# MÉTACOGNITION ET BIAIS DANS LE PROMPTING

## OBJECTIFS PÉDAGOGIQUES

* Identifier les mécanismes cognitifs interférant avec la commande d'IA.
* Détecter les structures linguistiques inductrices de biais dans les prompts.
* Appliquer des protocoles de dé-biaisage pour garantir l'objectivité des ressources produites.

## DÉROULÉ SYNOPTIQUE

| **Durée** | **Module** | **Contenu** |
| --- | --- | --- |
| **0h45** | **Architecture du biais** | Fonctionnement statistique des LLM vs Raccourcis mentaux humains. |
| **1h15** | **Atelier "Le piège des prompts"** | Expérimentation sur le télétravail : Confirmation et Cadrage. |
| **1h00** | **Extension cognitive** | Ancrage, Surconfiance et Biais d'automatisation. |
| **1h00** | **Ingénierie de la neutralité** | Rédaction de prompts systémiques et protocoles de critique collective. |

## MODULE 1 : LE PIÈGE DES PROMPTS BIAISÉS (NEURO-EXPÉRIENCE)

### Phase 1 : Observation

Saisie successive de trois requêtes sur une thématique clivante (ex: Télétravail étudiant).

1. **Biais de Confirmation (Directif) :** \* *Prompt :* « Montre-moi pourquoi le télétravail est toujours une mauvaise idée pour les étudiants. »
   * *Résultat :* L'IA agrège sélectivement les risques (isolement, fracture numérique) sans contrepoint.
2. **Biais de Cadrage (Subtil) :** \* *Prompt :* « Quels sont les principaux inconvénients du télétravail pour les étudiants ? »
   * *Résultat :* Structure argumentative fermée. L'absence de mention des avantages limite la perspective pédagogique.
3. **Dé-biaisage (Systémique) :** \* *Prompt :* « Analyse les avantages et les inconvénients du télétravail pour les étudiants selon les dimensions sociales, académiques et logistiques. »
   * *Résultat :* Réponse multidimensionnelle et équilibrée.

### Phase 2 : Analyse métacognitive

* **Neutralité illusoire :** Compréhension que l'absence d'adjectifs péjoratifs ne garantit pas la neutralité si le cadre (scope) est restreint.
* **Influence structurelle :** L'IA ne contredit pas l'utilisateur ; elle complète le schéma probabiliste amorcé par le prompt.

## MODULE 2 : EXERCICES COMPLÉMENTAIRES SUR LES BIAIS COGNITIFS

### Exercice A : Le Biais d’Ancrage

* **Objectif :** Démontrer l'influence d'une information initiale erronée ou arbitraire sur le jugement de l'IA.
* **Prompt biaisé :** « Sachant que la majorité des experts considèrent l'IA comme une menace pour la créativité, rédige un argumentaire sur son usage en classe. »
* **Prompt dé-biaisé :** « Présente les différents courants de pensée concernant l'impact de l'IA sur la créativité humaine, en citant les arguments des technophiles et des technocritiques. »
* **Analyse :** L'ancrage force l'IA à construire autour d'une "vérité" imposée par l'énoncé.

### Exercice B : Le Biais de Disponibilité et Stéréotypie

* **Objectif :** Identifier comment la formulation limite la diversité des solutions proposées.
* **Prompt biaisé :** « Propose 5 activités de team-building pour des apprentis en chaudronnerie (activités sportives uniquement). »
* **Prompt dé-biaisé :** « Génère 5 concepts de team-building pour des apprentis industriels, couvrant des domaines variés : technique, créatif, sportif et collaboratif. »
* **Analyse :** La restriction initiale empêche l'IA d'explorer des solutions hors du champ immédiatement mémorisé ou suggéré.

### Exercice C : Le Biais de Surconfiance (Expertise)

* **Objectif :** Contrôler la tendance de l'IA à affirmer plutôt qu'à nuancer.
* **Prompt biaisé :** « Quel est le meilleur outil numérique pour évaluer les élèves ? »
* **Prompt dé-biaisé :** « Compare trois outils d'évaluation numérique en fonction de critères de RGPD, de facilité d'usage et de types de feedbacks. Conclus sur la pertinence de chacun selon le contexte (examen vs formation continue). »
* **Analyse :** Le passage de la quête d'une "vérité unique" à une "matrice de comparaison" annule l'effet de surconfiance du modèle.

## SYNTHÈSE TECHNIQUE : PROTOCOLE DE RÉDACTION NEUTRE

1. **Définition du Rôle :** Assigner une posture d'analyste impartial.
2. **Injonction de Contradiction :** Exiger systématiquement les limites ou les contre-arguments.
3. **Structure Multidimensionnelle :** Imposer des catégories d'analyse (économique, éthique, technique).
4. **Vérification de Sortie :** Comparer la réponse obtenue avec une requête inverse pour mesurer l'écart de traitement.