirmation et sa preuve*. Et l'on ne peut pas compter sur l'auto-correction pour l'absorber : plus de 75% des articles rétractés restent cités l'année suivant leur retrait, et la quasi-totalité de ces citations l'ignorent. *La science amortit ; elle ne corrige pas toujours.* Une erreur entrée dans le corpus y opère indéfiniment à bas bruit, exactement le régime d'une référence fantôme.

Tout cela raisonne encore dans le paradigme documentaire : une affirmation, une référence, un document qu'on pourrait ouvrir. Les systèmes qui consomment la littérature l'ont déjà quitté. Dans une génération augmentée par la recherche, *personne ne lit les références, on interroge un espace vectoriel*. L'unité de preuve cesse d'être le document pour devenir un fragment, un passage retrouvé par proximité sémantique qu'aucun DOI ne désigne. *Le pipeline cite ; il ne lit pas.* Vérifier l'existence des références devient un contrôle qui porte sur l'objet d'hier.

Voilà la porte. Elle donne sur une question plus large que les références, et c'est cette question qui fait le reste du texte : de quoi cet effondrement est-il l'instance ?

II. La règle

L'existence d'une référence appartenait à une famille qu'il faut nommer avec précaution, car la nommer mal ruine l'argument. On serait tenté d'y joindre la calibration d'un modèle, le domaine d'applicabilité d'un système, la traçabilité d'un agent, et d'appeler le tout «invariants». Mais ces objets ne sont pas de même nature : la calibration est statistique, le domaine d'applicabilité épistémique, la traçabilité opérationnelle, l'existence d'une référence documentaire. Aucun théorème ne les réunit. Les appeler d'un même nom serait confondre une intuition avec une catégorie.

La catégorie correcte ne porte pas sur ce que ces propriétés *sont*, mais sur la façon dont elles étaient *tenues*. J'appelle *invariant présumé* non un type de propriété, mais un statut: **celui d'être tenu pour satisfait sans vérification active, parce que sa violation était trop coûteuse pour être produite à grande échelle.** *L'invariant n'est pas une propriété ; c'est un statut qu'une propriété perd.* Ce qui réunit la calibration et l'existence d'une référence n'est pas une essence commune, mais un même *mode de garantie* : la friction,

non le contrôle. La catégorie est fonctionnelle, et c'est pourquoi elle survit à l'hétérogénéité de ses membres.

De là une règle, que j'énonce sans la diluer : **une propriété critique garantie par une friction cesse d'être présumable dès que l'automatisation dissout cette friction.** Sa condition de réfutation est nette, qu'on exhibe une propriété dont la friction garante a disparu et qui est pourtant restée fiable sans qu'aucun contrôle ne la remplace, et c'est cette netteté qui lui donne sa valeur. Je n'y ajoute pas la procession de qualificatifs qui la rendrait inattaquable : une thèse qu'aucune observation ne pourrait démentir n'explique rien.

Un point, en revanche, doit être tenu fermement, parce que les premières formulations de ce raisonnement le manquaient. *Le besoin de contrôle n'implique pas l'instrumentation*. Quand la friction tombe, la propriété doit être garantie autrement, mais « autrement » se décline : par la réglementation, la responsabilité juridique, la réputation, la certification, la limitation organisationnelle, ou la mesure. L'instrumentation, mesurer la propriété pour vérifier qu'elle tient, a un seul avantage propre, décisif : *elle est la seule réponse qui passe à l'échelle de l'automatisation qu'elle corrige*. C'est ce qui la rend dominante, pas supérieure. Confondre « la plus scalable » avec « la bonne » est l'erreur que cette théorie doit refuser à chaque pas.

III. Détecter l'invariant suivant

Ici la théorie doit payer son écot. Telle qu'elle vient d'être posée, elle est *rétrospective* : elle explique magnifiquement une rupture une fois qu'elle a eu lieu, et ne dit rien de la prochaine. Pour un diagnostic, c'est satisfaisant ; pour une gouvernance, c'est insuffisant, un dirigeant ne veut pas qu'on lui explique l'incendie, il veut qu'on lui désigne les pièces sèches. La question utile n'est donc pas « comment un invariant se rompt-il ? » mais « comment repérer celui qui va se rompre, avant qu'il ne le fasse ? »

Je propose une signature en trois conditions. Une propriété est un *invariant à risque* lorsque, simultanément :

1. Elle est aujourd'hui garantie *principalement par un coût* et non par un contrôle,
2. Ce coût est *en cours de dissolution* par l'automatisation,
3. Et sa violation est *encore non mesurée*, donc invisible.

Là où les trois coïncident, la rupture n'est pas certaine, mais elle est *préparée* et, surtout, elle sera invisible jusqu'au jour où quelqu'un, comme Topaz, construira l'instrument qui la rend visible.

La troisième condition est la plus traître : un invariant à risque ne fait aucun bruit avant qu'on ne le mesure, ce qui signifie que l'absence d'incident documenté n'est jamais une

preuve de santé. C'est même l'inverse : *un invariant qui ne s'est pas encore manifesté est plus dangereux que celui qui s'est déjà rompu, parce qu'il n'a pas encore d'instrument*.

Le test se laisse appliquer aujourd'hui, et le lecteur en tirera ses propres candidats. La provenance des données d'entraînement, longtemps garantie par le coût de constituer un corpus, désormais triviale à brouiller, et encore mal mesurée. L'authenticité d'une signature d'auteur, jadis garantie par l'effort d'écrire. Le caractère humain d'un évaluateur, d'un répondant à une enquête, d'un compte. L'originalité d'une contribution, garantie par la difficulté de produire du plausible. Aucun de ces exemples n'est ici démontré ; ils ne valent que comme exercices d'application de la signature, et c'est précisément l'intérêt d'une signature : qu'elle s'applique avant la preuve. La théorie cesse d'être seulement diagnostique au moment exact où elle accepte de se tromper sur des cas particuliers.

IV. Trois familles d'instruments

Si l'instrumentation est la réponse dominante, encore faut-il ne pas la réduire à la vérification. Trois familles d'instruments tiennent un système de confiance, et elles ne font pas le même travail.

Les *instruments de vérification* constatent qu'un lien tient : un résolveur de DOI, une réplication, un audit de traçabilité. Ils répondent à la question « est-ce vrai du lien ? ».

Les *instruments de légitimation* confèrent l'autorité : le label d'une agence, le sceau d'un protocole, l'acceptation par une communauté. Ils répondent à « qui a le droit de faire foi ? ».

On aurait tort d'opposer ces seconds aux instruments tout court, comme si la légitimité échappait à toute machinerie : le peer review, l'essai randomisé, la pharmacovigilance *sont* des dispositifs. Les institutions ne s'opposent pas aux instruments, *elles en fabriquent*. La distinction n'est pas entre instrument et institution, mais entre *vérifier* et *adouber*.

Il manque une troisième famille, et son absence serait coupable pour qui travaille sur les ontologies et les systèmes de preuve : les *instruments de coordination*.

Les standards, les nomenclatures, les ontologies, les formats réglementaires, les protocoles ne vérifient rien et n'adoubent personne ; ils permettent à des acteurs hétérogènes d'*agir sur les mêmes objets*. Un identifiant pérenne, une terminologie partagée, un schéma d'échange ne disent ni le vrai ni l'autorisé, ils rendent le vrai *vérifiable par plusieurs* et l'autorisé *transférable d'une institution à l'autre*.

Vérifier, légitimer, coordonner : trois manières de tenir ensemble ce qu'aucune ne tient seule. La gouvernance d'un invariant rompu mobilise les trois, et une politique qui n'instrumente que la vérification, qui se contente de compter les DOI fantômes laisse les deux autres à découvert.

V. L'intégrité du lien, et ses deux bords

La famille de la vérification, appliquée au cas qui nous a servi de porte, se déploie en niveaux. On peut poser l'*intégrité du lien probatoire,* non de la preuve en général, en cinq degrés, ordonnés par visibilité décroissante, ce qui est le piège :

1. **E** (la référence désigne-t-elle un travail réel ?),
2. **Identité** (est-ce le bon travail, intact, non rétracté, question devenue ontologique quand préprints, versions et synthèses génératives brouillent *l'instance canonique d'une connaissance*),
3. **Pertinence** (la référence soutient-elle l'affirmation, ou seulement son thème ? Probablement le cœur du risque),
4. **Traçabilité documentaire** (peut-on remonter la chaîne ?),
5. **Traçabilité computationnelle** (peut-on relier une conclusion synthétisée aux fragments qui l'ont produite ?).

Le niveau 1 est visible parce qu'il est simple ; les suivants portent l'essentiel du risque et n'ont pas encore d'instrument à l'échelle. Cette grille n'a aucune prétention à l'exhaustivité : elle nomme les défaillances observables aujourd'hui.

Mais l'intégrité du lien n'est pas tout, et il faut nommer ce qu'elle ne touche pas. *Premier bord, la légitimité* : une référence ne devient preuve décisionnelle que parce qu'une institution l'accepte ; aucun instrument de vérification ne produit cette acceptation, il déplace seulement la question vers ceux qui décident quels instruments font foi. *Second bord, la validité* : une preuve peut franchir les cinq niveaux et rester *fausse*. Quand l'Open Science Collaboration n'a retrouvé que 36 % de réplications significatives là où 97 % des études originales l'étaient (chiffre lui-même disputé, ce qui dit l'état du terrain) elle ne mesurait pas un défaut de citation mais un défaut de vérité. *Vérifier que la preuve est bien reliée ne dit rien de ce qu'elle vaut.* Ces deux bords ne sont pas des oublis de la grille ; ils marquent la frontière au-delà de laquelle l'instrument de vérification n'a, par construction, aucune prise.

VI. Le problème n'a pas d'état stable

Reste la conséquence que les versions prudentes de ce raisonnement traitaient comme une réserve, et qui est en réalité son aboutissement. Si l'invariant *présumé* est fragile (il se rompt dès que sa friction tombe) et si l'invariant *instrumenté* devient manipulable (car une propriété mesurée et érigée en cible tombe sous la loi de Goodhart, et sous la loi de Campbell quand la cible sert à décider) alors la question n'est plus « comment réparer ? » mais « un état réparé existe-t-il ? ». Je soutiens qu'il n'existe pas, et je le soutiens comme thèse, non comme prudence.

Instrumenter la vérification d'existence des références produira des pipelines qui optimisent le taux de DOI résolus, et donc, à terme, des fabrications conçues pour le passer : des références qui existent, pointent vers un document réel, et ne soutiennent rien. *Le coût garantissait en silence ; l'instrument garantit en bruit et le bruit, on apprend à le produire*. L'instrument ne restaure pas la garantie perdue ; il déplace le mode de défaillance d'un cran (de la *présomption silencieuse* vers la *métrique manipulable*) et ce nouveau mode devient le prochain invariant à surveiller. La structure est cybernétique, et il faut l'accepter telle quelle : *tout mécanisme de contrôle engendre l'espace de son contournement*. Un système de confiance n'atteint pas un équilibre ; il *migre d'une vulnérabilité à la suivante*, et chaque instrument qui en ferme une en ouvre une autre, d'ordinaire moins visible parce que plus récente.

C'est la vulnérabilité que ce texte assume plutôt que de la colmater, parce qu'elle est sa thèse la plus utile. Elle interdit la promesse confortable, « instrumentons l'intégrité de la preuve et le problème sera résolu », et impose une posture différente : la gouvernance n'est pas un état terminal à atteindre mais une *maintenance sans terme*. On ne sécurise pas un système de confiance ; on l'entretient contre une dérive qui reprend dès qu'on cesse de lutter contre. Je ne peux pas prouver qu'aucun état stable n'existe (c'est l'honnête limite de la thèse). Je peux la mettre au défi : qu'on me montre un seul mécanisme de confiance qui, une fois instrumenté, ait cessé d'engendrer de nouveaux contournements. Je n'en connais pas.

VII. Conclusion

La bibliographie hallucinée n'était pas le sujet. Elle a été la porte : le premier endroit où une friction qui garantissait silencieusement une propriété critique s'est effondrée assez vite, et assez mesurablement, pour qu'on puisse en faire un chiffre.

Derrière cette porte, il n'y a pas un problème de citations, ni même un problème de preuve. Il y a un fait plus général, dont les références ne sont qu'une instance précoce : *l'automatisation dissout, dans de nombreux systèmes à la fois, les coûts qui*

garantissaient sans bruit des propriétés que personne n'avait jamais eu à vérifier. La calibration, la provenance, l'authenticité, l'autorité, l'existence d'une référence, autant de propriétés tenues pour acquises parce que les violer coûtait cher, et qui cessent de l'être à mesure que les violer ne coûte plus rien.

Le travail qui vient n'est pas de restaurer ces garanties (on ne restaure pas une friction disparue). Il est de choisir, pour chacune, comment la remplacer : par la mesure quand elle passe à l'échelle, par l'institution quand il faut adouber, par le standard quand il faut coordonner en sachant que chaque remplacement déplacera le problème sans le clore.

L'IA n'a pas révélé un problème de génération, ni même de vérification. Elle a révélé que la friction était la garantie et qu'elle disparaît. Ce qui reste à construire n'est pas un système qui n'a plus de faille. C'est une discipline qui sait que la prochaine est déjà en train de se former, sans bruit, là où plus rien ne coûte assez cher pour la rendre visible.

Sources

- Topaz M. et al., « Fabricated citations: an audit across 2·5 million biomedical papers », *The Lancet*, mai 2026. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(26)00603-3/fulltext
- « One in 277 PubMed-indexed papers in 2026 shows fabricated references », *Retraction Watch*, 2026-05-07. https://retractionwatch.com/2026/05/07/one-in-277-pubmed-indexed-papers-in-2026-shows-fabricated-references-says-analysis/
- « Study finds explosion of fraudulent AI citations in academic papers », *STAT News*, 2026-05-07. https://www.statnews.com/2026/05/07/lancet-study-finds-steep-rise-fraudulent-citations-academic-papers/
- « AI hallucinations are slipping past experts into papers and books to enter the permanent record », *Fortune*, 2026-05-24. https://fortune.com/2026/05/24/ai-hallucinations-scientific-research-authors-medical-journal-treatment/
- Simkin M.V., Roychowdhury V.P., « Read before you cite! » (≈20 % des citeurs lisent l'original), 2003/2005. https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2006.00202.x
- « Continued use of retracted papers: Temporal trends in citations and (lack of) awareness of retractions in biomedicine », *Quantitative Science Studies*, MIT Press. https://direct.mit.edu/qss/article/2/4/1144/107356/Continued-use-of-retracted-papers-Temporal-trends
- Open Science Collaboration, « Estimating the reproducibility of psychological science », *Science*, 2015. https://pubmed.ncbi.nlm.nih.gov/26315443/