Strasbourg, 12 février 2024

CEPEJ-GT-CYBERJUST(2023)5final

**COMMISSION EUROPÉENNE POUR L'EFFICACITÉ DE LA JUSTICE**

**(CEPEJ)**

**Groupe de travail de la CEPEJ sur la cyberjustice et de l'intelligence artificielle (CEPEJ-GT-CYBERJUST)**

# L'utilisation de l'intelligence artificielle (IA) générative par les professionnels de la justice dans un contexte professionnel

Note d'information préparée par le

Groupe de travail de la CEPEJ sur la cyberjustice et de l'intelligence artificielle (CEPEJ-GT-CYBERJUST)

## A. Introduction

L'intelligence artificielle (IA) générative est un système logiciel qui communique en langage naturel, capable de répondre à des questions relativement complexes et de créer un contenu (texte, image ou son) à la suite d'une question ou d'instructions formulées (invite). Parmi ces outils figurent OpenAI ChatGPT, Copilot, Gemini et Bard qui se développent rapidement.

L'objectif de cette Note[1] est de mener une réflexion préliminaire sur ce que les juges et les autres professionnels de la justice du secteur public peuvent attendre de l'utilisation d'outils d'IA générative dans un contexte judiciaire.

## B. Comment cela fonctionne-t-il ?

L'IA générative fonctionne en apprenant des modèles et des caractéristiques à partir de grandes collections de données. Elle est basée sur une compréhension statistique du langage : son but est de définir, avec la plus grande certitude possible, le mot suivant, sans avoir une connaissance propre.

Ainsi, lorsque le système écrit que J.F. Kennedy était président des Etats-Unis, ce n'est pas parce qu'il s'appuie sur une base de connaissances qui fait un lien direct entre ces deux informations, mais parce que, dans les cas qu'il a rencontrés (dans les données d'entraînement), l'association Kennedy et président des Etats-Unis a été faite très fréquemment. Il en déduit donc que cette association est susceptible d'être pertinente.

Les données d'entraînement sont généralement des informations trouvées sur internet, des ensembles de données sélectionnés et des informations introduites par d'autres utilisateurs dans la machine par le biais de messages-guides.

L'IA générative semble donner de bons résultats à l'intérieur d'un cadre clairement défini, tel que

- la traduction de textes (par exemple de l'anglais au français et vice versa),
- la production de textes, d'images ou de sons cohérents (mais pas nécessairement vrais),
- le résumé automatique de textes,
- l'analyse sémantique et la détection d'opinions,
- l'exploration de texte et l'accès au contenu.

## C. Quels sont les risques ?

- *Production potentielle d'informations factuellement inexactes (réponses fausses ou biaisées et "hallucinations")*

Les mauvaises réponses peuvent avoir pour origine des données de formation insuffisantes ou erronées. De fausses données conduisent à de fausses réponses.

Le terme « hallucination » est une expression plutôt amicale pour désigner l'observation selon laquelle certaines réponses sont tout simplement inventées. Si aucune réponse n'est trouvée, les algorithmes ont tendance à inventer une réponse « probable ». Une autre raison pourrait être l'établissement d'une fausse corrélation entre les données.

Plus important encore, toute intelligence artificielle est profondément déterminée par les données sur lesquelles elle a été entraînée. Elle n'est donc jamais neutre et, au contraire, elle incorpore tous les biais, inexactitudes, lacunes ou défaillances contenus dans la base

[1] La note est basée sur un projet de Manon Maus et Camille le Douaron et a été appuyée par le Bureau consultatif sur l'intelligence artificielle (AIAB) de la CEPEJ.

de données d'entraînement et/ou les biais culturels de ceux qui ont conçu le système et guidé son entraînement en (in)validant certaines de ses réponses. Il peut même arriver que des biais soient délibérément intégrés dans l'algorithme.

L'opacité de la programmation de l'algorithme et de la manière dont les données sous-jacentes sont connectées rend les réponses obtenues davantage incompréhensibles/inaccessibles et, par conséquent, difficiles à vérifier.

- *Divulgation possible de données sensibles et risque de confidentialité*

Les informations saisies sont transmises au fournisseur du système et peuvent être utilisées comme données d'entraînement pour de futurs utilisateurs et pour générer de futurs résultats. Cela peut entraîner une violation de la protection des données personnelles ou la divulgation involontaire d'informations classifiées ou sensibles.

La protection des données transmises par les systèmes n'est le plus souvent pas garantie. Ainsi, les conversations sont enregistrées sur les serveurs d'entreprises, souvent non européennes, et/ou revendues (voire récupérées via une attaque informatique, le niveau de sécurité de ces serveurs n'étant pas connu).

- *Absence de références pour les informations fournies et violation potentielle de la propriété intellectuelle et des droits d'auteur*

L'origine du matériel utilisé pour la base de données et les données d'entraînement utilisées manquent de transparence. La plupart des systèmes ne peuvent pas énumérer et créditer les textes utilisés pour créer les résultats. Cela peut non seulement entraîner des difficultés dans la vérification des résultats, mais aussi des violations des droits d'auteur.

Si les cadres réglementaires diffèrent d'un pays à l'autre, il n'en demeure pas moins qu'ils s'appliquent à l'utilisation de l'IA, ce qui signifie que le contenu créé pourrait être considéré comme du plagiat.

- *Capacité limitée à fournir la même réponse à une question identique*

La plupart des systèmes d'IA générative contiennent un degré d'aléa qui leur permet de proposer différentes réponses à une même question. Les réponses peuvent varier en fonction du moment où elles sont posées ou des nuances dans la formulation de la question. Il n'est donc pas possible de garantir toujours le même niveau de qualité de réponse.

- *Reproduction potentielle des résultats*

Le résultat de l'IA générative n'est en aucun cas unique et peut être identique ou similaire à celui généré pour un autre utilisateur, c'est pourquoi sa source ne doit pas être dissimulée. En outre, et particulièrement dans le cas de la justice, il est essentiel d'être transparent sur l'utilisation de l'IA : la relation avec le justiciable est basée sur la confiance.

- *Stabilité et fiabilité variables des modèles d'IA générative pour les processus critiques et sensibles au facteur temps*

Des variations ont été observées dans les temps de réponse et la disponibilité des services, ce qui devrait être pris en compte dans les processus sensibles au facteur temps.

- *Exagération des biais cognitifs*

La relation entre l'humain et la machine est intrinsèquement biaisée par nos capacités cognitives. La relation avec l'IA générative tend à exagérer ces biais, car la discussion qui peut s'établir avec la machine augmente sa perception comme "humaine". L'échange n'est en aucun cas neutre.

## D. Comment l'appliquer ?

1. Assurez-vous que l'utilisation de l'outil est autorisée et adaptée à l'objectif recherché.
2. Gardez à l'esprit qu'il ne s'agit que d'un outil et essayez de comprendre comment il fonctionne (soyez conscient des biais cognitifs humains).
3. Privilégiez les systèmes entraînés sur des données certifiées et officielles, dont la liste est connue, afin de limiter les risques de biais, d'hallucination et de violation des droits d'auteur.
4. Donnez à l'outil des instructions claires (messages-guides) sur ce que l'on attend de lui. C'est par la conversation que la machine obtiendra les instructions dont elle a besoin, n'hésitez donc pas à l'interpeller, contrairement à un moteur de recherche. Il est possible de demander des précisions, voire d'affiner ou de modifier la demande. Par exemple, donnez à la machine un contexte (pays, période), définissez une tâche (ex : rédiger un résumé en xx mots...), précisez à qui est destiné le résultat, comment il doit être produit et le ton que l'outil doit adopter. Demandez ensuite un format de présentation spécifique, vérifiez que les instructions ont bien été comprises (en demandant à la machine de les reformuler) et donnez des exemples de réponses attendues pour des questions similaires afin de permettre à l'outil d'en imiter la forme et le style.
5. Ne saisissez que des données non sensibles et des informations déjà disponibles dans le domaine public.
6. Vérifiez toujours l'exactitude des réponses, même si des références sont données (vérifiez en particulier l'existence de la référence).
7. Soyez transparent et indiquez toujours si une analyse ou un contenu a été généré par l'IA générative.
8. Reformuler le texte généré au cas où il serait utilisé dans des documents officiels et/ou juridiques.
9. Restez maître de votre choix et du processus de décision et examinez d'un œil critique les propositions qui vous sont faites.

## E. Quand ne faut-il pas l'appliquer ?

1. Si vous ne connaissez pas, ne comprenez pas ou n'acceptez pas les conditions d'utilisation.
2. Si cela est interdit ou contraire aux règles de votre organisation.
3. Dans le cas où vous ne pouvez pas évaluer le résultat en termes d'exactitude factuelle et de partialité.
4. Dans le cas où vous seriez amené à saisir et donc à divulguer des données personnelles, confidentielles, protégées par le droit d'auteur ou autrement sensibles.
5. Dans le cas où vous souhaiteriez savoir comment votre réponse a été obtenue.
6. Dans le cas où l'on attend de vous que vous produisiez une réponse authentiquement autodidacte.