Quels régimes de régulation des données pour entraîner les intelligences artificielles ?

Synthèse de conférence

Conférence d'ouverture des Dauphine Digital Days,
20-22 novembre 2023

Université Paris Dauphine-PSL, 20 novembre 2023

DEBATE

Dauphine | PSL
CHAIRE GOUVERNANCE
ET RÉGULATION

Conférence organisée par la Chaire Gouvernance et Régulation

Conférence d'ouverture des Dauphine Digital Days organisés par l'Université Paris Dauphine-PSL

Le 20 novembre 2023

Dauphine | PSL
CHAIRE GOUVERNANCE ET RÉGULATION

Synthèse n°85
Université Paris Dauphine-PSL

Quels régimes de régulation des données pour entraîner les intelligences artificielles ?

Ouverture

El Mouhoub Mouhoud | Président de l'Université Paris Dauphine-PSL

Isabelle Ryl | INRIA, Directrice de l'institut PR[AI]RIE et membre du Comité de l'intelligence artificielle

Intervenants

Guillaume Avrin | DGE

Adrien Basdevant | Conseil national du numérique (CNNum)

Anne Bouverot | Co-présidente de la Commission IA

Pierre-Carl Langlais | OPSCI

Bertrand Pailhès | CNIL

Karine Perset | OCDE

Giada Pistilli | Hugging Face et Sorbonne Université

Isabelle Ryl | INRIA, PR[AI]RIE

Benoît Sagot | INRIA, PR[AI]RIE

Modérateurs

Éric Brousseau | Chaire Gouvernance et Régulation, Université Paris Dauphine - PSL

Joëlle Toledano | Chaire Gouvernance et Régulation, Université Paris Dauphine – PSL, CNNum

Sommaire

IA ET SOCIÉTÉ : NOUVELLE DONNE, NOUVEAUX ENJEUX

Ouverture

El Mouhoub Mouhoud | Président de l'Université Paris Dauphine – PSL

L'irruption de l'IA change la donne à bien des égards et renouvelle de façon inédite les enjeux de société.

Promouvant un regard complexe sur des sujets qui le sont tout autant, les Dauphine Digital Days constituent un forum ouvert, dans lequel le dialogue sciences/société et le partage de connaissances trouvent pleinement leur place.

L'Université Paris Dauphine – PSL est pleinement engagée dans le domaine de l'IA et de ses interactions avec les sciences de l'homme et la société, au travers de son programme « Dauphine numérique » dont la stratégie de bi-disciplinarité et de bi-compétences de la licence au doctorat va dans le sens d'une division cognitive du travail et de la mise en complémentarité de différents blocs de compétences. En outre, grâce à la mise en place d'un cercle Numérique, « Dauphine numérique » accueille des membres du monde socioéconomique. Par ailleurs, la suite de l'institut PR[AI]RIE verra l'avènement de la Paris School of Artificial Intelligence. Enfin, une offre nouvelle de doubles masters sera mise sur les rails en partenariat avec l'ENS-PSL et d'autres initiatives seront annoncées prochainement.

Isabelle Ryl | INRIA, Directrice de l'institut PR[AI]RIE et membre du Comité de l'intelligence artificielle

Fondé il y a quatre ans dans le cadre de la stratégie nationale en IA, PR[AI]RIE est un institut interdisciplinaire d'intelligence artificielle (3IA). Les chaires de recherche qui le composent couvrent un grand nombre de volets de l'IA, jusqu'à la frontière des disciplines scientifiques (biologie, santé). Grâce à la dynamique engagée par l'institut, plusieurs programmes de formation ont été créés et des initiatives touchant l'IA dans des disciplines non informatiques ont vu le jour.

Pour la suite, dans le cadre de l'appel à manifestation d'intérêt « IA Cluster » lancé par l'État, il est envisagé de regrouper l'ensemble de ces actions et démarches au sein de la PRAIRIE – PSAI (Paris School of Artificial Intelligence). Ce nouveau projet s'inscrit dans la continuité de l'ancien tout en l'élargissant, dans son emprise d'action comme dans son emprise disciplinaire.

QUEL RÉGIME DE RÉGULATION DES DONNÉES POUR ENTRAÎNER LES IA ?

Introduction

Benoît Sagot | INRIA, PRAIRIE

Des modèles toujours plus grands et mobilisant un nombre croissant de paramètres et des corpus d'entraînement eux-mêmes toujours plus grands expliquent la rapidité des progrès des modèles de langue, et plus généralement de l'IA.

S'agissant des corpus, deux utilisations peuvent être distinguées. Les corpus de pré-entraînement, qui sont de grands corpus bruts classés ou non par langue, permettent d'entraîner des modèles dits de fondation, tandis que les corpus d'affinage (ou *fine-tuning*) incluent les données permettant de construire un modèle conversationnel à partir d'un modèle de fondation, comme ChatGPT. Ces corpus d'affinage sont souvent conçus avec l'aide d'annotateurs humains et soulèvent des questions éthiques importantes.

Je limiterai mon exposé aux corpus de pré-entraînement et à leurs très grands volumes de textes.

Plusieurs grands modèles de langue ont très vite été distribués sous des licences libres. C'est le cas de modèles par masquage comme BERT, avec les modèles pour le français comme CamemBERT et CamemBERTa, ainsi que des modèles génératifs comme LLaMA (1 et 2) publié par Meta, ou BLOOM développé par un consortium dirigé par Hugging Face. Cependant, depuis GPT-3, les plus gros modèles entraînés par des acteurs privés (dont Falcon, entraîné par des Français) sont rarement librement disponibles. Très souvent, les données d'entraînement ne sont ni connues ni décrites.

Parmi les grands corpus bruts primaires librement téléchargeables, peuvent être cités les Wikipédia dans toutes sortes de langues, précises, de bonne qualité, très homogènes linguistiquement et stylistiquement. Mais si ce corpus semblait encore récemment immense, il paraît aujourd'hui tout petit. Par ailleurs, un certain nombre de corpus ont été extraits à partir de *dumps* de Common Crawl et ses grandes collections de pages web. D'autres acteurs ont effectué des travaux similaires, pour les corpus CC100 ou le BookCorpus et sa grande collection de livres, conçue dès 2015 à partir d'une plateforme de publications plus ou moins librement disponibles. À l'exception de ce dernier, tous les corpus sont multilingues.

Émergent aussi de plus en plus de corpus agrégés, comme BLOOM qui a été entraîné à la fois à partir du corpus OSCAR développé à l'INRIA et d'autres données collectées. Certains d'entre eux sont librement téléchargeables, comme RedPajama, dont la première version a cherché à imiter les entraînements de LLaMA-1. Citons également ROOTS, le corpus d'entraînement de BLOOM, mais aussi The Pile ou RefinedWeb, utilisé pour entraîner le modèle Falcon, et plus récemment Glot500-c, fortement multilingue et cherchant à couvrir des langues moins communes. Ces corpus soulèvent un grand nombre d'enjeux et de problématiques comme les biais de représentation, les biais de représentativité, les contenus offensants, les fausses informations, les informations anachroniques, sans oublier les enjeux écologiques liés à leur très grande taille. Sur le plan juridique, il faut aussi évoquer la présence de données personnelles, de contenus illicites et de données protégées par le droit d'auteur.

Par ailleurs, les sources de données d'entraînement peuvent être appréhendées sous deux angles: celui de la légalité de l'accès gratuit à un jeu de données (les modèles actuels, dont ChatGPT, ont été entraînés sur des données qui n'étaient pas toutes légalement accessibles) et celui de légalité de l'utilisation de données accessibles librement pour faire de l'entraînement.

À ce stade, je propose la tripartition suivante, qui mérite certes d'être affinée :

- les « données blanches », dont l'accès est légal et dont l'utilisation pour entraîner des modèles l'est aussi (Wikipédia, transcription des minutes au Parlement européen…) ;

- les « données grises », dont l'accès est légal et dont l'utilisation pour entraîner des modèles est soit discutable, soit non autorisée (articles de presse accessibles gratuitement, mais protégés par le droit d'auteur, par exemple) ;

- les « données noires », dont l'accès est illégal. Celles-ci incluent la Library Genesis qui permet de télécharger illégalement des milliers de livres. Or, nous avons de bonnes raisons de penser que de très grands modèles de langue ont été entraînés sur des jeux de données issus de ce type de source.

Certains grands modèles de langue ont été entraînés uniquement sur des données blanches et grises, comme OSCAR 2019 pour CamemBERT et ROOTS pour BLOOM. Cependant, la présence de données noires chez d'autres explique pourquoi un certain nombre d'acteurs refusent de publier leurs données d'entraînement. Ainsi, si LLaMA-1 fournit une description relativement précise de son corpus, LLaMA-2 parle de « *new mix of publicly available online data* ».

Pourquoi les acteurs recourent-ils à des données dont la légalité est discutable ? Plus la taille des données est importante, plus la quantité de calculs effectués pendant l'entraînement l'est aussi. En outre, nous avons pu montrer qu'à taille égale, mieux vaut s'entraîner sur des données variées, extraites d'Internet, plutôt que sur des sources homogènes comme Wikipédia. Qualité et diversité sont donc les maîtres mots. Mais, si des données de type « manuels de référence » et « exercices » permettent d'améliorer la qualité des modèles, avoir beaucoup de données moins bien filtrées conduit à de meilleurs résultats que moins de données de meilleure qualité. Ainsi, autour de la collection de corpus de RefinedWeb constitué à partir d'une certaine quantité de calculs, l'équipe de Falcon parvient à réaliser un modèle plus performant qu'un autre modèle entraîné sur The Pile, qui est pourtant un corpus de meilleure qualité mais plus petit. D'où l'importance de distinguer entre les données d'affinage et les données de pré-entraînement.

Notre compréhension du rapport entre la taille des modèles et la quantité de données qu'il faut pour les entraîner a évolué au fil du temps, avec la notion de lois d'échelle selon laquelle, quand un modèle grandit, une quantité supplémentaire de données est nécessaire pour atteindre le niveau de performance souhaité.

En 2020, OpenAI a déclaré: « *Larger models are significantly more sample-efficient, such that optimally compute-efficient training involves training very large models on a relatively modest amount of data and stopping significantly before convergence.* » D'après cette entreprise, le plus important serait donc que le modèle soit grand, même si les données ne sont pas très nombreuses, et l'entraînement devrait s'arrêter avant la convergence. Mais il s'avère que ce n'était pas tout à fait exact. Une équipe de DeepMind a ainsi montré que plus le modèle grossit, plus il faut augmenter le nombre de données. En outre, il est nécessaire d'entraîner les modèles plus longtemps que ce que nous imaginions initialement. C'est d'ailleurs en suivant ces règles que LLaMA et Chinchilla ont été entraînés.

En somme, pour élaborer de meilleurs modèles, nous avons besoin de plus de données, de meilleure qualité, plus diverses, et si possible des données blanches dont l'utilisation est juridiquement acceptable. Nous voudrions aussi disposer de données récentes, ce qui soulève certaines difficultés. Pour les contourner, il convient d'une part d'éviter d'exploiter uniquement des données de pré-entraînement, d'autre part de coupler les modèles de langue avec des bases de connaissances ou des capacités d'accès à internet en temps réel. Nous avons également besoin de données pour les langues qui nous intéressent. Le français, par exemple, n'est pas la seule langue parlée en France. C'est pourquoi des initiatives seraient à déployer pour couvrir les autres langues de France en données de préapprentissage. Cela fait l'objet d'un projet que je coanime au sein de l'INRIA.

S'agissant des incertitudes juridiques qui demeurent concernant les données grises et noires, l'acceptabilité par la société de ces modèles de langue et des progrès en IA dépendra de la capacité à affirmer que les entraînements n'ont pas été réalisés sur des données personnelles et ne soulèvent pas de problèmes liés aux droits d'auteur. Pour répondre à ces problématiques, la « SACEMisation » des données grises et noires serait une fausse bonne idée, car inapplicable compte tenu du nombre d'ayants-droits. Un pays qui mettrait une telle législation en place étoufferait toute initiative de recherche et de développement.

Créer plus de données blanches peut se faire soit *ex nihilo*, soit grâce à des modèles génératifs, soit par la numérisation de documents existants. Une autre stratégie consisterait à rendre blanches des données qui ne le sont pas. En clarifiant les choses sur ces sujets, nous pourrions faire de toutes les données grises des données blanches. Dans tous les cas, un vaste travail d'harmonisation reste nécessaire.

À toutes fins utiles, je précise que lorsque je parle de données blanches, cela ne signifie pas que toute utilisation est permise : il s'agit bien du but spécifique d'entraîner des modèles de langue.

Par ailleurs, certaines données grises – notamment les articles de presse – ont une valeur qui décroît vite avec le temps. Peut-être faudrait-il élaborer un paradigme selon lequel elles deviendraient blanches après un certain délai. Enfin, certaines données grises ou noires sont de grande valeur et financées par de l'argent public, notamment les cours et les exercices du CNED. Les modèles gagneraient en qualité s'ils pouvaient s'entraîner sur ces données.

Pour finir, je propose d'ouvrir le débat sur la question suivante : les données grises sont-elles vraiment problématiques ? Après tout, utiliser un modèle qui produit un texte relevant du plagiat n'est pas un problème en soi. Seule l'utilisation de ce texte peut faire de l'utilisateur un plagiaire. Il est de la responsabilité de chacun de savoir si le texte produit est utilisable, même si cela reste difficile à détecter. Ainsi, les modèles pourraient être complétés par des algorithmes dédiés. C'est le cas de GPT-4 et de son « *content filter* ». Par définition, les données grises sont légalement lisibles et mémorisables par des humains. Aussi, pourquoi n'en irait-il pas de même pour les modèles de langue ? Je pose la question dans le contexte de l'exception de fouille de données et sous réserve de rester cohérent avec le droit de retrait des données personnelles. Au Japon, toute donnée textuelle même sous copyright peut être utilisée pour entraîner des modèles. Je ne dis pas qu'il faille aller jusque-là, mais la question mérite d'être posée.

Échanges

De la salle

Concernant le partenariat entre OpenAI et Microsoft, au-delà du soutien financier et de la puissance de calcul, y a-t-il un accès à des données détenues de façon privilégiée ou exclusive sur lesquelles ChatGPT aurait été entraîné ?

Benoît Sagot

Nous ne savons pas sur quoi ces modèles ont été entraînés. Cependant, nous pouvons supposer que la réponse est oui, y compris concernant les données noires que nous autres, académiques, ne nous autoriserions pas à utiliser. Cela a-t-il un impact, à moyen ou long terme, sur la viabilité de ces modèles et du modèle économique sous-jacent ? Je l'ignore. Je me contente simplement d'observer que le week-end a été agité à la tête d'OpenAI !

De la salle

L'idée d'une régulation des données sur lesquelles les modèles ont déjà été entraînés n'intervient-elle pas trop tard ? Nous n'allons pas les fermer.

Benoît Sagot

Nous ne savons pas sur quoi ces modèles ont été entraînés. Cependant, nous pouvons Ce serait très passéiste. Il ne faut pas simplement regarder les modèles déjà entraînés, mais penser également à ceux qui le seront dans le futur, lesquels seront vraisemblablement de meilleure qualité que les actuels.

De la salle

Dans la mesure où ces modèles mémorisent, ils reflètent une image à l'instant T des connaissances sur lesquelles ils ont été entraînés.

Benoît Sagot

En effet, mais les modèles d'aujourd'hui souffrent de plusieurs problèmes, comme leur grande taille et le coût en termes de données et de calcul de leur entraînement. Si, dans quelques années, nous savons entraîner des modèles en lien avec un écosystème d'agents de calcul, de raisonnement et d'accès à de la connaissance, ils seront moins axés sur la mémorisation et deviendront au moins en partie des intermédiaires avec cet écosystème, ce qui pourrait réduire la taille qu'ils devront avoir pour être performants. Des évolutions pourront rendre ces modèles plus facilement entraînables et utilisables, ce qui pourrait rebattre les cartes quant à la notion de modèle utile. Dans tous les cas, des progrès sont encore possibles, notamment pour la langue française. Nous aurons des modèles meilleurs que ceux d'aujourd'hui. Se poser des questions sur la nature des données sur lesquelles nous souhaitons entraîner ces modèles reste donc pertinent.

Cela dit, je suis d'accord avec vous concernant les modèles existants. Interdire les données grises impliquerait l'interdiction de LLaMA ou de GPT-3 ou GPT-4, ce qui reviendrait à se tirer une balle dans chaque pied !

De la salle

Si l'objectif est d'aboutir à de plus petits modèles, donc à de moins grands ensembles de données, les résultats ne manquent-ils pas de solidité, notamment dans les comparaisons ?

Benoît Sagot

Nos résultats sont solides. Nous commençons à avoir des séries de modèles de différentes tailles entraînés sur le même jeu de données. Nous progressons. En comparaison aux premières lois d'échelle, pour lesquelles nous avions très peu de points de contrôle, notre niveau de confiance dans les nouvelles lois d'échelle est plus élevé.

Qui plus est, avec 7 à 20 milliards de paramètres, les modèles actuels restent très gros. Les jeux de données sont également très gros, de l'ordre de milliers de milliards de tokens. Pour un humain, lire un tel texte sans dormir ni manger prendrait 20 000 ans. Si nous savons entraîner des modèles en lien avec des banques d'informations, cela n'implique pas seulement de faire des modèles plus petits tout en restant au moins aussi performants, mais possiblement des modèles toujours aussi gros, voire plus gros encore, et avec des performances encore supérieures.

Comment favoriser le développement d'IA éthiques et performantes ?

Isabelle Ryl | INRIA, PRAIRIE
Pierre-Carl Langlais | OPSCI
Giada Pistilli | Hugging Face & Sorbonne

Modérateur : Éric Brousseau | Chaire Gouvernance et régulation, Université Paris Dauphine - PSL

Isabelle Ryl

Les temps de transfert entre la recherche fondamentale et les startups valorisant ses résultats se sont extrêmement réduits. Cette accélération sans précédent change les relations entre le monde économique et la recherche.

Pour réussir en IA il faut trois choses : des ressources humaines, des ressources de calcul et des données. La France n'est pas si mal positionnée qu'on tend à l'affirmer, en la matière. Nous avons encore de très bonnes formations, de très bons ingénieurs et de très bons chercheurs. De nombreux grands groupes ont d'ailleurs installé leur laboratoire de R&D à Paris.

À l'instar des données, les ressources de calcul soulèvent un problème de disponibilité et de coût d'investissement. De fait, la recherche en IA requiert du « jus de cerveau », mais aussi énormément d'investissement et d'ingénierie. À titre d'exemple, la création de BLOOM a impliqué un consortium de nombreux chercheurs, mais aussi d'entreprises pour apporter l'ingénierie nécessaire. De façon générale, les moyens dont disposent les équipes et les organismes de recherche universitaires ou académiques ne sont pas comparables à ceux des entreprises. C'est pourquoi la recherche sur des sujets comme l'IA générative se fait dans des entreprises, ou souvent en collaboration avec ces dernières.

Par ailleurs, les débats sont nombreux entre les tenants d'une régulation et les autres. Longtemps parmi les très pro-régulation, la France a rejoint des pays plus ouverts depuis la prise de position très claire de Bruno Le Maire en novembre 2023. On a souvent plaisanté sur le fait que la France et l'Europe étaient plus aptes à réguler qu'à innover, mais aujourd'hui nous avons la volonté de rattraper le retard de l'Europe sur l'IA générative, après les réseaux sociaux et les moteurs de recherche. Pour une fois, nous ne sommes pas partis avec trop de retard et nous avons peut-être une chance de nous positionner. Mais si nous verrouillons tout, c'en sera fini.

Prenons garde à ne pas nous « emmêler les pinceaux » dans ce paysage très intriqué entre recherche publique, recherche privée, monde économique et monde académique, ni entre les aspects de progression des connaissances et de recherche pure – à la fois pour le bien de la connaissance universelle et pour la maîtrise de la technologie. Même si l'on refuse d'utiliser une technologie en considérant que son domaine ou son usage va trop loin, il reste primordial de maîtriser les choses. Par exemple, je pense et j'espère que la France refusera de faire des *deep fakes* à des fins de manipulation électorale. Mais si personne ici ne connaît cette technologie ni ne sait comment elle s'applique, elle devient très difficile à contrer. C'est pourquoi il faut faire la part des choses entre l'usage que l'on fait d'une connaissance et le fait de la posséder ou non.

Jusqu'ici, une classification par type d'usages avait été faite dans le règlement européen. Cette dynamique a ensuite été remise en question par l'arrivée des grands modèles de langue dont les usages sont multiples. Ce sujet inquiète tant les législateurs que les chercheurs et certaines entreprises, mais aussi la communauté du logiciel libre.

Récemment, la communauté de Software Heritage, archive ouverte de codes source de logiciels s'est interrogée quant à sa participation à l'effort pour entraîner des modèles. Elle a exigé le respect de plusieurs principes, parmi lesquels le droit d'un auteur à demander que son code ne soit pas utilisé pour les données. Ce droit existe déjà dans la régulation sur le *data mining*, mais il reste difficile à exercer et vérifier. C'est tout le problème de la transparence. Outre la question de droit, il s'agit de connaître ce que fait l'IA, pour savoir expliquer sur quoi un modèle de langue a été entraîné ou savoir détecter les biais.

Le risque existe de passer d'un règlement européen couvrant tout à peut-être rien, et de se retrouver dans un flou qui continuerait à encourager les pratiques soit peu respectueuses, soit trop prudentes. Nous pouvons aisément imaginer la raison pour laquelle une entreprise comme Meta ne diffuse pas les données sur lesquelles ses modèles sont entrainés, et les procès auxquels elle pourrait faire face. Le flou n'est pas bon pour l'innovation, et tout refermer conduirait probablement les entreprises à se développer au Japon et dans les pays où les données sont blanches.

Pour citer cet exemple, les entreprises sont assez bien positionnées en Europe concernant la voiture autonome. Mais l'utilisation des données réelles de circulation pour entraîner des IA est interdite. En revanche, si nous autorisons l'importation de véhicules extra-européens venant de pays dans lesquels l'entraînement sur ces données est autorisé, nous nous retrouverons potentiellement avec une introduction de produits plus performants que les produits européens. Cela reviendrait à tirer une balle dans le pied des entreprises européennes.

Enfin, le débat est souvent contradictoire. Ainsi, les industries de la culture comptent parmi elles les grands lobbies en faveur d'une réglementation plus stricte, avec la notion de droits d'auteur notamment, tout en déplorant le manque de représentation de la culture européenne et de la langue française dans les IA. Pour caricaturer, il n'est pas difficile de deviner quel serait le résultat d'un usage payant de toutes les données dans un monde francophone et européen, tandis que dans le monde anglo-saxon il serait possible.

Pierre-Carl Langlais

La question centrale de l'acclimatation des LLM peut être abordée sous l'angle de la diversité culturelle.

Les modèles open source sont en train de rattraper ChatGPT, avec l'émergence d'un écosystème qui a effectué un travail de conversion, de structuration et de consolidation du modèle de base grâce à l'intégration de centaines de milliers d'exemples d'instructions et de conversations. Au-delà de ce premier enjeu, le *fine-tuning* vise à faire de l'acclimatation y compris culturelle. C'est d'autant plus crucial que si ces modèles sont multilingues en théorie, tel n'est pas encore le cas en pratique. Llama, par exemple, est à 90 % en anglais, le reste étant réparti en 8 % de code et 2 % d'autres langues (dont 0,16 % de français – nécessairement standard, donc peu propice à la diversité).

Créer des modèles ouverts et développer des écosystèmes adaptables permet de se réapproprier l'IA, en passant d'une IA potentiellement ouverte à un véritable commun. Dans cette optique, le *fine-tuning* n'est une opération ni lourde ni coûteuse. Des formes de *fine-tuning* léger permettent ainsi de geler une partie du modèle, pour n'en modifier qu'un petit corpus. Elles sont peu coûteuses, mais ont un fort impact, notamment sur la mémorisation et sur l'agencement de la communication (pensée comme la réponse à un usager plutôt qu'à un client). Par ailleurs, la quantisation permet de compresser les modèles. En revanche, le *fine-tuning* demande un travail de réflexion et de design : à quoi servira le modèle, comment choisir les données… ? Ce savoir en construction n'est pas encore systématisé, mais il permet de répondre à la question « Comment favoriser une IA éthique et performante ? ». De fait, construire un tel modèle, c'est construire une vision culturelle et éthique. Or en matière de diversité culturelle, être éthique c'est aussi être performant.

Au total, le *fine-tuning* permet non seulement d'inventer autre chose, mais aussi de poser une autre image de l'IA, au profit de la diversité.

Dans ce contexte, la question des données est essentielle, car elle soulève un enjeu de reportabilité. À cet égard, si un modèle de base est un « blob » aux contours imprécis, un modèle adapté est assez précis – ce qui s'avère primordial pour la régulation et pour la sécurité. Il convient également de souligner le rôle des données dites synthétiques, c'est-à-dire produites directement par les modèles, dans le rattrapage accéléré des modèles en open source. En l'occurrence, les meilleurs modèles sont ceux dont les données sont entraînées par d'autres modèles, dans une logique d'autocorrection.

Enfin, ces modèles ouverts imposent de définir concrètement le modèle que l'on veut faire et à quelles fins. De fait, un LLM peut être bien plus qu'un chatbot.

Giada Pistilli

Hugging Face est une plateforme open source d'hébergement de modèles, de data sets et de spaces (applications complètes d'un modèle d'IA), dans une logique community-driven. En son sein, l'équipe Machine learning and Society à laquelle j'appartiens, regroupe des chercheurs scientifiques, des linguistes computationnels et des responsables de policy interne et externe. En tant que chercheuse en philosophie, je favorise la recherche interdisciplinaire à mi-chemin entre policy et régulation.

La modération de contenus, visant à trouver le juste équilibre entre liberté et sécurité, est un défi inédit s'agissant du machine learning.

Hugging Face applique une content policy stricte dont l'une des valeurs piliers est le consentement. À cet égard, une importante réflexion est en cours concernant les modèles entraînés avec des données non consenties, afin de mettre en place des mécanismes d'opt-out. Mais cela soulève une question écologique, puisque toute sortie d'un data set d'entraînement impose d'entraîner à nouveau le modèle, ce qui s'avère coûteux en énergie, en calcul et en ingénierie. Ainsi, l'opt-out n'est pas viable à terme, mais mérite d'être envisagé comme un premier pas vers un futur plus consenti. Hugging Face met aussi l'accent sur l'ouverture éthique (ethical openess), en proposant un stage-to-release : le déploiement d'un modèle ou d'une application ayant vocation à être hébergée sur la plateforme doit respecter différents mécanismes de sécurité.

Au-delà de Hugging Face, mes recherches portent sur la multiculturalité et sur le multilinguisme. Dans ce cadre, j'ai participé à BLOOM. Hugging Face a fourni l'infrastructure (Slack, Google Workplace, docs, meetings sur Zoom et Google Meet), mais Bloom résulte avant tout d'un effort collaboratif de sciences ouvertes, avec plus de 1 000 chercheurs. Dans le cadre de la gouvernance des données (Roots), des outils d'exploration de données ont été créés. Il convient également de noter que Bloom est un modèle de complétion. Contrairement à ChatGPT, il n'est donc pas fait pour répondre à des questions. Cela étant, des chercheurs indépendants ont développé un BloomChat, hébergé sur Hugging Face.

Échanges

Benoît Sagot

Le corpus d'entraînement de Bloom n'est pas consenti : 38 % au moins ont été récupérés sur Internet sans consentement – ce que j'ai appelé les données grises dans mon introduction. Il n'en reste pas moins que Roots est public, donc librement téléchargeable. Mais cela n'en fait en rien un corpus blanc.

Par ailleurs, même s'il y a 46 langues dans le corpus d'entraînement de Bloom, nombre d'entre elles sont peu représentées (0,00002 % pour la moins représentée d'entre elles).

Enfin, BLOOM est moins performant que ce qu'on attendait, car il a été entraîné en suivant les règles d'échelle publiées par OpenAI en 2020, sur pas assez de données ou trop de paramètres. Et, surtout, il n'a pas été entraîné assez longtemps. Ce n'est ni un problème structurel ni une erreur, mais conforme à ce qu'il fallait faire compte tenu des connaissances de l'époque.

Il arrive que le secteur privé aide la recherche publique, en apportant des moyens de calcul et l'expertise de certains chercheurs. Pour autant, il est vrai qu'il faut absolument que les données de pré-entraînement soient libres. C'est la condition de la reproductibilité et de la transparence, y compris pour la recherche scientifique. Encore faut-il vérifier que les données sont légalement utilisables pour entraîner un modèle de langue.

De la salle

Au Collège de France, le cours « Philosophie de l'esprit et du langage » de François Recanati montre les influences du langage sur la pensée et la façon dont la pensée structure le langage. Ces considérations s'appliquent-elles à l'IA ? À terme, les modèles de langage auront-ils un impact sur la structure mentale de la société ? Le fine tuning peut-il favoriser des orientations dans des biais particuliers – culturels, confessionnels… ? La régulation n'a-t-elle pas un rôle majeur à jouer dans ce domaine ?

Pierre-Carl Langlais

Les essais que l'on pense purement stylistiques – changer la manière dont un langage est utilisé, pour qu'il soit plus archaïque par exemple – ont un impact immédiat sur les représentations culturelles : on ne peut pas séparer les deux. Il semble que les grands modèles de langue sont en train de valider l'hypothèse structuraliste, selon laquelle le sens procède par association de concepts et la langue fait en partie la culture. De fait, ces modèles n'ont aucune connaissance du monde extérieur. Or avec l'élargissement du *word embedding*, on n'intègre des éléments qui ne sont plus seulement de l'ordre de la syntaxe, mais aussi de l'ordre de la culture. C'est à coup sûr, dans ce domaine que la régulation devrait aller. Au-delà de la régulation, il convient aussi de créer des incitations pour faire émerger des modèles différents.

Éric Brousseau

Outre les cultures nationales, les types de données ont toute leur importance.

Bertrand Pailhès

Des systèmes plus spécifiques que les grands modèles dits de fondation qui ont été présentés sont-ils envisageables ? Cela permettrait d'adapter les règles d'entraînement des données aux différents contextes.

Benoît Sagot

Aujourd'hui, on ne sait pas le faire. On a besoin des modèles de fondation qui accumulent de l'information et, surtout, de la structure avant d'être spécialisés sur des tâches, des corpus ou des contextes spécifiques.

Pierre-Carl Langlais

La notion de capabilité générale, qui n'implique pas seulement d'avoir une connaissance précise, mais de savoir-faire de la synthèse de documents ou de gérer une conversation en se souvenant des propos précédents, a besoin de modèles généralistes.

Par ailleurs, une question reste ouverte : celle des données du modèle d'origine qu'efface le *fine-tuning*. En effet, l'adaptation est à la fois de la mémorisation et de l'oubli – lequel règle peut-être certaines problématiques, comme celle des données non consenties qui ont permis l'apprentissage.

De la salle

Comment faire pour que les données de qualité dont un producteur de contenus est propriétaire servent à l'entraînement sans rester ouvertes aux concurrents ?

Benoît Sagot

Il importe de distinguer les données accessibles légalement et gratuitement (données grises) des données payantes. Par ailleurs, les données accessibles légalement moyennant finances mais qui sont aussi accessibles gratuitement illégalement deviennent, *de facto*, des données noires.

De la salle

Il me semble que dans le procès contre Midjourney et Stability, les débats ne sont pas allés dans le sens de l'ouverture totale, puisqu'il a été démontré que l'entreprise qui avait ouvert toutes ses données d'entraînement utilisait des images soumises à des droits d'auteur. En revanche, cela s'est avéré plus difficile pour celle qui n'a pas ouvert toutes ses données. Qu'en pensez-vous ? Les entreprises n'ont-elles pas intérêt à ne pas publier leurs données de façon complètement ouverte ?

Par ailleurs, dès lors que les modèles mémorisent les données, il devrait être facile de savoir, par une requête simple, si une donnée a été utilisée ou non.

Benoît Sagot

En effet. L'idée de cacher ses données de pré-entraînement ne tiendra d'ailleurs pas longtemps.

Concernant votre première question, si vous photocopiez une partition portant la mention « photocopie interdite », c'est vous qui être responsable, pas la photocopieuse. Il ne faut pas se tromper de cible.

Pierre-Carl Langlais

On pourrait aussi citer le procès Silverman vs Meta : sans accès aux données de Llama, il est impossible de prouver matériellement que le livre de cette humoriste américaine a été utilisé. Ces exemples posent de nombreuses questions qui ne peuvent pas être résolues dans l'immédiat. C'est la difficulté de penser une révolution sans voir concrètement ses conséquences. En l'occurrence, ces modèles absorbent des données mais l'on peine à concevoir à quel point ils régurgitent des probabilités sur les mots et pas un texte exact.

Aujourd'hui, la jurisprudence américaine va dans le sens de la jurisprudence Google Books d'il y a dix ans, considérant que les textes protégés peuvent être utilisés dès lors que l'output relevait du *fair use*. Pour autant, la situation est différente. Il est essentiel qu'un écosystème émerge, en accompagnement de la régulation qui est en train de se construire.

De la salle

Comment construire une régulation avant de savoir quoi réguler ? Que faudrait-il contrôler pour que tout ne puisse pas se passer ?

Isabelle Ryl

Il faut distinguer les données des usages. En l'occurrence, dans l'IA, la plupart des gens craignent les usages.

Vouloir fixer des limites, éthiques ou autres, est une bonne chose sous réserve de pouvoir techniquement le faire. Or, la plupart du temps, en faisant la loi avant l'innovation, on risque de rencontrer des cas dans lesquels on ne sait pas faire.

Benoît Sagot

Les données illicites qui figurent sur internet donc dans les données d'entraînement sont difficiles à détecter. Les modèles de langue n'y sont pour rien si ces données existent, mais ils peuvent servir de révélateur et justifier des contrôles *a posteriori*. C'est d'ailleurs le rôle de l'étape visant à transformer un modèle de base en modèle conversationnel.

Quelles modalités d'intervention publique ?

Guillaume Avrin | DGE
Karine Perset | OCDE
Adrien Basdevant | Conseil national du numérique (CNNum)
Bertrand Pailhès | Directeur de la technologie et de l'innovation, CNIL

Modératrice : Joëlle Toledano | Chaire Gouvernance et régulation, Université Paris Dauphine – PSL, CNNum

Guillaume Avrin

La mission de coordination nationale pour l'IA s'articule autour de trois composantes : l'investissement dans le cadre de France 2030 (1,5 milliard d'euros de budget), sous la forme de subventions, de commandes publiques ou d'investissement en capital ; la coordination interministérielle ; l'animation de l'écosystème pour maximiser l'effet réseau.

L'État dispose de plusieurs leviers pour favoriser l'établissement d'un cadre éthique et de confiance pour l'IA. Avec les *frontier models* par exemple, versions les plus avancées des giga-modèles génératifs, il a beaucoup été question de risques existentiels, dont la potentielle capacité à générer de manière automatique des armes biochimiques ou des virus informatiques. Pour répondre au besoin de maîtrise de ces risques existentiels, des réflexions sont engagées autour de la gouvernance internationale de l'IA – étant entendu qu'il est essentiel de ne pas confondre anticipation et spéculation. Le Partenariat mondial pour l'IA et l'OCDE sont les structures idoines pour nous permettre d'avancer sur ces sujets.

Un travail est également en cours concernant la sûreté de fonctionnement de l'IA, en lien avec les autorités européennes. Il s'agira notamment de définir les normes harmonisées que devront respecter les entreprises pour obtenir une présomption de conformité aux exigences de l'*AI Act*. Le Cen-Cenelec JTC21 et sa commission miroir à l'Afnor sont mobilisés sur ce sujet et devront maximiser l'implication des entreprises françaises et européennes dans la normalisation en 2024.

Doivent aussi être mentionnées les initiatives technologiques lancées pour contribuer à l'établissement d'un cadre de confiance, en Europe et en France. L'investissement européen, supérieur à 600 millions d'euros, avec notamment les *Testing and experimentation facilities* for AI, est le plus gros financement mondial dans l'IA de confiance. Il n'y a donc aucune raison que l'Europe n'assure pas un leadership international dans ce domaine. Sur le territoire national, nous pouvons par exemple citer le Grand défi IA de confiance qui reposait sur trois piliers : le programme « Confiance.ai », le projet Prisma pour l'évaluation des systèmes de mobilité autonomes et un projet dédié à la normalisation de l'IA.

Les dispositifs visant à créer des communs numériques seront aussi de nature à contribuer à l'établissement de ce cadre de confiance. En ce qui concerne les évaluations de conformité volontaires, nombre de labels et de certificats ont été créés en France – sans doute plus que nulle part ailleurs dans le monde, ce qui n'est pas nécessairement positif car cela témoigne d'une dispersion des efforts : pour que l'un de ces labels émerge au niveau international, il faut cesser de se faire concurrence sur le territoire national. À cet égard, le Safety Summit 2024 sera l'occasion d'articuler toutes les initiatives en cours pour présenter une copie globale et cohérente.

S'agissant de l'éthique, il faut également prendre en compte la frugalité de l'IA et plus globalement l'impact positif et négatif du numérique sur la transition écologique, développer les formations (plus de 700 millions d'euros ont déjà été investis) et d'accompagner la transformation de notre société par l'IA.

Karine Perset

L'OCDE est une organisation intergouvernementale composée de 44 États membres (tous des démocraties et des économies de marché) et de nombreux autres États partenaires. Le groupe de travail AIGO, consacré à la gouvernance de l'IA, permet de fixer des priorités communes pour avancer dans six domaines d'expertise, parmi lesquels la protection des données et de la vie privée – enjeu particulièrement accru avec l'IA générative et le scraping de données.

Par ailleurs, l'observatoire OECD AI conduit des recherches sur les politiques nationales d'IA et met à disposition de nombreux outils, notamment de suivi des incidents en temps réel ou de mesure des biais.

Des réflexions sont également en cours pour élaborer et harmoniser des conditions contractuelles standards (licences pour l'utilisation de données) – déployables plus rapidement que la réglementation –, des codes de conduite pour les entreprises, des solutions techniques et des dispositifs d'éducation.

Adrien Basdevant

Deux tendances sont à l'œuvre, au niveau européen : des régulations contraignantes et des codes de conduite. En tout état de cause, pour favoriser une innovation responsable, de la régulation contraignante sera nécessaire. Il faudra aussi comprendre comment avoir les bonnes règles pour accompagner l'innovation. Or tout le paradoxe d'une *general-purpose AI* est qu'elle peut être développée sans que son application soit connue. Qui plus est, l'étude d'impact conduite en amont de l'élaboration du EU IA Act n'a pas porté sur l'IA générative (*general-purpose AI* ou *foundation models*). Certes, les centaines de millions d'utilisateurs sont apparus entre décembre 2022 et janvier 2023. Mais l'article Attention is all you need sur ces technologies date de 2017.

Initialement, l'EU IA Act visait une approche par le risque : plus une application est risquée, plus elle doit être documentée. Désormais, on considère que, par défaut, il faudrait réguler une technologie. Pourtant, si une réglementation contraignante est nécessaire, on ne peut pas réglementer une technologie, mais seulement des applications et des use case. Des discussions apaisées entre les acteurs qui développent des algorithmes et ceux qui génèrent du contenu sont indispensables pour trouver les bonnes solutions.

S'agissant du scraping, il convient de préciser que cette pratique n'est pas illégale en soi. En revanche, certains usages peuvent être illégaux : extraction substantielle d'une base de données, non-respect des conditions générales d'utilisation d'un site, reproduction d'une œuvre originale protégée par le droit d'auteur… Les enjeux de traitement de données sont multiples, également. Le RGPD peut s'appliquer dans certains cas, en fonction des bases légales (consentement, intérêt légitime, cadre de recherche…). En outre, dans la mesure où ces traitements sont statistiques, il est primordial de bien comprendre la chaîne de valeur de l'IA. Cela nous incite à revoir notre doctrine sur la propriété intellectuelle, la protection des données personnelles et le scraping au sens large.

Il est également intéressant de noter que ceux qui développent des algorithmes veulent pouvoir scraper mais refusent d'être scrapés. De la même façon, les éditeurs souhaitent bénéficier de l'article 4 de la directive 2019-970, lequel autorise le *text and data mining* à des fins commerciales, mais refusent qu'il s'applique aux développeurs de LLM.

Par ailleurs, s'agissant de la volonté de blanchir les données – souvent qualifiées de nouvel or noir alors qu'il vaudrait mieux les comparer à l'eau, matière renouvelable –, il convient de rappeler que l'article 324-1 du code pénal punit le blanchiment, défini comme le fait de dissimuler l'origine des biens de façon mensongère. Mieux vaudrait, donc, ne pas utiliser cette expression !

Enfin, l'opposition à l'utilisation de ses données pour alimenter les algorithmes doit s'exprimer différemment selon les phases de la chaine de valeur (pré-entraînement, entraînement, affinage). Récemment, plusieurs éditeurs ou détenteurs de bases de données se sont retirés du Common Crawl, prônant une meilleure répartition de la valeur : c'est compréhensible, mais cela entraîne un risque de déperdition de la diversité du Common Crawl. Dans ce contexte, d'aucuns proposent de relancer le débat sur la propriété intellectuelle. Ce ne sera productif qu'en l'absence de guerre de chapelles. Éviter les oppositions stériles implique aussi de bien comprendre la valorisation des données.

La France aurait tout intérêt à renforcer sa politique industrielle sur les données, *a fortiori* les données de synthèse. Créer des données synthétiques de qualité nécessitera des acteurs de premier plan.

Bertrand Pailhès

La CNIL a lancé un service d'IA il y a un an, pour affiner sa lecture de l'application d'un cadre juridique – principalement le RGPD – aux questions de l'intelligence artificielle.

Concernant le scraping et la notion de données grises, nous avons publié des fiches très didactiques. Nous avons également régulé plusieurs activités, notamment Clearview qui scrape toutes les images disponibles en ligne pour faire des logiciels de reconnaissance faciale pour les forces de l'ordre. L'argument de cette entreprise est qu'aux États-Unis, le Premier Amendement de la Constitution l'autorise à réutiliser librement des données accessibles sur internet. En Europe, en revanche, la France, l'Italie et la Grèce ont une interprétation différente et l'ont condamnée à payer 20 000 000 euros d'amende – le Royaume-Uni aussi, mais la décision a été annulée depuis.

Par ailleurs, lorsque les moteurs de recherche ont été créés et notamment Google, celui-ci Google « scrapait » l'intégralité d'Internet (y compris des données personnelles et sous copyright) sans que le cadre légal soit clair. Puis, avec l'arrêt Droit à l'oubli en 2014 et deux nouveaux arrêts en 2019, la CJUE a finalement considéré que Google avait un intérêt légitime à le faire, lequel ne menace pas les droits et intérêts des personnes concernées par ces données à moins qu'elles ne fassent valoir l'inverse. Le cas échéant, Google doit supprimer le contenu visé. Il doit aussi supprimer les données de santé dès qu'il en est informé, ou obtenir le consentement explicite des personnes concernées. Le droit s'adapte donc aux nouveaux produits et aux nouvelles technologies, avec un certain temps certes, mais ce sera aussi le cas pour les outils qui émergent aujourd'hui. Globalement, les droits et libertés ne sont pas massivement mis en question, pour ce qui concerne l'entraînement de ces grands modèles de langage. La CNIL n'avait d'ailleurs été saisie d'aucune plainte relative à ChatGPT avant la procédure de son homologue italien. Une régulation reste toutefois à inventer, ne serait-ce que pour la gestion des droits des personnes. En effet, ce modèle statistique crée du nouveau contenu, ce qui soulève de nouvelles questions.

S'agissant de la réutilisation des données, la CNIL propose de déterminer si la base de données réutilisée est « manifestement illégale » – cette solution procure une marge d'interprétation, pour du cas par cas – et de s'appuyer sur la notion de finalité au sens du RGPD qui précise que la finalité de la réutilisation doit être compatible avec la finalité de la collecte. Le RGPD prévoit de plus que si la finalité de réutilisation est la recherche, celle-ci est présumée compatible avec la finalité initiale, ce qui constitue un fort élément d'assurance pour les acteurs de la recherche. Dans le cas d'OpenAI ou Meta, la difficulté vient du fait que ces acteurs réutilisent les données à la fois pour une finalité de recherche et pour une finalité commerciale. Des travaux sont en cours pour savoir s'il faut distinguer ces deux activités.

Ainsi, si nous parvenons à dégager des principes généraux, la difficulté viendra du traitement des cas particuliers. En 2023, la CNIL a par exemple engagé une expérimentation avec la loi sur les Jeux olympiques, qui crée un cadre spécifique directement inspiré de l'AI Act permettant d'entraîner des systèmes de caméras intelligentes sur des données de voie publique et à destination des forces de l'ordre – donc un système d'IA à haut risque au sens de l'AI Act. Nous accompagnons plusieurs entreprises vis-à-vis des nouvelles exigences de la loi : représentativité des données, mesure et correction des erreurs et des biais dans le système d'IA.

Dans tous les cas, la logique visant à se concentrer sur les usages semble la bonne.

Enfin, pour réagir sur le propos liminaire de Benoît Sagot, je me permets de souligner que de mon point de vue, trois « cordes de rappel » pourraient permettre d'éclaircir les « données grises » : la transparence de la documentation, les droits des personnes (consentement, droit d'opposition) et enfin l'encadrement des usages autorisés pour ces modèles entraînés sur des données grises.

Échanges

De la salle

Quelle place pour la prospective ?

Guillaume Avrin

La prospective a, par nature, toute sa place dans la Stratégie nationale pour l'IA. Aussi faudrait-il renforcer les contacts avec les cellules de prospective de l'État (France Stratégie, Haut-Commissariat au plan, SGDSN).

De la salle

Concernant la valorisation des données, n'y a-t-il pas un risque de spoliation ou d'appropriation, *a fortiori* si l'on va jusqu'aux données synthétiques. L'exemple des ressources naturelles ne devrait pas être oublié.

Adrien Basdevant

Les données ne sont pas le nouvel or noir : si le baril de pétrole est fongible avec la donnée de mobilité, la donnée de santé ne l'est pas. Mieux vaudrait utiliser la métaphore de l'air ou de l'eau, car c'est l'interaction entre les données qui fait leur valeur.

Par ailleurs, arrivera le moment où les données de synthèse seront plus nombreuses que celles créées par l'humain. Or, comme le montrent les papiers sur le Model Collapse, du *reinforcement learning* est indispensable dans la boucle de rétroaction pour avoir des données de synthèse de qualité, soit avec *Human in the loop* soit avec *AI synthetic composant in the loop*. Ainsi, faire de la France et de l'Europe un producteur de données de synthèse de grande qualité est clé pour les années à venir.

Benoît Sagot

Les données de synthèse ont un rôle important à jouer, en particulier lorsque créer de la donnée réelle est coûteux ou lorsque les problématiques de données personnelles sont proéminentes, mais aussi pour les données textuelles. Les équivalents de ChatGPT en open source ont ainsi été produits grâce à des conversations en partie synthétiques, y compris par ChatGPT lui-même. Un équilibre est à trouver.

De la salle

S'agissant des LLM, le droit à l'oubli sera sans doute la seule régulation possible. Une recherche est d'ailleurs en cours pour l'implémenter par le *machine unlearning*, qui ne nécessite pas de réentraîner tout le modèle. La CNIL a un rôle à jouer, y compris pour mettre à disposition des outils permettant de savoir si des données ont été utilisées.

Bertrand Pailhès

Pour le moment, ce domaine académique reste très pointu. Il est également envisagé de placer des filtres sur les outputs. Par ailleurs, il semble qu'un droit à la correction ne pourra exister sur le modèle lui-même, en raison de la nature statistique du système.

De la salle

L'AI Act arrive après d'autres réglementations (DGA, Data Act), et ne prend pas en compte l'IA générative qui n'existait pas encore. N'y a-t-il pas là une totale inadéquation ?

Guillaume Avrin

L'IA générative est couverte par l'AI Act, depuis la première version de la Commission européenne. En outre, aucun État membre ne défend l'idée que l'IA générative ne devrait pas être couverte par ce règlement dans les domaines d'application à haut risque. Le débat porte davantage sur les IA d'usage général et les modèles de fondation, génératifs ou non. L'enjeu consiste à traiter l'ensemble de ces sujets dans le bon cadre, en évitant de disperser le traitement de certains sujets dans de nombreux textes réglementaires (par exemple concernant les données personnelles, les droits d'auteur, la cybersécurité, etc.), pour éviter de perdre les acteurs qui doivent se mettre en conformité.

Adrien Basdevant

Quand on parle de data en Europe, on parle de données personnelles. Le périmètre est pourtant bien plus large et la frontière entre les deux est l'anonymisation. Or, celle-ci dépend d'un standard datant de 2014 qui n'est donc plus à l'état de l'art. Comment traiter les ensembles de données inextricablement liées, avec des données industrielles et des données personnelles, et pour lesquelles plusieurs textes parfois incompatibles s'appliquent ? Cela pose la question de la compatibilité des textes et des autorités administratives indépendantes qui devront les superviser, chapeautées par la Commission ou non.

Benoît Sagot

Si l'on regarde les choses à moyen terme (trois mois), il apparaît que l'importance des outils avec lesquels les LLM interagiront de plus en plus est partiellement sous-estimée.

Éric Brousseau

L'histoire a montré que l'on n'arrive jamais à prévoir les usages d'une technologie sur la base des intentions de ses développeurs. De fait, la technologie est d'abord appropriée par les utilisateurs, puis les développeurs créent des applications répondant à leurs besoins. Dès lors, tenter de réguler en fonction de l'anticipation des risques est en général voué à l'échec. En revanche, il est indispensable de se doter de capacités ex-ante pour être réactif. Que serait une force de réaction rapide pour répondre aux défis de l'IA ?

Guillaume Avrin

Les réflexions sont en cours autour de la création d'un AI Office européen. La France considère qu'il devrait être composé d'acteurs représentatifs, comme les administrations publiques mais aussi des acteurs privés. Quelles que soient les exigences contenues dans l'AI Act, ce sont les normes harmonisées qui font présomption de conformité. Or ces normes sont principalement définies par des acteurs privés.

Par ailleurs, il serait intéressant de positionner l'AI Office sur des sujets d'anticipation, par exemple l'adaptation de l'annexe 3 de l'AI Act. À l'inverse, il convient d'éviter de multiplier les autorités notifiantes et organismes notifiés, donc les interlocuteurs pour les entreprises qui devront se mettre en conformité.

Karine Perset

La prolifération des standards internationaux est un vrai sujet. L'OCDE a également pu observer qu'il existe de très nombreux Risk Management Frameworks, similaires mais suffisamment différents pour être aisément suivis par les non-experts de la gestion de risque, notamment les PME opérant à l'international. Aussi l'OCDE tente-t-elle de rapprocher ces cadres et d'établir une cartographie.

Bertrand Pailhès

Ma position est ambivalente vis-à-vis des standards techniques, qui me semblent à la fois utiles, mais trop généraux pour une approche au cas par cas. Il faut donc les adopter sans en faire un ensemble suffisant de règles.

De la salle

Quid des communs de données ? Les travaux en cours sont-ils connectés les uns avec les autres ? L'Europe a besoin d'une stratégie claire.

Guillaume Avrin

Ce sujet est intégré dans l'EDIC (*European Digital Infrastructure Consortium*), qui a vocation à regrouper l'ensemble des ressources d'entraînement des IA génératives au niveau européen. Nous réfléchissons aussi à la façon de concilier les positions sur l'innovation et la protection des droits d'auteur.

Conclusion

Anne Bouverot | Co-présidente de la Commission IA

La Commission IA (initialement Comité de l'intelligence artificielle générative), lancée par la Première ministre en septembre, est chargée de rendre un rapport et des recommandations début mars. En termes de méthode, nous nous réunissons tous les vendredis et nous conduisons des auditions sur cinq thèmes : impacts économiques et sur le travail, souveraineté technologique, impacts éthiques et sociétaux, IA et culture, impacts sur les services publics. Notre objectif est d'aider à apprécier ce qui peut être fait et à prioriser les actions. À cet égard, même si la situation de la France est différente, l'*Executive Order* des États-Unis est inspirant. Il présente une vision d'ensemble du rôle de chaque partie face au développement de l'IA.

Le développement de l'IA est très rapide et son champ d'application est très large – bien plus que ceux des précédentes révolutions technologiques. L'ambition est d'accompagner ce développement, pour qu'il soit aussi bénéfique et responsable que possible.

En France et en Europe, la première bataille vise non pas à réguler, mais à être dans l'action, donc à développer des services d'IA. De ce point de vue, l'annonce récente de la création d'un laboratoire privé de recherche en IA est une bonne nouvelle, de même que l'existence de plusieurs startups dans ce domaine. Le secteur public a aussi son rôle à jouer, *a fortiori* en France grâce à l'excellence de la formation, de l'enseignement supérieur et de la recherche, et à des experts mondialement reconnus. Outre ces capacités, le besoin de financement est réel. Le plan France 2030 est à saluer, mais il en faudrait beaucoup plus !

Les autres enjeux concernent le développement d'une IA responsable, la prise en compte des risques (désinformation, biais, discriminations, modification de nombreux emplois et tâches, etc.), la compréhension des peurs et les conditions de la confiance.

Par ailleurs, si la régulation ne doit pas être notre première réaction, elle est nécessaire. En la matière, la réflexion doit s'inscrire dans un continuum avec des travaux sur la gouvernance des entreprises, sur les chartes et les labels, sur la formation et sur l'expérimentation.

La récente actualité d'OpenAI souligne l'importance de la gouvernance. Créée en 2015 comme une société à but non lucratif pour faire de la recherche ouverte sur le développement de l'IA, elle est parvenue à lever 100 millions de dollars en trois ans. Ce montant n'étant pas suffisant pour entraîner son modèle, il a été décidé de créer une filiale commerciale chargée de passer des accords (principalement avec Microsoft), de développer des produits et de disposer de moyens financiers et de capacités de calcul. Or cette partie a crû de manière très rapide, avec 13 milliards de dollars investis par Microsoft, et engendré une dichotomie et une discordance avec un conseil d'administration ayant peu évolué dans sa composition et ses missions. Ce désalignement entre un conseil d'administration et sa filiale devenue beaucoup plus grosse est l'une des raisons de la « saga » qu'il faudra continuer à suivre.

Chaire Gouvernance et Régulation
Fondation Paris-Dauphine
Place du Maréchal de Lattre de Tassigny - 75016 Paris (France)
https://chairgovreg.fondation-dauphine.fr/