

LE PETIT LIVRE DE L'IA · CHAPITRE 0.3 · FICHE DE RÉVISION, STYLE EXAMEN

La vigilance critique : vérifier avant de croire

Méthode · LPD 6 L'IA répond vite et avec assurance, même quand elle se trompe. Le risque ne tient pas seulement à l'outil, mais à notre tendance à le croire : on retient ce qui confirme nos idées, on fait confiance à ce qui a l'air savant. La vigilance critique se travaille en quatre temps : comprendre pourquoi on se laisse piéger, disposer d'une boîte à outils de vérification, s'exercer sur des cas réels, puis savoir défendre sa conclusion.

1 L'essentiel à savoir par cœur

- **Pourquoi** : le biais de confirmation et le biais d'autorité (ou d'automatisation) sont amplifiés par l'IA ; elle peut inventer des faits (hallucination) et fabriquer des voix, des images ou des vidéos crédibles (hypertrucage).
- **Quoi** : les **5 pourquoi** pour remonter à la cause, le **test C.R.A.P.** pour juger une source (actualité, pertinence, autorité, but), la **triangulation** (plusieurs sources indépendantes, dont une officielle), la **recherche d'image inversée**.
- **Comment** : s'exercer sur des contenus produits par l'IA de l'entreprise : un résumé d'arrêt, un chiffre de salaire, une photo reçue, un appel urgent.
- **Et maintenant** : conclure en trois temps, **affirmation, preuve, justification**, pour décider et convaincre.

4	5	2+	3
questions du test C.R.A.P.	pourquoi jusqu'à la cause	sources indépendantes, dont une officielle	temps : affirmation, preuve, justification

2 La base légale, article par article

- Qui traite des données personnelles doit s'assurer qu'elles sont exactes (LPD 6 al. 5).
- Une décision fondée exclusivement sur un traitement automatisé, qui a des effets importants pour la personne, doit lui être annoncée ; elle peut demander qu'une personne physique la revoie (LPD 21).
- Le travailleur exécute avec soin le travail confié (CO 321a) : relayer sans contrôle une réponse fausse n'est pas diligent.
- Usurper l'identité d'autrui pour lui nuire ou se procurer un avantage est punissable (CP 179decies) ; tromper pour obtenir un virement, aussi (CP 146).

3 Les règles, cas par cas

Situation	Règle	Article
Une réponse de l'IA cite un arrêt du Tribunal fédéral	Ouvrir l'arrêt sur bger.ch avant de s'en servir : il peut ne pas exister	Triangulation
Un salaire minimum, un taux ou un plafond	Vérifier à la source officielle de l'année (canton, OFAS)	C.R.A.P. : actualité
Une réponse qui dit exactement ce qu'on espérait	Chercher activement l'avis contraire	Biais de confirmation
Une réponse au ton très sûr, pleine de jargon	L'assurance n'est pas une preuve : demander la source	Biais d'autorité
Une photo ou une vidéo choquante d'un collaborateur	Recherche d'image inversée, origine, date ; ne rien décider dans l'urgence	Hypertrucage
Un appel du « directeur » qui exige un virement urgent	Rappeler par un numéro connu ; double validation	CP 146 · CP 179decies
Une erreur qui revient dans les décomptes	Les 5 pourquoi jusqu'à la cause, puis corriger la cause	Analyse structurée

4 La méthode, dans l'ordre

1. Douter d'une réponse trop sûre.
2. Demander la source et la date.
3. Recouper avec une source officielle.
4. Vérifier l'origine d'une image.
5. Conclure en trois temps.

> **« Ce que l'IA affirme est une piste ; ce qu'une source officielle confirme est une preuve. »**

6 Les pièges

Croire une réponse parce qu'elle est bien écrite, précise et chiffrée : une IA peut inventer un numéro d'arrêt ou un article de loi qui a l'air vrai.

7 Le cas type du chapitre

Le résumé trop beau. Une collègue demande à l'IA de l'entreprise : « Peut-on réduire les vacances d'une employée après son congé de maternité ? » La réponse, très assurée, dit oui, cite un arrêt du Tribunal fédéral avec son numéro et propose une formule de calcul.

a. Quels biais sont en jeu ?

b. Comment vérifier ?

c. Rédigez la conclusion en trois temps.

Corrigé en fin de fiche.

8 Le coin de l'expert

- Le biais d'automatisation pousse à suivre la recommandation d'une machine même quand d'autres indices la contredisent ; il augmente avec la fatigue et l'urgence.
- Les détecteurs de contenus produits par l'IA ne sont pas fiables : un doute sur une image se lève par son origine (qui l'a publiée, quand, où) plutôt que par un détecteur.
- Les 5 pourquoi s'arrêtent à une cause sur laquelle on peut agir ; noter chaque réponse avec sa preuve évite de deviner.
- Dans une équipe, une règle simple aide : toute affirmation qui sert à une décision porte sa source et sa date.

10 Questions d'examen

1. Vrai ou faux ? Une réponse de l'IA qui cite un numéro d'arrêt peut être reprise telle quelle.

2. Vrai ou faux ? Le biais de confirmation consiste à retenir ce qui confirme ce qu'on pense déjà.

3. Vrai ou faux ? Le test C.R.A.P. examine l'actualité, la pertinence, l'autorité et le but d'une source.

4. Vrai ou faux ? Une recherche d'image inversée aide à retrouver l'origine d'une photo.

5. Vrai ou faux ? Une seule source suffit si elle vient d'une IA réputée.

6. Vrai ou faux ? Affirmation, preuve, justification : c'est une façon de structurer un argument.

7. Vrai ou faux ? Les détecteurs de texte produit par l'IA sont assez fiables pour fonder une sanction.

8. Cas express : L'IA cite un arrêt pour justifier une réduction des vacances après une maternité : on l'applique ?

9. Résolvez le cas type de la section 7, « Le résumé trop beau », questions a à c.

10. Citez les étapes de la méthode, dans l'ordre, et la base légale du chapitre.

11 Corrigés

1. Faux : Il faut ouvrir l'arrêt : l'IA peut l'inventer.

2. Vrai : D'où l'intérêt de chercher l'avis contraire.

3. Vrai : Quatre questions avant de croire.

4. Vrai : Elle montre où et quand l'image a déjà été publiée.

5. Faux : Trianguler : plusieurs sources indépendantes, dont une officielle.

6. Vrai : On dit quoi, on montre la source, on explique le lien.

7. Faux : Ils ne le sont pas.

8. Non. On vérifie : CO 329b al. 3 let. b l'interdit.

9.

1. Biais d'autorité (ton sûr, numéro d'arrêt) et, si la collègue espérait cette réponse, biais de confirmation.

2. Test C.R.A.P. : ni date ni source vérifiable. Triangulation : ouvrir le Code des obligations sur Fedlex (art. 329b al. 3) et chercher l'arrêt sur bger.ch ; ne jamais reprendre un numéro qu'on n'a pas lu.

3. Affirmation : les vacances ne peuvent pas être réduites à cause du congé de maternité. Preuve : CO 329b al. 3 let. b, relu sur Fedlex. Justification : la loi l'exclut expressément ; la réponse de l'IA était fausse.

10. 1) Douter d'une réponse trop sûre. ; 2) Demander la source et la date. ; 3) Recouper avec une source officielle. ; 4) Vérifier l'origine d'une image. ; 5) Conclure en trois temps. Base : LPD 6 et 21 ; CO 321a, 329b ; CP 146, 179decies.