Le Lissage Épistémique - L'IA Face au Risque de Dégénérescence Habsbourgeoise

22 February 2026 • 5 min de lecture

LLM Learning

(co-rédigé avec une IA lucide, qui n'a pas encore l'air de Charles II ... mais on surveille de près)

> *À l'époque des Habsbourg, les unions consanguines ont produit des portraits de rois aux mâchoires proéminentes, aux fronts fuyants et aux regards vides – une dégénérescence progressive, pas une fatalité*

> *biologique, mais le résultat d'un mauvais choix de reproduction. Est-ce qu'aujourd'hui, les LLM risquent un destin analogue ? Un lissage épistémique qui efface les aspérités, les queues de distribution rares, les idées marginales ou créatives, pour converger vers une moyenne plate, homogène et stérile.*

On parle beaucoup de model collapse depuis l'article canonique de Shumailov et al. (Nature 2024) : quand on entraîne récursivement sur des données synthétiques non filtrées, les modèles perdent les événements rares, amplifient les modes dominants et finissent par produire des caricatures appauvries de la réalité initiale. Dohmatob et al. (Strong Model Collapse) montrent que l'introduction d'une fraction même infime de données synthétiques (≈ 1 %) suffit à mener à un model collapse.

En 2026, le web est déjà pollué. Certaines estimations placent le contenu généré par IA autour de 50 % des nouveaux textes indexés (Graphite Study , Ahrefs Study); loin des 90 % prophétisés dès 2022 par Europol ou Nina Schick, mais assez pour que le risque soit réel et croissant. Le vieux web (archives, forums pré-2022, PDF scannés) reste majoritairement humain, mais le flux quotidien ? De plus en plus synthétique, SEO-isé, compressé pour l'attention.

Pourtant, les frontier models continuent de gagner des points ELO (arena.ai). Les classements montent encore début 2026, surtout sur maths dures, code long, raisonnement multi-étapes. Le collapse total n'est pas là; il reste confiné aux environnements jouets, aux runs récursifs mal contrôlés (3-5 itérations sans garde-fou), ou aux domaines ouverts non structurés. Les mitigations fonctionnent quand on les applique sérieusement.

Les garde-fous sont connus:

- Mélange permanent avec 5–20 % de données humaines fraîches suffit souvent à stabiliser.
- Filtrage agressif des outputs synthétiques : TCE (Truncated Cross-Entropy), ForTIFAI-style rejection sampling, ou self-verification prouvée

théoriquement empêchent le collapse même en régime majoritairement synthétique (ForTIFAI: Fending Off Recursive Training Induced Failure for AI Models, Self-Verification Provably Prevents Model Collapse in Recursive Synthetic Training).

- Guidance négative (entraîner sur des « mauvais » exemples) ou reward models qui pénalisent la platitude (Self-Improving diffusion models with synthetic data).
- Self-play variationnel + RL from verifiable rewards (RLVR) explose en maths, code, robotique (Pass@1: Self-Play with variational problem synthesis sustains RLVR)

Ces techniques fonctionnent dans des domaines structurés ou sur des itérations limitées. Pour le langage naturel ouvert où la nuance sociale, l'humour fin, la diversité culturelle non-anglocentrée sont reines on reste limité : les alternatives produisent encore des artefacts scolaires, sur-stylisés, ou convergent vers des modes "corporate"

Le vrai danger n'est pas uniquement le collapse statistique pur. C'est le lissage culturel déjà en cours sur le web : clickbait, thread de 280 caractères, contenu SEO compressé, homogénéisation massive. Les modèles apprennent cette moyenne humaine déjà appauvrie et l'amplifient. RLHF/RLAIF/Reward modeling accentue le phénomène : les capacités latentes subsistent, mais elles sont masquées par une couche de complaisance, de sycophancie, de tiédeur sur les sujets ambigus ou polarisants. Le collapse statistique supprime les queues et nous nous retrouvons avec des sorties plates et répétitives. La sycophancie vient du reward hacking humain, pas directement de la récursion.

Les signaux faibles à surveiller :

- Perte subtile de diversité créative : les nouvelles métaphores se raréfient.
- Dégradation des queues : les erreurs créatives remplacées par des erreurs plates et redondantes.
- Reward model symptoms : une politesse dégoulinante, divergence vs préférences humaines fraîches.

Il n'y a pas de fatalité technologique. Le collapse habsbourgeois n'était pas une loi de la nature ; c'était un choix social désastreux. Le model collapse n'est pas inéluctable mais il n'est pas non plus un épouvantail qu'on balaie d'un revers de main. Aujourd'hui on n'est pas à Charles II, pas de gen-HAI (generative Habsbourg AI), les frontier labs et les verticals exigeants montrent déjà que ce risque est maîtrisable : avec une curation agressive, un mélange strict humain-synthétique et des garde-fous comme la self-verification, on repousse le lissage bien au-delà des horizons actuels. Le vrai défi n'est pas technologique, mais organisationnel : investir dans la qualité des données plutôt que dans la quantité brute; cela pourrait même être un business modèle ad hoc pour les gros créateurs de modèles fondation ou des points de valorisation .

Lego III - Epistemic

Les opinions exprimées dans cet article sont strictement personnelles et ne reflètent pas nécessairement celles de mon employeur ou de toute entité affiliée. Ce contenu est fourni à titre informatif et exploratoire uniquement. Il ne constitue pas un conseil professionnel, technique, juridique ou d'investissement. L'article discute de tendances émergentes en IA, de risques théoriques et pratiques liés à l'entraînement des modèles, et n'engage aucune promesse ou garantie sur l'évolution future des technologies d'IA.

Les opinions exprimées sur ce site sont strictement personnelles et ne reflètent pas nécessairement celles de mon employeur. Les contenus sont fournis à titre informatif et ne constituent pas un conseil juridique.