cepej

European
Commission
for the Efficiency
of Justice

Commission
européenne
pour l'efficacité
de la justice

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE

Strasbourg, 12 February 2024

CEPEJ-GT-CYBERJUST(2023)5final

EUROPEAN COMMISSION FOR THE EFFICIENCY OF JUSTICE

(CEPEJ)

CEPEJ Working group on Cyberjustice and Artificial Intelligence
(CEPEJ-GT-CYBERJUST)

Use of Generative Artificial Intelligence (AI) by judicial professionals in a work-related context

Information note prepared by the

CEPEJ Working group on Cyberjustice and Artificial Intelligence
(CEPEJ-GT-CYBERJUST)

A. Introduction

Generative Artificial Intelligence (AI) are software systems that communicate in natural language, able to give answers to relatively complex questions and can create content (provide a text, picture, or sound) following a formulated question or instructions (prompt). These tools include OpenAI ChatGPT, Copilot, Gemini, and Bard, all of which are developing rapidly.

Aim of this Note[1] is to give some preliminary thought to what judges and other public sector justice professionals can expect from the use of generative AI tools in a judicial context.

B. How does it work?

Generative AI works by learning patterns and characteristics from large collections of data. It is based on a statistical understanding of language: its purpose is to define, with the greatest possible certainty, the next word, without knowledge of its own.

So, when the system writes that J.F. Kennedy was President of the United States, it is not because it is relying on a knowledge base that makes a direct link between these two pieces of information, but because, in the cases it encountered (in the training data), the association Kennedy and President of the United States was very often made. It therefore deduced that this association was likely to be relevant.

Mostly, the training data is the information found on the internet, selected datasets, and information fed by other users into the machine through prompts.

Generative AI appears to provide good results within a clearly defined frame, such as

- translation of texts (for example English to French and vice versa),
- the generation of coherent (but not necessarily true) text, images or sounds,
- automatic summary of texts,
- semantic analysis and opinion detection,
- text mining and content access.

C. What are the risks?

❖ *Potential production of factually inaccurate information (false answers, “hallucinations” and bias)*

Wrong answers might have their origin in insufficient or wrong training data in the first place. False data leads to false answers.

“Hallucination” is a rather friendly expression for the observation that some answers are simply invented. If no answer is found, the algorithms tend to invent a “probable” answer. Another reason might be the establishment of a false correlation between the data.

Most importantly, all artificial intelligence is profoundly determined by the data on which it has been trained: it is therefore never neutral and, on the contrary, embeds all biases, inaccuracies, gaps or failures contained in the training database and/or the cultural biases of those who designed the system and guided its training by (in)validating some of its answers. There might be even cases where bias has been built deliberately into the algorithm.

Opacity of how the algorithm is programmed and how the underlying data is connected leads to further incomprehensibility and thus difficulties in the verification of the given answers.

[1] The note is based on a draft by Manon Maus and Camille le Douaron and has been endorsed by the CEPEJ Artificial Intelligence Advisory Board (AIAB).

- ❖ *Possible disclosure of sensitive data and risk of confidentiality*

Entered information is transmitted to the provider of the system and potentially used as training data for future users and to generate future outputs. This might result in a breach of personal data protection or unintended disclosure of classified or otherwise sensitive information.

Protection for the data transmitted through the systems is mostly not guaranteed. As a result, conversations are recorded on the servers of companies, often non-European, and/or resold (or even recovered via a computer attack, as the level of security of these servers is not known).

- ❖ *Lack of references for the information provided and potential violation of intellectual property and copyright*

There is a lack of transparency in the origin of the material used for the data base and used training data. Most systems cannot list and credit the texts used for creating the output. This might not only cause difficulties in verifying the outputs but also result in copyright infringements.

While the regulatory frameworks differ from country to country, they nonetheless apply to the use of AI, meaning that the content created could be considered plagiarism.

- ❖ *Limited capability of providing the same answer to an identical question*

Most generative AI systems contain a degree of randomness that allows them to propose different answers to the same question. Answers can vary depending on the point of time they are asked, or nuances in the formulation of the question. It is therefore not possible to always guarantee the same level of response quality.

- ❖ *Potential replication of outputs*

The outcome of generative AI is by no means unique and may be identical or similar to that generated for another user, therefore its source should not be concealed. In addition, and particularly in the case of justice, it is essential to be transparent about the use of AI: the relationship with the litigant is based on trust.

- ❖ *Varying stability and reliability of Generative AI models for critical and time-sensitive processes*

There are observed variations in response times and availability of the services, which should be considered in time sensitive processes.

- ❖ *Exaggeration of cognitive biases*

The relationship between human and machine is inherently biased by our cognitive capacities. The relationship with generative AI tends to exaggerate these biases, as the discussion that can be established with the machine increases its perception as "human". The exchange is by no means neutral.

D. How should it be applied?

1. Make sure that the tool’s use is authorised and appropriate for the desired purpose.
2. Bear in mind that it is only a tool and try to understand how it works (be aware of human cognitive biases).
3. Give preference to systems that have been trained on certified and official data, the list of which is known, to limit the risks of bias, hallucination, and copyright infringement.

4. Give the tool clear instructions (prompts) about what is expected of it. It is through conversation that the machine will obtain the instructions it needs, so do not hesitate to engage it, unlike a search engine. Asking for clarification or even refining or modifying the request is possible. For example, give the machine a context (country, period of time), define the task (e.g. write a summary in xx words), specify who the output is intended for, how it is to be produced and the tone the tool should adopt, ask for a specific presentation format, check that the instructions have been properly understood (by asking the machine rephrasing them), provide examples of the answers expected for similar questions to enable the tool to imitate their form and style.
5. Enter only non-sensitive data and information which is already available in the public domain.
6. Always check the correctness of the answers, even in case references are given (especially check the existence of the reference).
7. Be transparent and always indicate if an analysis or content was generated by generative AI.
8. Reformulate the generated text in case it shall feed into official and/or legal documents.
9. Remain in control of your choice and the decision-making process and take a critical look at the made proposals.

E. When should it not be applied?

1. In case you are not aware of, do not understand or do not agree to the terms and conditions of use.
2. In case it is forbidden/against your organisational regulations.
3. In case you cannot assess the result for factual correctness and bias.
4. In case you would be required to enter and thus disclose personal, confidential, copy right protected or otherwise sensitive data.
5. In case you must know how your answer was derived.
6. In case you are expected to produce a genuinely self-derived answer.