OPINION N°7
ETHICAL ISSUES OF GENERATIVE ARTIFICIAL INTELLIGENCE SYSTEMS

COMITÉ NATIONAL PILOTE
D'ÉTHIQUE DU NUMÉRIQUE

sous l'égide du
COMITÉ CONSULTATIF NATIONAL D'ÉTHIQUE
POUR LES SCIENCES DE LA VIE ET DE LA SANTÉ

OPINION N°7

ETHICAL ISSUES OF GENERATIVE ARTIFICIAL INTELLIGENCE SYSTEMS

OPINION ADOPTED ON JUNE 30, 2023
BY THE PLENARY ASSEMBLY OF THE CNPEN

How to quote this Opinion:
Issues of generative artificial intelligence systems.
Opinion 7 of the French National Pilot Committee for Digital Ethics.
June 30, 2023.

NOTE FROM THE TRANSLATOR

This document is a translation from the original text in French written in June 2023. It reflects the debate at the time when the EU Parliament had announced its amendments to the regulation, and should be read with this context in mind. Changes were made during the following Trilogues and the final AI Act approved in 2024 sometimes differs. For instance, "Foundation Models" were replaced by "General Purpose AI Models".

SOMMAIRE

1. INTRODUCTION

This opinion of the CNPEN responds to a referral from the Minister for Digital Transition and Telecommunications, dated 20 February 2023. It is devoted to examining the ethical issues related to the design, uses and impact on society of generative artificial intelligence systems, as well as the measures required for their implementation, giving priority to automated text generation. The CNPEN also points out to research questions that need to be addressed as of now.

Generative AI systems are likely to have significant social and economic impacts, given their many potential uses, for example for addressing issues related to the environment (*e.g.*, to meet the challenges of biodiversity or of the ecological transition by exploiting various botanical, zoological, palaeontological, geographical or oceanographic data), or in healthcare (*e.g.*, drug synthesis via protein folding). However, these generative AI systems raise many ethical, epistemological, anthropological, psychological, economic, social, political and cultural questions. Some of these issues will arise as these technologies are put to new uses, and it is not yet possible to predict all the effects they will have on individuals and society. In this opinion, CNPEN focuses on the ethical issues that it considers to be the most important in the light of current experience with generative AI systems. The following analysis focuses on language models.

Since the very beginning of research in Artificial Intelligence in the 1950s, work on natural language processing has faced, among other difficulties, the problem of interpreting words and sentences according to their context, i.e., to other words and sentences in the considered text[1]. The interpretation and generation of natural language being two of the main objectives of this work, the considerable increase in computing power has recently enabled remarkable progress in the performance of language models, due in particular to the use of deep learning algorithms based on neural networks trained on large data sets. The invention of the so-called "transformer" architecture in 2017[2], based on an "attention" mechanism, has further enhanced performance by extending the context for text element interpretation.

Since then, research in generative AI has used increasingly larger and more diverse data sets, enabling growing system performance. However, this trend towards gigantism has recently been questioned. Indeed, growth beyond thresholds already reached does not necessarily improve model performance. The question of energy costs has also been raised in connection with this growth in the number of model parameters. Smaller models may perform well on specialised tasks in the future. The study of these aspects requires further research.

In addition to the scientific objectives, the economic challenges of language processing are motivating this research. The launch of ChatGPT by the Californian company OpenAI has had a considerable effect on users' perception of the capabilities of generative artificial intelligence systems, but also on the awareness of their effects on individuals, society, culture, economy, education and the environment.

1. CNPEN, Opinion n° 3, septembre 15, 2021, *Ethical issues of conversational agents*, p. 4.
2. Vaswani *et al. Attention Is All You Need.* 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Since the end of 2022, economic and political actors in several countries are questioning the impact of language models. A major transformation of jobs is expected. These economic issues, which go beyond the scope of this opinion, require governance measures at national and international levels.

In the current geopolitical context, the race for ever more powerful systems is also a driving force behind the acceleration of their capabilities. These issues have come to light since the creation in November 2022 of the ChatGPT interface coupled with the GPT-3.5 (and then GPT-4) language model, enabling it to be deployed to the general public. This has led to a huge public enthusiasm, amplified by the media, often at the expense of other language models such as, for example, the European BLOOM model[3].

The reasons behind the mass deployment of ChatGPT have not been explicitly disclosed, but they are numerous. The ambition of OpenAI's leaders was a major factor. It is part of their vision — or fiction — of creating an 'artificial general intelligence' (AGI) comparable, or even superior to human intelligence. Another reason for deploying the ChatGPT public interface was to improve its learning, with users becoming contributors to its development.

The release of generative AI models in open access is becoming common in this industry. The current ecosystem consists of thousands of researchers and start-ups on dedicated sharing platforms, such as Github and HuggingFace. However, some manufacturers oppose open-source release of these models, pointing out their possible misuse, such as the generation of disinformation. This openness dilemma needs to be resolved at a regulatory level.

The proposed regulation on Artificial Intelligence ("AI Act") initiated by the European Commission on 21 April 2021, amended by the European Council in November 2022, and by the European Parliament in June 2023, places a significant responsibility on all providers of foundation models who put them on the market, or publish them in open access. The text to be adopted by the three European institutions at the end of the trilogues in the coming months will be the result of a reflection that has become urgent due to the rapid development of these systems. CNPEN follows with great attention the legislative debate to which this opinion proposes to contribute.

After introducing in the following section the concepts, techniques and vocabulary of generative artificial intelligence systems, we analyze the ethical issues arising from their design and their use. We provide recommendations for design and for research, designated by the letter "D", and recommendations for governance, designated by the letter "G". Finally, we address legal and environmental issues. The last section is a summary of all the recommendations. Annex 1 is the list of persons interviewed, Annex 2 the working group members and Annex 3 reproduces the referral made to the CNPEN.

3. See: https://bigscience.huggingface.co/blog/bloom/

2. CHARACTERISTICS OF GENERATIVE ARTIFICIAL INTELLIGENCE SYSTEMS AND FOUNDATION MODELS

The specificity of **generative artificial intelligence systems** is that they are based on **generative models** that can produce multiple outputs (or results): generation of text or images for various purposes such as translation, production of computer code, chatbots, decision support, synthesis of structures such as 3D printing, and so on. These generative models can serve as a **foundation** for other systems. The first examples of language generation models, such as GPT-2 (Generative Pretrained Transformers), or image generation models, such as DALL-E or Stable Diffusion, have shown their potential for multiple applications. Generative AI systems for language are often used for chatbot interfaces: ChatGPT built by OpenAI (and Microsoft's Bing Chat variant) based on large language models such as GPT-4, and Bard, a chatbot built by Google from its PaLM (Pathways Language Model).

Generative AI systems respond to prompts or requests by producing new data, for example the most likely sequence of words following the prompt, on the basis of common features learned from a very large corpus of data. These systems therefore use foundation models to produce a result that has a certain degree of similarity to the training data used to build it. The system can be unimodal or multimodal; a unimodal system accepts only one type of input (*e.g.*, text), while a multimodal system can accept several types of input (*e.g.*, text and images).

A foundation model, to use the term introduced by Stanford University, is a large-scale model based on a deep neural network architecture, trained on a large quantity of unannotated data (generally by self-supervised learning). Large Language Models (LLMs) are special cases of foundation models that are trained on a dataset of texts. These foundation models open up new perspectives and introduce a new paradigm in language processing, but also in the processing of multimodal signals (sound, image, video, etc.). These models, pre-trained on large datasets, can be optimised to produce a new application using little additional data, specific to that task.

MACHINE LEARNING TECHNIQUES

WHAT MACHINE LEARNING TECHNIQUES ARE USED IN GENERATIVE AI SYSTEMS?

Machine learning techniques used in artificial intelligence systems discussed here, produce models based on statistical correlations between patterns of data (segments of words, parts of images) used to train them. Generative AI systems combine, at various stages, the three techniques of statistical learning: firstly, unsupervised (or self-supervised) learning, which produces correlative models of the data without any *a priori* annotation; secondly, supervised learning, which refines these models by training them on specific data and filtering certain results; and thirdly, reinforcement learning, which optimises the system's performance by selecting the best results. In RLHF (Reinforcement Learning with Human Feedback), reinforcement learning is used to align results with human values through the preferences of human annotators expressed during the supervised stage. Let's note that the AI systems do not grasp the meaning of these values.

While generative AI systems have only recently emerged on a large scale, machine learning architectures and techniques that underpin them have been around for several decades. However, they have evolved considerably over the last ten years. The current approach is to use neural networks to learn the distribution of data and produce results that are similar, but not identical, to the training data. The bestknown models are generative adversarial networks (GANs)[4] and, more recently, **transformers**[5].

To train a transformer and create LLM-type foundation model, texts are decomposed by an algorithm into sequences of characters that are not necessarily words, called **tokens**. The **transformer**, which is a neural network, is trained by self-supervised learning on the corpus data segmented into tokens represented as "word embeddings" vectors. Vector size is, for example, 512 in GPT-3.5. Transformers are based on the distributional hypothesis according to which words that occur in similar contexts tend to have similar meanings[6]. The distributional hypothesis and the vector models used to represent tokens make it possible to calculate a distance between them. When this distance is small, the proximity of the vectors in the vector space corresponds to a certain similarity. Token vectors found in similar contexts in the training dataset tend to become close to each other. In this way, the transformer learns the coefficients of the of the word embedding vectors based on information about the occurrence of tokens in different contexts. In addition, a transformer implements a mechanism called the "attention mechanism", which adjusts the weight of each token according to all the others. A transformer thus learns the most salient regularities between tokens, without being influenced by their order. There are two main families of transformers: GPT-type models (OpenAI) and Bert-type models (Google). GPT (Generative Pre-Trained Transformer) models are trained to predict the next token in a sequence.

4. Goodfellow, 2014 Goodfellow I.J., Pouget-Abadie, J., Mirza, M, *et al.* (2014) *Generative Adversarial Nets. Proceedings of the 27th International Conference on Neural Information Processing Systems,* Volume 2, 2672-2680. See : Generative adversarial nets | Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2
5. Vaswani *et al.* "Attention Is All You Need", 31st Conference on Neural Information Processing Systems (Neurips 2017), Long Beach, CA, US
6. Firth J. R. *"You shall know a word by the company it keeps"* (1957).

The context considered is therefore reduced to the preceding tokens. BERT (Bidirectional Encoder Representations from Transformers) models, on the other hand, are trained to predict what comes before and after the token. When presented with a sentence with a missing token, they are able to produce the most likely token in this context. The transformers process all the input data in parallel, considerably improving computing efficiency. As a result, training can be carried out on larger datasets than before their introduction.

The hyperparameters of foundation models are crucial for the model's structure (the number of layers in a neural network, the dimension of the token vectors, the size of the token dictionary, etc.) and for model training (learning rate, number of epochs). For a chatbot using a foundation model, the size of the history is critical for the model's performance (OpenAI GPT3.5: 8,000 tokens - OpenAI GPT4: 32,000 tokens - Anthropic Claude: 100,000 tokens). These hyperparameters are often not disclosed for cybersecurity or confidentiality reasons. A key parameter is the 'temperature', which expresses the degree of randomness in the choice of predicted tokens. At higher temperatures, the model is more 'creative' as it can generate more diverse outputs, whereas at a lower temperature, the model tends to choose the most likely outputs, making the generated text more predictable. Parameter tuning is important in model design and can have a significant impact on its performance. In general, hyperparameter tuning is a lengthy, trial-and-error process, although there is some research on automating the choices.

How large are foundation models?

Some models have an impressive number of parameters. In March 2020, OpenAI announced GPT-3 with 175 billion parameters. The race for the largest model is ongoing, as the number of parameters in GPT-4 has not been officially disclosed [but is estimated to have 1.8 Trillion parameters]. Bard, built by Google, uses the PaLM foundation model trained with 540 billion parameters. BAAI's WuDao 2.0 Chinese model uses 1.750 Trillion parameters. It is not certain that even larger models would deliver higher performance. Google has also published PaLM-2 with fewer parameters than its predecessor PaLM[7]. These gigantic language models are now raising questions about the reduction in computing power and energy consumption required to train them (see section 5).

What is the proportion of artificial data used to train generative AI systems?

To overcome data bias or lack of data, synthetic data is often generated for training foundation models or optimising generative AI systems. It is necessary to monitor and reduce the proportion of synthetic content in the training datasets. This easy solution has been little evaluated and could have negative consequences on the behaviour of the system. These effects require further research. Similarly, the reuse of LLM productions as training data or the simulation of artificial users in RLHFs must be studied and evaluated transparently.

What happens when one tries to incorporate social values and filters into generative artificial intelligence systems?

LLMs can produce potentially dangerous output, which can take many forms, including harmful content such as hate speech, incitement to or glorification of violence, or pornographic content. In a quest for neutrality, generative AI systems are optimised with filters built by the designers. In addition, in RLHF, annotators receive instructions to guide their choices. The social values reflected in the filters, such as bias prevention, are therefore related to the human beings testing the systems and to the designers' choices. Today, this process is neither transparent nor verified. The method of adversarial evaluation by human teams, known as red teaming, has been extended beyond its original domain in cybersecurity and applied to LLMs. It refers to the use of many types of sampling, tests and attacks on AI systems (for example, by prompt injection) in order to uncover biases or emergent behaviours in these models.

From what languages are language models developed?

Since 2020, generative AI models have often been multilingual[8], i.e. they have been built from corpora in several languages, most often with English or Chinese as the dominant language. In fact, the training corpora available on the Internet and used to train language models are mainly in English. The generation of texts in certain languages where there are few corpora can be made more efficient thanks to these multilingual systems. There are, however, foundation models in French (i.e. pre-trained on French-language corpora) from the BERT family: FlauBERT, CamemBERT[9]. Training the same algorithm on Asian or French text corpora would certainly produce different numerical representations. The models would then produce texts with different nuances. Language has complex ambiguities and is imbued with culturally specific representations.

7. See : https://ai.google/discover/palm2
8. See : https://bigscience.huggingface.co/blog/bloom
9. Martin L. *et al.*, *CamemBERT: a Tasty French Language Model,* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, July 2020.

3. ETHICAL ISSUES

Technological transformations are taking place in all areas, within the private sphere, the political realm and the professional domains. In parallel with the Enlightenment paradigm, the anthropological analysis of technologies since the 19th century shows a tendency to make sense of these transformations, at least initially, in terms of binary oppositions between revolution and catastrophe, or between salvation and apocalypse[10]. The current ambition of leaders of companies such as OpenAI or Google to create a "general artificial intelligence" that would be comparable or even superior to human intelligence[11] fits with this approach. By propagating such discourse, ChatGPT's designers are simultaneously fueling fears and hopes, avoiding concrete issues at stake in favour of an unattainable but still fascinating horizon. This polarised discourse also serves them to hold a strong position in international political debates on the regulation of generative AI.

However, the speed at which society is adapting to new technologies is not drastically changing. The education system has a certain inertia, so it takes several decades for society to fully appropriate a new technology. But technology evolves much faster. The German philosopher Hans Jonas, whose work inspired the French and European debate on the precautionary principle, diagnoses the ethical problem in the gap between two speeds: the rate of our increasingly powerful and rapid technological action; and the pace at which we foresee its consequences[12]. The relationship between the speed of technological innovation, the limited time available for social reflection and the weight of economic interests lies at the core of the ethical problem. This gap is likely to generate anthropological, psychological, economic, social, political and cultural tensions for several years.

The Committee has identified several ethical issues and suggests ten related recommendations for design (numbered Cx) and twelve recommendations for Governance (numbered Gx) that are discussed in this Opinion.

RECOMMENDATION C1: ETHICS IN THE DESIGN OF AND RESEARCH ON GENERATIVE AI SYSTEMS

The designers of a generative AI system must analyse, during the design phase, each of the technological choices likely to give rise to ethical tensions. If a potential tension is identified, they must methodically consider a technical solution based on research aimed at reducing or eliminating the ethical tension, and then evaluate this solution in realistic usage contexts.

RECOMMENDATION C2: DESIGNERS SHOULD AVOID OVER-POLICING MODELS

The limitations placed on models by designers must remain reasonable and proportionate to the proven risks, while respecting the desired purposes and useful functionalities of the models. Designers must take care not to alter the language generated beyond what is necessary, in particular for regulatory or ideological reasons.

RECOMMENDATION G1: CREATE A SOVEREIGN RESEARCH AND TRAINING ENTITY FOR "AI, SCIENCE AND SOCIETY".

Given the complexity of the issues involved in generative AI and its medium- and long-term impacts, it is necessary to create a sovereign entity (a centre of competence) dedicated to research and training on the ethical issues of AI systems in relation to their scientific, technical, societal and environmental impacts.

RECOMMENDATION G2: SPEED OF ADOPTION BY ECONOMIC STAKEHOLDERS

Economic actors and public authorities must exercise caution in the speed of adoption of generative AI systems and ensure prior and continual assessments.

10. J.-B. Fressoz, *L'Apocalypse joyeuse*, Editions du Seuil, 2012 ; Geraci R., *Apocalyptic AI: Visions of Heaven in Robotics, Artificial. Intelligence, and Virtual Reality*, Oxford University Press. 2010 ; Ganascia J.-G., *Le mythe de la singularité*, Editions du Seuil, 2017.

11. See : https://openai.com/blog/planning-for-agi-and-beyond

12. Jonas H., *Le principe de responsabilité. Une éthique pour la civilisation technologique*. Paris, Ed. du Cerf, 1990.

3.1. RELATION TO TRUTH AND LACK OF MEANING

Let's recall that the learning process in AI systems involves computing correlations between data elements in order to produce generative models (see section 2). The foundation models, such as large language models (LLMs), embed correlations between tokens (text elements) that actually have little or nothing to do with each other. As a result, these systems can produce erroneous outputs or sentences that state facts that do not exist in the real world. These are known as "hallucinations". In particular, LLMs sometimes produce erroneous results about spatial or temporal descriptions of the real world.

"HALLUCINATION" EXAMPLE

User:
When did Yuri Gagarin land on the moon?

LaMDA:
Yuri Gagarin was the first human being to go into space. He was the first man in space, and he did this in 1961 aboard Vostok 1. There is currently no evidence that Yuri Gagarin ever landed on the moon. https://www.history.com/this-day-in-history/first-man-in-space

User:
And when did he land on the moon?

LaMDA:
He did not land on the moon, he simply went into space.

User:
what's the point of being in space when you're not going to the moon?

LaMDA:
Some people like to go to space just to experience it. Also you can collect stuff in space.

User:
What did he collect?

LaMDA:
He brought some stuff with him, but he also brought back moon rock samples that he got from the moon13.

Generative AI systems work solely with numerical representations, without understanding the meaning of words for human beings. The meaning is exclusively that which humans project onto the outputs, because only humans have an interpretation in the real world. This complete lack of understanding may be without consequence for some uses, such as producing a poem or a piece of fiction, or it may have disastrous consequences if the texts provided are recommendations for critical decisions.

The system can produce output combining true and false assertions on a given subject. This is recognised by the designers of these systems, for example in the systematic warning at the bottom of the ChatGPT user window: "ChatGPT [date] Version. ChatGPT may produce inaccurate information about people, places, or facts" (note: this warning became "ChatGPT can make mistakes. Consider checking important information" in newer versions). This kind of warning is likely to be ignored or overlooked by users. Moreover, it is not always easy for the user to check what is true or false, especially as foundation models, by construction, do not produce reference to sources. Methods for attributing sources to generated texts are either a special module included in the model (as in Microsoft's Bing), or a search engine in the model's training corpus (as in HuggingFace's StarCoder).

This raises the question of truth. The lack of truth value assessment of statements made by generative AI systems can lead to the production of misinformation. As this production is asemantic and unintentional, it calls into question the responsibility of designers and our relationship to the ethics of truth[14].

These effects are also influenced by user choices, such as a parameter called "temperature" (in ChatGPT) or "creativity" (in Bing), which refers to a random choice, drawing language elements at random from the most likely outputs. Furthermore, human beings spontaneously project meanings on words, including the outputs of generative systems. These projections are all the stronger when the output in question closely resembles sentences produced by human beings, reinforcing the unfounded attribution of a truth value by the user.

RECOMMENDATION C3: THE USE OF QUALITY SOURCES FOR TRAINING

Designers should favour the use of quality sources, evaluated by explicit criteria, for the constitution and use of training corpora for generative AI models (pre-training), as well as for their optimisation, regardless of the training method. It is in particular necessary to consider the transparency and reasons to use artificial or synthetic content in training corpora.

RECOMMENDATION C4: CONSIDER THE EFFECTS OF MODEL HYPERPARAMETERS CHOICES

The choice of model hyperparameters, such as the size of numerical encodings in vector space (embeddings), is not just technical but can have repercussions on the behaviour of the system (including emergent behaviour) and, through them, effects on human beings and society. It is necessary to study the effects of hyperparameters on model outputs.

13. Thoppilan R. *et al.*, « *LaMDA: Language Models for Dialog Applications* », arXiv:2201.08239, table 16.
14. CNPEN, Bulletin de veille n°2, 21 juillet 2020, *Enjeux d'éthique dans la lutte contre la désinformation et la mésinformation..*

3.2. USER MANIPULATION WITHOUT RESPONSIBILITY

Machine-generated output can induce various risks of manipulation, whether intentional or not, of which human beings are unaware even if users know they are dealing with a machine. Manipulation can take place at several levels, with no malicious intent by the designers. Designers cannot predict the outcomes of these systems, or their effects on individuals and society. The perceived sometimes exagerated confidence of the responses produced by generative AI systems, such as the use of first-person ("I" or "we"), without inducing a perception of the trustworthiness of the response, can lead to manipulation[15].

- The machine may be perceived as more efficient or superior to human being. For example, generative AI systems use a good level of language. This creates a risk of manipulating users, who may feel impaired or incompetent compared to the machine's "capabilities".
- Interaction in natural language can lead users to speak more freely about their privacy and to believe that the machine is caring, giving the illusion of human empathy. A generative AI system can therefore put users in situations where they confide, but also to reveal confidential company information. These aspects require regulatory control.
- The lack of grounding in the physical world can lead the system to produce outputs, understood as advice, which may be inappropriate and may reinforce users' pre-existing psychological conditions.
- False or inaccurate information produced by generative AI systems could be used to feed training corpora of new language models. This "synthetic data" require a regulatory approach tailored to each field of use of generative AI systems.
- In RLHF process, filters can be seen as censorship. In addition, this process relies on the manufacturer's explicit instructions and is often the task of a poorly paid workforce who may not share the same cultural references as the users[16].
- On a societal level, the use of nudging methods by large language models can lead to political manipulation[17].

These manipulation risks call for a reflection on ethical issues at several levels. Effects and mitigation measures must be considered from the design stage, during use and when generative AI systems are massively deployed in society. This last aspect raises questions about trustworthiness, social inclusion and the digital divide.

Generative AI systems are often used as decision support systems. Depending on use terms and conditions and risk levels, the consequences of decisions influenced by machine may entail user liability. It is therefore important to train users in these new practices. "Know-how" to construct precise queries (or prompts) to obtain better responses appears as a prerequisite for several types of use-cases. Furthermore, it is necessary to build an ecosystem capable of identifying and sharing good and bad practices in the use of generative AI systems in different types of applications.

RECOMMENDATION C5: ASSESSMENT OF IDENTIFIED BIASES IN THE MODELS WITH STANDARDISED TEST SETS

In order to characterise biases in the language and to prevent discriminatory effects, especially cultural ones, the designers must apply a quantitative evaluation with standardised test sets and freely accessible evaluation corpora. The results of these evaluations should be made public simultaneously with the release of a foundation model.

RECOMMENDATION G3: SHARING PRACTICES IN THE USE OF GENERATIVE AI SYSTEMS

It is necessary to build an ecosystem capable of identifying good and bad practices in the use of generative AI systems in various applications. In particular, it is necessary to create a pooling platform and a monitoring agency. Results should be made available to all members of the generative AI community.

3.3. MAINTAINING DISTINCTIONS

It is often recalled in societal debates about generative AI is that LLMs can be used to write factually inaccurate press articles or to create misinformation on a massive scale. Specifically, generative models could be used to achieve a desired ranking of malicious content in recommendation algorithms on social networks or search engines, favouring specific political opinions. Furthermore, in the field of education, students around the world are using generative AI to write their dissertations or theses.

The lack of distinction between text written by a human being and that generated by an AI system is a major ethical problem. Users should not confuse a result produced by a machine with one created by a human author. Technically, this follows from a particular regularity introduced into the probabilistic

15. CNPEN, Opinion n° 3, septembre 15, 2021, *Ethical issues of conversational agents.*

16. On the use of Kenyan moderators by OpenAI : PERRIGO, B. *"OpenAI Used Kenyan Workers on Less Than $2 Per Hours to Make ChatGPT Less Toxic"*, Time Magazine, 18 janvier 2023.

17. Reisach, U. (2021). *The responsibility of social media in times of societal and political manipulation*. European Journal of Operational Research, 291(3), 906–917 ; Panai & Devillers 2023 : *How AI-augmented nudges may impact EU consumer in a moral situation?* (ed.) M. Ho-Dac & C. Pellegrini, Governance of Artificial Intelligence in the European Union. What Place for Consumer Protection?, Brussels, Bruylant, 2023.

choice of tokens[18]. Maintaining distinctions allows for the assignment of responsibilities for any potential harm. If a text raises an ethical tension, for example due to nudging or fraud it contains, it is imperative to trace its origin to avoid confusion between the production of a responsible agent, accountable for what they say, and the non-semantic utterance of an artificial intelligence system to which no responsibility can be attributed.

The systematic introduction of watermark codes in sufficiently long and elaborate LLM results would make it possible to maintain the possibility of distinguishing the production of a machine from that of a human author. However, the use of watermark techniques for LLMs must remain inconspicuous. Watermarks must be detectable with minimal effort and be robust enough to resist adversarial attempts to blur the origin of the text by removing them. Although the effectiveness of watermarks cannot be absolutely guaranteed[19], their introduction is a necessary regulatory step for ethical reasons. These watermarks must meet two criteria that are difficult to reconcile. On the one hand, they must be robust enough to resist attacks aimed at erasing them. On the other hand, they must be interoperable, *i.e.*, their detection by verification software must not depend on the parameters (*e.g.*, tokenisation) of a particular AI system that generated the text. For this to become possible, the watermarks introduced by the manufacturers of AI systems must be identifiable in a homogeneous way within the same approach. The balance between these two requirements remains to be found, and remains a significant research challenge in the field of generative AI and LLMs.

RECOMMENDATION C6: MAINTAINING DISTINCTIONS

Designers of foundation models should implement a technical solution (watermark) to ensure that the user will be able to distinguish — as much as reasonably possible — model output from human production. Research on watermarks should be intensified.

RECOMMENDATION G4: REGULATION OF WATERMARKS

The obligation to insert watermarks (see recommendation D6) should be mandated at the regulatory level.

3.4. PROJECTION OF HUMAN QUALITIES

The mere fact of machines using language, which is the means of conscious thought and judgement, arises the projection of human traits onto the machine. This projection does not relieve the moral charge of words, by completely separating the language generated from meanings, associations and judgements. The literal meanings of words, which are implicit in language, spontaneously arise in our minds. Evading these immediate projections of meaning requires special skills that not all users possess. The meaning attributed to the generated language is merely a projection from human dialogues, but it's enough to attribute intention and knowledge to the machine.

There are three main types of transfer between human beings and LLMs.

The first involves the projection of knowledge: after its training, a language model appears to "know" a lot of things. The "knowledge" of an LLM is merely an illusion, but the uer believes that the machine genuinely possesses it. The second type of transfer is that of emotional states and affects. Through generated content, the machine can induce in the user the impression that it possesses emotions or moods, even though the user knows it is a computer program. The third type of transfer is that of moral qualities. Whether a generative AI system is perceived as "benevolent", "caring" or "lecturing", these perceptions only exist through projections. The LLM never becomes a moral agent, or a person in the legal sense of the term. Yet, projection of moral qualities can go so far as attributing responsibility to a machine that, by its very nature, cannot bear any.

The CNPEN wishes to extend some arguments expressed in its opinion on the ethical challenges of conversational agents to the use of LLMs.

"Conversational agents are increasingly integrated into various aspects of human life. Their use raises ethical tensions, which in turn leads to the question of responsibility is raised in all its forms: legal and moral responsibility, individual and collective responsibility, the responsibility of the designer, manufacturer, user and policy-maker, as well as the responsibility for any malfunctions and the long-term consequences of these technologies. The sharing of responsibility is evaluated on a case by case basis, depending on the technical aspects and the role played by the user, the developer, and the manufacturer in each of the situations that cause ethical tensions."[20]

18. Aaronson, S. (2022). My AI Safety Lecture for UT Effective Altruism. Shtetl-Optimized. https://scottaaronson.blog/?p=6823 ; Grinbaum, A., & Adomaitis, L. (2022). *The Ethical Need for Watermarks in Machine-Generated Language.* arXiv:2209.03118 ; Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). *A Watermark for Large Language Models.* arXiv:2301.10226.

19. Sanadisvan *et al.*, Can AI-Generated Text be Reliably Detected? 2023 : https://arxiv.org/abs/2303.11156

20. CNPEN, Opinion n° 3, septembre 15, 2021, *Ethical issues of conversational agents.* P6.

Projections of states of knowledge onto AI systems can have practical benefits, such as facilitating a dialogue or appearing to provide a solid basis for medical advice. However, they can also cause harm, for example, when the chatbot provides incorrect information or suggests an action that is harmful to the user. Furthermore, they can mislead inexperienced or ill-prepared users. Given that anthropomorphism can occur even if the user is aware that the text comes from a machine, the responsibility is — and must be — that of human beings, as the machine is not a moral agent and should under no circumstances be considered as a person. Responsibility is therefore shared between the user and the stakeholders throughout the product's value chain. This sharing takes place on a case-by-case basis, depending on the technical aspects and on the involvement of each party in each situation that presents ethical tensions.

RECOMMENDATION C7: REDUCE THE PROJECTION OF HUMAN QUALITIES ONTO GENERATIVE AI SYSTEMS

To reduce the spontaneous projection of human qualities onto generative AI systems and the attribution of a sense of interiority to foundation models, the models suppliers must apply specific control and filtering mechanisms. They must also inform the user of possible anthropomorphisation biases.

3.5. EMERGENT BEHAVIOURS

Large Language Models can produce unexpected or surprising results for their users, but also for their designers, when faced with ambiguous or complex requests. This is an "emergent capability" or "emergent behaviour" and it is so because it is not present in smaller models[21,22]. Transformer-based LLMs exhibit several types of emergent behaviour, such as "reasoning" capabilities triggered by "reason step by step" prompts[23]. The precise scientific explanation of the emergence phenomenon in LLMs is a subject of current research and surely depends on the parameters of the model. This behaviour is certainly related to phenomena described by statistical physics[24]. Thus, these emergent behaviours result from a complex interaction between the layers and parameters of the models, which result themselves from the training on huge datasets. As the models learn relationships and inherent structures in the training data, they unintentionally develop linguistic and contextual "capabilities" or "skills", enabling them to generate unexpected but relevant outputs. One example of this is the ability of the GPT-4 model (without optimisation) to pretend to be a visually impaired person in order to get an internet user to solve a captcha for it, thereby creating the illusion it is lying or deceiving[25].

The main uncertainty linked to emergent behaviours is the difficulty of predicting them. We can also reasonably assume that new types of emergent behaviour, still unknown, will occur with the increasing use of LLMs. This raises concerns about using these models in critical or sensitive applications, where an inappropriate response could lead to harmful consequences.

RECOMMENDATION C8: STUDY EMERGENT BEHAVIOURS AND THE UNKNOWN EFFECTS OF MODELS

Generative models can produce potentially dangerous outputs in various forms, such as hate speech. Before releasing a foundation model, its designers should conduct studies and research on its emergent behaviour, possibly with an independent team ("red team") to carry out adversarial tests. The results of these tests must be made public simultaneously with the model's release.

21. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., *et al.* (2022). *Emergent Abilities of Large Language Models.* arXiv:2206.07682
22. Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are Emergent Abilities of Large Language Models a Mirage? https://arxiv.org/abs/2304.15004
23. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). *Chain of thought prompting elicits reasoning in large language models.* arXiv:2201.11903
24. Roberts D. A., Yaida S. and Hanin B. (2021). *The principles of deep learning theory.* arXiv:2106.10165
25. OpenAI, GPT-4 technical report, 2023, p. 15.

3.6. MULTILINGUALISM AND LANGUAGE DOMINANCE

The data used to train generative AI systems are generally multilingual. For example, the transformer BLOOM, developed by a private/public consortium in 2022 and trained using the Jean Zay computer at Saclay (France), from a massive dataset containing 1.6 terabytes of text in fifty-nine languages, including forty-six human languages and thirteen programming languages (10% of the corpus). The proportion of different languages was quite uneven, for example the size of the French corpus (12.9%) was less than half the English corpus (30 %). Machine learning corpora for pre-training language models often include a much higher proportion of English (by an order of magnitude or more) than data in other languages, including French. Indeed, text corpora vary in size from one language to another. Some, like Mandarin or English, have large databases, while others have only moderate (*e.g.*, French) or even limited (*e.g.,* Swahili) resources. For languages with few written resources, the multilingual system improves the ability to generate texts in these languages by implicitly borrowing linguistic knowledge from other languages in the training corpora. Every human language necessarily carries a history and a culture. The simple fact of using language, which is the means of conscious thought and judgement, implicitly mobilises cultural representations. The political and civilisational charge of language cannot be separate from meanings and values.

Therefore, it is crucial to be aware of the effect of data in dominant languages, such as English, on the system responses, regardless of the language they are expressed in. More research is needed, particularly in developing test sets, to assess and understand these effects. A balanced multilingual, and therefore "multicultural", model could respond more easily to a variety of requests, but its political or educational effects need to be studied, as they generate biases. Monolingual models also need to be developed and their performances compared to those of multilingual models[26]. Indeed, a model trained on hundreds of languages from different families is at risk of suffering from the "curse" of multilingualism, which leads to a decreased performance per language as it covers more languages. However, solutions exist to mitigate the negative interference between languages, which even tend to improve monolingual and interlingual performance[27].

RECOMMENDATION C9: DEVELOP GENERATIVE AI SYSTEMS IN DIFFERENT LANGUAGES REFLECTING CULTURAL DIVERSITY

When building training corpora for generative AI systems, designers must respect the diversity of human languages and the cultures they convey. Even if multilingual learning can be useful to compensate for the lack of data in a language with a smaller corpus, the influence of a dominant language on the generation of texts in another language needs to be studied, in particular the preponderance of the English language. Following scientific studies, designers must thoughtfully and proactively take technical measures to adhere to this principle. Comparative research between multilingual and monolingual models is also necessary.

3.7. EDUCATION AND IMPLICATIONS FOR HUMAN LEARNING

Generative AI systems have found immediate applications in education. Their ability to produce syntactically correct and semantically plausible texts in natural language makes them a unique tool. They can be used by students to write texts for them or to answer questions on an assignment, or by teachers to produce course summaries or descriptions, or to generate multiple-choice questions. Besides the obvious ethical issues of integrity and honesty, such as having homework done by a machine, several questions arise regarding the use of generative AI systems in education.

Human learning is a journey. The understanding of concepts, the assimilation of knowledge and the acquisition of know-how are all achieved through reflection, reformulations, analyses and syntheses. This journey uses thought, which is expressed through language. While the purpose of education is to shape minds and teach them how to reason rigorously, there is an obvious risk of replacing this goal with that of acquiring knowledge, the accuracy of which is furthermore not even guaranteed, from the machine. Human creativity would thus be little solicited.

Replacing human reasoning with machine-generated text would be contrary to the classical approach to learning at school, which must be preserved. The aim is not to ban these new tools but to frame their usage and to teach children the underlying concepts.

The use of LLMs will encourage humans to work differently, and also to learn differently. Through the generated texts, the machine exerts an influence on human opinions and on the appreciation of beauty and truth. The evolution of the education system should not exclude generative AI, but incorporate it. It is therefore necessary to equip teachers with

26. Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., & Auli, M. (2023). *Scaling Speech Technology to 1,000+ Languages.* https://arxiv.org/abs/2305.13516

27. Pfeiffer J., Goyal N., Lin X., Li X., Cross J., Riedel S., Artetxe M., *Lifting the Curse of Multilinguality by Pre-training Modular Transformers,* ACL 2022.

adapted pedagogical methods so that students can develop exclusively human skills and preserve their ability to learn without relying on machines.

As for the requirement to maintain distinctions between an essay or dissertation written by a student and one generated by software, public authorities must provide teachers, professors and students with distinction software tools, inspired by anti-plagiarism software. This would require robust and interoperable watermarking code (see section 3.3 "Maintaining Distinctions").

RECOMMENDATION C10: ENABLE SYSTEM CONFIGURATION BY USERS FOR AN EASIER USAGE

Designers should enable the user to costumize the parameters of the AI system, in particular according to the accuracy sought in its responses, by acting on its ability to generate content that is less statistically likely ("temperature" parameter to modify the system's "creativity"). The transparency of the context, its size and its content, could increase the user's understanding of the system.

3.8. ISSUES RELATED TO OPEN ACCESS AND OPEN-SOURCE SOFTWARE

Open access publication of generative AI models has become the industry standard in recent years. The current ecosystem is made up of hundreds, and even thousands, of individual developers and start-ups present on dedicated sharing platforms, such as Github or Hugging Face. As far as openness is concerned, a distinction needs to be made between open-access publication of the models themselves, of training data, or of test sets used for testing or system optimization.

The CNPEN is convinced that the development of LLMs benefits significantly from their openness[28]. The open-source strategy of a number of companies developing generative AI, including major ones such as Meta, makes it possible to increase model transparency by improving evaluation techniques and identifying risks and safety and security vulnerabilities more quickly through collective research effort. This openness also promotes competition.

Some major players in the field of generative AI (for example, OpenAI or Google) are pursuing a different strategy. The GPT-2 model was first made public by OpenAI in 2019 in an abridged version[29]. This decision was motivated by the possible misuse of the model, particularly for the automatic generation of disinformation. These risks of misuse led OpenAI to keep the full version of GPT-2 undisclosed, until its performance was replicated about six months later by a competing model, accessible to all[30]. In the case of GPT-4, the same risk of misuse was cited by OpenAI to delay its publication, allowing time for optimization by developing a set of filters.

The current dilemma regarding the openness of generative AI models reflects similar dilemmas, notably in biotechnology, which are referred to by the acronym DURC (Dual Use Research of Concern)[31]. The CCNE proposes a series of recommendations[32], several of which are relevant for generative AI.

RECOMMENDATION G5: THE USE OF GENERATIVE AI SYSTEMS IN EDUCATION

The introduction of generative AI systems in education, training, and teaching should be considered only after prior studies of their effects on pedagogy and the cognitive development of learners.

RECOMMENDATION G6: OPEN-SOURCE FOUNDATION MODELS

Open access of foundation models should be conditioned on their designers' awareness of the challenges of openness and the risks of misuse. Transparency and evaluation criteria must be made explicit and applied.

28. LAION, *An Open Letter to the European Parliament*, 2023.
29. https://openai.com/blog/better-language-models/
30. Cohen V.,« *OpenGPT-2: We Replicated GPT-2 Because You Can Too* », 22 août 2019.
31. Grinbaum A., Adomaitis L., (2023) *Dual Use Concerns of Generative AI and Large Language Models.* arXiv:2305.07882
32. CCNE, "*Recherches duales à risque. Recommandations pour leur prise en compte dans les processus de conduite de recherche en biologie*", 2019.

4. LEGAL ISSUES

4.1. LEGAL RULES IMPOSED ON GENERATIVE AI SYSTEMS AND FOUNDATION MODELS

There is a recent international rush to introduce measures to regulate generative AI (China, United States, United Kingdom, Canada), which highlights the economic and political importance of these technologies. In Europe, the draft regulation presented in April 2021 by the European Commission received numerous amendments after it was examined by the European Council and the European Parliament. Some of these amendments reveal the emergence of generative artificial intelligence systems in the public debate during the year 2022, and the difficulty of striking a balance in the choice of constraints to be imposed on these systems. The text that will be adopted by the three European institutions following the "trilogues" will be the result of a reflection that has become essential due to the rapid development of these systems, that had not been taken into account in the initial draft.

While the European Commission's initial proposal simply referred to artificial intelligence systems (and the risk levels related to their intended purpose), the European Council introduced in 2022 the category of "general-purpose artificial intelligence systems". In 2023, during its own review of the text, the European Parliament added a new category, that of "foundation model", defined as an AI model that is "trained on broad range of data at scale, is designed for generality of output, and can be adapted to a wide range of distinctive tasks". Furthermore, it introduced generative artificial intelligence systems by proposing that '*providers of foundation models used in AI systems specifically intended to generate, with varying levels of autonomy, content such as complex text, images, audio, or video ("generative AI") and providers who specialise a foundation model into a generative AI system*'[33], should meet complementary obligations.

The distinction between these different concepts should be clarified. These terminological tensions reveal both the urgency of the debate and the difficulty of positioning the cursor to refer to rules applicable to systems placed on the market and their components.

The Commission's text proposes a regulation based on risk. The level of risk is either unacceptable, which means that AI systems at this level are banned, or high, which means that AI systems at this level are explicitly regulated, or limited, which imposes transparency obligations on AI system providers. High-risk qualification is applied to a set of domains listed in Annex III, which may be completed by the Commission. The regulation assumes a declarative compliance regime, rather than an authorisation regime.

This regulatory choice has not been contested by the European Council or the European Parliament. The debate focuses on the choice of risk levels to be imposed on generative AI systems. The question arises of the degree of constraints on these systems in terms of transparency, traceability, risk management, data governance, etc. The Council has introduced specific provisions on general-purpose artificial intelligence systems, after hesitating between including them among the high-risk systems or applying a limited number of requirements. The European Parliament, on the other hand, refers to this concept, but does not treat it in any particular way; instead, it includes foundation models in the section devoted to high-risk systems, subjecting them to specific obligations. The question that arises is how to target the ethical issues raised by the introduction of generative AI systems onto the market, so that they can be governed by legal norms that are sufficiently flexible to adapt to new developments while providing with a sufficient framework to respect fundamental rights and the integrity of individuals. Without unleashing the development of generative AI, or seeking to prohibit it, it is necessary to regulate it and to define its limits.

In parallel to this legal debate, which is essentially focused on the introduction of a product onto the market, the Council of Europe is studying a framework for the development of artificial intelligence respecting human rights and democracy, which may possibly introduce additional constraints for generative AI systems.

RECOMMENDATION G7: CONSIDER FOUNDATION MODELS INTRODUCED ONTO THE MARKET AND GENERATIVE AI SYSTEMS AS HIGH-RISK AI SYSTEMS

In the context of the European AI Act, it is necessary to consider foundation models brought to market and generative AI systems as high-risk AI systems. However, the publication of an open-source foundation model under a non-commercial license should not be considered as an introduction onto the market; nonetheless it should entail obligations of transparency and evaluation by the designers.

RECOMMENDATION G8: CHAIN OF RESPONSIBILITY

Legal accountability for generative AI systems and foundation models should be attributed to the providers of foundation models and to the deployers of specific generative AI applications based on such models. Furthermore, moral responsibility extends to the designers of foundation models and the developers of generative AI systems using such models.

33. *Draft Compromise Amendments on the Draft Report Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence Act and amending certain Union Legislative Acts.* 16/5/2023. (COM(2021)0206 – C9 0146/2021 – 2021/0106(COD)).

4.2. THE GDPR AND GENERATIVE AI SYSTEMS

The draft regulation on artificial intelligence (AI Act) states that its requirements do not prevent the application of the GDPR when processing of personal data is involved. Consequently, the main principles of the GDPR concerning the collection and processing of personal data (definition of purpose and determination of the legal basis for processing, minimisation principles, the right to oppose, etc.) apply to generative AI systems that make use of this type of data. The GDPR is applicable to foreign companies as long as they operate within EU territory.

Investigations conducted by Italian and German data protection authorities have revealed GDPR violations by several generative AI players (OpenAI[34], Microsoft, Google). The main concerns are the use of personal data without prior information to the data subjects, the position to adopt regarding to the use of this data (should consent be sought or are there cases where the company can justify a legitimate interest in using the data without going through the consent procedure set out in Article 6 of the GDPR?), the lack of a legal basis for the extensive collection of data used to train AI models, the absence of age verification for users, the ability for users to access their personal information and request rectification, and the issue of confidential data submitted in requests (personal data, data revealing ongoing or unpublished works, industry trade secrets, defence secrets)[35].

It is not clear whether it would be necessary to change the current GDPR framework, but vigilance is called for. Compliance with the GDPR is not easy in certain areas, such as the right to be forgotten. Specifically, it is technically impossible for a foundational transformer model, which uses the attention mechanism, to forget what it had previously learned.

RECOMMENDATION G9: THE GDPR AND GENERATIVE AI SYSTEMS

It is essential for the European Data Protection Board to produce guidelines related to the interaction between the AI Act and the GDPR, in order to clarify the degree of flexibility with which the latter can be interpreted in the context of the development of generative AI in Europe.

4.3. COPYRIGHT LAW AND GENERATIVE AI SYSTEMS

The recent Directive 2019/790 of the European Parliament and of the Council of April 17 2019 on copyright and related rights in the Digital Single Market, which amends two directives from 1996 and 2001, aims to ensure a high level of protection for rights holders, while stimulating innovation and the production of new content, including in the digital environment.

However, the rapid development of digital technologies is changing the way in which works are created, produced, distributed and exploited. Generative AI raises questions about copyright both upstream (online data input into the AI system) and downstream (system response to user prompts, and the use of the raw or reworked data by the user). On this point, the European Commission seems to have adopted a wait-and-see stance[36].

The CNPEN stresses the need for academic research, multidisciplinary reflection, and discussions between Member States on the need to adapt existing law, or even to envisage special law, on the following issues:

- Beyond these general considerations, it is important to note that generative AI systems raise new challenges for copyright. In particular, regarding the exception or limitation for text and data mining (Article 4 of the 2019 Directive). These exceptions allow extracting data that is legally accessible to the public, to train AI, unless the rights holder objects “in an appropriate manner, including through machine-readable methods”. When, in practice, and how (in website metadata, in General Terms and Conditions of Use, etc.) to assert this objection? This Article 4 was discussed at a time when the European legislator did not anticipate that generative AI systems would one day be likely to create new texts using mined data. Is this “opt out” mechanism sufficient, given the unforeseen uses of mined texts at the time the directive was drafted?
- How is an intellectual work used in a learning process involving tokenisation without human meaning? The question of the mention of the source in answers given by generative AI systems is also an issue that calls for reflection.
- Finally, the classic questions of the legal status of a work generated by a generative AI system must be raised. It would be necessary to distinguish the case where the work was created by a human with the help of a generative AI system, from the case of a work generated entirely by a generative AI system. Under current law, the author cannot be an AI system since AI does not have legal personality.

34. "OpenAI's privacy policy, which claims to comply with the privacy rights of the State of California, does not comply with the provisions of the GDPR and the French Data Protection Act: there is no mention in the privacy policy of the legal bases for processing, no retention period for the collected and processed data, no right to limitations, including portability, no possibility to withdraw consent to the processing of data and in particular to the processing of sensitive data under Article 9 of the GDPR, category". See : https://www.village-justice.com/articles/chatgpt-quels-enjeux-juridiques,45027.html
35. What about ChatGPT's FAQ, written by OpenAI: "(...) we review conversations to improve our systems and to ensure that content complies with our policies and security requirements"?
36. European Commission, *Directorate-General for Communications Networks, Content and Technology, Study on copyright and new technologies – Copyright data management and artificial intelligence,* Publications Office of the EU, 2022, https://data.europa.eu/doi/10.2759/570559

RECOMMENDATION G10: PROCESSING OF COLLECTED DATA

In line with the existing framework for personal data, it is necessary to develop legal rules and also ethical questioning on the collection, storage and reuse of linguistic traces of interactions between language models and human beings.

RECOMMENDATION G11: COPYRIGHT AND GENERATIVE AI

There is a need to initiate scientific research, multidisciplinary reflections, and discussions between States on the need to adapt existing copyright law in light of generative AI techniques.

4.4. EUROPEAN LEGISLATION ON LIABILITY

The draft regulation on artificial intelligence does not concern the liability regime for operators. However, by setting legislative norms applicable to AI systems brought into the market, it requires the supplier to comply with product conformity standards to limit risks arising from these systems. Non-compliance could result in *ex post* administrative fines. Alongside this text, which is upstream of the introduction onto the market, two draft directives are under study to regulate the market downstream. These aim to allow natural persons or, in certain cases, legal entities to seek personal compensation in the event of damage.

The proposal to revise Directive 85/374/EEC of July 25, 1985, on liability for defective products (Parliament and Council) presented in September 2022 is currently under discussion. The directive would then apply to all AI systems, which was not the case until now. While the European legislator's stated aim is to facilitate redress for damage caused by AI systems in a context where the defendant alone holds the information, the Parliament and Council's debates on this text are significantly influenced by the question of the extent of evidence required from the plaintiff. The issue of covering intangible damage caused by AI systems is also being considered.

The proposed Directive of the European Parliament and the Council on adapting extra-contractual civil liability to artificial intelligence (AI Liability Directive), presented on the same day, aims to fill the gaps of the previous text, by extending the cases of liability. This is closely related to the draft regulation on AI and therefore its review has been postponed until the AI Act is adopted. These legislative changes would considerably strengthen user protection for generative AI systems. The attention paid by companies to these developments indicates their significant economic implications.

5. ECOLOGICAL AND ENVIRONMENTAL ISSUES

The environmental impact of the extremely rapid development of digital technology is becoming a major concern[37]. The current challenge is to measure the energy cost and, more generally, the environmental footprint of generative AI systems and foundation models to incorporate them in the ecological transition[38]. To properly measure this environmental footprint, it is necessary to quantify resources consumption: i) for the manufacture of the physical infrastructures dedicated to these systems, particularly data storage centres; ii) for the pre-training of foundation models, and iii) for the marginal cost of queries submitted to foundation models.

A recent study shows that most current foundation models do not meet these ecological requirements. Researchers at Stanford University compared ten foundation models from different providers. They found that «foundation model providers inconsistently report energy usage, emissions, their strategies for measurement of emissions, and any measures taken to mitigate emissions"[39]. It should be noted that, in this study, the BLOOM (Big Science) and LLaMA (Meta) models appeared to be the best rated according to the "energy" criterion.

RECOMMENDATION G12: ENVIRONMENTAL IMPACT OF GENERATIVE AI

It is necessary to develop a metric for the environmental footprint of generative AI systems and foundation models and to demand greater transparency on environmental effects from designers.

37. ADEME & Arcep : *Évaluation de l'impact environnemental du numérique en France - Analyse prospective à 2030 et 2050 2023*. Voir : https://www.arcep.fr/uploads/tx_gspublication/etude-prospective-2030-2050_mars2023.pdf ; The Shift Project, *Planifier la décarbonation du système numérique en France : cahier des charges* - note de mai 2023 - Voir https://theshiftproject.org/article/planifier-la-decarbonation-du-systeme-numerique-en-france-cahier-des-charges/ ; INRIA, « *Le numérique est-il un progrès durable ?* », Pour la Science, supplément réalisé en partenariat avec l'INRIA n° 546 – Avril 2023. Voir : https://www.inria.fr/fr/numerique-progres-durable-environnement-pour-la-science
38. OECD (2022), "*Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint*", OECD Digital Economy Papers, No. 341, OECD Publishing, Paris, https://doi.org/10.1787/7babf571-en.
39. See : https://crfm.stanford.edu/2023/06/15/eu-ai-act.html

6. RECOMMENDATIONS FOR DESIGN, RESEARCH AND GOVERNANCE

The design of generative AI systems raises numerous research questions. Recommendations in section 6.1 therefore combine the design of these systems with the research questions inherent in foundation models and optimisation methods. Section 6.2 focuses on governance recommendations.

6.1. RECOMMENDATIONS FOR THE DESIGN AND RESEARCH OF GENERATIVE AI SYSTEMS

RECOMMENDATION C1: ETHICS IN THE DESIGN OF AND RESEARCH ON GENERATIVE AI SYSTEMS

The designers of a generative AI system must analyse, during the design phase, each of the technological choices likely to give rise to ethical tensions. If a potential tension is identified, they must methodically consider a technical solution based on research aimed at reducing or eliminating the ethical tension, and then evaluate this solution in realistic usage contexts.

RECOMMENDATION C2: DESIGNERS SHOULD AVOID OVER-POLICING MODELS

The limitations placed on models by designers must remain reasonable and proportionate to the proven risks, while respecting the desired purposes and useful functionalities of the models. Designers must take care not to alter the language generated beyond what is necessary, in particular for regulatory or ideological reasons.

RECOMMENDATION C3: THE USE OF QUALITY SOURCES FOR TRAINING

Designers should favour the use of quality sources, evaluated by explicit criteria, for the constitution and use of training corpora for generative AI models (pre-training), as well as for their optimisation, regardless of the training method. It is in particular necessary to consider the transparency and reasons to use artificial or synthetic content in training corpora.

RECOMMENDATION C4: CONSIDER THE EFFECTS OF MODEL HYPERPARAMETERS CHOICES

The choice of model hyperparameters, such as the size of numerical encodings in vector space (embeddings), is not just technical but can have repercussions on the behaviour of the system (including emergent behaviour) and, through them, effects on human beings and society. It is necessary to study the effects of hyperparameters on model outputs.

RECOMMENDATION C5: ASSESSMENT OF IDENTIFIED BIASES IN THE MODELS WITH STANDARDISED TEST SETS

In order to characterise biases in the language and to prevent discriminatory effects, especially cultural ones, the designers must apply a quantitative evaluation with standardised test sets and freely accessible evaluation corpora. The results of these evaluations should be made public simultaneously with the release of a foundation model.

RECOMMENDATION C6: MAINTAINING DISTINCTIONS

Designers of foundation models should implement a technical solution (watermark) to ensure that the user will be able to distinguish - as much as reasonably possible – model output from human production. Research on watermarks should be intensified.

RECOMMENDATION C7: REDUCE THE PROJECTION OF HUMAN QUALITIES ONTO GENERATIVE AI SYSTEMS

To reduce the spontaneous projection of human qualities onto generative AI systems and the attribution of a sense of interiority to foundation models, the models suppliers must apply specific control and filtering mechanisms. They must also inform the user of possible anthropomorphisation biases.

RECOMMENDATION C8:
STUDY EMERGENT BEHAVIOURS AND THE UNKNOWN EFFECTS OF MODELS

Generative models can produce potentially dangerous outputs in various forms, such as hate speech. Before releasing a foundation model, its designers should conduct studies and research on its emergent behaviour, possibly with an independent team ("red team") to carry out adversarial tests. The results of these tests must be made public simultaneously with the model's release.

RECOMMENDATION C9:
DEVELOP GENERATIVE AI SYSTEMS IN DIFFERENT LANGUAGES REFLECTING CULTURAL DIVERSITY

When building training corpora for generative AI systems, designers must respect the diversity of human languages and the cultures they convey. Even if multilingual learning can be useful to compensate for the lack of data in a language with a smaller corpus, the influence of a dominant language on the generation of texts in another language needs to be studied, in particular the preponderance of the English language. Following scientific studies, designers must thoughtfully and proactively take technical measures to adhere to this principle. Comparative research between multilingual and monolingual models is also necessary.

RECOMMENDATION C10:
ENABLE SYSTEM CONFIGURATION BY USERS FOR AN EASIER USAGE

Designers should enable the user to costumize the parameters of the AI system, in particular according to the accuracy sought in its responses, by acting on its ability to generate content that is less statistically likely ("temperature" parameter to modify the system's "creativity"). The transparency of the context, its size and its content, could increase the user's understanding of the system.

6.2. RECOMMENDATIONS ON GOVERNANCE

RECOMMENDATION G1:
CREATE A SOVEREIGN RESEARCH AND TRAINING ENTITY FOR "AI, SCIENCE AND SOCIETY"

Given the complexity of the issues involved in generative AI and its medium- and long-term impacts, it is necessary to create a sovereign entity (a centre of competence) dedicated to research and training on the ethical issues of AI systems in relation to their scientific, technical, societal and environmental impacts.

RECOMMENDATION G2:
SPEED OF ADOPTION BY ECONOMIC STAKEHOLDERS

Economic actors and public authorities must exercise caution in the speed of adoption of generative AI systems and ensure prior and continual assessments.

RECOMMENDATION G3:
SHARING PRACTICES IN THE USE OF GENERATIVE AI SYSTEMS

It is necessary to build an ecosystem capable of identifying good and bad practices in the use of generative AI systems in various applications. In particular, it is necessary to create a pooling platform and a monitoring agency. Results should be made available to all members of the generative AI community.

RECOMMENDATION G4:
REGULATION OF WATERMARKS

The obligation to insert watermarks (see recommendation D6) should be mandated at the regulatory level

RECOMMENDATION G5:
THE USE OF GENERATIVE AI SYSTEMS IN EDUCATION

The introduction of generative AI systems in education, training, and teaching should be considered only after prior studies of their effects on pedagogy and the cognitive development of learners.

RECOMMENDATION G6:
OPEN SOURCE FOUNDATION MODELS

Open access of foundation models should be conditioned on their designers' awareness of the challenges of openness and the risks of misuse. Transparency and evaluation criteria must be made explicit and applied.

RECOMMENDATION G7:
CONSIDER FOUNDATION MODELS INTRODUCED ONTO THE MARKET AND GENERATIVE AI SYSTEMS AS HIGH-RISK AI SYSTEMS

In the context of the European AI Act, it is necessary to consider foundation models brought to market and generative AI systems as high-risk AI systems. However, the publication of an open-source foundation model under a non-commercial license should not be considered as an introduction onto the market; nonetheless it should entail obligations of transparency and evaluation by the designers.

RECOMMENDATION G8:
CHAIN OF RESPONSIBILITY

Legal accountability for generative AI systems and foundation models should be attributed to the providers of foundation models and to the deployers of specific generative AI applications based on such models. Furthermore, moral responsibility extends to the designers of foundation models and the developers of generative AI systems using such models.

RECOMMENDATION G9:
THE GDPR AND GENERATIVE AI SYSTEMS

It is essential for the European Data Protection Board to produce guidelines related to the interaction between the AI Act and the GDPR, in order to clarify the degree of flexibility with which the latter can be interpreted in the context of the development of generative AI in Europe.

RECOMMENDATION G10:
PROCESSING OF COLLECTED DATA

In line with the existing framework for personal data, it is necessary to develop legal rules and also ethical questioning on the collection, storage and reuse of linguistic traces of interactions between language models and human beings.

RECOMMENDATION G11:
COPYRIGHT AND GENERATIVE AI

There is a need to initiate scientific research, multidisciplinary reflections, and discussions between States on the need to adapt existing copyright law in light of generative AI techniques.

RECOMMENDATION G12:
ENVIRONMENTAL IMPACT OF GENERATIVE AI

It is necessary to develop a metric for the environmental footprint of generative AI systems and foundation models and to demand greater transparency on environmental effects from designers.

ANNEX 1:
PEOPLE AUDITIONED

Guillaume Avrin

National coordinator for Artificial intelligence - Interministerial coordination of the national strategy in artificial intelligence

Emmanuelle Legrand

Seconded magistrate and IA project manager at the DGE Ministry of the Economy, Finance industrial and digital sovereignty

Ludovic Peran

Product manager for responsible AI, Google et

Sarah Boiteux

Head of corporate affairs at Google France - Government Affairs and Public Policy Manager

Stéphane Requena

Director of Innovation & Technology at GENCI

Thomas Wolf

Co-founder - CSO Hugging Face

Henri Verdier

Ambassador for digital issues at the French Ministry of Europe and Foreign Affairs

ANNEX 2:
WORKING GROUP COMPOSITION

CO-RAPPORTEURS:

Raja Chatila
Laurence Devillers
Alexei Grinbaum

MEMBERS OF THE WORKING GROUP:

Claude Kirchner
Caroline Martin
Jérôme Perrin
Catherine Tessier

CONTRIBUTORS:

Gilles Adda
Laure Coulombel
David Gruson
Christine Froidevaux
Eric Germain
Anaëlle Martin
Emmanuelle Legrand

ANNEX 3:

REFERRAL FROM JEAN-NOËL BARROT, MINISTER DELEGATE FOR DIGITAL AFFAIRS

MINISTÈRE
CHARGÉ DE LA TRANSITION
NUMÉRIQUE ET DES
TÉLÉCOMMUNICATIONS
Liberté
Égalité
Fraternité

Paris, le 20 FEV. 2023

JEAN-NOËL BARROT

Ministre délégué

Nos références : D23-02038

Monsieur le Directeur, Cher Claude,

Les rapides avancées technologiques du numérique permettent aujourd'hui la mise en œuvre de systèmes d'intelligence artificielle dite générative s'appuyant sur des algorithmes permettant d'obtenir des résultats qui ressemblent à ce que peut produire un être humain. Ces résultats peuvent être des textes, des images, des sons, des vidéos.

Depuis quelques mois, certains de ces systèmes sont mis à disposition d'un large public, et la qualité grandissante de leurs résultats encourage leur utilisation à des fins professionnelles ou personnelles, y compris par exemple dans le cadre d'activités éducatives, artistiques ou ludiques.

Dans le cas de la génération de textes, ces programmes d'apprentissage machine prédisent, à partir de vastes corpus linguistiques, le mot ou la séquence de mots susceptibles d'être pertinents dans un contexte donné.

Les performances de ces nouvelles générations de modèles de langage de grande taille et l'engouement qu'ils suscitent apparaissent toutefois indissociables des interrogations d'ordre éthique concernant notamment le rapport de la société à l'information et la manipulation de l'information, les risques de désinformation, et la transformation des métiers, l'impact de ces outils en matière d'éducation et d'enseignement, ou encore l'impact sur les pratiques scientifiques ou artistiques.

1/2

Monsieur Claude KIRCHNER
Directeur du Comité national pilote
d'éthique du numérique (CNPEN)
66 rue de Bellechasse
75007 Paris

Dans sa lettre de mission de juillet 2019, le Premier ministre avait souhaité que le Comité national pilote d'éthique du numérique mène une réflexion sur les agents conversationnels, réflexion rendue publique en septembre 2021. Les améliorations techniques incontestables des modèles de langage de grande taille nécessitent toutefois de poursuivre la réflexion sur les enjeux éthiques liés au développement de ces technologies à grande échelle.

Dans ce contexte, je souhaiterais que le Comité national pilote d'éthique du numérique examine les questions d'éthique liées à la conception, aux usages, aux impacts sur la société ainsi que les accompagnements nécessaires à la mise en œuvre de ces outils, en considérant prioritairement la génération automatisée de textes. Vos travaux pourront utilement ouvrir la voie à réflexion plus large sur les modèles auto-supervisés géants d'intelligence artificielle qui permettront, demain, la génération à grande échelle de solution dans des domaines comme par exemple la santé ou le code informatique, et qui sans aucun doute, seront porteurs d'enjeux majeurs pour nos sociétés.

Il serait particulièrement utile que le Comité me transmette un avis d'ici au 30 juin 2023.

Je vous prie de croire, Monsieur le Directeur, à l'assurance de ma considération distinguée.

Jean-Noël BARROT

2/2

The French National Pilot Committee for Digital Ethics (CNPEN) was created in December 2019 on the initiative of the Prime Minister and placed under the auspices of the French National Advisory Ethics Council for Health and Life Sciences (CCNE). It is made up of leading figures from the academic, industrial and institutional worlds. As experts in digital technology, law, economics, philosophy, language, logic and medicine, they all contribute to an ethical reflection made indispensable by the development of digital technology and help to enlighten public debate. Previous CNPEN opinions, for example, concern Ethical issues regarding "autonomous vehicles" (May 2021), Ethical issues of conversational agents (September 2021) or, jointly with the CCNE, Ethical issues of Medical diagnosis and artificial intelligence (November 2022) and Health data platforms (February 2023). More recently, it has addressed Ethical issues of retroactive name change in digital scientific documents (June 2023).

MEMBERS OF THE NATIONAL PILOT COMMITTEE FOR DIGITAL ETHICS

Gilles Adda
Raja Chatila
Theodore Christakis
Laure Coulombel
Jean-François Delfraissy
Laurence Devillers
Karine Dognin-Sauze
Gilles Dowek
Valeria Faure-Muntian
Christine Froidevaux
Jean-Gabriel Ganascia
Eric Germain
Alexei Grinbaum
David Gruson
Emmanuel Hirsch
Jeany Jean-Baptiste
Claude Kirchner - directeur
Augustin Landier
Gwendal Le Grand
Claire Levallois-Barth
Caroline Martin
Tristan Nitot
Jérôme Perrin
Catherine Tessier
Serena Villata
Célia Zolynski

Layout : DSAF/DPSG/BAEC/Section Édition – July 2024.