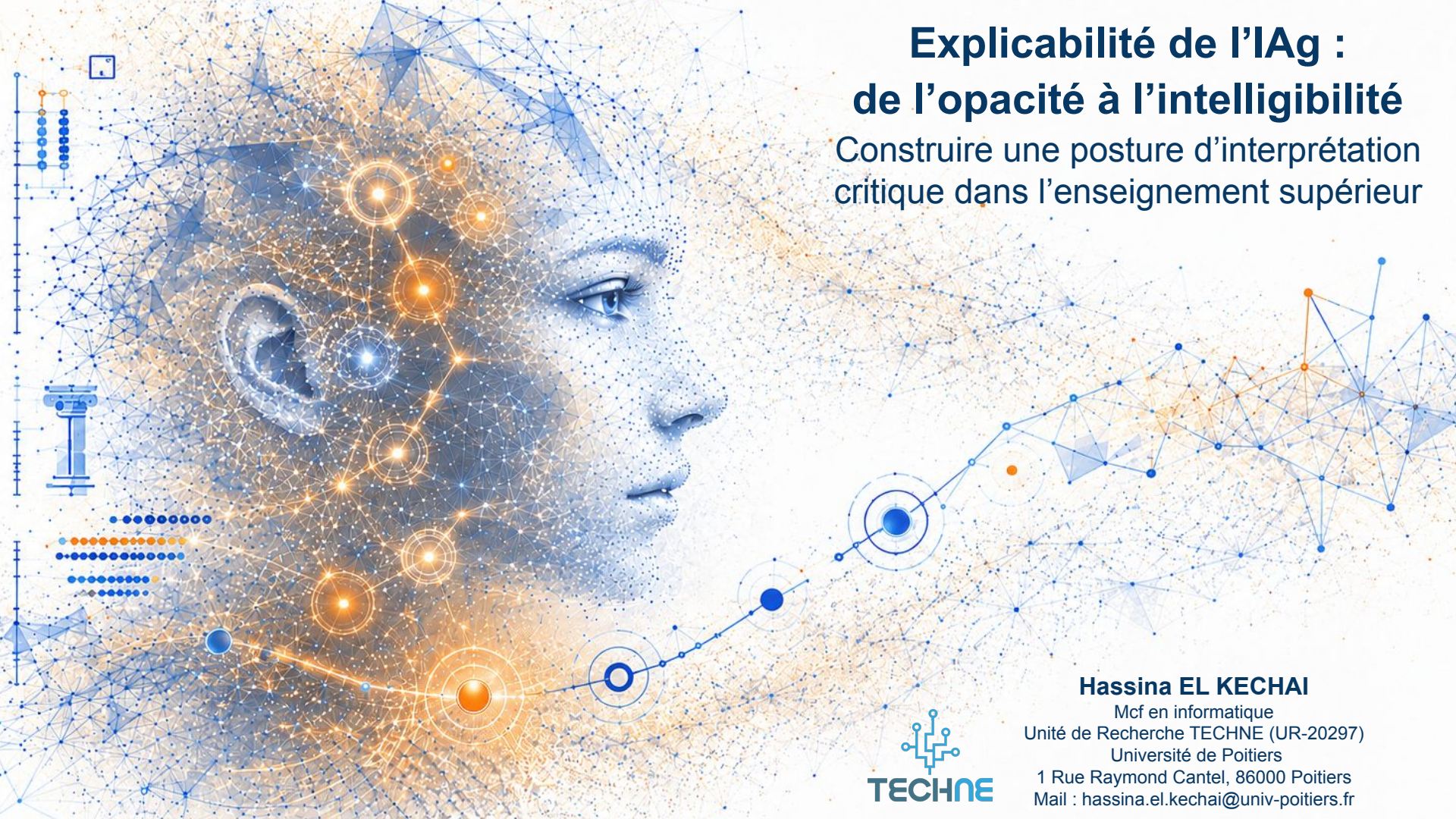


Explicabilité de l'IAg : de l'opacité à l'intelligibilité

Construire une posture d'interprétation critique dans l'enseignement supérieur

Hassina EL KECHAI
Mcf en informatique
Unité de Recherche TECHNE (UR-20297)
Université de Poitiers
1 Rue Raymond Cantel, 86000 Poitiers
Mail : hassina.el.kechai@univ-poitiers.fr

Pourquoi rendre l’IAg intelligible dans l’enseignement supérieur ?

1. Usage avant compréhension

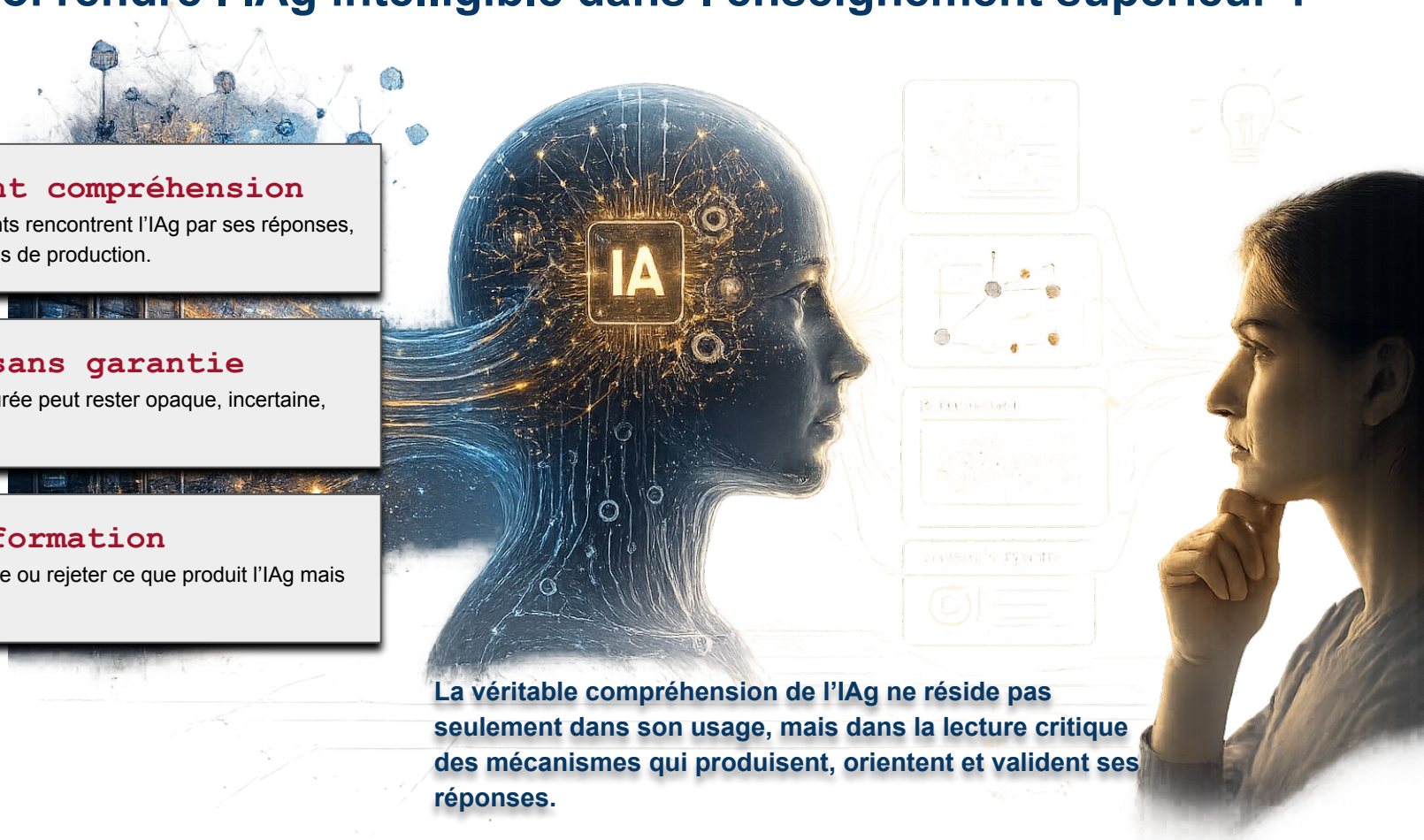
Les étudiants et enseignants rencontrent l’IAg par ses réponses, rarement par ses conditions de production.

2. Fluidité sans garantie

Une réponse claire et assurée peut rester opaque, incertaine, biaisée ou non vérifiée.

3. Enjeu de formation

L’objectif n’est pas de croire ou rejeter ce que produit l’IAg mais d’apprendre à l’interpréter.



La véritable compréhension de l’IAg ne réside pas seulement dans son usage, mais dans la lecture critique des mécanismes qui produisent, orientent et valident ses réponses.

Objectif : Passer d’une confiance intuitive à une vigilance outillée grâce à l’explicabilité

Positionner l'explicabilité critique de l'IAg dans cette présentation

- Les approches classiques de la XAI visent surtout à rendre les modèles plus transparents basées notamment sur des traces calculatoires.

- Les critiques contemporaines montrent que ce type de XAI bien qu'utile n'est pas à la portée de la compréhension humaine, ni l'interprétation critique.

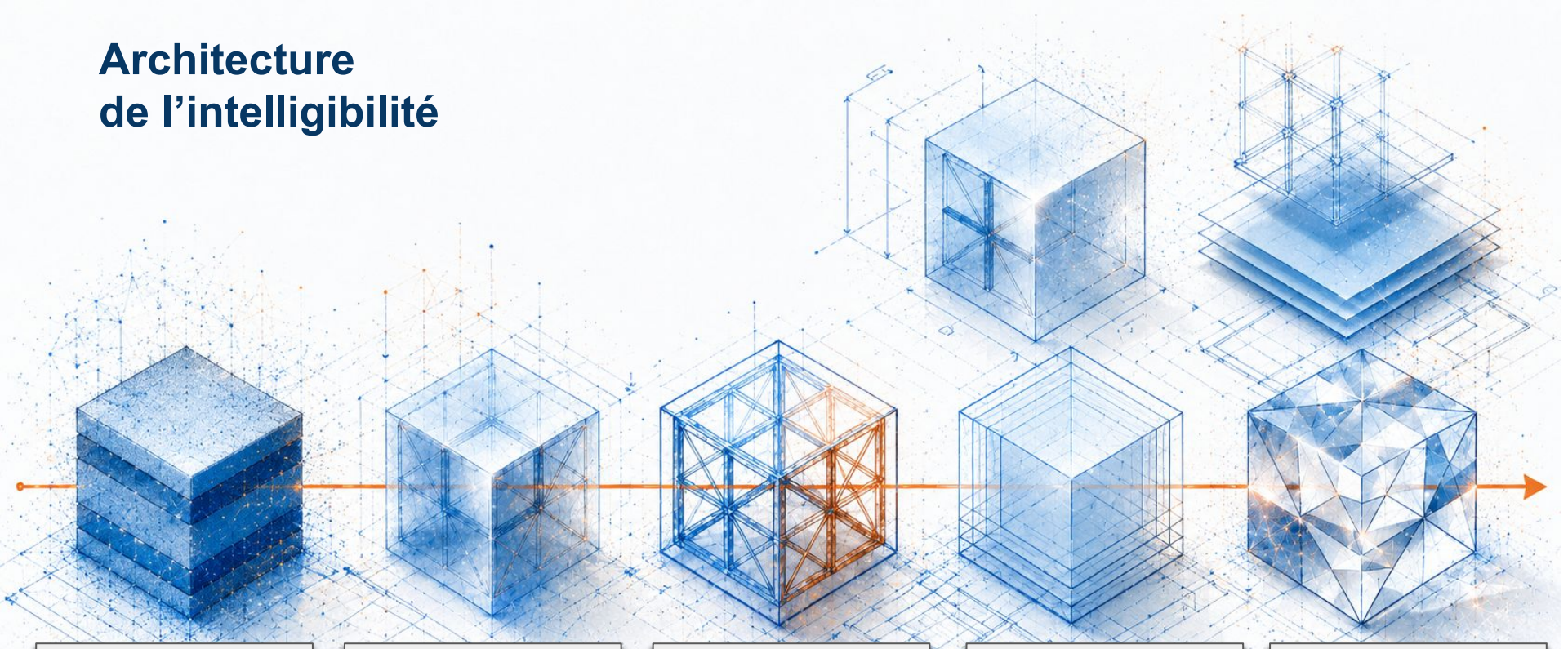
- Dans la perspective de Tim Miller (2019), l'explicabilité ne relève pas seulement d'un problème computationnel, mais d'une interaction humain–machine.

- c'est une réponse située, sélective, contrastive et sociale, construite pour permettre à un humain de comprendre et d'agir.

- L'explicabilité devient alors un artefact interprétable :
  - non pour révéler totalement le fonctionnement interne du modèle,
  - mais pour permettre une enquête critique sur les conditions de production, d'interaction et de validation des réponses générées.

Miller, T. *Explanation in Artificial Intelligence: Insights from the Social Sciences*

Architecture de l’intelligibilité



01

L’illusion d’intelligence .

Pourquoi avons-nous l’impression que l’IAg comprend ?

02

Explicabilité structurelle.

Comment l’IAg produit-elle ses réponses ?

03

Explicabilité interactionnelle.

Pourquoi les réponses varient-elles ?

04

Explicabilité cognitive.

Comment l’IAg semble raisonner et que vaut un raisonnement généré ?

05

L’agentivité humaine

Comment former une posture critique ?

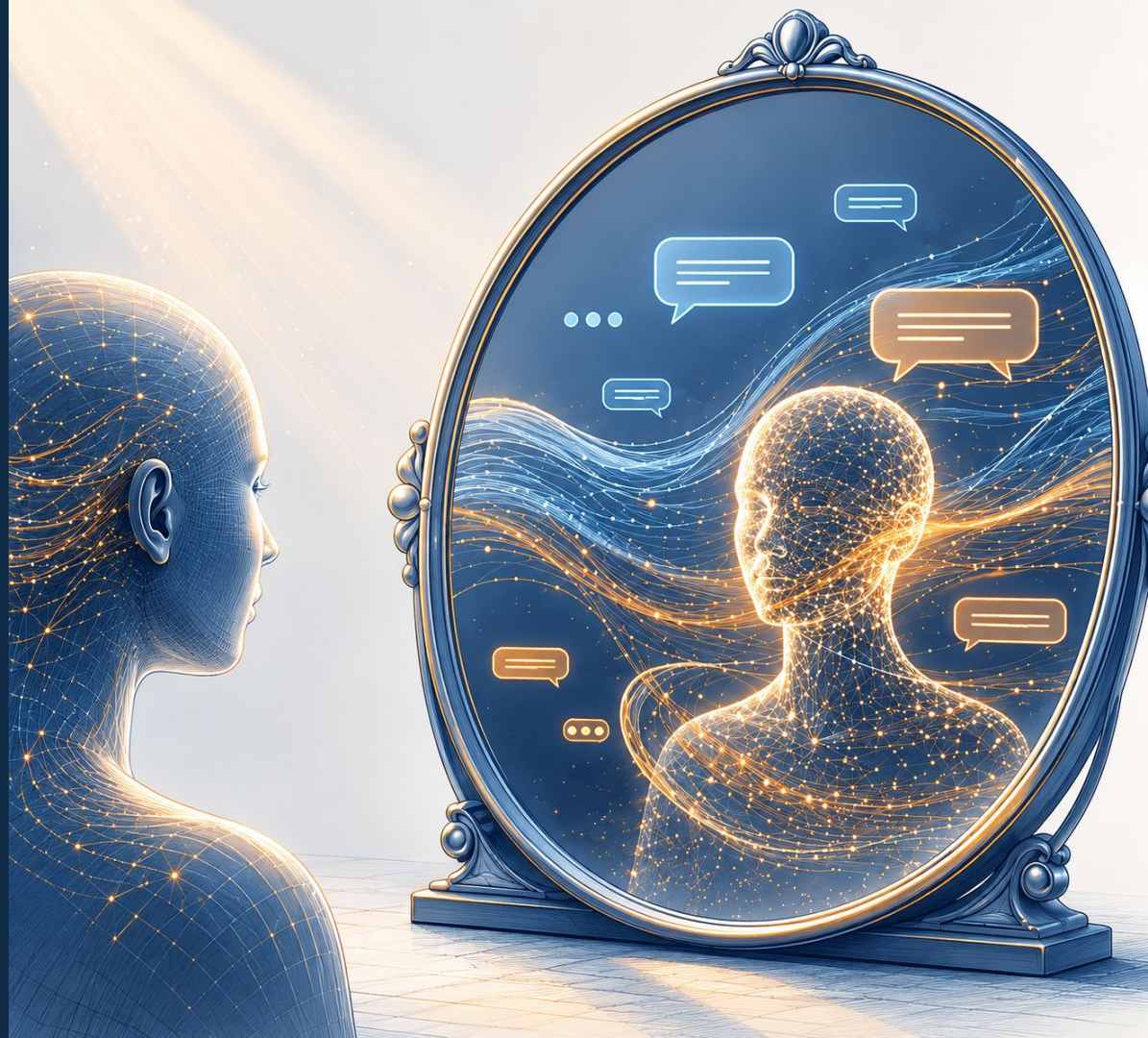
01

Le miroir : L'illusion d’intelligence

“Une réponse cohérente est-elle pour autant une réponse compréhensible ? ”

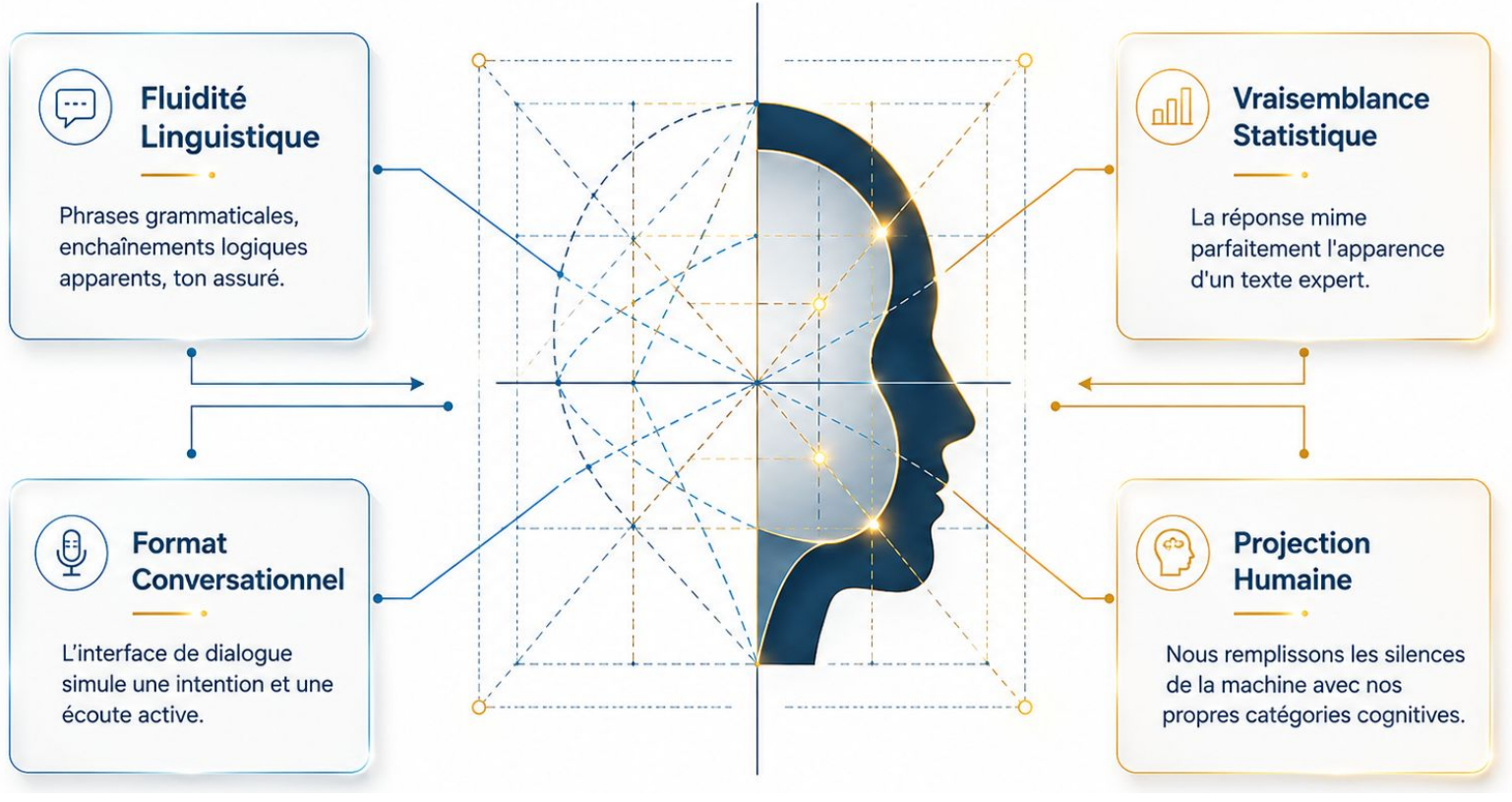
La fluidité conversationnelle produit une impression de compréhension, mais elle masque une opacité structurelle profonde.

Performance ≠ Intelligibilité.



L’anatomie d’une illusion socio-technique

L’illusion nait de la rencontre entre probabilité mathématique et désir humain de sens

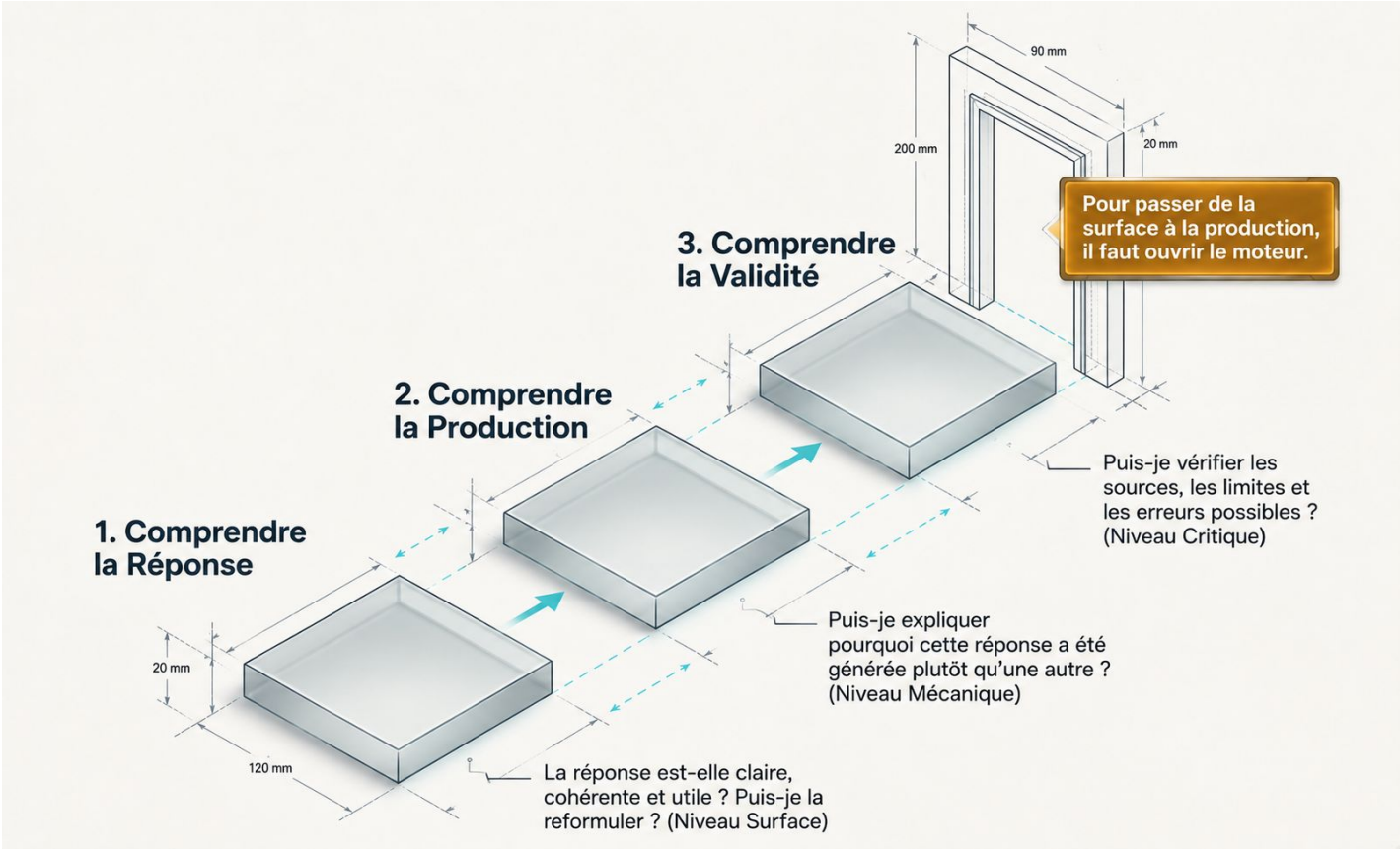


L’illusion naît de la rencontre entre la probabilité mathématique et notre désir de sens.

point critique : l’impression de compréhension est produite autant par nous que par la machine

3 niveaux d’interrogation face à la boîte noire

Une réponse peut être claire sans que son processus ni sa validité soient compris



Pour dépasser la surface, il faut ouvrir la machine de production

La phase cachée de la machine de production

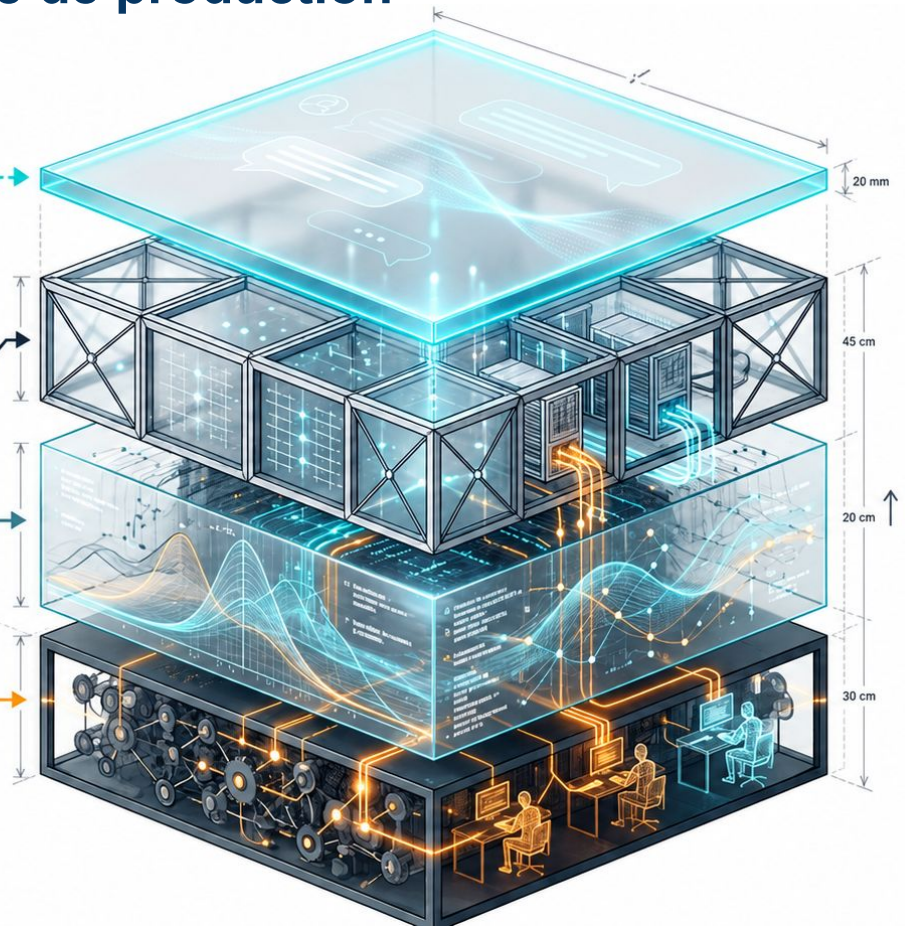
Entre l'interface visible et la réponse générée, il y a une chaîne socio-technique

L'Interface
Fluidité conversationnelle et apparence de conscience

L'Infrastructure (Base matérielle)
Fermes de GPU, puissance de calcul massive, impact environnemental

Les Données (Carburant)
Corpus massifs aspirés sur Internet, choix éditoriaux invisibles, biais encodés

Le Travail Humain
L'invisibilisation des « clic-workers » (annotateurs, modérateurs, RLHF) qui corrigent et alignent le modèle.



Ouvrir la boite noire, c'est relier la surface aux conditions matérielles, aux données et aux conditions humaines

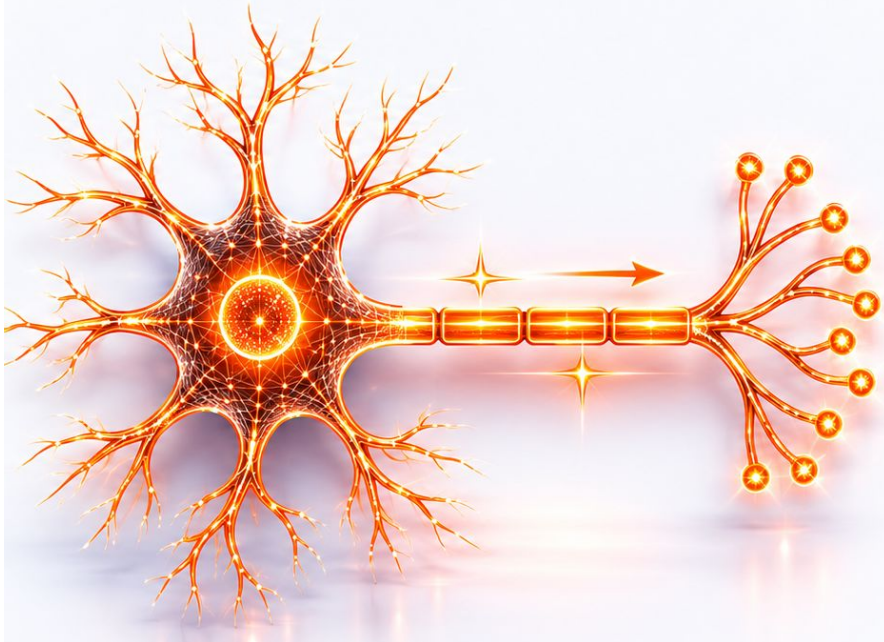
02

Les engrenages : Explicabilité structurelle

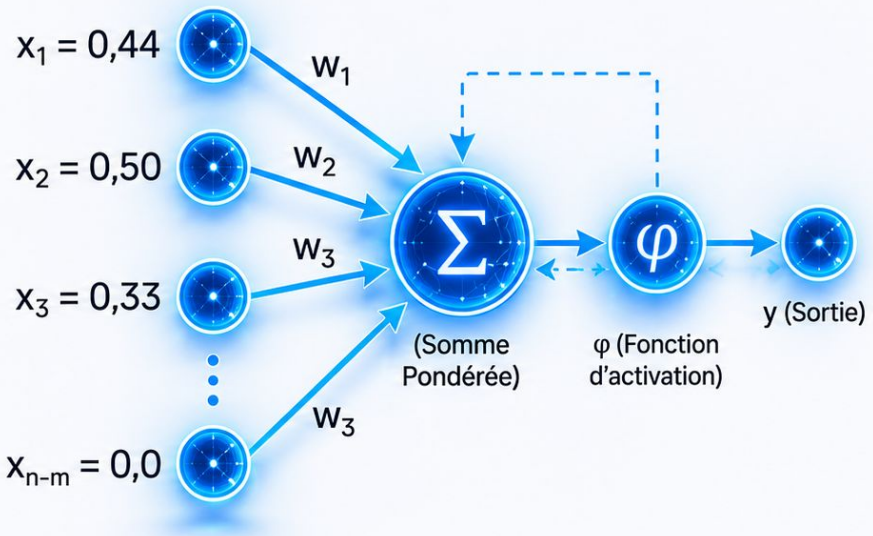
“Les notions techniques deviennent des instruments d’intelligibilité.”

Comprendre comment l’IAg produit en transformant le langage en espace mathématique

L'héritage connexionniste : du neurone biologique au calcul de probabilités



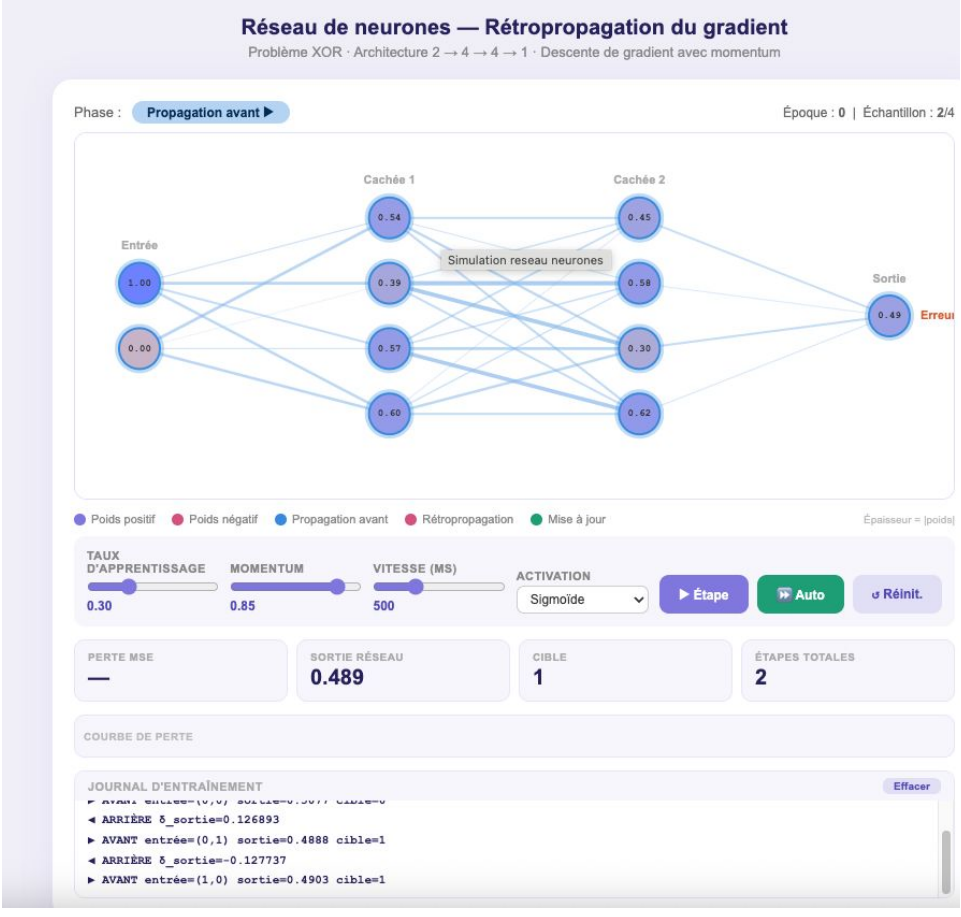
Plasticité, connexions synaptiques, adaptation par l'expérience



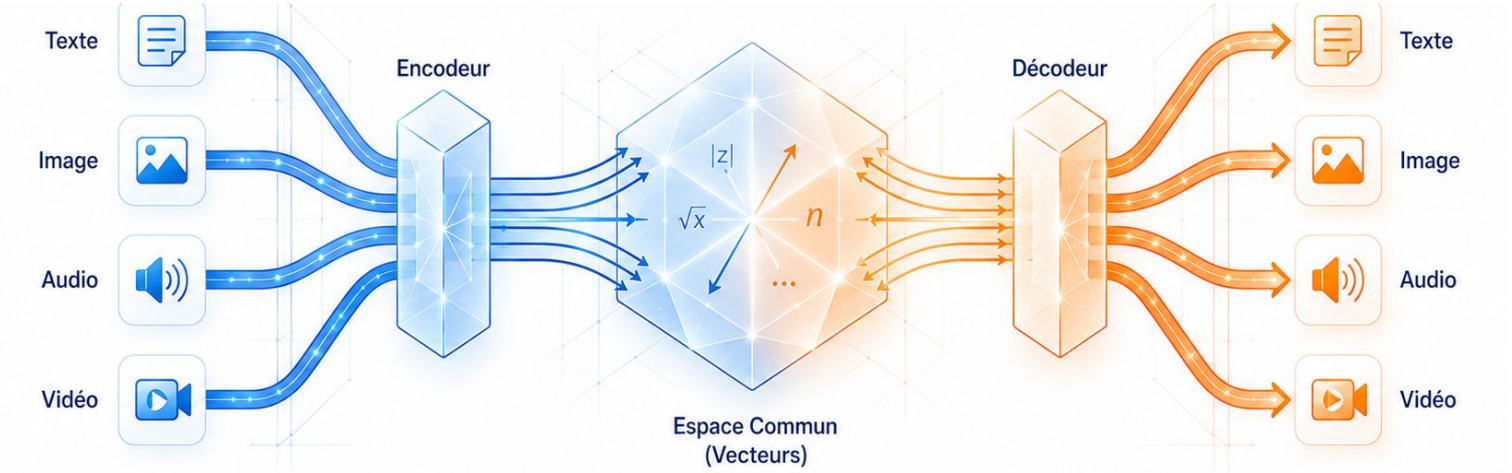
poids synaptiques (valeurs mathématiques ajustées) & rétropropagation (algorithme de correction des erreurs)

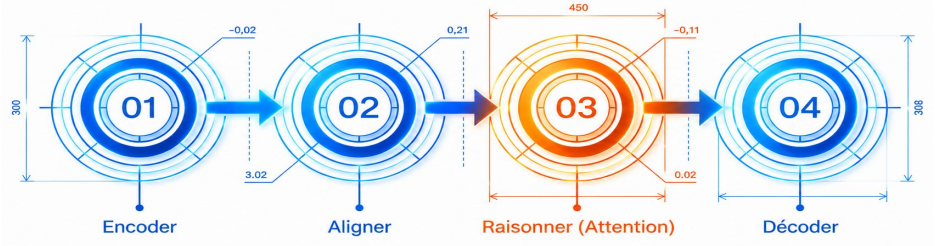
Le changement de paradigme : l'IA connexionniste ne stocke pas de bases de données de faits explicites (IA symbolique) ; elle apprend des motifs et des régularités statistiques. Nous sommes passés du connexionnisme aux architectures génératives contemporaines.

Comment le réseau de neurones apprend par erreur : simulation de la plasticité cérébrale



Le prisme multimodal : la convergence de tous les médias vers un langage mathématique unique



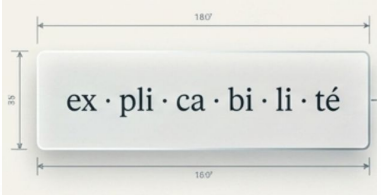


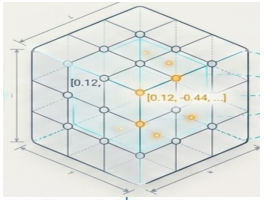
Encoder	Aligner	Raisonner (Attention)	Décoder
Transformer la requête (texte, image, audio) en vecteurs numériques	Projeter ces vecteurs dans un espace latent commun	Pondérer l'importance de chaque token par rapport au contexte global (le mécanisme d'attention)	Prédire et générer le prochain élément (token par token) selon la plus forte probabilité conditionnelle

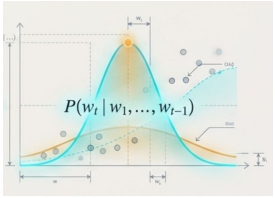
L'architecture sous-jacente est universelle : Tout n'est que chiffres

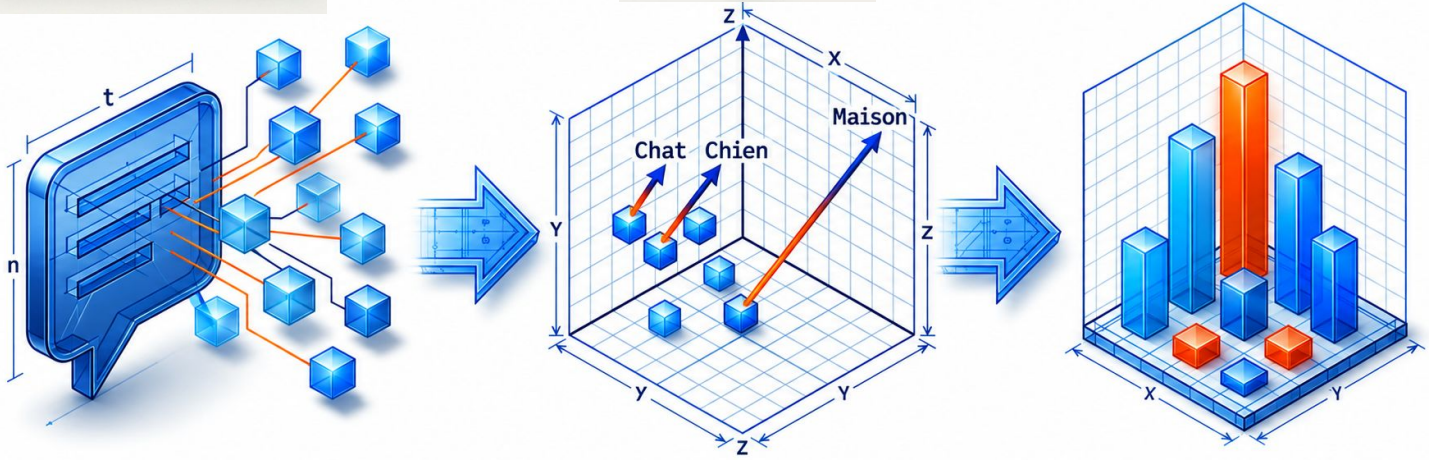
Le sens réduit à une proximité statistique dans un espace multidimensionnel

Tokens, embeddings et probabilités donnent une entrée accessible dans l'explicabilité structurelle









Texte -> Tokens

Tokenisation : Le texte est fragmenté en unités mathématiques (mots, syllabes).

Embeddings

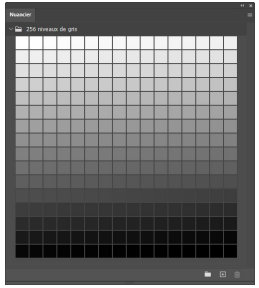
Proximité sémantique : Chaque token est projeté dans un espace multidimensionnel. Le 'sens' est une distance spatiale.

Probabilités

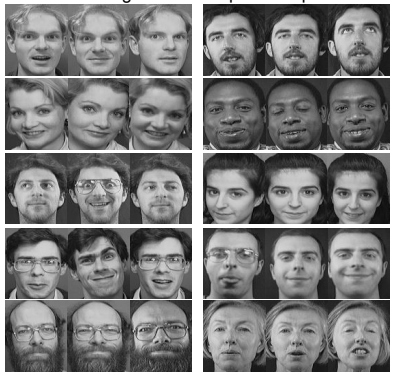
Calcul : Choix du token suivant le plus plausible. P(wt | w1, ..., wt-1).

Avertissement épistémologique : le modèle ne manipule jamais d'idées. Il calcule des relations spatiales entre des unités numériques.

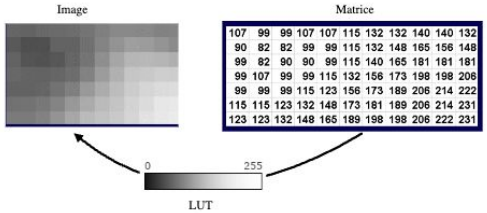
Exemple concret sur la vectorisation d’image et simulation de classification

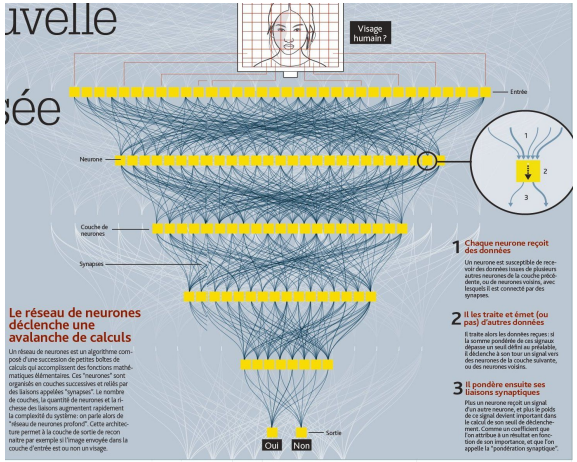


Source de l’image : nuancier photoshop

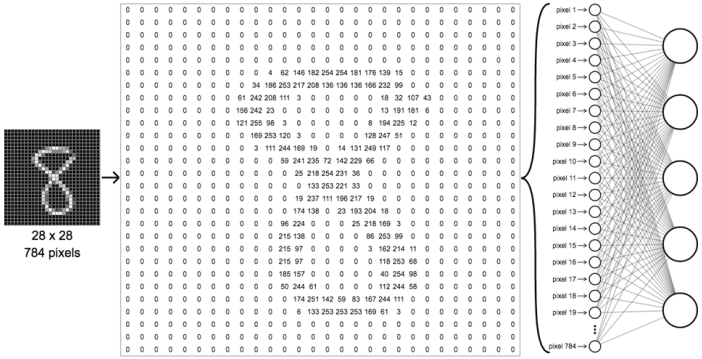


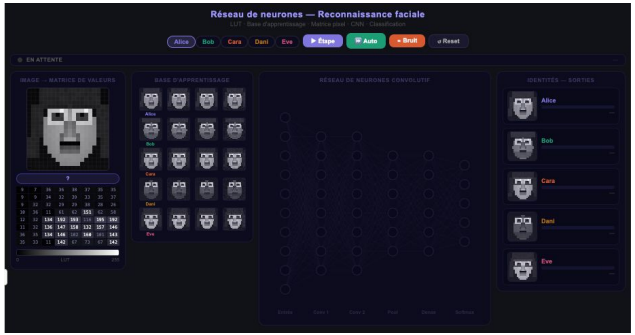
Base d’apprentissage





Source de l’image :
https://science-et-vie.com/article-magazine/une-nouvelle-forme-de-pensee





Le côté obscur de la force statistique : L'IAg est un miroir probabiliste de nos prores données

La force statistique de l'IAg est aussi sa principale fragilité

La logique de l'hallucination

L'IA poursuit une continuité statistique. Elle produit une suite plausible, imitant les codes du discours fiable (ton assertif, structure logique), sans aucune capacité de vérification externe

L'hallucination n'est ni un caprice ni un bug, c'est l'essence même de la génération probabiliste.

Les biais sont le résultat de l'encodage de données d'entrainement qui ne sont pas intègres ⇒ exemple

Reproduction des Biais

L'IA nourrtie aux corpus historiques reproduit intrinsèquement les biais sexistes, racistes ou complotistes présents dans ses données d'entraînement.

Affabulations (Hallucinations)

Face à un manque de données, le LLM ne dit pas « je ne sais pas ». Il invente statistiquement la suite la plus probable, produisant des faits erronés avec aplomb.

La Question du Droit

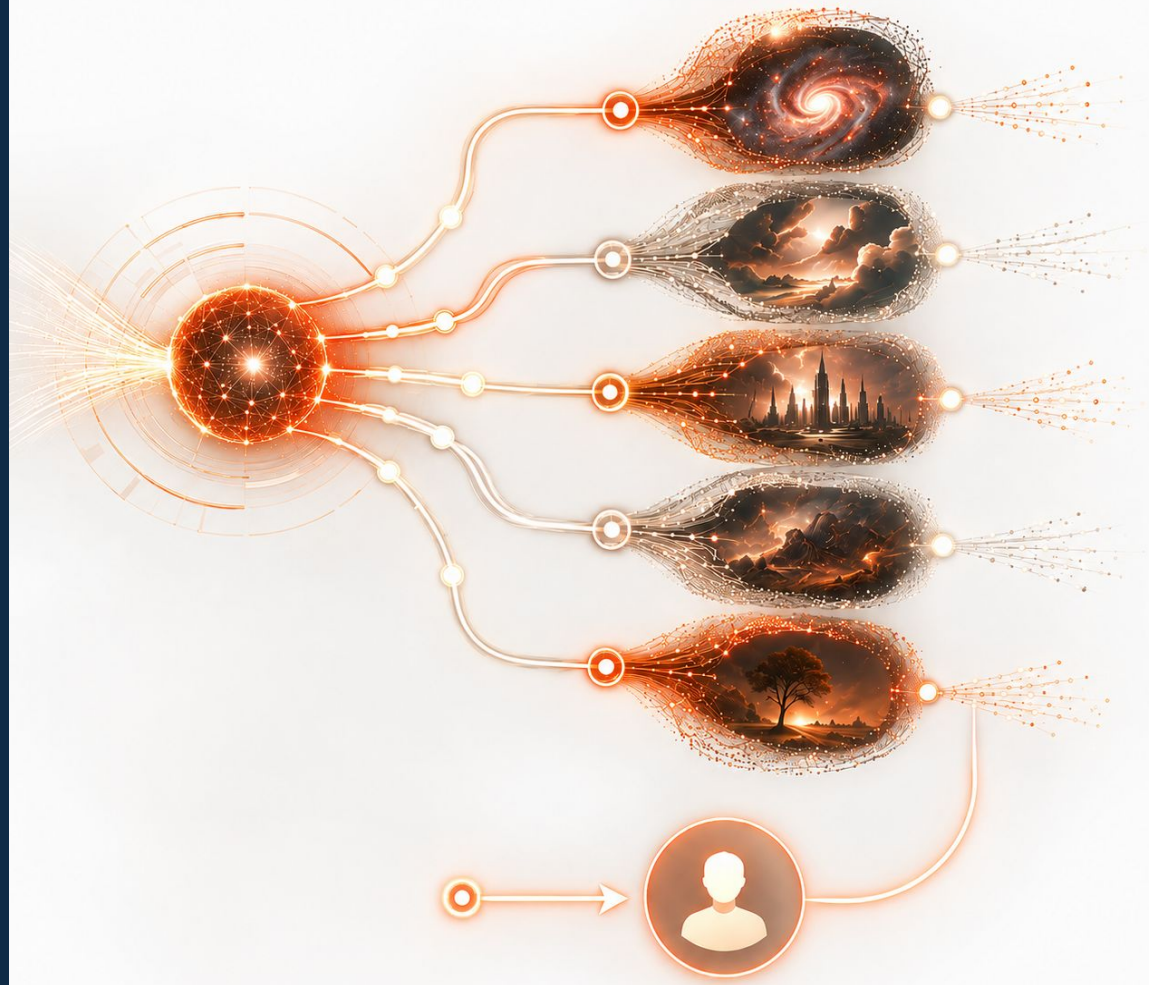
L'ingestion non consentie d'œuvres sous droits d'auteur pose le problème fondamental de la propriété intellectuelle et du plagiat automatisé.

03

Le volant : Explicabilité interactionnelle

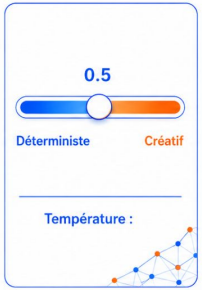
“La réponse n’est pas seulement produite par le modèle, elle émane d’une configuration d’interaction”

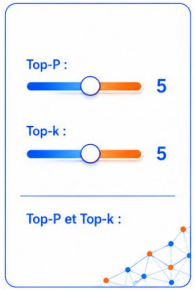
Architecture du prompt, contexte et paramètres de performance du LLM



Le tableau de bord du LLM : les dimensions de la performance interactionnelle

Ce que le modèle peut traiter : Quelles sont ses capacités et ses contraintes ?





Ce que l'utilisateur configure : Comment oriente-t-il la réponse ?

Ce que le système ajoute : Comment construit-il le contexte de travail ?

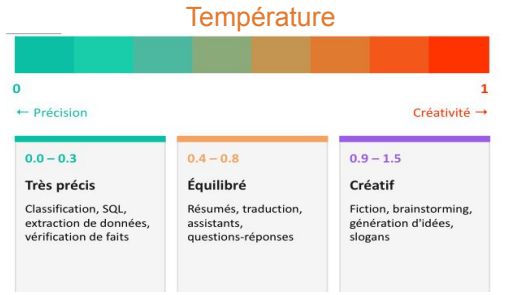
Prompt et réglage de génération : qualité interactionnelle

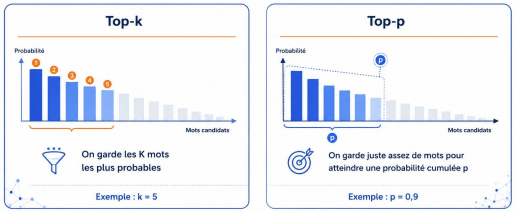
Conception naïve (l'ordre magique)

- Une simple question à répondre
- Un ordre à exécuter
- Exemple : explique les hallucinations de l'IAg
- (Niveau implicite, format non contrôlé)

Conception avancée (le cadrage épistémologique)

- Une architecture d'orientation probabiliste
- Contient : Rôle → Tâche → Contexte → Critères → Format
- Exemple : Agis comme un enseignant. Explique les hallucinations à des étudiants. Distingue les causes probabilistes. Termine par 3 questions de vigilance





Top-k fixe un nombre ; top-p fixe un seuil de probabilité.

- Un prompt bien conçu peut multiplier par 10 la qualité de la réponse, sans changer de modèle.
- Point de vigilance : il ne transforme pas une sortie en preuve.
- Le prompt devient un pilote cognitif

L'utilisateur n'est pas un simple demandeur, il est le co-architecte des conditions de production.

04

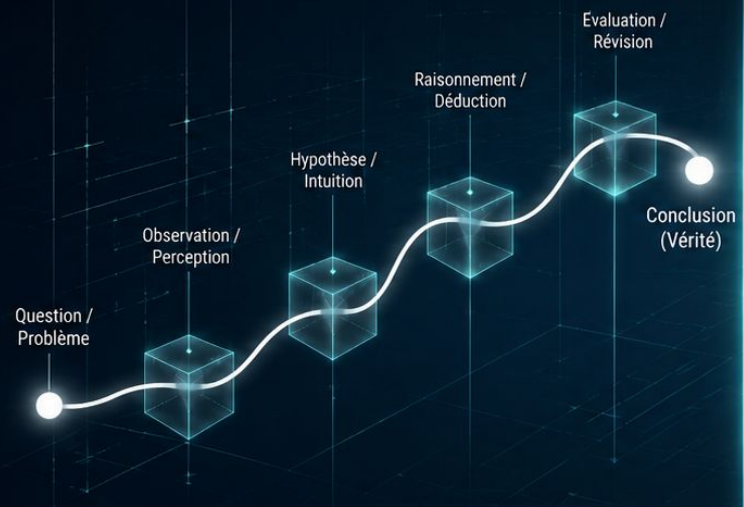
Le mirage : Explicabilité cognitive

“Une chaîne de raisonnement visible n’est pas nécessairement une trace fidèle du modèle ”

Squelettes, arbre de pensée et rationnalisation post-hoc

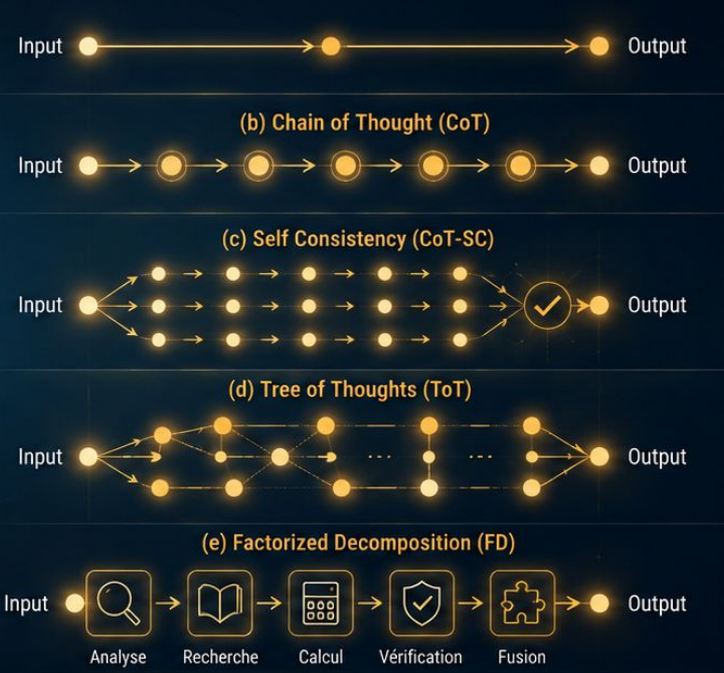
L’explicabilité cognitive : Déjouer l’illusion du raisonnement interne

Processus de Pensée Humaine



La pensée humaine est itérative, non-linéaire et révisable. Elle utilise l’intuition, l’expérience et l’évaluation continue pour converger progressivement vers la vérité.

Modèles de Raisonnement de l’IA



Le raisonnement explicite de l’IAg:
Certaines architectures demandent au modèle une trace intermédiaire de raisonnement (étapes, branches, votes, sous-tâches…)

Le piège épistémologique:
Ces traces ne révèlent pas le mécanisme causal interne du modèle. Elles constituent une reconstruction statistique plausible, produite après génération

L’utilité réelle:
Ces méthodes améliorent souvent la cohérence, la vérifiabilité et les performances sur les tâches complexes, sans garantir la vérité ni la compréhension réelle.

Exemple d'application du CoT

Hassina a 8 patates. Elle en donne deux à Jean, puis deux autres à Marie. Marie lui en rend une et donne une à Jean. Combien de patates ont Hassina et Marie ?

Hassina : 7
Marie : 1

Hassina a 8 patates. Elle en donne deux à Jean, puis deux autres à Marie. Marie lui en rend une et donne une à Jean. Combien de patates ont Hassina et Marie ? Donne juste la réponse. Raisonne étape par étape

1. Hassina commence avec 8 patates.
2. Elle donne 2 patates à Jean, il lui reste **6 patates**.
3. Elle donne encore 2 patates à Marie, il lui reste **4 patates**.
4. Marie lui rend 1 patate, Hassina a maintenant **5 patates**.
5. Marie donne 1 patate à Jean, il lui reste **1 patate**.

Réponse :
Hassina : 5 patates
Marie : 1 patate

Utilisons la **technique des chaînes de raisonnement** pour résoudre ce problème.

Étape 1 : Situation initiale

- Hassina commence avec **8 patates**.
- Marie et Jean n'ont aucune patate au départ.

Étape 2 : Hassina donne 2 patates à Jean

- Hassina : $8 - 2 = 6$
- Jean : $0 + 2 = 2$

Étape 3 : Hassina donne 2 patates à Marie

- Hassina : $6 - 2 = 4$
- Marie : $0 + 2 = 2$

Étape 4 : Marie rend 1 patate à Hassina

- Hassina : $4 + 1 = 5$
- Marie : $2 - 1 = 1$

Étape 5 : Marie donne 1 patate à Jean

- Marie : $1 - 1 = 0$
- Jean : $2 + 1 = 3$

Conclusion

- Hassina : **5 patates**
- Marie : **0 patate**

Réponse finale : Hassina a 5 patates et Marie a 0 patate.

Message ChatGPT

Transformer le raisonnement généré en activité critique

L’explicabilité cognitive devient utile lorsqu’elle ouvre une discussion outillée

Faire expliciter	Demander au modèle ses hypothèses, ses critères et ses limites
Faire comparer	Comparer deux prompts, deux réponses ou deux raisonnements produits
Faire vérifier	Identifier ce qui doit être sourcé, calculé, testé ou contredit
Faire décider	Faire assumer à l’humain la validation finale et la responsabilité

Comment lire un raisonnement généré ?

Les méthodes avancées (Chain-of-Though, Tree-of-Though…) obligent l'IAg à décomposer son problème. Mais comment interpréter ce texte ?

Comme Indice : OUI

Indique comment le système formule le problème et organise une solution possible par le langage.

Comme support : OUI

Offre des prises pour discuter, isoler une erreur ou comparer des démarches différentes.

Comme preuve : NON

Danger d'illusion : le raisonnement affiché est une rationnalisation post-hoc, une narration linguistique plausible, pas une trace causale du calcul vectoriel interne.

Traiter le raisonnement de l'IAg comme artefact utile, jamais comme une preuve interne.
En formation : s'en servir pour questionner, jamais pour conclure que l'IA a compris

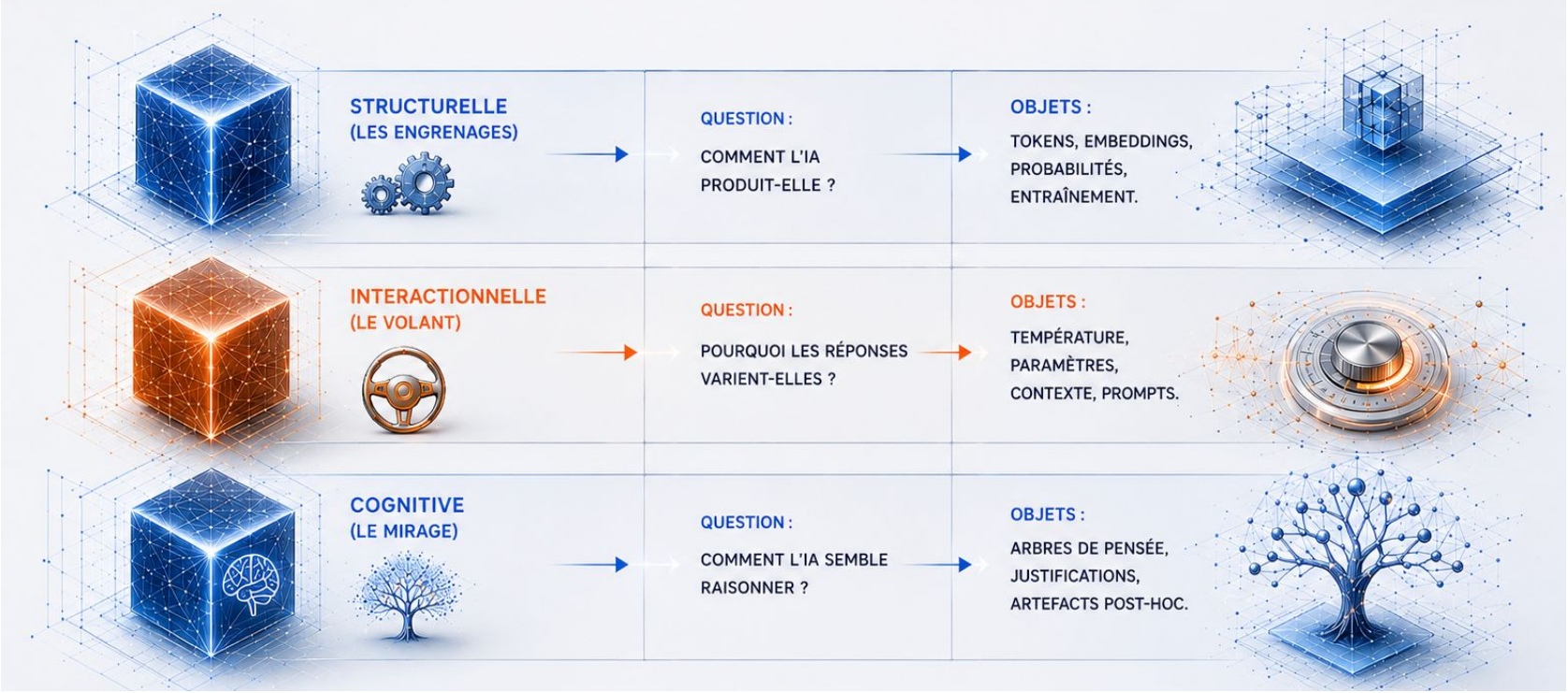
05

Conclusion et perspectives

Intelligibilité de l’IAg et agentivité humaine : vers une alliance critique

Former enseignants et étudiants à développer leur autonomie critique, pour une adoption responsable, éthique et émancipatrice de l’IAg

Le modèle tripartite de l’intelligibilité



LITTÉRATIE DE L’IAg

L’enjeu n’est pas seulement de produire des IAg trasparentes, mais de former l’autonomie critique des étudiants et des enseignants capables de les interpréter.

L’agentivité humaine : l’IAg comme prothèse probabiliste

LA MACHINE

Probabilités,
Génération,
Vraisemblance.

L’HUMAIN

Sujet connaissant,
Évaluation,
Cadrage éthique,
Validation.

LE DISCERNEMENT NE S’IMPROVISE PAS :

Il repose sur la compréhension technique détaillée des mécanismes de l’IA (le capot).

RECONSTRUCTION STATISTIQUE VS VÉRITÉ :

L’IA génère du plausible.
L’enseignant construit et valide le savoir.
L’étudiant développe son esprit critique.

L’IMPÉRATIF ACADÉMIQUE :

Le système n’est qu’un accélérateur de flux : la responsabilité de la preuve incombe exclusivement à l’enseignant et à l’étudiant.

Merci pour votre attention.

Hassina EL KECHAI
Mcf en informatique
Unité de Recherche TECHNE (UR-20297)
Université de Poitiers
U.F.R Lettres et Langues
1 Rue Raymond Cantel, 86000 Poitiers
Mail : hassina.el.kechai@univ-poitiers.fr

TECHNE