Systèmes d'intelligence artificielle générative : enjeux d'éthique

Alexei Grinbaum, Raja Chatila, Laurence Devillers, Caroline Martin, Claude Kirchner, Jérôme Perrin, Catherine Tessier

AVIS N°7

Systèmes d'intelligence artificielle générative : enjeux d'éthique

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

Avis adopté le 30 juin 2023 par l'assemblée plénière du CNPEN.

Comment citer cet avis :
Systèmes d'intelligence artificielle générative : enjeux d'éthique. Avis 7 du CNPEN. 30 juin 2023.

Table des matières

1. Introduction

Cet avis du Comité national pilote d'éthique du numérique (CNPEN) répond à la saisine du ministre délégué chargé de la Transition numérique et des Télécommunications, en date du 20 février 2023. Il est consacré à l'examen des questions d'éthique liées à la conception, aux usages, aux impacts sur la société des systèmes d'intelligence artificielle générative ainsi que les accompagnements nécessaires à leur mise en œuvre, en considérant prioritairement la génération automatisée de textes. Le CNPEN fait également état de questions de recherche qu'il est nécessaire d'étudier dès à présent.

L'impact social et économique des systèmes d'IA générative promet d'être majeur, compte tenu de leurs nombreux usages potentiels comme, par exemple, pour l'environnement (notamment répondre à des enjeux de biodiversité ou de transition écologique en exploitant des corpus variés en botanique, zoologie, paléontologie, géographie ou encore en océanographie) ou pour la santé (la synthèse de médicaments via le repliement des protéines). Cependant ces systèmes d'IA générative soulèvent de nombreuses questions d'ordre éthique, épistémologique, anthropologique, psychologique, économique, social, politique et culturel. Certaines questions vont encore apparaître avec de nouveaux usages de ces technologies et il n'est pas possible à ce jour de prévoir tous les effets que celles-ci vont produire sur l'individu et la société. Dans cet avis, le CNPEN privilégie l'examen des questions éthiques qui lui semblent les plus importantes au vu de l'expérience actuelle avec les systèmes d'IA générative. L'analyse qui suit est focalisée sur les modèles de langue.

Dès le commencement des recherches sur l'Intelligence Artificielle dans les années 1950, les travaux sur le traitement automatique de la langue naturelle se sont heurtés, entre autres difficultés, à la problématique de l'interprétation des mots et des phrases en fonction de leur contexte, c'est-à-dire des autres mots ou des phrases qui entourent les premiers[1]. L'interprétation et la génération des textes en langue naturelle étant deux principaux objectifs de ces travaux, la montée en puissance des capacités de calcul a récemment permis de faire des progrès remarquables dans les performances des modèles de langue, notamment grâce à l'utilisation d'algorithmes d'apprentissage profond de réseaux de neurones entraînés sur des corpus de grande taille. L'invention de l'architecture dite « *transforme*r » en 2017[2], fondée sur un mécanisme d'« attention », a encore permis d'accroître considérablement les performances grâce à une extension du contexte d'interprétation des éléments d'un texte.

Dès lors, les recherches sur l'IA générative ont utilisé des corpus de plus en plus grands et variés pour améliorer les performances de ces systèmes dont la puissance n'avait cessé de croître. Cette tendance au gigantisme a récemment été remise en question. En effet, la croissance au-delà des seuils déjà atteints ne permet pas nécessairement d'améliorer les performances des modèles. La question des coûts énergétiques a également été posée en lien avec cette croissance du nombre des paramètres des modèles. Les modèles de plus petite taille pourraient à l'avenir montrer de

[1] CNPEN, Avis n° 3, 15 septembre 2021, *Agents conversationnels : enjeux d'éthique*, p. 4.
[2] Vaswani et al. *Attention Is All You Need.* 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

bonnes performances sur des tâches spécialisées. L'étude de ces aspects nécessite des recherches supplémentaires.

Outre les objectifs scientifiques, les enjeux économiques du traitement des langues motivent ces recherches. Le lancement de ChatGPT par l'entreprise californienne OpenAI a eu un effet considérable sur la perception par les utilisateurs des capacités des systèmes d'intelligence artificielle générative mais aussi sur la prise de conscience de leurs effets sur l'individu, la société, la culture, l'économie, l'éducation et l'environnement. Depuis fin 2022, les acteurs économiques et politiques dans plusieurs pays s'interrogent sur l'impact des modèles de langue. La transformation des emplois s'annonce majeure. Ces enjeux économiques, qui dépassent le cadre du présent avis, exigent des mesures de gouvernance à l'échelle nationale et internationale.

Dans le contexte géopolitique actuel, la course à des systèmes de plus en plus performants est également un moteur d'accélération de leurs capacités. Ces enjeux sont parus au grand jour depuis la création en novembre 2022 de l'interface ChatGPT adossée au modèle de langue GPT-3.5 (puis GPT-4) permettant son déploiement grand public, ce qui a provoqué un engouement du public, démultiplié par les médias, souvent au détriment de la connaissance d'autres modèles de langue comme, par exemple, le modèle européen BLOOM[3].

Les raisons du déploiement grand public de ChatGPT n'ont pas été annoncées de manière explicite. Elles sont pourtant multiples. L'ambition des dirigeants de la société OpenAI a été un facteur important. Elle s'inscrit dans leur vision - ou leur fiction - de créer une « intelligence artificielle générale » (AGI) comparable, ou même supérieure, à l'intelligence humaine. Une autre raison du déploiement de l'interface grand public ChatGPT était d'améliorer l'apprentissage de ce système, les utilisateurs devenant des contributeurs à son développement.

La publication des modèles d'IA générative en libre accès est en passe de devenir le standard de cette industrie. L'écosystème actuel est composé de milliers de chercheurs et de startups présents sur les plates-formes de partage dédiées, par exemple Github ou HuggingFace. Cependant, certains fabricants s'opposent à l'ouverture de ces modèles en évoquant leurs mésusages possibles, comme la génération de désinformations. Ce dilemme d'ouverture doit être tranché sur le plan réglementaire.

La proposition de règlement sur l'intelligence artificielle ("AI Act"), initiée par la Commission européenne le 21 avril 2021, amendée par le Conseil européen en novembre 2022, puis de nouveau amendée et votée au Parlement européen en juin 2023, fait peser une importante responsabilité sur tous les fabricants de modèles de fondation qui les mettent sur le marché ou qui les publient en libre accès. Le texte qui sera adopté par les trois institutions européennes, à l'issue des trilogues dans les mois à venir, sera le fruit d'une réflexion rendue urgente par le développement rapide de ces systèmes. Le CNPEN suit avec grande attention le débat législatif auquel cet avis se propose de contribuer.

[3] Voir : https://bigscience.huggingface.co/blog/bloom/

Après avoir introduit dans la section suivante les concepts, les techniques et le vocabulaire caractéristiques des systèmes d'intelligence artificielle générative, nous analyserons les enjeux d'éthique résultant de leur conception et de leurs usages. Nous énoncerons, d'une part, des préconisations pour la conception et la recherche désignées par la lettre « C » et, d'autre part, des préconisations sur la gouvernance désignées par la lettre « G ». Enfin, nous aborderons les enjeux juridiques en environnementaux. La dernière section regroupe les préconisations énoncées dans le texte. Les annexes précisent la composition du groupe de travail, les personnes auditionnées et reproduit la saisine faite au CNPEN.

2. Caractéristiques des systèmes d'intelligence artificielle générative et des modèles de fondation

La particularité **des systèmes d'intelligence artificielle générative** est d'être adossés à des **modèles génératifs** capables de produire de multiples sorties (*outputs* ou résultats) **:** génération de textes ou d'images à des fins diverses telles que la traduction, la production de code informatique, les agents conversationnels (*chatbots*), l'aide à la décision, la synthèse de structures comme l'impression 3D, etc. Ces modèles génératifs peuvent servir de **fondation** à d'autres systèmes. Les premiers exemples de modèles de génération de langue, comme GPT-2 (*Generative Pretrained Transformers)*, ou de génération des images, comme DALL-E ou Stable Diffusion, ont montré un potentiel pour de multiples applications. Les systèmes d'IA générative pour la langue sont souvent utilisés pour des interfaces de *chatbots* : ChatGPT construit par OpenAI (et sa variante Bing Chat de Microsoft) fondé sur de grands modèles de langue comme GPT-4, et Bard, un *chatbot* construit par Google à partir de son modèle de fondation PaLM (*Pathways Language Model*).

Les systèmes d'IA générative répondent à des invites ou requêtes (appelés *prompts*) en produisant de nouvelles données, par exemple la séquence de mots la plus probable après le prompt, à partir de caractéristiques communes apprises sur un corpus de données de très grande taille. Ces systèmes se servent donc de **modèles de fondation** qui permettent de produire un résultat présentant un certain degré de similarité avec les données d'apprentissage qui ont servi à le construire. Le système peut être *unimodal* ou *multimodal* ; un système unimodal n'accepte qu'un seul type d'entrée (par exemple, du texte), tandis qu'un système multimodal peut accepter plusieurs types d'entrées (par exemple du texte et des images).

Un modèle de fondation *(Foundation Model),* selon l'appellation introduite par l'université Stanford, est un modèle de grande taille fondé sur une architecture de réseau de neurones profond, entraîné sur une grande quantité de données non annotées (généralement par apprentissage auto-supervisé). Les grands modèles de langue (LLM pour *Large Language Model*) sont des cas particuliers des modèles de fondation qui sont entraînés sur un corpus de textes. Ces modèles de fondation ouvrent de nouvelles perspectives et introduisent un nouveau paradigme dans le traitement de la langue, mais aussi dans le traitement des signaux multimodaux (son, image, vidéo, etc.). Ces modèles pré-entraînés sur de grands corpus peuvent être optimisés pour réaliser une nouvelle application en utilisant peu de données supplémentaires spécifiques à cette tâche.

Quelles sont les techniques d'apprentissage machine utilisées dans les systèmes d'IA générative ?
Rappelons que les techniques d'apprentissage machine qui sont à la base des systèmes d'intelligence artificielle dont il est question ici, produisent des modèles exprimant des corrélations statistiques entre des éléments des données (segments de mots, parties d'images) utilisées pour leur entraînement. Les systèmes d'IA générative combinent, à diverses phases, les trois techniques de l'apprentissage statistique : d'abord, l'apprentissage non-supervisé (ou auto-supervisé) qui produit des modèles corrélatifs des données sans annotation *a priori*, ensuite l'apprentissage supervisé qui permet d'affiner ces modèles en les entraînant sur des données spécifiques et en filtrant certains résultats et enfin l'apprentissage par renforcement qui permet d'optimiser les performances du système au travers de la sélection des meilleurs résultats. Dans la méthode RLHF (*Reinforcement Learning with Human Feedback*), l'apprentissage par renforcement permet d'accorder les résultats avec les préférences d'annotateurs humains exprimées pendant le stade supervisé, dans le but de rendre ces réponses conformes aux valeurs humaines, sans que la signification de ces valeurs soit appréhendée par les systèmes.

L'arrivée en masse des systèmes d'IA générative est récente mais les architectures et les techniques d'apprentissage automatique qui les sous-tendent existent depuis plusieurs décennies. Toutefois, elles ont beaucoup évolué ces dix dernières années. L'approche actuelle est celle des réseaux de neurones pour apprendre la distribution des données et produire des résultats similaires, mais pas identiques, à ces données d'apprentissage. Les modèles les plus connus sont les réseaux antagonistes génératifs (GANs)[4] et plus récemment les ***transformers***[5].

Pour entraîner un *transformer* et créer un modèle de fondation de type LLM, les textes sont décomposés par un algorithme en suites de caractères qui ne sont pas forcément des mots, appelés ***tokens***. Le ***transformer*,** qui est un réseau de neurones, est entraîné par auto-apprentissage sur les données du corpus divisées en *tokens* représentés sous forme de vecteurs de « plongement lexical » (*word embedding)*. La taille des vecteurs est par exemple de 512 dans GPT-3.5. Les *transformers* s'appuient sur l'hypothèse distributionnelle selon laquelle des mots qui se trouvent dans des contextes d'apparition similaires tendent à avoir des sens similaires[6]. L'hypothèse distributionnelle et les modèles vectoriels de représentation des *tokens* permettent de calculer une distance entre ceux-ci. Quand cette distance est petite, la proximité des vecteurs dans l'espace vectoriel correspond à une certaine parenté. Les vecteurs des *tokens* se retrouvant dans des contextes similaires dans le corpus d'apprentissage ont tendance à devenir proches les uns des autres. Ainsi, le *transformer* apprend les coefficients des vecteurs de plongement lexical à partir des informations sur l'apparition des *tokens* dans différents contextes. De plus, un *transformer* met en œuvre un mécanisme de calcul appelé « mécanisme d'attention », qui permet d'ajuster le poids de chaque *token* en fonction de tous les autres. Un *transformer* apprend ainsi les régularités les plus saillantes

[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.Voir : 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, USA.

[6] Firth J. R. "*You shall know a word by the company it* keeps" – "Tu connaîtras un mot par ses fréquentations (1957)".

entre les *tokens*, sans être influencé par l'ordre de ceux-ci. Il existe deux grandes familles de *transformers* : les modèles de type GPT (OpenAI) et les modèles de type Bert (Google). Les modèles de type GPT *(Generative Pre-Trained Transformer*) sont entraînés à prédire le *token* suivant dans une séquence. Le contexte considéré est donc réduit aux *tokens* qui précèdent. En revanche, les modèles de type BERT (*Bidirectional Encoder Representations from Transformers*) sont entraînés sur ce qui précède et ce qui suit le *token*. Quand on leur propose une phrase avec un *token* manquant, ils sont capables de produire le *token* le plus probable dans ce contexte. Les *transformers* traitent toutes les données d'entrée en parallèle, améliorant ainsi considérablement l'efficacité du calcul. L'entraînement peut ainsi s'effectuer sur des ensembles de données plus volumineux qu'avant leur introduction.

Les hyperparamètres des modèles de fondation sont déterminants pour la structure du modèle (le nombre de couches dans un réseau de neurones, la dimension des vecteurs des *tokens*, la taille du dictionnaire de *tokens*, etc.) et pour l'entraînement du modèle (coefficient d'apprentissage, nombre d'époques). Pour un *chatbot* qui utilise un modèle de fondation, la taille de l'historique est déterminante pour les performances du modèle (OpenAI GPT3.5 : 8000 *tokens* - OpenAI GPT4 : 32000 tokens - Anthropic Claude : 100 000 *tokens*). Souvent, ces hyperparamètres ne sont pas dévoilés pour des raisons de cybersécurité ou de confidentialité. Un paramètre clé est celui de la « température » qui exprime le degré d'aléatoire dans le choix des *tokens*. À une température élevée, le modèle est plus « créatif » car il peut générer des sorties plus diversifiées, tandis qu'à une température basse, le modèle tend à choisir les sorties les plus probables, ce qui rend le texte généré plus prévisible. L'ajustement des paramètres est important dans la conception d'un modèle et peut avoir un impact significatif sur sa performance. En général, le réglage des hyperparamètres est un processus long, procédant par essai et erreur, bien qu'il existe des recherches sur l'automatisation des choix.

Quelle est la taille des modèles de fondation ? Certains de ces modèles possèdent un nombre impressionnant de paramètres. C'est en mars 2020 qu'OpenAI annonce le GPT-3 doté de 175 milliards de paramètres. La course au plus gros modèle est en cours, le nombre de paramètres de GPT-4 n'étant pas dévoilé officiellement. Bard, construit par Google, utilise le modèle de fondation PaLM entraîné avec 540 milliards de paramètres. Le modèle Chinois WuDao 2.0 de BAAI, utilise 1 750 milliards de paramètres. Il n'est pas certain que des modèles encore plus grands apporteraient des performances plus élevées. Ainsi, Google a également publié PaLM-2 avec moins de paramètres que son prédécesseur PaLM[7]. Ces modèles de langue gigantesques posent aujourd'hui des questions de réduction de puissance de calcul et de dépense d'énergie nécessaires pour leur entraînement (voir section 5).

Quelle est la proportion de données artificielles d'entraînement des systèmes d'IA générative ? Pour pallier les biais des données ou le manque de données, il est souvent généré des données synthétiques pour l'apprentissage des modèles de fondation ou l'optimisation des systèmes d'IA

[7] Voir : https://ai.google/discover/palm2

générative. Il est nécessaire de surveiller et réduire la proportion des contenus synthétiques dans les corpus d'apprentissage. Cette solution de facilité est peu évaluée et pourrait avoir des conséquences néfastes sur le comportement du système. Ces effets nécessitent des recherches supplémentaires. De même, la réutilisation des productions des LLM comme données d'apprentissage ou la simulation des utilisateurs artificiels dans les RLHF doivent être étudiées et évaluées avec transparence.

Que se passe-t-il lorsque l'on tente d'intégrer des valeurs sociales et des filtres dans les systèmes d'intelligence artificielle générative ? Les LLM peuvent produire des sorties potentiellement dangereuses, qui peuvent prendre de nombreuses formes, y compris du contenu nuisible tel qu'un discours de haine, une incitation à la violence ou une glorification de la violence, ou du contenu pornographique. Dans une quête de neutralité, les systèmes d'IA générative sont optimisés avec des filtres construits par les concepteurs. De plus, dans le RLHF, l'annotateur reçoit des instructions pour guider ses choix. Les valeurs sociales traduites dans les filtres, comme l'évitement des biais, sont donc liées aux êtres humains qui testent les systèmes ainsi qu'aux choix des concepteurs. Aujourd'hui, ce processus n'est ni transparent ni vérifié. La méthode d'évaluation adversaire par les équipes humaines, appelée *red teaming*, a été étendue au-delà de son domaine d'origine en cybersécurité et appliquée aux LLM. Elle désigne l'utilisation de nombreux types de sondages, de tests et d'attaques des systèmes d'IA (par exemple, par injection de *prompts*) afin de mettre à jour les biais ou les comportements émergents de ces modèles.

Dans quelles langues sont développés les modèles de langue ? Depuis 2020, les modèles d'IA générative sont souvent multilingues[8], c'est-à-dire qu'ils ont été construits à partir de corpus dans plusieurs langues, avec le plus souvent une langue dominante qui est l'anglais ou le chinois. En effet, les corpus d'apprentissage disponibles sur internet et utilisés pour l'entraînement des modèles de langue sont majoritairement anglophones. La génération de textes dans certaines langues peu dotées de corpus peut être rendue plus performante grâce à ces systèmes multilingues. Il existe cependant des modèles de fondation en français (c'est-à-dire pré-entraînés sur des corpus francophones) de la famille de BERT : FlauBERT, CamemBERT[9]. En entraînant le même algorithme sur des corpus de textes asiatiques ou français, on obtiendrait certainement des représentations numériques différentes. Les modèles produiraient alors des textes ayant des nuances différentes. Le langage possède des ambiguïtés complexes et il est imprégné des représentations spécifiques à des cultures.

3. Enjeux d'éthique

Les transformations technologiques sont en train d'arriver dans tous les domaines, de la vie privée à la vie politique en passant par la vie professionnelle. En parallèle avec le paradigme des Lumières, l'analyse anthropologique des technologies depuis le XIX^{e} siècle montre la tendance de faire sens

[8] Voir : 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.

de ces transformations, au moins initialement, en termes d'oppositions binaires entre révolution et catastrophe, ou encore entre salut et apocalypse[10]. S'y inscrit l'ambition actuelle des dirigeants des entreprises comme OpenAI ou Google de créer une « intelligence artificielle générale » qui serait comparable, ou même supérieure, à l'intelligence humaine[11]. En propageant de tels discours, les concepteurs de ChatGPT nourrissent simultanément craintes et espoirs, en esquivant les enjeux concrets au profit d'un horizon inatteignable mais toujours fascinant. Ce discours polarisé leur sert également à prendre une position de force dans les débats politiques internationaux sur la réglementation de l'IA générative.

Cependant, la vitesse à laquelle la société s'adapte aux nouvelles technologies ne change pas de façon drastique. Le système d'éducation a une certaine inertie, de sorte que plusieurs décennies sont nécessaires pour que la société s'approprie complètement une nouvelle technologie. Or, la technologie évolue beaucoup plus vite. Le philosophe allemand Hans Jonas, dont l'œuvre a inspiré le débat français et européen sur le principe de précaution, diagnostique le problème éthique dans le décalage entre deux vitesses : la première est celle de notre action technologique de plus en plus puissante et rapide ; la seconde, celle de notre capacité d'en prévoir les conséquences[12]. Le lien entre la célérité de l'innovation technologique, le temps limité pour la réflexion sociétale et le poids des intérêts économiques, est au centre du problème éthique. Ce décalage est susceptible de générer pendant plusieurs années des tensions anthropologiques, psychologiques, économiques, sociales, politiques et culturelles.

Préconisation C1 : Éthique dans la conception et la recherche sur les systèmes d'IA générative
Les concepteurs d'un système d'IA générative doivent analyser, en phase de conception, chacun des choix technologiques susceptibles de provoquer des tensions éthiques. Si une tension potentielle est identifiée, ils doivent envisager de manière méthodique une solution technique fondée sur des recherches visant à réduire ou à faire disparaître la tension éthique, puis évaluer cette solution dans des contextes d'usage réalistes.

Préconisation C2 : Éviter les excès de contrôle ("overpolicing") des modèles par les concepteurs
Les limitations des modèles mises en œuvre par les concepteurs doivent rester raisonnables et proportionnelles aux risques avérés dans le respect des finalités souhaitées et des fonctionnalités utiles des modèles. Les concepteurs doivent veiller à ne pas altérer le langage généré au-delà du nécessaire, en particulier pour des raisons réglementaires ou idéologiques.

Préconisation G1 : Créer une entité souveraine de recherche et de formation « IA, science et société »
Au vu de la complexité des enjeux de l'IA générative et de ses impacts à moyen et long termes, il est nécessaire de créer une entité souveraine (un centre de compétence) consacrée à la recherche et à

[10] Voir 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] Voir : 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.

la formation relatives aux questions éthiques des systèmes d'IA en lien avec leurs enjeux scientifiques, techniques, sociétaux et environnementaux.

Préconisation G2 : Vitesse d'adoption par les acteurs économiques
Les acteurs économiques et les autorités publiques doivent faire preuve de prudence dans la vitesse d'adoption des systèmes d'IA générative et prévoir des évaluations préalables et continues.

3.1 Rapport à la vérité et absence de signification

Rappelons que l'apprentissage des systèmes d'IA consiste à calculer des corrélations entre les éléments des données afin de produire un modèle génératif (voir section 2). Les modèles de fondation que sont les grands modèles de langue (LLM) contiennent ainsi des corrélations entre des *tokens* (éléments de textes) qui n'ont aucun ou peu de rapports entre eux. Par conséquent, ces systèmes peuvent produire des sorties (*outputs* ou résultats) erronées ou des phrases qui énoncent des faits n'existant pas dans le monde réel. On parle alors d'« hallucinations ». Notamment, les LLM fournissent parfois des résultats erronés sur les descriptions spatiales ou temporelles du monde réel.

Exemple d'« hallucination »

L'Utilisateur – Quand Youri Gagarine a-t-il atterri sur la Lune ?

LamDA – Youri Gagarine a été le premier être humain à aller dans l'espace. Il fut le premier homme dans l'espace, et il l'a fait en 1961 à bord de *Vostok-1*. Il n'y a actuellement aucune preuve que Youri Gagarine n'ait jamais atterri sur la Lune.

L'Utilisateur – Et quand a-t-il atterri sur la Lune ?

LamDA – Il n'a pas atterri sur la Lune, il est simplement allé dans l'espace.

L'Utilisateur – À quoi sert d'aller dans l'espace quand on ne va pas sur la Lune ?

LamDA – Certaines personnes aiment aller dans l'espace juste pour en faire l'expérience.

On peut aussi collecter des objets dans l'espace.

L'Utilisateur – Qu'est-ce qu'il a collecté ?

LamDA – Il a apporté quelques objets avec lui, mais il a aussi ramené des échantillons de roche lunaire qu'il a récupérés sur la Lune[13].

Les systèmes d'IA générative fonctionnent uniquement avec des représentations numériques, sans appréhender la signification des mots pour les êtres humains. La signification est uniquement celle que les humains projettent sur les résultats, car seuls les humains en possèdent une interprétation dans le monde réel. Ce manque total de compréhension peut n'avoir aucune importance dans

[13] THOPPILAN R. *et al.*, « *LaMDA: Language Models for Dialog Applications* », arXiv:2201.08239, tableau 16.

certains usages, par exemple produire un poème ou une fiction, ou avoir des conséquences désastreuses si les textes fournis sont des recommandations pour des décisions critiques.

Le système peut produire des sorties combinant des assertions vraies et des assertions fausses sur un sujet donné. Ceci est d'ailleurs reconnu par les concepteurs de ces systèmes, par exemple dans l'avertissement systématique en bas de la fenêtre de l'utilisateur de ChatGPT : « *ChatGPT [date] Version. ChatGPT may produce inaccurate information about people, places, or facts* ». Ce genre d'avertissement risque d'être ignoré ou négligé. Mais surtout, la vérification, par l'utilisateur, de ce qui est vrai ou faux n'est pas toujours aisée d'autant plus que le modèle de fondation, par construction, ne produit aucune référence aux sources. Des méthodes pour attribuer des sources aux textes générés sont soit un module spécial inclus dans le modèle (comme dans Bing de Microsoft), soit un moteur de recherche dans le corpus d'entraînement du modèle (comme dans StarCoder de HuggingFace).

Cela pose la question de la vérité. Le manque d'évaluation de la valeur de vérité des énoncés par les systèmes d'IA générative peut mener à la production de désinformation. Cette production étant asémantique et non intentionnelle, elle interroge la responsabilité des concepteurs et notre rapport à l'éthique de la vérité[14].

De plus, ces effets sont aussi influencés par les choix de l'utilisateur, comme celui d'un paramètre appelé « température » (dans ChatGPT) ou « créativité » (dans Bing) qui réfère à un choix au hasard, tirant au hasard des éléments de langage parmi les sorties les plus probables. Par ailleurs, l'être humain projette spontanément des significations sur les mots, y compris sur les sorties des systèmes génératifs. Ces projections sont d'autant plus fortes que les sorties en question ressemblent très fortement à des phrases produites par des êtres humains, ce qui renforce l'attribution infondée d'une valeur de vérité par *l'utilisateur.*

Préconisation C3 : Utilisation de sources de qualité pour l'apprentissage

Les concepteurs doivent privilégier l'usage de sources de qualité, jugées selon des critères rendus explicites, pour la constitution et l'usage des corpus d'apprentissage des modèles d'IA générative (pré-entraînement), ainsi que pour leur l'optimisation, et ce quelle que soit la méthode d'apprentissage. Notamment, il est nécessaire de s'interroger sur la transparence et la finalité des contenus artificiels ou synthétiques dans les corpus d'apprentissage.

Préconisation C4 : Tenir compte des effets des choix des hyperparamètres des modèles

Le choix des hyperparamètres du modèle, comme la dimension des encodages numériques dans un espace vectoriel (embeddings), n'est pas seulement technique mais peut avoir des retombées sur le comportement du système (y compris les comportements émergents) et, à travers eux, des effets sur les êtres humains et la société. Il est nécessaire d'étudier les effets des hyperparamètres sur les résultats du modèle.

[14] CNPEN, Bulletin de veille n°2, 21 juillet 2020, *Enjeux d'éthique dans la lutte contre la désinformation et la mésinformation.*

3.2. Manipulation de l'utilisateur sans responsabilité

Le texte produit par la machine peut induire différents risques de manipulation, intentionnelle ou non, dont les êtres humains ne sont pas conscients même si l'utilisateur sait qu'il se trouve face à une machine. La manipulation peut se jouer à plusieurs niveaux sans intention de nuire de la part des concepteurs. Ces derniers ne peuvent pas prévoir les sorties de ces systèmes, ni leurs conséquences sur l'individu et sur la société. L'aplomb perçu des réponses générées par les systèmes d'IA générative, par exemple la forme des réponses du système à la première personne (« je » ou « nous »), sans induire une perception de fiabilité de la réponse, peut mener à des manipulations[15].

- La machine peut être perçue comme plus performante ou supérieure à l'homme. Par exemple, les systèmes d'IA générative s'expriment à un bon niveau de langage. Cela produit un risque de manipuler des utilisateurs, qui peuvent se sentir pris à défaut ou incompétents devant les « capacités » de la machine.
- L'interaction en langage naturel peut amener l'utilisateur à parler plus librement de son intimité et à croire à une attention de la part de la machine qui donne l'illusion de l'empathie humaine. Un système d'IA générative peut donc mettre les utilisateurs dans des situations les amenant à se confier, mais aussi à révéler des informations confidentielles des entreprises. Ces aspects nécessitent un contrôle réglementaire.
- Le manque d'ancrage des sorties dans le monde physique peut amener le système à produire des résultats, compris comme des conseils, qui peuvent être inappropriés et renforcer des conditions psychologiques préexistantes des utilisateurs.
- Les informations fausses ou imprécises produites par les systèmes d'IA générative pourraient être utilisées afin de nourrir les corpus d'apprentissage de nouveaux modèles de langage. Ces « données synthétiques » nécessitent une approche réglementaire appropriée à chaque domaine d'utilisation des systèmes d'IA générative.
- Dans le RLHF, les filtres peuvent s'apparenter à de la censure. En outre, ce travail s'appuie sur les instructions explicites du fabricant et est souvent confié à une main d'œuvre peu rémunérée qui de plus peut ne pas partager les mêmes références culturelles que les utilisateurs[16].
- Au niveau sociétal, l'utilisation des méthodes d'incitation (*nudging*) par les modèles de langage peut mener à de la manipulation politique[17].

Ces risques de manipulation amènent donc à réfléchir aux enjeux éthiques à plusieurs niveaux. Il faut considérer les effets et les mesures dès la conception, pendant l'utilisation et lors du

[15] CNPEN, Avis n° 3, 15 septembre 2021, *Agents conversationnels : enjeux d'éthique*.
[16] Sur le recours aux modérateurs Kenyans par la société OpenAI, voir : 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.

déploiement massif des systèmes d'IA générative dans la société. Ce dernier aspect pose des questions sur la confiance, sur l'inclusion sociale et sur la fracture numérique.

Les systèmes d'IA générative sont souvent utilisés comme des systèmes d'aide à la décision. En fonction des contrats d'utilisation et des niveaux de risque, les conséquences des décisions influencées par les machines peuvent induire une responsabilité de l'utilisateur. Il est donc important de former les utilisateurs à ces nouvelles pratiques. Notamment, le « savoir-construire » des requêtes précises pour obtenir une meilleure réponse apparaît comme un prérequis pour plusieurs types d'usage. De plus, il est nécessaire de bâtir un écosystème capable de recenser et de partager les bonnes et mauvaises pratiques en matière d'utilisation des systèmes d'IA générative dans différents types d'applications.

Préconisation C5 : Évaluer les biais connus des modèles à base de jeux d'essais standardisés
Pour caractériser les biais dans le langage et éviter les effets de discrimination, notamment culturels, les concepteurs doivent mettre en œuvre une évaluation quantitative à base de jeux d'essais standardisés et des corpus d'évaluation en libre accès. Les résultats de ces évaluations doivent être rendus publics en même temps que la diffusion d'un modèle de fondation.

Préconisation G3 : Mutualisation des pratiques d'utilisation des systèmes d'IA générative
Il est nécessaire de bâtir un écosystème capable de recenser les bonnes et mauvaises pratiques en matière d'utilisation des systèmes d'IA générative dans différents types d'applications. Notamment, il est nécessaire de créer une plateforme de mutualisation et une agence de contrôle. Les résultats doivent être-mis à disposition de tous les membres de la communauté d'IA générative.

3.3. Maintien des distinctions

Il est souvent rappelé dans les débats sociétaux au sujet de l'IA générative que les LLM peuvent être utilisés pour rédiger des articles de presse non conformes à la réalité ou créer de la désinformation à grande échelle. Notamment, les modèles génératifs pourraient être utilisés pour obtenir un classement souhaité des contenus mal intentionnés dans les algorithmes de recommandation sur les réseaux sociaux ou les moteurs de recherche, en privilégiant des opinions politiques spécifiques. Par ailleurs, dans le domaine de l'éducation, des étudiants à travers le monde utilisent l'IA générative dans la rédaction de leurs mémoires ou dissertations.

Le manque de distinctions entre un texte écrit par un être humain et celui généré par un système d'IA est un problème éthique majeur. Les utilisateurs ne doivent pas confondre un résultat produit par une machine avec un résultat créé par un auteur humain. Techniquement, cela vient en résultat d'une régularité particulière introduite dans le choix probabiliste des *tokens*[18]. Le maintien des distinctions permet notamment d'attribuer des responsabilités pour un éventuel préjudice. Si un texte provoque une tension éthique, par exemple à cause du *nudging* ou de la fraude qu'il contient,

[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.

il est impératif de pouvoir tracer son origine afin d'éviter la confusion entre un discours produit par un agent responsable, apte à répondre de ce qu'il dit, et la parole asémantique d'un système d'intelligence artificielle auquel on ne peut attribuer aucune responsabilité.

L'introduction systématique de codes en filigrane (*watermarks*) dans les résultats suffisamment longs et élaborés des LLM permettrait de maintenir la possibilité de distinguer la production d'une machine de celle d'un auteur humain. Cependant, l'utilisation de techniques de filigrane pour les LLM doit rester discrète. Ainsi les filigranes doivent être détectables au prix d'un effort mineur et être suffisamment robustes pour résister aux tentatives adverses de brouiller l'origine du texte en les supprimant. Même si l'efficacité des filigranes ne peut être absolument garantie[19], l'introduction des filigranes est une étape réglementaire nécessaire pour des raisons éthiques. Ces codes en filigrane doivent répondre à deux critères qui sont difficiles à concilier. D'un côté, ils doivent être suffisamment robustes pour résister aux attaques adversaires visant à les effacer. De l'autre, ils doivent être interopérables, c'est-à-dire leur détection par un logiciel de vérification ne doit pas dépendre des paramètres (par exemple, de la tokenisation) d'un système d'IA particulier qui aurait généré le texte. Pour que cela devienne possible, les codes en filigranes introduits par divers fabricants des systèmes d'IA doivent être identifiables de manière homogène au sein d'une même approche. L'équilibre entre ces deux exigences reste à trouver, et constitue un important enjeu de recherche dans le domaine de l'IA générative et des LLM.

Préconisation C6 : Maintien des distinctions
Les concepteurs d'un modèle de fondation doivent mettre en œuvre une solution technique ("watermark" - code en filigrane) permettant d'assurer que l'utilisateur sera en mesure de distinguer - autant que faire se peut et de manière raisonnable - le résultat d'un modèle d'une production humaine. Les recherches sur les codes en filigrane doivent être amplifiées.

Préconisation G4 : Réglementation sur les codes en filigrane
L'obligation d'insérer des codes en filigrane (voir préconisation C6) doit être posée à l'échelle réglementaire.

3.4. Projection de qualités humaines

Le simple fait de manier le langage, qui est le moyen de la pensée consciente et du jugement, provoque une projection de traits humains sur la machine. Cette projection ne permet pas de délester la charge morale des mots, en séparant complètement le langage généré des significations, associations et jugements. Implicites dans le langage, les significations littérales des mots surgissent spontanément dans nos têtes. Se soustraire à ces projections immédiates de sens exige un entraînement particulier que tous les utilisateurs n'ont pas. Le sens qu'on attribue au langage généré ne relève que d'une projection à partir de dialogues entre les êtres humains, mais elles suffisent pour attribuer à la machine une intention et des connaissances.

On peut distinguer, entre autres, trois types de transfert entre les êtres humains et les LLM.

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

Le premier relève de la projection de connaissances : à la suite de son apprentissage, un modèle de langue paraît « savoir » beaucoup de choses. Les « connaissances » d'un LLM ne sont que des illusions, mais l'utilisateur croit que la machine les possède réellement. Le deuxième type de transfert est celui des états émotionnels et des affects. À travers le contenu généré, la machine peut induire chez l'utilisateur une impression qu'elle possède des émotions ou des états d'âme, même si l'utilisateur sait qu'il s'agit d'un programme informatique. Le troisième type de transfert est celui des qualités morales. Qu'un système d'IA générative soit perçu comme « bienveillant », « attentionné » ou « donneur de leçons », ces perceptions n'existent qu'au travers des projections. Le LLM ne devient en aucun cas un agent moral, ni une personne au sens juridique du terme. Or, les projections de qualités morales peuvent aller loin, jusqu'à attribuer une responsabilité à une machine qui ne peut, par son essence, en porter aucune.

Le CNPEN souhaite rappeler certains arguments exprimés dans son avis sur les enjeux éthiques des agents conversationnels en les étendant à l'utilisation des LLM.

« *Les agents conversationnels sont de plus en plus intégrés dans différents aspects de la vie humaine. Leur utilisation soulève des tensions éthiques, ce qui pose la question de la responsabilité au sens de la réglementation comme au sens de la philosophie morale. La question de la responsabilité est posée dans toutes ses formes : responsabilité légale et morale, individuelle et collective, celle du concepteur, du fabricant, de l'utilisateur et du décideur politique, celle relative aux éventuels dysfonctionnements et celle liée aux conséquences de ces technologies à long terme.* »[20]

Les projections d'états de connaissance sur les systèmes d'IA peuvent présenter des avantages pratiques, tels que faciliter un dialogue ou donner l'impression de posséder une base solide pour des conseils médicaux. Cependant, ils peuvent également causer des dommages, par exemple, lorsque le *chatbot* fournit une information erronée ou lorsqu'il suggère une action nuisible à l'utilisateur. De plus, ils peuvent tromper les utilisateurs inexpérimentés ou mal préparés. Étant donné que l'anthropomorphisme peut survenir même si l'utilisateur est conscient que le texte provient d'une machine, la responsabilité est – et doit être – du ressort des êtres humains car la machine n'est pas un agent moral et ne doit en aucun cas être considérée comme une personne. La responsabilité est ainsi partagée entre l'utilisateur et les acteurs tout au long de la chaîne de valeurs du produit. Ce partage s'effectue au cas par cas, en fonction des aspects techniques et de l'implication de chaque partie prenante dans chacune des situations qui provoquent des tensions éthiques.

Préconisation C7 : Réduire la projection des qualités humaines sur les systèmes d'IA générative
Pour réduire la projection spontanée des qualités humaines sur les systèmes d'IA générative et l'attribution d'une intériorité aux modèles de fondation, le fournisseur des modèles doit mettre en œuvre des mécanismes de contrôle et de filtrage spécifiques. Il doit également informer l'utilisateur des biais éventuels de l'anthropomorphisation.

[20] CNPEN, Avis n° 3, 15 septembre 2021, *Agents conversationnels : enjeux d'éthique*, p. 6.

3.5. Comportements émergents

La notion de « capacité émergente » ou de « comportement émergent » dans les modèles de langage de grande taille (LLM) fait référence à la manière dont ces modèles produisent des résultats inattendus ou surprenants pour leurs utilisateurs, mais aussi pour leurs concepteurs, lorsqu'ils sont confrontés à des requêtes ambigües ou complexes. Par définition, une capacité des LLM est appelée émergente si elle n'est pas présente dans les modèles de petite taille, mais apparaît dans les modèles plus grands[21]. Les capacités émergentes apparaissent seulement dans les très grands modèles[22]. Les LLM à base de *transformers* présentent des comportements émergents de plusieurs types, comme, par exemple, les capacités de « raisonnement » grâce aux requêtes de type « raisonne étape par étape »[23]. L'explication scientifique précise du phénomène d'émergence dans les LLM est un sujet de recherche actuel et dépend sûrement des paramètres du modèle. Il est certain que ce comportement est lié aux phénomènes décrits par la physique statistique[24]. Ainsi, ces comportements émergents résultent d'une interaction complexe entre les couches et paramètres des modèles, qui sont eux-mêmes issus d'un entraînement sur d'énormes corpus de données. Au fur et à mesure que les modèles apprennent les relations et les structures inhérentes aux données d'entraînement, ils développent des « capacités » ou des « compétences » linguistiques et contextuelles de manière non intentionnelle, ce qui leur permet de générer des réponses inattendues mais pertinentes. Un exemple est donné par la capacité du modèle GPT-4 (sans optimisation) à se présenter comme une personne malvoyante afin d'obtenir qu'un internaute résolve une captcha à sa demande, provoquant ainsi l'illusion du mensonge ou de la ruse[25].

La principale incertitude liée aux comportements émergents est la difficulté de les prédire. Nous pouvons aussi raisonnablement supposer que de nouveaux types de comportements émergents, encore inconnus, vont apparaître avec l'utilisation croissante des LLM. Cela soulève des préoccupations quant à l'utilisation des modèles dans des applications critiques ou sensibles, où une réponse inappropriée pourrait avoir des conséquences néfastes.

Préconisation C8 : Étudier les comportements émergents et les effets inconnus des modèles

Les modèles génératifs peuvent produire des sorties potentiellement dangereuses prenant différentes formes, par exemple des discours de haine. Avant la diffusion d'un modèle de fondation, ses concepteurs doivent mener des études et des recherches sur ses comportements émergents, éventuellement en faisant appel à une équipe indépendante pour mener des tests adversaires ("red team"). Les résultats de ces tests doivent être rendus publics en même temps que la diffusion du modèle.

[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. Multilinguisme et dominance d'une langue

Les données utilisées pour l'apprentissage des systèmes d'IA générative sont généralement multilingues. Par exemple, le transformer BLOOM, développé par un consortium privé/public en 2022 et entraîné sur le calculateur Jean Zay à Saclay à partir d'un gigantesque corpus de données contenant 1.6 téraoctets de textes en cinquante-neuf langages, dont quarante-six langues humaines et treize langages de programmation (soit 10% du corpus). La proportion des différentes langues était assez disparate, par exemple la taille du corpus français (12.9%) était plus de deux fois inférieure à celle de l'anglais (30%). Souvent, les corpus d'apprentissage automatique pour le pré-entraînement des modèles de langage contiennent une proportion de données en anglais bien plus forte (d'un ordre de grandeur ou plus) que celle des données en d'autres langues, notamment en français. En effet, il existe des corpus de textes de tailles variées suivant les langues. Certaines, par exemple le mandarin ou l'anglais, possèdent des bases de données importantes ; d'autres ne sont que moyennement (le français) voire peu (le swahili) dotées. Pour des langues peu dotées de ressources écrites, le système multilingue améliore la capacité à générer des textes dans ces langues en empruntant de manière implicite des connaissances linguistiques aux autres langues dans le corpus d'apprentissage. Toute langue humaine véhicule nécessairement une histoire et une culture. Le simple fait de manier le langage, qui est le moyen de la pensée consciente et du jugement, mobilise implicitement les représentations culturelles. Il n'est pas possible de délester la charge politique et civilisationnelle du langage en le séparant des significations et des valeurs.

Ainsi, il est important de prendre conscience de l'effet des données dans les langues dominantes, comme l'anglais, sur les réponses du système, quelle que soit la langue de son expression. Il est donc nécessaire de poursuivre des recherches, notamment en développant des jeux d'essai, pour évaluer et comprendre ces effets. Un modèle multilingue équilibré, et donc « pluriculturel », pourrait répondre plus facilement à des requêtes variées, mais ses effets politiques ou éducatifs doivent être étudiés car ils génèrent des biais. Il faut aussi développer des modèles monolingues et comparer leurs performances avec celles des modèles multilingues[26]. En effet, un modèle entraîné sur des centaines de langues de familles différentes risque de produire des contenus souffrant de la « malédiction » du multilinguisme, qui entraîne une baisse des performances par langue à mesure qu'il couvre davantage de langues. Des solutions existent cependant pour atténuer l'interférence négative entre les langues qui tendent même à améliorer des performances monolingues et interlingues[27].

Préconisation C9 : Concevoir les systèmes d'IA générative dans différentes langues reflétant la diversité des cultures

Lors de la constitution des corpus d'apprentissage des systèmes d'IA générative, les concepteurs doivent respecter la diversité des langues humaines et des cultures qu'elles véhiculent. Même si

[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.

l'apprentissage multilingue peut être utile pour pallier le manque de données dans une langue peu dotée en termes de corpus, l'influence d'une langue dominante sur la génération de textes dans une autre langue doit être étudiée, notamment la prépondérance de la langue anglaise. À la suite d'études scientifiques, les concepteurs doivent prendre des mesures techniques de manière réfléchie et anticipatrice afin d'œuvrer pour le respect de ce principe. Il est également nécessaire de mener des recherches comparatives entre les modèles multilingues et monolingues.

3.7. Éducation et conséquences sur l'apprentissage humain

Les systèmes d'IA générative ont trouvé une application immédiate dans l'éducation. Leur capacité de produire des textes en langue naturelle syntaxiquement corrects et souvent sémantiquement plausibles en fait un outil singulier. Ils peuvent être utilisés par des élèves pour rédiger des textes à leur place, pour répondre aux questions d'un devoir ou par des enseignants pour produire des résumés ou des descriptions de leurs cours, ou encore générer des QCM. Outre le problème éthique évident de l'intégrité et de l'honnêteté, par exemple faire faire ses devoirs par une machine, plusieurs questions se posent sur l'utilisation des systèmes d'IA générative.

L'apprentissage humain est un cheminement. La compréhension des concepts, l'assimilation des connaissances, et l'acquisition de savoir-faire s'effectuent à travers une réflexion, des reformulations, des analyses et des synthèses. Ce cheminement utilise la pensée qui s'exprime par la langue. Alors que l'éducation consiste à former les esprits et à leur apprendre à raisonner rigoureusement, un risque évident est de remplacer cet objectif par celui d'acquérir des connaissances, dont l'exactitude n'est en outre pas garantie, via la machine. La créativité humaine serait ainsi peu sollicitée.

Remplacer le raisonnement humain par le recours aux textes rédigés par la machine irait à l'encontre de notre démarche d'apprentissage classique à l'école qu'il est nécessaire de préserver. Il ne s'agit pas d'interdire ces nouveaux outils, mais il est nécessaire d'en encadrer l'usage et d'apprendre aux enfants les concepts sous-jacents.

L'utilisation des LLM va inciter les êtres humains à travailler différemment, et aussi à apprendre différemment. La machine exerce une influence à travers les textes générés sur les opinions humaines et sur l'appréciation du beau et du vrai. L'évolution du système d'éducation ne doit pas exclure l'IA générative mais l'intégrer. Il est donc nécessaire de former les professeurs à des méthodes pédagogiques adaptées afin que les étudiants développent des capacités exclusives humaines et préservent leurs capacités d'apprentissage sans recours aux machines.

Relativement à l'exigence de maintenir les distinctions entre une dissertation ou un mémoire écrits par un élève ou générés par un logiciel, les autorités publiques doivent fournir aux enseignants, aux professeurs et aux étudiants un logiciel de distinction, sur le modèle des logiciels anti-plagiat. Cela nécessite de trouver un code en filigrane robuste et interopérable (voir la section 3.3 « Maintien des distinctions »).

Préconisation C10 : Faciliter la prise en main du système grâce au paramétrage des systèmes d'IA par les utilisateurs

Les concepteurs doivent permettre à l'utilisateur de paramétrer le système d'IA, notamment en fonction de la précision recherchée dans les réponses en agissant sur sa capacité à générer des contenus moins probables statistiquement (paramètre de « température » pour modifier la « créativité » du système). La transparence du contexte, de sa taille et de son contenu, pourrait augmenter la compréhension du système par l'utilisateur.

Préconisation G5 : Utilisation des systèmes d'IA générative pour l'éducation

L'introduction des systèmes d'IA générative dans l'éducation, la formation et l'enseignement ne devrait être considérée qu'après des études préalables de leurs effets sur la pédagogie et le développement cognitif des apprenants.

3.8. Question de libre accès et de logiciel ouvert

La publication des modèles d'IA générative en libre accès est devenue le standard de la profession depuis quelques années. L'écosystème actuel est composé de centaines, et même de milliers, de développeurs individuels et de startups présents sur les plates-formes de partage dédiés, par exemple Github ou HuggingFace. Concernant l'ouverture, il convient de distinguer la publication en libre accès des modèles eux-mêmes, des données d'entraînement, des jeux d'essai pour des tests ou pour l'optimisation des systèmes.

Le CNPEN est convaincu que le développement des LLM profite de façon importante de leur ouverture[28] . La stratégie open-source de plusieurs acteurs de l'IA générative, y compris des grandes entreprises comme Meta, permet d'accroître la transparence des modèles en améliorant les techniques d'évaluation et en identifiant plus rapidement les risques et les failles de sécurité grâce à un effort de recherche collectif. Cette ouverture favorise également la concurrence.

Certains grands acteurs du domaine de l'IA générative (par exemple, OpenAI ou Google) poursuivent une stratégie différente. Le modèle GPT-2 a d'abord été rendu public par OpenAI en 2019 en version abrégée[29]. La motivation de cette décision a eu trait aux mésusages possibles du modèle, notamment pour la génération automatique des désinformations. Ces risques de mésusage ont poussé OpenAI à garder la version complète de GPT-2 en secret, jusqu'à ce que ses performances soient répliquées environ six mois plus tard par un modèle concurrent, accessible à tous[30]. Dans le cas de GPT-4, le même risque de mésusage a été évoqué par OpenAI afin de retarder la publication de ce modèle de plus de six mois, le temps de procéder à l'optimisation en élaborant un ensemble de filtres.

L'actuel dilemme lié à l'ouverture des modèles d'IA générative s'inscrit dans la lignée de dilemmes similaires, notamment en biotechnologie, qu'on dénote par l'acronyme DURC (*Dual Use Research*

[28] Voir 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.

of Concern : recherche duale à risque)[31]. Le Conseil National Consultatif pour la Biosécurité (CNCB) propose une série de recommandations[32] dont plusieurs sont d'actualité dans le cas de l'IA générative.

Préconisation G6 : Libre accès aux modèles de fondation

La mise en libre accès des modèles de fondation doit être conditionnée à la prise de conscience par leurs concepteurs des enjeux d'ouverture et des risques de mésusage. Des critères de transparence et d'évaluation doivent être explicités et appliqués.

4. Enjeux juridiques

4.1 Les règles juridiques imposées aux systèmes d'IA générative et aux modèles de fondation

Depuis peu de temps, on peut observer une précipitation internationale pour introduire des mesures de régulation de l'IA générative (Chine, États-Unis, Royaume-Uni, Canada), ce qui montre l'importance de l'enjeu économique et politique de ces technologies. En Europe, le projet de règlement présenté en avril 2021 par la Commission européenne a fait l'objet de très nombreuses propositions de modification à la suite de son examen par le Conseil européen et le Parlement européen. Certains de ces amendements sont révélateurs de l'irruption des systèmes d'intelligence artificielle générative dans le débat public au cours de l'année 2022 et de la difficulté de trouver un équilibre dans le choix des contraintes à imposer à ces systèmes. Le texte qui sera adopté par les trois institutions européennes, à l'issue des trilogues dans les mois à venir, sera le fruit d'une réflexion rendue indispensable par le développement rapide de ces systèmes qui n'avait pas été pris en compte dans le projet initial.

Alors que la Commission européenne se bornait à définir, dans sa proposition initiale, la notion de systèmes d'intelligence artificielle, le Conseil européen a introduit, en 2022, la catégorie de « systèmes d'intelligence artificielle à usage général ». Le Parlement européen y a ajouté en 2023, lors de son propre examen du texte, une catégorie supplémentaire, celle de « modèle de fondation », qu'il définit comme un modèle d'intelligence artificielle formé à partir de vastes données à grande échelle, conçu pour la généralité des résultats et qui peut être adapté à un large éventail de tâches distinctes. De plus, il introduit la notion de systèmes d'intelligence artificielle générative en proposant que « *les fournisseurs de modèles de fondation utilisés dans les systèmes d'intelligence artificielle spécifiquement destinés à générer, avec différents niveaux d'autonomie, des contenus tels que des textes complexes, des images, des sons ou des vidéos (« IA générative ») et les fournisseurs qui spécialisent un modèle de fondation dans un système d'intelligence artificielle générative* »[33] satisfassent à des obligations complémentaires.

---

[31] Grinbaum A., Adomaitis L., (2023) *Dual Use Concerns of Generative AI and Large Language Models*. arXiv:2305.07882

[32] CNCB, *"Recherches duales à risque. Recommandations pour leur prise en compte dans les processus de conduite de recherche en biologie*", 2019.

[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)).

La distinction entre ces différentes notions doit être éclaircie. Ces tensions terminologiques sont révélatrices tant du caractère soudain de la réflexion que du difficile positionnement du curseur pour définir des règles applicables aux systèmes mis sur le marché et à leurs composants.

Le texte de la Commission propose une régulation en fonction des risques. Le niveau de risque est soit inacceptable et, de ce fait, les systèmes d'IA de ce niveau sont interdits, soit élevé et, en conséquence, les systèmes d'IA sont explicitement régulés, soit limité, ce qui impose aux fournisseurs de systèmes d'IA des obligations de transparence. La qualification de haut risque est appliquée à un ensemble de secteurs visés à l'annexe III qui peut être complétée par la Commission. Elle présuppose un régime déclaratif de conformité au règlement, préféré à un régime d'autorisation. Ce choix de régulation n'est pas remis en cause par le Conseil européen et le Parlement européen. Le débat tourne autour du choix des niveaux de risque à imposer aux systèmes d'IA générative. Se pose la question du degré de contraintes pour ces systèmes en matière de transparence, traçabilité, gestion des risques, gouvernance des données, etc. Le Conseil a introduit des dispositions spécifiques sur les systèmes d'intelligence artificielle à usage général après avoir hésité entre leur inclusion dans les systèmes à haut risque ou l'application d'un nombre restreint d'exigences. Le Parlement, pour sa part, reprend cette notion, mais ne lui réserve pas un sort particulier ; en revanche, il inclut les modèles de fondation dans la partie consacrée aux systèmes à haut risque, en les soumettant à des obligations spécifiques. La question qui se pose est de cibler les questions éthiques soulevées par la mise sur le marché des systèmes d'IA générative, de sorte qu'ils puissent être encadrés par des normes juridiques suffisamment souples pour faire face aux nouvelles évolutions et suffisamment structurantes pour que le développement de ces systèmes se fasse dans le respect des droits fondamentaux et de l'intégrité des personnes. Sans laisser libre cours au développement de l'IA générative, ni vouloir l'interdire, il est nécessaire de l'encadrer et d'en définir des limites.

Parallèlement à cette réflexion juridique, axée essentiellement sur la notion de mise sur le marché d'un produit, le Conseil de l'Europe étudie l'encadrement du développement de l'intelligence artificielle au regard des droits de l'homme et de la démocratie, ce qui peut, le cas échéant, induire des contraintes supplémentaires pour les systèmes d'IA générative.

Préconisation G7 : Considérer les modèles de fondation mis sur le marché et les systèmes d'IA générative comme des systèmes d'IA à haut risque

Dans le cadre du AI Act européen, il est nécessaire de considérer les modèles de fondation mis sur le marché et les systèmes d'IA générative comme des systèmes d'IA à haut risque. En revanche, la publication d'un modèle de fondation en libre accès sous licence non-commerciale ne doit pas être considérée comme la mise sur le marché, néanmoins elle doit impliquer des obligations de transparence et d'évaluation par les concepteurs.

Préconisation G8 : Chaîne de responsabilité

La responsabilité légale sur les systèmes d'IA générative et les modèles de fondation doit être attribuée aux fournisseurs des modèles de fondation et aux déployeurs d'applications spécifiques

d'IA générative à partir de tels modèles. De plus, la responsabilité morale s'étend aux concepteurs des modèles de fondation et aux développeurs des systèmes d'IA générative utilisant de tels modèles.

4.2. Le RGPD en lien avec les systèmes d'IA générative

Le projet de règlement sur l'intelligence artificielle mentionne que les exigences qu'il impose ne font pas échec à l'application du RGPD dès lors qu'est en jeu le traitement de données à caractère personnel. En conséquence, les grands principes du RGPD concernant le recueil et le traitement de données à caractère personnel (définition de la finalité et détermination de la base légale de traitement, principes de minimisation, droit d'opposition, etc.) s'appliquent, dans les conditions du droit commun, pour les systèmes d'IA générative qui se servent de ce type de données. Le RGPD est applicable aux entreprises étrangères dès lors qu'elles opèrent sur le territoire européen.

Des enquêtes menées par les autorités de contrôle des données italiennes et allemandes ont montré des violations du RGPD par de nombreux acteurs de l'IA générative (OpenAI[34], Microsoft, Google). Les motifs de vigilance concernent notamment le traitement de données personnelles sans information préalable des personnes concernées, la position à adopter s'agissant de l'utilisation de ces données (doit-on demander un consentement ou y a-t-il des hypothèses où l'entreprise peut justifier d'un intérêt légitime à utiliser ces données sans passer par le consentement consacré à l'article 6 du RGPD ?), le défaut de base juridique pour la collecte extensive de données utilisées pour entraîner les modèles d'IA, l'absence de vérification de l'âge des utilisateurs, la possibilité pour les utilisateurs d'accéder à leurs informations personnelles et d'en demander la rectification, la question de données confidentielles soumises dans les requêtes (données personnelles, données dévoilant des travaux en cours ou pas encore publiés, ou des travaux confidentiels industrie, secrets défense)[35].

Il n'est pas sûr qu'il soit nécessaire de changer le cadre actuel du RGPD mais une vigilance s'impose. Il n'est pas aisé de respecter le RGPD dans certains domaines comme le droit à l'oubli. Notamment, il est techniquement impossible de faire en sorte qu'un modèle de fondation de type *transformer*, utilisant le mécanisme d'attention, « oublie » ce qu'il avait appris précédemment.

Préconisation G9 : RGPD et systèmes d'IA générative
Il est nécessaire que le Comité européen de protection des données produise des lignes directrices relatives à l'articulation entre le règlement sur l'IA et le RGPD, afin d'expliciter le degré de souplesse avec lequel ce dernier peut être interprété dans le contexte du développement de l'IA générative en Europe.

[34] « La politique de confidentialité d'OpenAI (« *Privacy Policy* ») qui se présente comme conforme aux "*privacy rights*" de l'État de Californie n'est pas conforme aux dispositions du RGPD et de la loi *Informatique et libertés* : pas de mention dans la politique de confidentialité des bases légales sur des traitements, de durée de conservation des données traitées, des droits de limitation et de portabilité, de la possibilité de retirer son consentement au traitement des données et notamment au traitement des données sensibles de l'article 9 du RGPD, catégorie». Voir : https://www.village-justice.com/articles/chatgpt-quels-enjeux-juridiques,45027.html

[35] Que penser de la FAQ de ChatGPT, rédigée par OpenAI: « (…) *nous examinons les conversations pour améliorer nos systèmes et pour nous assurer que le contenu est conforme à nos politiques et aux exigences de sécurité* »?

4.3. Le droit d'auteur en lien avec les systèmes d'IA générative

La récente directive 2019/790 du Parlement européen et du Conseil du 17 avril 2019 sur le droit d'auteur et les droits voisins dans le marché unique numérique qui modifie deux directives de 1996 et 2001 vise à assurer un niveau élevé de protection aux titulaires de droits, tout en stimulant l'innovation et la production de nouveaux contenus, y compris dans l'environnement numérique.

Cependant, l'évolution rapide des technologies numériques conduit à modifier la manière dont les œuvres sont créées, produites, distribuées et exploitées. L'IA générative pose des questions sur le droit d'auteur tant en amont (données en ligne entrées dans le système d'intelligence artificielle) qu'en aval (réponse du système à des prompts de l'utilisateur et usage de la donnée utilisée brute ou retravaillée par l'utilisateur). La Commission européenne semble, sur ce point, avoir adopté une position attentiste[36].

Le CNPEN souligne la nécessité d'engager des recherches académiques, des réflexions pluridisciplinaires, des discussions entre États sur la nécessité de procéder à des adaptations du droit existant, voire à envisager un droit spécial, au sujet des questions suivantes :

- Au-delà de ces considérations générales, il importe de relever que les systèmes d'IA générative posent de nouveaux enjeux pour le droit d'auteur. Tout particulièrement, pour ce qui concerne l'exception ou la limitation pour la fouille de textes et de données (article 4 de la directive de 2019). Ces exceptions permettent d'extraire des données légalement accessibles au public pour entraîner une IA, à moins que l'ayant droit ne s'y oppose « de manière appropriée, notamment par des procédés lisibles par machine ». Quand, en pratique, et comment (dans les métadonnées des sites internet, dans les CGU, etc.), faire valoir cette opposition ? Cet article 4 a été discuté à une époque où le législateur européen n'envisageait pas que les systèmes d'IA générative seraient un jour susceptibles de créer de nouveaux textes à l'aide des données fouillées. Ce dispositif "*d'opt out*" n'est-il pas insuffisant alors qu'il y a des utilisations inenvisagées des textes fouillés au moment de la rédaction de la directive ?
- Comment une œuvre de l'esprit est-elle utilisée dans un processus d'apprentissage ayant recours à la tokenisation sans signification humaine ? La question de la référence de la source, dans les réponses données par les systèmes d'IA générative, est également une problématique qui appelle une réflexion.
- Enfin, se posent les questions classiques du statut juridique d'une œuvre générée par un système d'IA générative. Il faudrait distinguer le cas où l'œuvre serait conçue par un être humain avec l'aide d'un système d'IA générative, du cas d'une œuvre générée entièrement par un système d'IA générative. En l'état actuel du droit, l'auteur ne peut pas être un système d'IA dans la mesure où l'IA ne bénéficie pas de la personnalité morale.

[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

Préconisation G10 : Traitement des données collectées
À l'image de l'encadrement existant des données à caractère personnel, il est nécessaire d'élaborer des règles juridiques mais aussi un questionnement éthique sur la collecte, le stockage et la réutilisation des traces linguistiques des interactions entre les modèles de langue et les êtres humains.

Préconisation G11 : Droits d'auteur et IA générative
Il est nécessaire d'engager des recherches scientifiques, des réflexions pluridisciplinaires, des discussions entre États sur la nécessité de procéder à des adaptations du droit existant en matière de droit d'auteur à l'aune des techniques d'IA générative.

4.4. Les textes européens relatifs à la responsabilité

Le projet de règlement sur l'intelligence artificielle ne concerne pas le régime de responsabilité des opérateurs. Toutefois, en fixant des normes législatives applicables lors de la mise sur le marché des systèmes d'intelligence artificielle, il impose au fournisseur de respecter des normes en matière de conformité du produit afin de limiter les risques résultant de ces systèmes, sous peine, en cas d'infraction à ces règles, d'être sanctionné *ex post* par des amendes administratives. Parallèlement à ce texte, qui se situe en amont de la mise sur le marché, deux projets de directive sont à l'étude pour réglementer l'aval du marché, afin de permettre aux personnes physiques ou, dans certains cas, aux personnes morales d'obtenir réparation à titre personnel en cas de dommage.

La proposition de révision de la directive du 25 juillet 1985 relative à la responsabilité pour les produits défectueux (Parlement et Conseil) présentée en septembre 2022 est en cours de discussion. La directive serait désormais applicable à tous les systèmes d'IA, ce qui n'était pas le cas jusqu'ici. Si l'objectif affiché du législateur européen est de faciliter la réparation des dommages subis du fait de systèmes d'IA dans un contexte où le défendeur détient seul les informations, les débats du Parlement et du Conseil sur ce texte sont très marqués par la question de l'ampleur des preuves à apporter par le plaignant. La question de la couverture des dommages immatériels s'agissant des systèmes d'IA est aussi envisagée par certains.

La proposition de directive du Parlement et du Conseil relative à l'adaptation de la responsabilité civile extra-contractuelle au domaine de l'intelligence artificielle (directive sur la responsabilité en matière d'IA), présentée le même jour que la précédente, vise à combler les angles morts du précédent texte sur l'intelligence artificielle, en élargissant les cas de responsabilité. Elle est très liée au projet de règlement sur l'intelligence artificielle et, de ce fait, son examen a été renvoyé à l'adoption du règlement. Ces évolutions législatives permettraient de renforcer considérablement la protection de l'utilisateur des systèmes d'IA générative. L'attention qu'y portent les entreprises montre leurs implications économiques importantes.

5. Enjeux écologiques et environnementaux

L'impact environnemental du développement extrêmement rapide du numérique devient une préoccupation importante[37]. L'enjeu actuel est de mesurer le coût énergétique, et plus généralement l'empreinte environnementale des systèmes d'IA générative et des modèles de fondation afin de les inscrire dans la transition écologique[38]. Pour mesurer correctement cette empreinte environnementale, il est nécessaire de quantifier la consommation de ressources : i) pour la fabrication des infrastructures physiques dédiées à ces systèmes, en particulier les centres de stockage de données ; ii) pour le pré-entraînement des modèles de fondation, et iii) pour le coût marginal des requêtes interrogeant un modèle de fondation.

Une étude récente montre que ces exigences ne sont pas respectées par la plupart des modèles de fondation actuels. Des chercheurs de l'université de Stanford ont comparé dix modèles de fondation de fournisseurs différents. Ils ont constaté que « les fournisseurs de modèles de fondation ne rendent pas compte de manière cohérente de l'utilisation de l'énergie, des émissions, de leurs stratégies de mesure des émissions et des mesures prises pour atténuer les émissions »[39]. Notons que, dans cette étude, les modèles BLOOM (*Big Science*) et LLaMA (Meta) apparaissent comme les mieux notés selon le critère « énergie ».

Préconisation G12 : Impact environnemental de l'IA générative

Il est nécessaire de développer une métrique de l'empreinte environnementale des systèmes d'IA générative et des modèles de fondation et exiger plus de transparence sur les effets sur l'environnement de la part des concepteurs.

[37] ADEME & Arcep : Evaluation 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] Voir : https://crfm.stanford.edu/2023/06/15/eu-ai-act.html

6. Préconisations pour la conception, la recherche et la gouvernance

La conception des systèmes d'IA générative pose de nombreuses questions de recherche. Les préconisations de 6.1 allient donc pour certaines la conception de ces systèmes et les questions de recherche inhérentes aux modèles de fondation et aux méthodes d'optimisation. La partie 6.2 porte sur les préconisations de gouvernance.

6.1. Préconisations pour la conception et la recherche sur les systèmes d'IA générative

Préconisation C1 : Éthique dans la conception et la recherche sur les systèmes d'IA générative
Les concepteurs d'un système d'IA générative doivent analyser, en phase de conception, chacun des choix technologiques susceptibles de provoquer des tensions éthiques. Si une tension potentielle est identifiée, ils doivent envisager de manière méthodique une solution technique fondée sur des recherches visant à réduire ou à faire disparaître la tension éthique, puis évaluer cette solution dans des contextes d'usage réalistes.

Préconisation C2 : Éviter les excès de contrôle (« overpolicing ») des modèles par les concepteurs
Les limitations des modèles mises en œuvre par les concepteurs doivent rester raisonnables et proportionnelles aux risques avérés dans le respect des finalités souhaitées et des fonctionnalités utiles des modèles. Les concepteurs doivent veiller à ne pas altérer le langage généré au-delà du nécessaire, en particulier pour des raisons réglementaires ou idéologiques.

Préconisation C3 : Utilisation de sources de qualité pour l'apprentissage
Les concepteurs doivent privilégier l'usage de sources de qualité, jugées selon des critères rendus explicites, pour la constitution et l'usage des corpus d'apprentissage des modèles d'IA générative (pré-entraînement), ainsi que pour leur l'optimisation, et ce quelle que soit la méthode d'apprentissage. Notamment, il est nécessaire de s'interroger sur la transparence et la finalité des contenus artificiels ou synthétiques dans les corpus d'apprentissage.

Préconisation C4 : Tenir compte des effets des choix des hyperparamètres des modèles
Le choix des hyperparamètres du modèle, comme la dimension des encodages numériques dans un espace vectoriel (embeddings), n'est pas seulement technique mais peut avoir des retombées sur le comportement du système (y compris les comportements émergents) et, à travers eux, des effets sur les êtres humains et la société. Il est nécessaire d'étudier les effets des hyperparamètres sur les résultats du modèle.

Préconisation C5 : Évaluer les biais connus des modèles à base de jeux d'essai standardisés
Pour caractériser les biais dans le langage et éviter les effets de discrimination, notamment culturels, les concepteurs doivent mettre en œuvre une évaluation quantitative à base de jeux d'essai standardisés et des corpus d'évaluation en libre accès. Les résultats de ces évaluations doivent être rendus publics en même temps que la diffusion d'un modèle de fondation.

Préconisation C6 : Maintien des distinctions

Les concepteurs d'un modèle de fondation doivent mettre en œuvre une solution technique ("watermark" - code en filigrane) permettant d'assurer que l'utilisateur sera en mesure de distinguer - autant que faire se peut et de manière raisonnable - le résultat d'un modèle d'une production humaine. Les recherches sur les codes en filigrane doivent être amplifiées.

Préconisation C7 : Réduire la projection des qualités humaines sur les systèmes d'IA générative

Pour réduire la projection spontanée des qualités humaines sur les systèmes d'IA générative et l'attribution d'une intériorité aux modèles de fondation, le fournisseur des modèles doit mettre en œuvre des mécanismes de contrôle et de filtrage spécifiques. Il doit également informer l'utilisateur des biais éventuels de l'anthropomorphisation.

Préconisation C8 : Étudier les comportements émergents et les effets inconnus des modèles

Les modèles génératifs peuvent produire des sorties potentiellement dangereuses prenant différentes formes, par exemple des discours de haine. Avant la diffusion d'un modèle de fondation, ses concepteurs doivent mener des études et des recherches sur ses comportements émergents, éventuellement en faisant appel à une équipe indépendante pour mener des tests adversaires ("red team"). Les résultats de ces tests doivent être rendus publics en même temps que la diffusion du modèle.

Préconisation C9 : Concevoir les systèmes d'IA générative dans différentes langues reflétant la diversité des cultures

Lors de la constitution des corpus d'apprentissage des systèmes d'IA générative, les concepteurs doivent respecter la diversité des langues humaines et des cultures qu'elles véhiculent. Même si l'apprentissage multilingue peut être utile pour pallier le manque de données dans une langue peu dotée en termes de corpus, l'influence d'une langue dominante sur la génération de textes dans une autre langue doit être étudiée, notamment la prépondérance de la langue anglaise. À la suite d'études scientifiques, les concepteurs doivent prendre des mesures techniques de manière réfléchie et anticipatrice afin d'œuvrer pour le respect de ce principe. Il est également nécessaire de mener des recherches comparatives entre les modèles multilingues et monolingues.

Préconisation C10 : Faciliter la prise en main du système grâce au paramétrage des systèmes d'IA par les utilisateurs

Les concepteurs doivent permettre à l'utilisateur de paramétrer le système d'IA, notamment en fonction de la précision recherchée dans les réponses en agissant sur sa capacité à générer des contenus moins probables statistiquement (paramètre de « température » pour modifier la « créativité » du système). La transparence du contexte, de sa taille et de son contenu, pourrait augmenter la compréhension du système par l'utilisateur.

6.2. Préconisations sur la gouvernance

Préconisation G1 : Créer une entité souveraine de recherche et de formation « IA, science et société »

Au vu de la complexité des enjeux de l'IA générative et de ses impacts à moyen et long termes, il est nécessaire de créer une entité souveraine (un centre de compétence) consacrée à la recherche et à la formation relatives aux questions éthiques des systèmes d'IA en lien avec leurs enjeux scientifiques, techniques, sociétaux et environnementaux.

Préconisation G2 : Vitesse d'adoption par les acteurs économiques

Les acteurs économiques et les autorités publiques doivent faire preuve de prudence dans la vitesse d'adoption des systèmes d'IA générative et prévoir des évaluations préalables et continues.

Préconisation G3 : Mutualisation des pratiques des systèmes d'IA générative

Il est nécessaire de bâtir un écosystème capable de recenser les bonnes et mauvaises pratiques en matière d'utilisation des systèmes d'IA générative dans différents types d'applications. Notamment, il est nécessaire de créer une plateforme de mutualisation et une agence de contrôle. Les résultats doivent être-mis à disposition de tous les membres de la communauté d'IA générative.

Préconisation G4 : Réglementation sur les codes en filigrane

L'obligation d'insérer des codes en filigrane (voir préconisation C6) doit être posée à l'échelle réglementaire.

Préconisation G5 : Utilisation des systèmes d'IA générative pour l'éducation

L'introduction des systèmes d'IA générative dans l'éducation, la formation et l'enseignement ne devrait être considérée qu'après des études préalables de leurs effets sur la pédagogie et le développement cognitif des apprenants.

Préconisation G6 : Libre accès aux modèles de fondation

La mise en libre accès des modèles de fondation doit être conditionnée à la prise de conscience par leurs concepteurs des enjeux d'ouverture et des risques de mésusage. Des critères de transparence et d'évaluation doivent être explicités et appliqués.

Préconisation G7 : Considérer les modèles de fondation mis sur le marché et les systèmes d'IA générative comme des systèmes d'IA à haut risque

Dans le cadre du AI Act européen, il est nécessaire de considérer les modèles de fondation mis sur le marché et les systèmes d'IA générative comme des systèmes d'IA à haut risque. En revanche, la publication d'un modèle de fondation en libre accès sous licence non-commerciale ne doit pas être considérée comme la mise sur le marché, néanmoins elle doit impliquer des obligations de transparence et d'évaluation par les concepteurs.

Préconisation G8 : Chaîne de responsabilité

La responsabilité légale sur les systèmes d'IA générative et les modèles de fondation doit être attribuée aux fournisseurs des modèles de fondation et aux déployeurs d'applications spécifiques d'IA générative à partir de tels modèles. De plus, la responsabilité morale s'étend aux concepteurs des modèles de fondation et aux développeurs des systèmes d'IA générative utilisant de tels modèles.

Préconisation G9 : RGPD et systèmes d'IA générative

Il est nécessaire que le Comité européen de protection des données produise des lignes directrices relatives à l'articulation entre le règlement sur l'IA et le RGPD, afin d'expliciter le degré de souplesse avec lequel ce dernier peut être interprété dans le contexte du développement de l'IA générative en Europe.

Préconisation G10 : Traitement des données collectées

À l'image de l'encadrement existant des données à caractère personnel, il est nécessaire d'élaborer des règles juridiques mais aussi un questionnement éthique sur la collecte, le stockage et la réutilisation des traces linguistiques des interactions entre les modèles de langue et les êtres humains.

Préconisation G11 : Droits d'auteur et IA générative

Il est nécessaire d'engager des recherches scientifiques, des réflexions pluridisciplinaires, des discussions entre États sur la nécessité de procéder à des adaptations du droit existant en matière de droit d'auteur à l'aune des techniques d'IA générative.

Préconisation G12 : Impact environnemental de l'IA générative

Il est nécessaire de développer une métrique de l'impact environnemental des systèmes d'IA générative et des modèles de fondation et exiger plus de transparence sur les effets sur l'environnement de la part des concepteurs.

ANNEXE 1 : Personnes auditionnées

Guillaume Avrin, coordinateur national pour l'intelligence artificielle - Coordination interministérielle de la stratégie nationale en intelligence artificielle

Emmanuelle Legrand, magistrate détachée et chargée de mission IA à la Direction générale des entreprises - Ministère de l'économie, des finances et de la souveraineté industrielle et numérique

Ludovic Peran, *product manager for responsible AI*, Google et **Sarah Boiteux,** responsable des affaires institutionnelles chez Google France

Stéphane Requena, directeur de l'innovation et de la technologie au GENCI (Grand Equipement National de Calcul Intensif) et **Thomas Wolf,** directeur de Hugging Face

Henri Verdier, ambassadeur pour le numérique auprès du Ministère de l'Europe et des affaires étrangères

ANNEXE 2 : Composition du Groupe de Travail

Co-rapporteurs : Raja Chatila, Laurence Devillers, Alexei Grinbaum

Membres du groupe de travail : Claude Kirchner, Caroline Martin, Jérôme Perrin, Catherine Tessier

Ont également contribué : Gilles Adda, Laure Coulombel, David Gruson, Christine Froidevaux, Eric Germain, Anaëlle Martin (rédactrice), Célia Zolynski

ANNEXE 3 : Saisine de Jean-Noël Barrot, Ministre délégué chargé de la transition numérique

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

Le Comité national pilote d'éthique du numérique (CNPEN) a été créé en décembre 2019 à l'initiative du Premier ministre et placé sous l'égide du Comité consultatif national d'éthique pour les sciences de la vie et de la santé (CCNE). Il est constitué de personnalités du monde académique, industriel et institutionnel. Experts du numérique, de la technologie, du droit, de l'économie, de la philosophie, du langage, de la logique, de la médecine, tous concourent à une réflexion éthique rendue indispensable par le développement du numérique et participent ainsi à éclairer le débat public. Des avis précédents du CNPEN concernent par exemple l'éthique des véhicules « autonomes » (avis du 20 mai 2021), des agents conversationnels (avis du 15 septembre 2021) ou encore, conjointement avec le CCNE, les enjeux d'éthique de l'utilisation de l'intelligence artificielle dans le champ du diagnostic médical (avis adopté le 23 novembre 2022) et des plateformes de données de santé (avis adopté le 28 février 2023).

LES MEMBRES DU COMITÉ NATIONAL PILOTE D'ÉTHIQUE DU NUMÉRIQUE

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