Cultur'IA

Pour une intelligence de confiance au service de la sécurité

n°18

Janvier - Février 2024

2024

MEILLEURS VOEUX

Gendarmerie nationale

Cap IA

Edito

2023 se termine avec le sentiment d'avoir vécu un bouleversement marquant par l'invasion des IA génératives de type "transformers". Des Big Tech aux institutions publiques, tout le monde s'y est mis au point d'occulter toute innovation dans le domaine. 2024 devrait voir se prolonger l'engouement pour les larges modèles de langage mais aussi pour des solutions plus frugales en énergie et plus rentables économiquement. 2024 sera aussi l'année d'une nouvelle appréhension du champ cyber en sécurité intérieure par la création du Comcyber-MI au sein duquel l'IA constituera un enjeu stratégique majeur. Alors, ouvrons cette nouvelle année sans plus attendre avec une scientifique de renom, une entrepreneure audacieuse et une auteure prolixe, qui nous a fait le plaisir de répondre à nos questions : Aurélie Jean.

Cultur'IA vous adresse ses meilleurs voeux pour 2024

Général Patrick Perrot
Coordonnateur pour l'intelligence artificielle
Conseiller IA auprès du ComCyber-MI
Gendarmerie nationale

Aurélie JEAN, docteure en sciences, entrepreneure, autrice, spécialiste en modélisation algorithmique.

L'intelligence artificielle est-elle une discipline qui peut bouleverser les équilibres de souveraineté entre les Etats et les Big Tech?

Les Big Tech possèdent un chiffre d'affaires de l'ordre d'un budget étatique On comprend alors le pouvoir économique mais aussi politique de ces géants. Le professeur Mark Coeckelbergh étudie entre autres les mécanismes de ses pouvoirs politiques obtenus par la technologie. Cela étant dit, il faut comprendre les relations entre ces Big Tech, ainsi que les relations qu'elles entretiennent avec les États. La souveraineté technologique d'un pays passant par sa capacité à concevoir, créer et innover, mais aussi à travers les rapports qu'il entretient avec les grandes sociétés technologiques qu'il a fait naître. Rappelons que des sociétés comme Google ou SpaceX ont vu le jour grâce entre autres à la DARPA qui est une agence de financement et de soutien de la recherche et le développement public et privé pour entre autres maintenir la souveraineté des États-Unis.

Faut-il avoir peur de l'intelligence artificielle ?

Il ne faut pas avoir peur de l'intelligence artificielle, il faut au contraire la dompter en en comprenant les mécanismes de base, et en maîtrisant les principaux concepts même dans les grandes lignes. Il faut craindre et interroger ceux qui les possèdent, qui les conçoivent, qui les vendent et qui les utilisent. Plus généralement, quand on parle d'intelligence artificielle, c'est important de distinguer la technologie de l'application au risque de rejeter une technologie pour la simple raison qu'elle est utilisée au sein d'une application controversée voire inacceptable.

Comment l'intelligence artificielle peut-elle aider à une meilleure protection la société des citoyens ?

L'intelligence artificielle est déjà utilisée par la Défense ou la gendarmerie par exemple, que ce soit pour comprendre des phénomènes afin d'en prévenir l'apparition ou encore de prédire un évènement pour l'appréhender. Dans la Défense, l'usage de drones dits "intelligents", de systèmes de recherche, d'analyse de comportements suspects sur vidéo ou encore d'aide à la logistique et à la planification existent. Sur ce sujet de la protection des citoyens, il faut prêter une attention particulière à la frontière fine qui existe entre protection et surveillance de masse des citoyens. Comme je le dis souvent, si on mettait des caméras dans tous les foyers on mettrait sans aucun doute fin aux violences conjugales, pour autant il ne faudrait pas le faire car nous violerions le droit à une vie privée.

Comment démystifier l'intelligence artificielle pour la rendre à la fois accessible et compréhensible au plus grand nombre ?

Il faut en effet sortir des nombreux mythes qui abiment la science algorithmique et ses acteurs. J'ai un conseil simple mais efficace, ne jamais utiliser un terme qu'on ne sait pas expliquer clairement. On tend à croire qu'on comprend un mot ou un concept car on l'entend, on le lit et le dit souvent. L'écrasante majorité des individus utilisent le mot algorithme sans en donner une stricte et juste définition, c'est ainsi qu'ils font des raccourcis concernant leur soi-disante responsabilité dans tous les maux de notre temps, qui, lorsqu'on comprend leur fonctionnement, n'a plus aucun sens. Un autre conseil, est d'utiliser davantage l'apprentissage intergénérationnel, qui présente de nombreux bénéfices sur une science comme l'IA, que les jeunes manipulent avec aisance et dont ils peuvent enseigner les fonctionnalités aux plus anciens. En retour les plus anciens qui s'interrogent davantage (ayant connu une autre époque) chercherons des réponses avec les plus jeunes qui (vous l'avez compris) ne s'interrogent pas assez étant nés avec ces technologies.

L'être humain peut-il être asservi aux algorithmes ou à ceux qui les conçoivent ?

Le risque est bien réel à travers l'affaiblissement de notre libre arbitre au regard des technologies algorithmisées que nous utilisons aujourd'hui pour s'informer et communiquer. Cela étant dit, si asservissement il y a, seuls ceux qui possèdent, conçoivent, et utilisent ces technologies sont responsables. Même si une technologie est bien conçue, elle peut être utilisée à mauvais escient. Encore une fois, comprendre ce qu'est l'IA, ce qu'elle n'est pas, ce qu'elle sera et ce qu'elle ne sera jamais, permettra à chacun de s'interroger, de faire l'effort de comprendre afin de défier ceux qui sont à l'origine de leurs mauvais emplois.

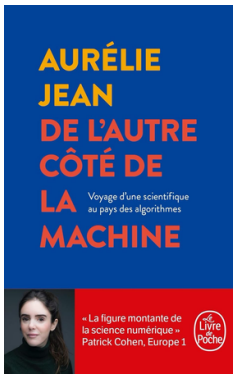
Qui est Aurélie Jean ?

Docteure en sciences, entrepreneuse et auteure, Aurélie Jean participe aujourd'hui à rendre accessible à tous la compréhension de l'IA en en présentant les opportunités et en proposant une vision réaliste et plutôt positive de ce qu'est réellement l' IA. Elle est à l'origine de différents ouvrages qui replacent l'IA à la fois dans sa complexité scientifique et son acceptabilité sociale. Aurélie est aussi éditorialiste scientifique pour Le Figaro, Le Point, Les Echos ou Elle Internationale.

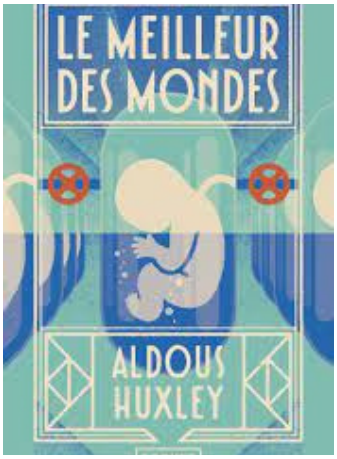
De formation universitaire, titulaire d'une thèse en science et génie des matériaux, relative à l'étude d'un élastomère chargé, de sa nanostructure à son comportement macroscopique, elle s'engage en postdoc au sein de l'université de Pensylvanie et du renommé Massachussetts Institute of Technology (MIT) dans le domaine médical. Ses travaux portent alors sur l'élaboration d'un tissu cardiaque plus flexible afin de réparer les tissus malades du myocarde après un infarctus. À l'issue de ses travaux universitaires, elle entre chez Bloomberg puis est nommée consultante senior par le Boston Consulting Group en 2018. Aujourd'hui, elle partage sa vie entre entre les États-Unis et la France, entre le conseil, la recherche et l'enseignement dans le cadre de la formation des cadres, principalement au Sloan MIT. Sa polyvalence l'a conduit également à être conseillère technique et scientifique pour plusieurs entreprises, et mentor au Frontier Development Lab de la NASA. Elle est enfin entrepreneuse et Chief AI Officer, co-fondatrice d'une startup AI deep tech dans la médecine de précision et prédictive appliquée au cancer du sein. Boulimique ? Non, plutôt passionnée et investie.

Particulièrement impliquée pour rendre accessible l'IA au plus grand nombre, elle est auteure d'ouvrages qui détaillent et expliquent de manière très pédagogique les fondements théoriques et algorithmiques de l'IA mais aussi les applications très concrètes et opérationnelles qui émanent de travaux de recherche. Et, elle a même co-écrit un roman d'anticipation avec Amanda Sthers.

Bibliographie d'Aurélie JEAN: pour aller plus loin



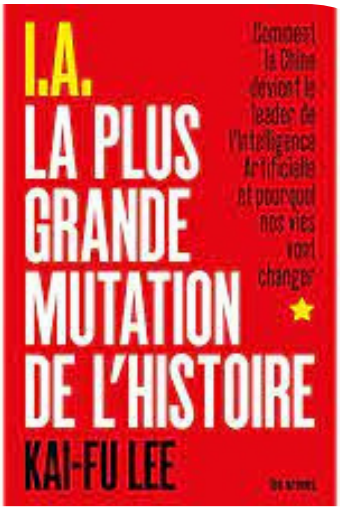
À l'image de l'être artificiel, les structures machiniques de la dystopie acquièrent suffisamment d'autonomie pour pouvoir, elles aussi, attenter à la souveraineté de l'individu. L'humain, noyé au sein de ces structures qui le submergent, n'est alors plus qu'une composante insignifiante dont la spécificité se fait systématiquement gommer par la machine sociale. En définitive, qu'il s'agisse d'un être ou d'un système, la figure de l'intelligence artificielle apparaît comme un sujet agissant, qui incarne la propension de la dystopie à réduire l'individu à l'état de pion.



Aldous Huxley s'adonne au genre littéraire de la dystopie en 1932 et révèle les problématiques sociétales liés à la suprématie de la science et du progrès. Un classique du genre!

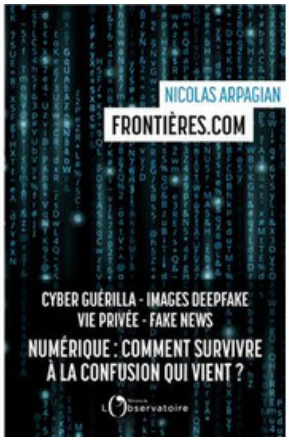
Un podcast qui nous plonge dans la France de 2050 & qui reprend, à travers un scénario bien ficelé, les différentes questions et craintes que nous aurions sur l'IA et la technologie en général. Un format différent qui propose un autre regard sur la puissance technologique!

PODCAST & LITTERATURE



Une course technologique (ou guerre froide 2.0) serait lancée entre les USA & la Chine. L'occasion de se plonger dans cette lecture afin de comprendre les projets chinois en matière d'IA. Une lecture utile pour éviter de tomber dans le prêt à mâcher! Une vision et une appropriation de l'IA différente et...instructive!

Plongez dans l'univers de Julien, professeur de piano qui, le temps d'un été, décide de passer ses vacances dans le... métavers. Nathan Devers, philosophe, utilise le genre romanesque pour nous proposer une réflexion sur les liens entre notre vie réelle et les projections que nous faisons dans l'espace numérique.
Un plongeon dans nos vies connectées absolument passionnant et riche en questionnements sur ce que nous souhaitons
pour l'avenir avec nos technologies !
À mettre entre toutes les mains!



Nicolas Arpagian propose un voyage dans l'espace numérique sans frontière et dominé par les entreprises privées. Une bataille 2.0 s'y livre, entre souveraineté des États, pouvoirs des sociétés civiles et puissances des BigTech. Il met en perspective le rôle et la place de chaque acteur dans l'écosystème numérique.
Ça se lit comme un roman et s'analyse comme un bon documentaire.
Une lecture nécessaire pour une première plongée dans l'océan bleu.

2024 est là, regardons un instant dans le rétroviseur. 2023 restera comme une année exceptionnelle en IA par l'émergence de l'usage des **G**enerative **P**re-trained **T**ransformers. Ils se sont imposés et ont été largement adoptés par les citoyens entrainant une nouvelle orientations des Big Tech. Revenons sur quelques évènements de l'année.

JANVIER

- **Microsoft investit près de 10 milliards de dollars dans OpenAI.**

FEVRIER

- **Présentation de BARD par Google**
- **Chat-GPT intègre BING (Microsoft)**

MARS

- **Sortie de GPT-4**
- **Appel des BIG Tech pour une pause en IA**
- **L'Italie interdit Chat-GPT**

MAI

- **Un moteur de recherche IA pour Google**
- **1000 milliards de valorisation pour NVDIA**

AVRIL

- **Des chercherus proposent un projet d'IA sans précédent sur le changement climatique**
- **la NASA annonce que son drone (piloté par une IA) a effectué son 50ème vol sur Mars**

JUIN

- **Mistral AI lève 105 millions d'euros.**

JUILLET

- **Elon Musk lance xAI, sa startup d'intelligence artificielle**
- **Grève à Hollywood contre les robots-écrivains**
- **Publication de LLAMA2 en open source**
- **Lancement par la Chine du 1er satellite d'observation équipé d'une IA**

AOUT

- Hugging Face lève 235 millions de dollars.
- Elon Musk lance son IA générative Grok (pause de 6 mois !!!)

SEPTEMBRE

- Onclusive remplace 217 employés français par l'IA.
- Le gouvernement français se dote d'un comité interministériel sur l'IA.
- Amazon investit quatre milliards de dollars dans Anthropic
- Instagram, Messenger et WhatsApp reçoivent des assistants d'IA.

NOVEMBRE

- Xavier Niel (Iliad), Rodolphe Saadé (CMA CGM) et Eric Schmidt créent Kyutai, un OpenAI à la Française
- début du feuilleton Sam Altman chez OpenAI

DECEMBRE

- Google dégaine Gemini, son IA tant attendue.
- Publication d'un accord de l'UE sur l'AI Act
- Mistral lève 385 millions d'euros et devient un géant de l'IA.
- Apple sort IA Ferret, son Nouveau Modèle Open Source

COMCYBER-MI

L’intelligence artificielle en cyber sécurité

Le 23 novembre 2023, par décret publié au Journal officiel (J.O.) entré en vigueur le 1er décembre, le Commandement du ministère de l'Intérieur dans le cyberespace, placé sous l'autorité du Directeur général de la gendarmerie nationale (DGGN) a été créé. Ce nouvel organe de lutte contre la cybercriminalité est articulé autour de trois axes majeurs:

- l'élaboration de la stratégie de lutte contre la cybercriminalité
- la police judiciaire du haut du spectre à partir de compétences rares
- la formation à destination du personnel du ministère.

L'intelligence artificielle s'inscrit naturellement dans la stratégie du Comcyber-MI par sa capacité à évaluer et contrer les menaces. Il est essentiel aujourd'hui d'exploiter tout le potentiel de l'IA au profit de la protection du citoyen dans l'espace cyber comme la Gendarmerie le fait dans l'espace physique. L'intelligence artificielle permet en effet d'être bien plus efficace que l'être humain dans la célérité à détecter les attaques comme les vulnérabilités des systèmes d'information.

Intéressons nous tout d'abord aux potentialités criminelles de l'IA dans le cyber espace.

Il est difficile de rapporter de l'actualité des éléments concrets d'attaques majeurs ayant fait essentiellement appel à l'IA. Pour autant, la menace est là, notamment autour des opportunités suivantes:

- **Génération de “deepfakes”:** depuis 2014 et l'apparition des réseaux génératifs adverses, il est possible de produire des fausses vidéo, de faux audios ou de faux textes et même de combiner l'ensemble. Dès lors, les applications en pédo pornographie, en usurpation d'identité numérique en développement de théories complotistes sont facilement mise en oeuvre par une criminalité organisée comme une délinquance d'opportunité.

- **Génération de logiciels et codes malveillants:** l'IA générative version “Transformers” est en mesure de proposer des codes et créer des logiciels malveillants et cela bien entendu sans que l'utilisateur n'ait à maitriser un quelconque langage de programmation. La diffusion de ces logiciels ou codes malveillants peut ensuite s'effectuer de manière ciblée ou totalement hasardeuse en fonction de l'objectif des attaquants.

- **Le “phishing automatisé”:** automatiser la production de campagnes de phishing performantes est parfaitement accessible par l'utilisation de l'IA. Aujourd'hui grâce à la disponibilité des données, il est aisé aux attaquants même très amateurs de développer des phishing ciblés et d'en automatiser la diffusion.

COMCYBER-MI

- **Réseaux de zombies (botnets)** - L'IA est également une solution pour analyser le comportement d'un réseau et reconfigurer les schémas d'attaque afin de s'adapter aux cyberdéfenses.

Face à ces différentes possibilités d'attaques malheureusement non exhaustives, l'IA constitue également le remède dans bien des cas. Elle est en effet en mesure de proposer une protection dans les champs suivants:

- **détection des deepfakes:** les réseaux génératifs adverses à l'origine des deepfakes sont aussi l'arme de détection de ces impostures. La Gendarmerie a développé des travaux en ce sens.

- **détection des logiciels malveillants**: L'IA peut analyser différentes caractéristiques des fichiers, le comportement du système ou encore la structure du code pour évaluer si un nouveau fichier introduit présente une menace. Elle peut également modéliser les logiciels malveillants à partir de leur signature et ainsi être en mesure de proposer une classification.

- **détection des activités malveillantes:** Par sa capacité à analyser les réseaux en temps réel, l'IA est en mesure d'alerter sur des situations qui peuvent être considérée comme inhabituelle voire dangereuse. Cette aspect de l'IA d'analyse du comportement des entités comme des utilisateurs est connu sous l'acronyme UBEA (User and Entity Behavior Analytics). Cette technique permet une veille continue en temps réel des systèmes.

- **Gestion et priorisation des menaces :** L'IA permet de déterminer l'occurence d'une menace en baissant le niveau de faux positifs par la corrélation de données provenant de sources multiples. Ce travail en optimisant le point de fonctionnement de la remontée d'alertes permet de prioriser les menaces les plus prégnantes afin de les traiter dans les meilleurs délais.

- **Recherche de vulnérabilités:** L'IA peut automatiser le processus de recherche de vulnérabilités et de menaces potentielles. Par un apprentissage automatique, il est possible de surveiller en permanence le trafic réseau, et d'appliquer des règles et des décisions de cybersécurité pour s'assurer que les menaces sont détectées et résolues avant qu'elles ne causent des problèmes.

Comme dans l'espace physique, l'IA est à la fois le remède et le poison dans l'espace cyber. Ne pas utiliser l'IA à des fins de protection serait une erreur stratégique majeure.

Que retenir de la réglementation européenne en IA ?

Après quelques années d'élaboration et de débats, les institutions européennes ont réussi **en décembre 2023 à trouver un accord politique autour d'une première réglementation sur l'intelligence artificielle**. Cet accord est consécutif à un trilogue réunissant la Commission, le Conseil et le parlement européen. **Un accord politique ne signifie pas que le texte est définitif**. Des travaux d'optimisation de rédaction vont encore se dérouler mais les mesures principales sont en place. Le texte deviendra une loi après une adoption formelle du Parlement et du Conseil durant l'année 2024. Il intègre même une partie relative aux modèles de fondation à la base des IA générative même si ceux-ci ne constituent pas un usage de l'IA mais une technique de modélisation. Nous pouvons d'ors et déjà nous interroger sur l'évolution de la réglementation à la suite de nouvelles avancées scientifique !

La réglementation proposée est élaborée à partir des usages des systèmes d'IA auxquels est associé un niveau de risque, allant du risque limité aux usages interdits. **Plus le risque est élevé, plus les règles sont strictes.**

Les systèmes d'IA classés comme présentant un risque élevé (en raison des dommages potentiels importants qu'ils peuvent causer à la santé, à la sécurité, aux droits fondamentaux, à l'environnement, à la démocratie et à l'État de droit), sont soumis à des obligations strictes qui peuvent aller théoriquement jusqu'à en dissuader l'utilisation. Les obligations sont par exemple : les évaluations de l'impact sur les droits fondamentaux, les évaluations de conformité, les exigences en matière de gouvernance des données, les systèmes de gestion des risques et de gestion de la qualité, la transparence, la surveillance humaine, l'exactitude, la robustesse, la formation et la cybersécurité. Parmi les exemples de ces systèmes figurent certains dispositifs médicaux, les outils de recrutement, de gestion des ressources humaines et des travailleurs, ainsi que la gestion des infrastructures critiques (par exemple, l'eau, le gaz, l'électricité, etc.) ou encore nombre de systèmes utilisés en sécurité intérieure.

Les usages interdits sont les suivants:

- les systèmes de catégorisation biométrique qui utilisent des caractéristiques sensibles (par exemple, les convictions politiques, religieuses et philosophiques, l'orientation sexuelle, la race) ;
- l'extraction non ciblée d'images faciales d'Internet ou d'images de vidéosurveillance pour créer des bases de données de reconnaissance faciale ;

- la reconnaissance des émotions sur le lieu de travail et dans les établissements d'enseignement ;
- la notation sociale basée sur le comportement social ou les caractéristiques personnelles ;
- les systèmes d'IA qui manipulent le comportement humain pour contourner son libre arbitre ;
- l'IA utilisée pour exploiter les vulnérabilités des personnes (en raison de leur âge, de leur handicap, de leur situation sociale ou économique).

L'utilisation de systèmes d'IA par les services répressifs à des fins institutionnelles sera soumise à des garanties/contraintes spécifiques.

La réglementation va également générer la mise en place de nouvelles structures administratives comme:

- Un bureau de l'IA au sein de la Commission: l'objectif annoncé est de superviser les modèles d'IA les plus avancés, de contribuer à la promotion de nouvelles normes et pratiques de test, et de faire respecter les règles communes dans tous les États membres de l'UE. Il est probable qu'il devienne l'équivalent des instituts de sécurité de l'IA dont la création a été récemment annoncée au Royaume-Uni et aux États-Unis ;
- Un groupe scientifique d'experts indépendants, qui conseillera le bureau de l'IA sur les modèles d'IA générative générale à fort impact, contribuera au développement de méthodologies d'évaluation des capacités des modèles de fondations et surveillera les éventuels risques de sécurité matérielle liés aux modèles de fondations ;
- Un conseil de l'IA, composé de représentants des États membres de l'UE, comme organe consultatif de la Commission, tout en contribuant à la mise en œuvre de la loi sur l'IA (par exemple, en concevant des codes de pratique)
- Un forum consultatif pour les parties prenantes sera mis en place pour fournir une expertise technique à la Commission de l'IA.

Parmi les éléments à venir est **la déclinaison de l'AI Act au niveau national**. En effet, chaque Etat membre adoptera une approche spécifique pour mettre en place les structures locales en charge de la mise en application de cette réglementation. L'enjeu est aujourd'hui de savoir si le DPO en charge des données pourrait être l'acteur au sein des entreprises en charge de l'IA. les choix sont ouverts tant la dynamique de l'IA est différente de l'immobilisme de la donnée. Aujourd'hui, **différentes institutions tentent de se positionner sur ce créneaux car c'est un enjeu de pouvoir mais aussi économique pour certains.**

Enfin, la réglementation prévoit des sanctions pour le non-respect des règles. Ces pénalités seront des amendes allant de 7,5 millions d'euros ou 1,5 % du chiffre d'affaires mondial à 35 millions d'euros ou 7 % du chiffre d'affaires mondial, en fonction de l'infraction et de la taille de l'entreprise.

IA & Impact environnemental: l'avenir des"Smart langage Models"

Les larges modèles de langage (LLM) ont envahi la planète IA pour être l'évènement majeur de 2023 et l'être encore, sans nul doute, en 2024. Ainsi, il est temps de s'interroger sur l'impact environnemental de cette course a toujours plus de données. En effet, la quantité de donnée ne doit pas être l'unique solution à l'amélioration de la performance. Nous sommes convaincus qu'il existe d'autres possibilités plus frugales en énergie. Les SML (Small Langage Model) seront certainement une tendance pour l'année 2024. Ainsi, alors que les LLM sont en plein essor,

Les LLM fonctionnent à partir de grands ensembles de données textuelles pour proposer une génération de texte, de codes, de synthèse de documents ou encore de traductions. Ces LLM semblent constituer la solution à tous les problèmes mais est-il véritablement utile de mobiliser autant de données pour toutes les applications ? Les SML ont démontré une capacité à satisfaire aux même enjeux que les LLM mais en exploitant des corpus beaucoup moins importants. Qu'est ce qui distingue alors les LLM des SML ? Une lettre, me direz-vous ?

Il n'existe pas de limite précise entre SLM et LLM mais en général, des modèles de moins de 100 millions de paramètres sont considérés comme petits. Les LLM atteignent, eux, plusieurs milliards de paramètres que ce soit GPT-3, BARD ou LLAMA. Il est indéniable que la réduction de paramètres réduit considérablement l'impact environnemental mais pour être efficace, il faut aussi que le gain des SML s'illustre au niveau économique et de la performance. Ce dernier point est essentiel car la capacité des modèles de langage est liée à leur taille et donc au nombre de corrélations possibles. L'emploi des SLM doit alors se concevoir à partir d'un compromis entre l'impact environnemental, la flexibilité, la personnalisation et la performance.

Quels sont les atouts des SLM ?

1. <u>Complexité</u>

Le gain en terme de célérité de calcul est évident du fait du nombre réduit de paramètres. **Les SML sont supérieurs aux LLM en ce qui concerne la vitesse d'inférence et de débit avec un impact direct sur la complexité calculatoire**.

2. <u>Coût</u>

Les LLM mobilisent des capacités de calcul considérables et particulièrement couteuses A titre d'illustration, GPT-3 a nécessité des dizaines de millions de dollars en coût de matériel et d'ingénierie. La rentabilité d'un LLM n'est donc pas immédiate. L'adoption de SLM permet aux entreprises comme aux institutions un déploiement de solutions opérationnelles à moindre coût à partir de ressources informatiques parfois même existantes. **L'enjeu est alors de bien cibler ses applications et son seuil de rentabilité.**

3. <u>Personnalisation</u>

Les SML sont aisément adaptables à différents types de corpus et présentent une capacité de personnalisation bien plus importante que celle des LLM. Avec des cycles d'itération plus rapides, les SML permettent d'expérimenter l'adaptation des modèles à des types de données specifiques et **répondent ainsi particulièrement bien à des applications très spécialisées**

4. Performance

La quantité de données dans la construction des modèles à partir des “Transformers“ est corrélée à la performance. Néanmoins, de récentes études (2023), ont montré que le niveau de performance augmentait significativement pour des tâches génératives de 60 millions de paramètres mais stagnait autour de 300 millions dès lors que les données étaient bien choisies. Au delà les performances croissent de manière marginale. Ainsi, **les modèles linguistiques de taille moyenne atteignent un niveau de compétence raisonnable dans de nombreuses applications** de traitement du langage, à condition qu'ils soient exposés à un nombre suffisant de données d'entraînement adéquates.

Les petits modèles de langage devraient en cette année 2024 connaître un essor considérable pour deux raisons principales:

- **leur impact environnemental**
- **leur capacité à être embarquée notamment dans des smartphones.**

L'engouement marketing pourrait alors profiter à l'environnement !

SLM (Small Langage Model) vs SLM (Spécialized langage Model)

Parce que l'acronyme SLM, des petits modèles de langage est le même que celui des modèles de langage spécialisé, ces deux notions sont souvent confondues. Arrêtons nous un instant pour y voir plus clair.

L'objectif des modèles spécialisés est de satisfaire à une tâche particulière et parfaitement ciblée avec la recherche d'un très haut niveau de performance comme dans le domaine du diagnostic médical.

Les petits modèles de langage sont d'abord caractérisés par leur faible nombre de paramètres, ceux-ci pouvant tout de même aller jusqu'à plusieurs millions. Ils peuvent se définir par:

- le nombre de paramètres
- la taille de leur empreinte
- la quantité de données nécessaire à leur apprentissage

alors que les modèles spécialisés se caractérisent plutôt par la tâche particulière à accomplir indépendamment du nombre de paramètres.

Ainsi les petits modèles de langage ne sont pas tous spécialisés même si certains peuvent l'être et les modèles spécialisés peuvent exploiter de large quantité de données.

Mais qu'est ce donc qu'une hallucination ?

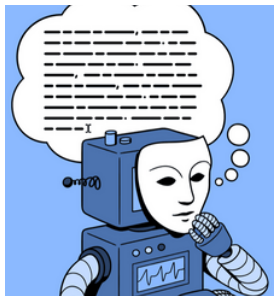
Le terme d'hallucination a pris, depuis l'émergence des modèles génératifs à base de LLM (Large langage Model) un sens particulier. Les LLM sont particulièrement performants pour de nombreuses applications. Ils peuvent résumer, une grande quantité d'information, trouver et extraire une information utile dans une grande quantité de texte mais aussi produire et corriger du code. Les LLM peuvent encore planifier des vacances, des projets ou des programmes technologiques d'envergure.

Mais **toutes ces fonctionnalités courent le risque d'être entachées par des hallucinations.** Essayons de mieux comprendre ce terme arrivé dans le jargon de l'IA générative?

L'hallucination est un résultat proposé par une IA qui est soit absurde, soit totalement erronée. Interrogez une IA générative telles que celles aujourd'hui en vogue sur votre propre biographie et vous pourriez être témoin d'hallucinations par la création de parcours de carrière comme de diplôme relevant de la pure invention. Des cas retentissants ont même fait la une de la presse comme celui de cet avocat new-yorkais qui s'est appuyé sur une jurisprudence "générative" inventée de toute pièce pour défendre son client. Le juge fédéral s'est rapidement aperçu que les faits relatés étaient faux. Ces hallucinations qualifiées par les scientifiques de perroquets stochastiques ne font en réalité que construire des phrases sans en comprendre le sens malgré l'impression que cela donne.

Mais d'où viennent ces hallucinations ?

Il est difficile en réalité de savoir exactement comment sont formées les hallucinations même s'il n'est guère difficile d'en deviner les facteurs déclenchants.



Comme pour tout processus d'apprentissage, la donnée est essentielle et constitue la source de performance des LLM. Si cette donnée est erronnée ou en quantité insuffisante, le résultat produit par l'IA générative risque fort de proposer un résultat loin de toute vérité. La difficulté réside dans la forme de la réponse car celle-ci paraît tout à fait correcte et limpide avec une bonne formulation syntaxique et une tournure très affirmative.

En intelligence artificielle, il est un phénomène bien connu des scientifiques qui peut également être **à l'origine de certaines hallucinations: le sur-apprentissage.** Il s'agit de l'incapacité d'un système à généraliser un résultat au delà des données apprises. Lorsque l'apprentissage est bien réalisé, il existe une forme d'élasticité des modèles capables de classer ou prédire des données inconnues du système. Or quand le modèle est formé jusqu'à apprendre par coeur ses données d'apprentissage, c'est-à-dire à associer une entrée à une sortie, il n'est plus capable de bien analyser des données qu'il n'a jamais vu. Le niveau de performance chute alors considérablement et les réponses du système sont totalement fausses.

Une autre cause possible d'hallucination peut être liée à l'encodage entre les textes et les prompts. En effet, les LLM associent les termes à un ensemble de nombres - un processus connu sous le nom d'encodage vectoriel - et ces encodages présentent certains avantages clefs par rapport au travail direct à partir des mots.

L'intérêt est par exemple de revenir sur les ambiguïtés lorsqu'un mot possède plusieurs sens. Il existera un encodage pour chaque sens et pas seulement pour chaque mot. Nombreux sont les mots polysémiques comme "baguette" qui peut être un pain ou un objet magique ou encore "bouton", "sommet", "échelle"...

Les représentations vectorielles agissent comme des converstisseurs de sens en opérations mathématiques par la recherche de corrélations. Mais, cette conversion peut mal se passer et générer des hallucinations dans la phase d'encodage-décodage entre le texte et les représentations.

Une autre source d'hallucination se situe au niveau de la pertinence des données apprises. Si ces dernières ne sont pas exactes ou mal adaptées à l'enjeu, le système renvoie des résultats erronées. Ce risque doit être pris en compte lors de la constitution des modèles d'apprentissage comme d'adaptation. C'est un véritable sujet du fait notamment de la méconnaissance des sources d'information alimentant les LLM. Par exemple, produire des rapports juridiques dans le domaine financier nécessitent de travailler sur des données du même domaine. Des données juridiques trop généralistes entraineraient des erreurs hallucinantes sans nul doute.

Comment se prémunir des hallucinations ?

Face à ces diverses hallucinations, il existe des possibilités de prévention qui vont du bon sens à la mise en oeuvre de concepts mathématiques.

La première façon particulièrement simple est d'**exploiter les LLM sur des sujets que l'on connaît et maitrise**. Il s'agit d'être en mesure de disposer d'un regard critique sur la réponse proposée.

D'autres possibilités simples de prévenir des hallucinations est de **réaliser des prompts précis, donc de comprendre ce que l'on demande, d'éviter de mélanger des notions, d' utiliser du vocable scientifique à son juste sens, d' éviter de mélanger des réalités différentes** ou encore de veiller à ne pas mentir dans son prompt.

Parmi les méthodes scientifiques pour identifier la sortie hallucinée l'une consiste à **utiliser des textes synthétiques hallucinés pour créer un ensemble de données qui peut être utilisé comme base pour les futurs filtres** et mécanismes qui pourraient éventuellement devenir une partie essentielle des architectures de traitement du langage naturel..

Ainsi, supprimer les hallucinations relève à la fois du bons sens au niveau des utilisateurs et du développement d'approches scientifiques complexes au niveau de la recherche. c'est à n'en pas douter un enjeu majeur de 2024.

START'UP INFO

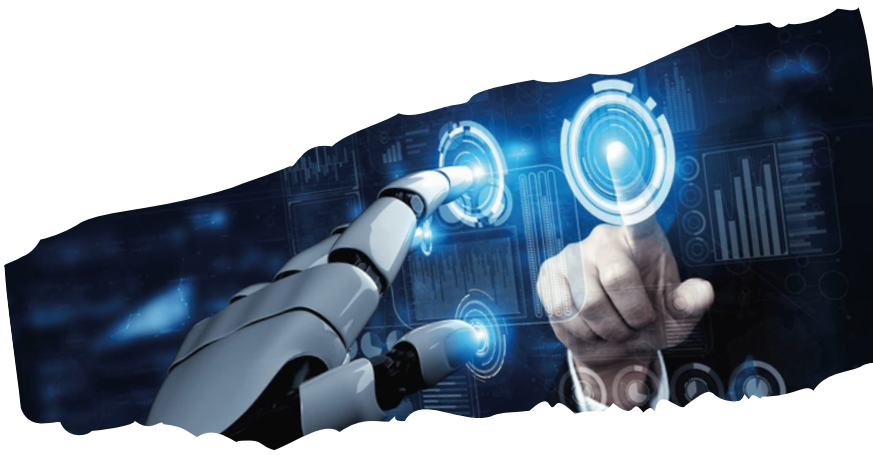
explore le champ cyber

La start up française Gatewatcher est un des leaders dans la détection des cybermenaces, Nombreuses sont les grands sociétés et les institutions qui font confiance à Gatewatcher pour protéger les réseaux et systèmes d'information. L'objectif est de détecter les intrusions et de répondre rapidement aux diverses formes d'attaque. Par l'IA, Gatewatcher propose une vision à 360° en temps réel des cybermenaces sur l'ensemble du réseau.

Anozr Way est une start-up bretonne, basée à Cesson-Sévigné (Ille-et-Vilaine), créée en 2019, dont l'objectif est la protection des personnes face aux risques cyber. Anozr Way propose deux logiciels complémentaires, l'un à destination des responsables de la sécurité pour suivre l'évolution de la menace, et l'autre pour les collaborateurs. L'ambition est d'identifier les personnes vulnérables dans les organisations.

Fondée en novembre 2017, GitGuardian, une start-up française développe une solution de détection de secrets inscrits en dur dans le code présent dans des dépôts Git. En 2021, Gitguardian a détecté plus de 6 millions de secrets, soit deux fois plus qu'en 2020. La startup cherche à avoir un haut taux de précision et un haut taux de rappel, c'est-à-dire un faible nombre de fausses alertes, et un faible nombre de secrets non identifiés.

En 5 ans, la start-up rouennaise "Yes we hack" est devenue leader européen de "bug bounty". Le "bug bounty", ou chasse aux bugs, consiste à mettre en relation une communauté de hackers éthiques avec des entreprises et des organisations pour tester leur système de sécurité. Les premiers hackers qui trouvent une faille perçoivent une prime définie par l'entreprise ou l'organisation. « La prime va de 50 € à 230 000 €, la moyenne est de 500 € », précise Guillaume Vassault-Houlière, le fondateur de Yes we hack.



Bienvenue

2024

Cultur'IA

Au sommaire du prochain numéro

- l'IA aux Jeux Olympiques
- les différentes formes d'apprentissage
- l'IA en cybersécurité: le hacking éthique
- Comment expliquer l'IA aux plus jeunes ?

Gendarmerie nationale

CEND 20.24 Cap IA