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

Paris, le 5 juillet 2023

COMMUNIQUE DE PRESSE

AVIS 7

Systèmes d'intelligence artificielle générative : enjeux d'éthique

Plus un jour sans un sujet sur l'intelligence artificielle (IA) dans les médias. Pour les entreprises, les scientifiques, les éducateurs, les élèves, les producteurs de contenu, les gouvernants… pour tous, l'heure est aux grandes interrogations, aux enthousiasmes, aux fantasmes, aux inquiétudes. Dans la continuité de son Avis (n°3) sur les enjeux éthiques des agents conversationnels (« chatbots ») fin 2021, le comité national pilote pour l'éthique du numérique (CNPEN) publie un nouvel avis « *Systèmes d'intelligence artificielle générative : enjeux d'éthique* », en réponse à une saisine du ministre Jean-Noël Barrot, chargé de la Transition numérique et des Télécommunications.

Entre ces deux avis du CNPEN, ChatGPT est entré dans les bureaux et les foyers (novembre 2022), Science Po en a interdit l'usage à ses étudiants (janvier 2023), certains des fondateurs emblématiques de ces systèmes ont demandé une pause dans leur développement (mars 2023), le président Emmanuel Macron annonce un financement de 500 millions d'euros pour rester dans la course (juin 2023) et le Parlement européen peaufine le texte de l'*AI ACT* pour poser un cadre règlementaire nécessaire d'ici à fin 2023. Tout va très vite, le changement d'échelle est bien là.

Le CNPEN a mené cette réflexion en un temps record, la saisine ministérielle datant du 20 février. Un temps court à l'image du sentiment d'urgence de la réponse à apporter à l'emballement technologique. Sa mission est d'« *examiner les questions d'éthique liées à la conception, aux usages et aux impacts sur la société ainsi que les accompagnements nécessaires à la mise en œuvre de ces outils.* » Le présent avis s'intéresse aux systèmes d'IA générative- spécifiquement sur les modèles de langue- dans l'état actuel

des connaissances scientifiques. Il émet 22 préconisations à destination des concepteurs et de la recherche scientifique (10) d'une part, et des gouvernances (12) d'autre part. Ce travail vise à contribuer aux débats législatif et environnemental européens.

Gigantisme et accélération

Si les techniques d'apprentissage automatique existent depuis plusieurs décennies, c'est l'arrivée en 2017 de l'architecture de réseaux de neurones entraînés par auto-apprentissage- les transformers- qui a accéléré la puissance de calcul des modèles. Ces systèmes sont aujourd'hui entrainés sur des corpus de données de plus en plus larges. Leur mode d'apprentissage consiste à établir des corrélations statistiques entre des éléments de données appelés « *tokens* » (segments de mots, parties d'images) utilisés pour leur entraînement. Pour générer une réponse, le programme procède donc en « devinant » la suite de chaque élément de son texte d'après les modèles de son encyclopédie personnelle. Ainsi, chaque utilisateur qui interroge l'outil par le biais d'un *prompt* (énoncé ou question posée) contribue à augmenter les connaissances de la machine en enrichissant son corpus des données saisies lors de l'échange.

Le déploiement gratuit de la version 3.5 du chatbot générateur de texte ChatGPT par Open AI (Californie) il y a 7 mois est l'illustration de cette accélération des capacités des systèmes d'intelligence générative. Mais les résultats émis par l'outil peuvent être source d'erreur et mener à la désinformation si l'utilisateur ne fait pas preuve d'esprit critique. En démocratisant l'usage de l'IA générative de texte, ChatGPT a eu un effet considérable sur la perception par les utilisateurs des performances de ces systèmes mais aussi sur la prise de conscience de leurs effets sur l'individu, la société, la culture, l'économie, l'éducation et l'environnement.

Responsabilisation dès la conception, transparence des sources et des processus

Pour éviter des tensions technologiques puis sociétales, le CNPEN invite les concepteurs de modèles de langage à faire preuve d'éthique dès la phase de conception et d'exigence dans la prise en compte des biais. **Ainsi, le Comité appelle à l'analyse systématique de chacun des choix technologiques** (C1), à la pondération en évitant les excès de contrôle des modèles pour ne pas appauvrir le langage généré (C2), et à l'utilisation de sources de qualité pour l'apprentissage du système *vs* données synthétiques (C3). Il préconise également la vigilance quant aux effets des choix des hyperparamètres, qui au-delà de l'aspect technique peuvent générer des comportements « émergents », c'est-à-dire imprévisibles (C4). En effet, le système d'IA générative n'a aucune logique ni compréhension des mots qu'il emploie. Son mode d'apprentissage par corrélation numérique statistique, sans notion du sens, peut générer des erreurs ou « hallucinations » ce qui pose la question de la vérité. La production d'erreurs et l'illusion de la vérité ne pouvant être attribué à la machine, c'est le concepteur du programme qui doit en assumer la responsabilité et avertir l'utilisateur des risques encourus, afin d'éviter le risque de la désinformation.

Par ailleurs, le CNPEN est convaincu que le développement des systèmes profite de façon importante de leur ouverture en libre accès, comme c'est le cas actuellement dans l'écosystème. Il préconise cependant que cette ouverture soit soumise à la prise de conscience par les concepteurs des enjeux et des risques de mésusage *via* des critères de transparence et d'évaluation explicites (G5).

Distinguer l'homme et la machine

Le changement radical de l'usage des systèmes d'IA générative survient avec le maniement de la langue par la machine. Le dialogue avec l'agent conversationnel peut donner l'illusion à l'utilisateur d'interagir avec un être doté de conscience. **Le CNPEN rappelle que cette situation nouvelle peut induire différents risques de manipulation, intentionnelle ou pas, et de projection de qualités humaines sur la machine.** Il met en garde contre l'anthropomorphisation, contre les risques de transfert ou éventuels risques psychologiques pour l'utilisateur (perte de repères, confiance aveugle, dévoilement de l'intimité, manipulation politique).

Afin de lutter contre ces impacts négatifs, le Comité prône un maintien des distinctions d'une production humaine des résultats issus de la machine par l'utilisation de code en filigrane (C6), la mise en œuvre d'une évaluation quantitative des biais connus (C5) et celle de mécanismes de contrôle de filtrage spécifique (C7). **Le CNPEN plaide pour la construction d'un écosystème vertueux capable de recenser et de partager les bonnes et mauvaises pratiques en matière d'utilisation des systèmes d'IA générative** (G3), qui passe notamment par l'obligation réglementaire d'insérer des codes en filigrane (G10).

La langue, véhicule de culture et le chemin de l'apprentissage humain

La langue utilisée pour l'apprentissage des systèmes d'IA générative n'est pas anodine. Si les données utilisées sont généralement multilingues, on note une forte prédominance du corpus de ressources en anglais. Or, une langue est indissociable des représentations culturelles qui l'accompagnent. Ainsi, selon le CNPEN, **les concepteurs de systèmes d'IA générative doivent respecter la diversité des langues humaines et donc des cultures** (C9).

Dès leur mise à disposition, les systèmes d'IA générative ont trouvé une application dans l'éducation. Au-delà du problème évident de l'intégrité et de l'honnêteté (faire faire ses devoirs par une machine), l'enjeu sociétal est de préserver l'apprentissage humain qui passe par la compréhension des concepts, la réflexion et la créativité, sans avoir recours aux machines. Si le système d'éducation ne peut et ne doit pas exclure l'IA générative, il doit l'intégrer, en encadrer l'usage. Il s'agit d'apprendre aux enfants et étudiants les concepts sous-jacents pour augmenter leur compréhension du système et faciliter sa prise en main (C10). **Le Comité recommande de conditionner l'utilisation des systèmes d'IA générative à des études préalables de leur effet sur le développement cognitif des jeunes cerveaux** (G4).

La nécessité de marquer des limites par une régulation juridique

Le CNPEN et observe une certaine précipitation internationale pour introduire des mesures de régulation de l'IA générative et suit avec grande attention le débat législatif européen. Ils sont le signe de l'importance de l'enjeu économique et politique du développement de ces technologies. Les questions soulevées par la mise sur le marché des systèmes d'IA générative résident dans la nécessité de poser les limites via des normes juridiques suffisamment souples pour faire face aux nouvelles évolutions et suffisamment structurantes pour répondre au respect des droits fondamentaux et de l'intégrité des personnes. **Dans le cadre du *AI Act* européen, le CNPEN tend à considérer les modèles de fondation et les systèmes d'IA générative mis sur le marché comme des systèmes d'IA à haut risque (G7).**

Par ailleurs, il lui semble 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 d'interprétation de ce dernier dans le contexte du développement de l'IA générative en

Europe (G9). Quant à la question du traitement des données collectées, le CNPEN préconise l'élaboration de règles juridiques complétées d'un questionnement éthique sur la collecte, le stockage et la réutilisation des traces linguistiques des interactions entre machine et êtres humains (G8). Le comité plébiscite par ailleurs des recherches scientifiques et pluridisciplinaires sur l'adaptation du droit en matière de droit d'auteur, dans le cadre de discussions entre Etats (G11).

La question des responsabilités individuelle et collective est cruciale pour prévenir les abus et instaurer la confiance. Selon le Comité, 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. La responsabilité morale s'étend aux concepteurs des modèles de fondation et aux développeurs des systèmes d'IA générative utilisant ces modèles (G8).

De la conscience dans l'impact écologique, de la confiance dans une gouvernance souveraine

Enfin, il apparait crucial de pouvoir mesurer le coût énergétiques des systèmes d'IA générative et des modèles de fondation pour les inscrire dans la transition écologique. Les contraintes actuelles semblent difficiles à respecter pour la plupart des modèles. Les systèmes d'IA générative doivent se montrer plus conscients et transparents à propos de leur utilisation énergétique, de leurs émissions et des mesures prises pour atténuer ces dernières. Afin d'envisager un développement vertueux de ces technologies, les CNPEN propose la mise en place d'une métrique de l'impact environnemental des systèmes d'IA générative (G12).

Pour conclure, Le CNPEN juge que le développement des systèmes d'IA générative, leurs applications et impacts doivent faire l'objet d'une attention spécifique en termes de gouvernance. Le Comité préconise la création d'une entité souveraine de recherche et de formation « IA, science et société » (G1). Il recommande également la prudence dans la vitesse d'adoption de ces systèmes et la mise en place d'évaluations par les acteurs économiques et les autorités publiques (G2).

Les travaux du CNPEN font état des réflexions éthiques menées par les experts consultés sur la base des connaissances scientifiques actuelles. D'autres sujets viendront émailler le débat mondial autour de la conception, des usages et des impacts de l'intelligence artificielle dans un avenir très proche.

-