GOUVERNEMENT
Liberté
Égalité
Fraternité

Pôle d'expertise de la régulation numérique

Éclairage sur…

Avril 2024

#07

Open source et IA : des synergies à repenser ?

L'intelligence artificielle générative (IAG), a connu tout récemment une percée fulgurante plaçant désormais la production automatique de contenus synthétiques (texte, sons, images, vidéos) à la portée de tous.

Les modèles d'apprentissage profond sont la pierre angulaire de ce secteur d'activité en ébullition. Créés à partir de très nombreuses données aux statuts variés, ils représentent des briques logicielles aux multiples applications et dont le format de diffusion ou de réutilisation éventuelle soulève de nombreuses questions.

Aujourd'hui, les acteurs principaux de ce nouvel écosystème s'appuient directement ou indirectement sur des ressources en licences *open source*. Afin de comprendre comment ces innovations ont pu s'appuyer sur un tel socle et évaluer si ce cadre reste adapté à leur développement, le PEReN explore dans ce numéro d'« Éclairage sur… » les formats de licence et en quoi ils peuvent favoriser une dynamique d'innovation durable et équilibrée entre acteurs de cet écosystème.

L'IAG EN QUELQUES MOTS…

L'IAG, ou intelligence artificielle générative, désigne un ensemble d'objets techniques permettant de générer ou d'éditer des contenus (textes, sons, images, vidéos). Le plus souvent, ces objets sont des **modèles** d'apprentissage profond, ou **réseaux de neurones**. La structure, le type et le nombre de neurones contenus dans le modèle se nomme « **architecture** ».

Un réseau de neurones contient **des poids ou paramètres,** c'est-à-dire des nombres qui caractérisent le comportement d'un modèle en fonction des données fournies en entrée, en pondérant notamment certains motifs (ensemble d'attributs communs) au sein de ces données. Une fois l'architecture déterminée, les modèles sont entraînés sur de grands volumes de données pour optimiser la **fonction objectif** qui leur a été assignée. En pratique, il s'agit d'ajuster progressivement les poids afin de maximiser la performance du modèle à partir du jeu de données d'entraînement. Par exemple, les grands modèles de langage (ou LLM, pour *Large Language Models*, sur lesquels le PEReN a déjà eu l'occasion d'écrire un précédent numéro d'Éclairage sur[1]…) ont pour objectif de prédire le mot suivant d'une séquence de mots. Afin de se tromper le plus rarement possible, le modèle va tendre à prédire le mot le plus fréquent au regard de la séquence, même lorsque d'autres mots plus rares auraient pu être corrects.

Une fois ces réseaux de neurones entraînés, ils peuvent être vendus ou rendus disponibles contre paiement grâce à des API, mais d'autres acteurs choisissent de les distribuer publiquement. Leur publication repose alors sur des licences, qui créent un lien contractuel entre les utilisateurs et les fournisseurs du modèle.

QUELLES ARTICULATIONS ENTRE IAG ET LICENCES *OPEN-SOURCE* ?

Licences *open source, kesako ?*

Les licences sont un élément primordial de la création et du partage en ligne. Dans le domaine du droit d'auteur, elles constituent un contrat entre l'auteur et l'utilisateur (ou un acte unilatéral de l'auteur lorsqu'il renonce à certains droits).

Dans le domaine informatique, le mouvement *open-source* vise à ouvrir le code source au public. Pour accompagner cette dynamique, des **définitions de référence de cette ouverture ont été érigées par des organisations dévouées à la démarche :**

- ***l'Open Knowledge Foundation*** **(OKF)[2] pour ce qui est des données ;**
- ***l'Open Source Initiative*** **(OSI)[3] et la *Free Software Foundation* (FSF) pour ce qui concerne le code source[4].**

Les licences qui respectent ces définitions *d'open source* doivent librement permettre l'utilisation, la redistribution (y compris commerciale) et la modification du contenu concerné. Lorsque les permissions accordées sont soumises à des conditions relatives aux utilisateurs ou au but poursuivi par ces derniers (par exemple, la licence *Creative Commons*[5] CC BY-NC interdit l'utilisation commerciale

[1] https://www.peren.gouv.fr/actualites/2023-04-06_eclairage_sur_chatgpt/

[2] https://opendefinition.org/od/2.1/en/

[3] https://opensource.org/osd/

[4] Une liste de référence des licences et des évaluations FSF et OSI est disponible à l'adresse suivante : https://spdx.org/licenses/. La DINUM publie également une liste de licences libres pour les administrations : https://code.gouv.fr/fr/doc/licences-libres-dinum/.

[5] https://creativecommons.org/share-your-work/cclicenses/

du contenu visé), la licence ne peut alors pas être qualifiée de libre ou *open source* et sera qualifiée de « licence de libre diffusion ».

Les licences *open-source* se déclinent en deux catégories :

- les licences *copyleft* ou licences à réciprocité : des conditions relatives à la distribution peuvent être ajoutées aux licences *open source*, notamment l'attribution (crédit des auteurs) et le partage à l'identique (sous la même licence). Ces conditions caractérisent les licences *copyleft* qui se diffusent de contenu en contenu. Autrement dit, lorsqu'une œuvre est dérivée d'un contenu publié sous une telle licence, cette œuvre doit être publiée sous la même licence ou une licence compatible, c'est ce qui explique que ces licences puissent être appelées « contaminantes ». Les licences CC-BY-SA (*Share Alike*, partage à l'identique), *Open Database License*[6] (ODbL) et *GNU General Public License*[7] entrent par exemple dans cette catégorie.
- les licences permissives : autorisant l'utilisation, la modification et la diffusion, elles ne nécessitent pas la préservation de la licence initiale.

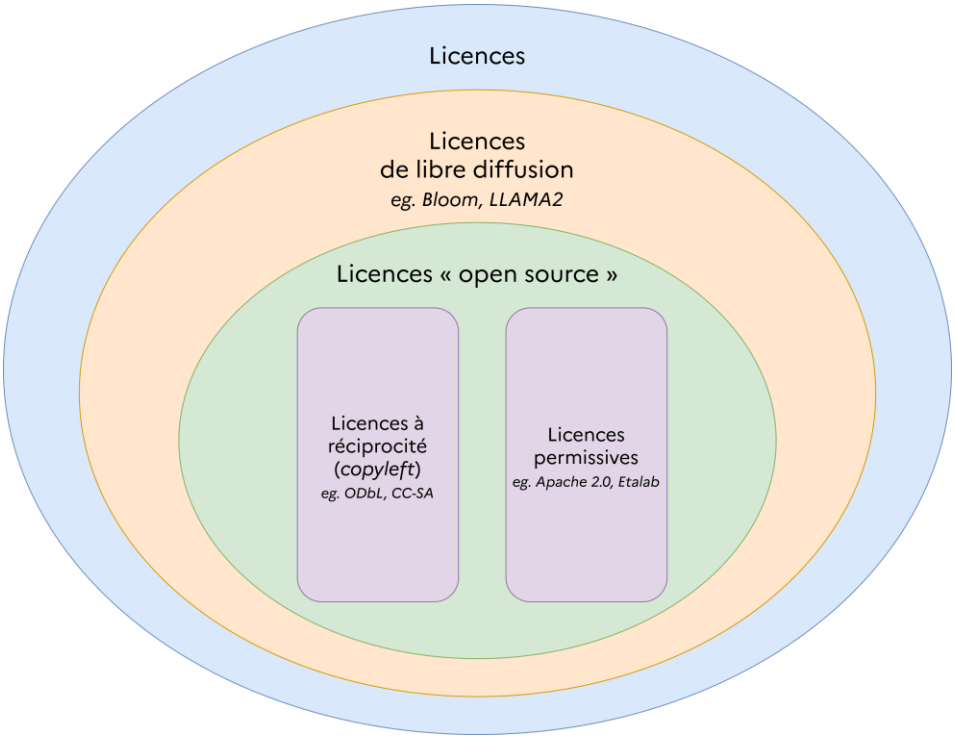

Figure 1 : Écosystème des licences

Ouverture des modèles d'IA : quelles licences accorder au code et aux données ?

D'une part, **les jeux de données d'entraînement (*datasets*)** sont disponibles via différentes licences, et certains peuvent même comporter des données ayant des licences différentes. À titre d'exemple, *The Pile*[8], un jeu de données destiné à l'entraînement des modèles de langage, compile des données de plus d'une dizaine de sources (dont le *Common Crawl*, Wikipedia, ArXiv et Github). Le jeu de données intégral *The Pile* n'étant pas disponible au téléchargement, le code qui permet de le

[6] https://opendatacommons.org/licenses/odbl/
[7] https://www.gnu.org/licenses/gpl-3.0.en.html
[8] https://arxiv.org/pdf/2101.00027.pdf

reproduire est accessible sous la licence permissive MIT[9]. Toutes les données qui le composent sont disponibles via leurs propres licences. Certaines sources sont disponibles sous licence à réciprocité : c'est le cas du contenu de Wikipedia, disponible sous licence CC-BY-SA[10] ou de nombreux codes sources hébergés sur GitHub sous licence GNU GPL[11].

D'autre part, **les différentes ressources de code nécessaires à l'entraînement des modèles de langage** ont chacune leur licence. Pour citer plusieurs exemples populaires, CUDA (utilisé pour optimiser l'entraînement sur processeurs graphiques Nvidia) est disponible sous licence non-libre (ou dite propriétaire)[12] ; les bibliothèques logicielles Pytorch et Xformers de Meta sont disponibles sous licence permissive BSD[13,14] ; les librairies Tensorflow (Google) et Transformers (HuggingFace) sont disponibles sous licence permissive Apache[15,16]. Ces exemples montrent que pour l'instant, bien que le développement du code nécessaire à l'entraînement et à l'utilisation des modèles d'IAG soit concentré au sein d'une poignée d'entreprises, le format *open source* demeure dominant.

L'entraînement d'un modèle d'intelligence artificielle repose sur ces données et ce code source pour produire un modèle, qui peuvent chacun être rendus disponibles (sous licence), ou au contraire ne pas être diffusés. Le code permet de définir l'architecture du modèle ainsi que les objectifs et la méthode d'entraînement. L'entraînement consiste à ajuster un très grand nombre de fois les poids du modèle au regard des données fournies et de l'objectif poursuivi. Les poids obtenus à l'issue de l'entraînement concentrent la majeure partie de la valeur ajoutée du modèle.

Les licences à réciprocité les plus populaires (par exemple ODbL et GNU GPL) ne traitent pas explicitement des questions de réciprocité liées au statut des modèles, et il n'est pas certain qu'elles permettent de contraindre leur publication. En effet, si les poids des modèles constituent bien des « données » (ce sont en pratique des chiffres arrangés dans une architecture) et bien qu'ils dépendent entièrement des données d'entraînement, ni ces poids, ni les données obtenues en sortie des modèles ne constituent, à proprement parler, une version modifiée des données initiales. De même, ces modèles ne constituent pas du code, et la mise à disposition du code source nécessaire à l'entraînement n'inclut donc pas les poids finaux. Ainsi, en l'absence de décision de justice sur ce sujet, **il n'est pas certain que les conditions de réciprocité des licences les plus utilisées pour les données d'entraînement ou le code source s'appliqueront aux poids ou aux sorties (*outputs*) des modèles. En ce sens, cette réciprocité pourrait s'avérer inopérante sur les modèles d'IA.**

[9] https://github.com/EleutherAI/the-pile/blob/master/LICENSE

[10] https://en.wikipedia.org/wiki/Wikipedia:Reusing_Wikipedia_content

[11] L'utilisation de données distribuées sous licence à réciprocité afin d'entraîner des modèles d'IAG est au cœur d'une plainte instruite en ce moment aux États-Unis. La décision qui s'ensuivra sera cruciale pour la constitution des futures bases d'entraînement et la pertinence de certaines licences à réciprocité pour protéger certains travaux : https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data

[12] https://docs.nvidia.com/cuda/eula/index.html

[13] https://github.com/pytorch/pytorch/blob/main/LICENSE

[14] https://github.com/facebookresearch/xformers/blob/main/LICENSE

[15] https://github.com/huggingface/transformers/blob/main/LICENSE

[16] https://github.com/tensorflow/tensorflow/blob/master/LICENSE

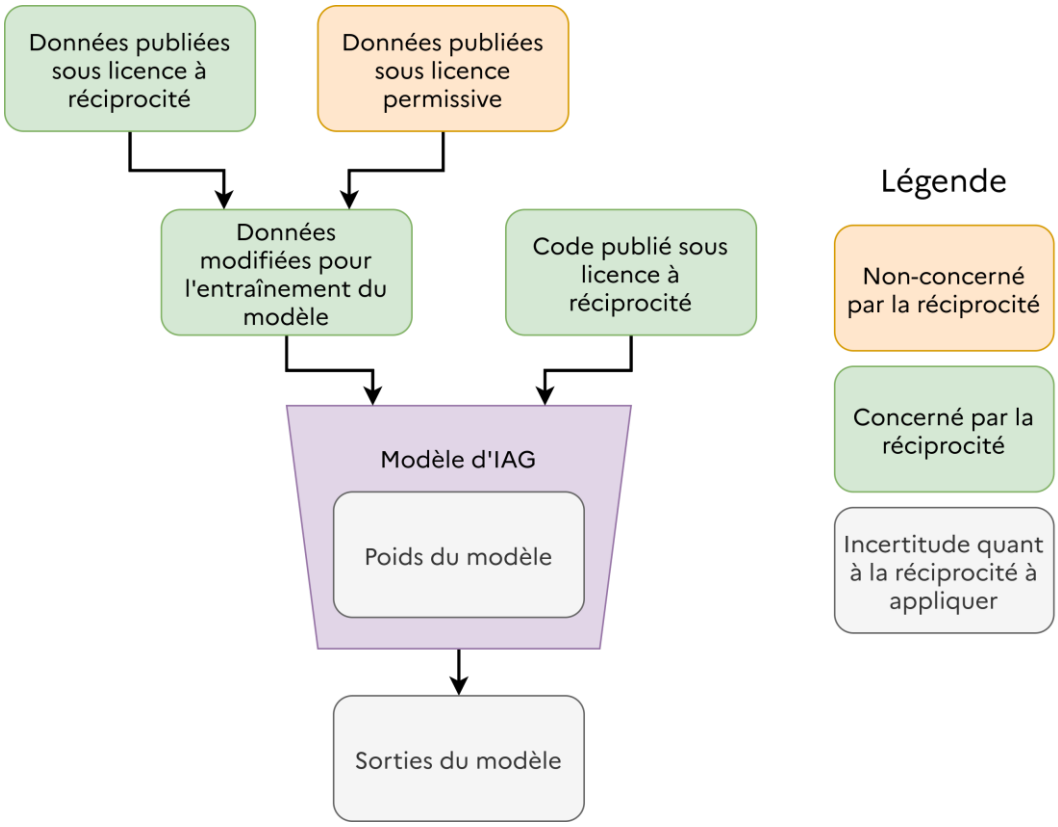

Figure 2 : Illustration des enjeux de réciprocité des licences liées aux données ou au code source lors du développement d'un modèle d'IA. En l'absence de décision juridique sur le sujet, les licences sont potentiellement inopérantes sur le modèle d'IA et ses sorties.

Certaines organisations ayant fait le choix de l'ouverture ont publié leurs modèles via des licences permissives[17] ou créé leurs propres licences portant spécifiquement sur les modèles, c'est-à-dire les poids. La licence du modèle LLAMA 2[18] permet sa réutilisation, sa modification et sa redistribution (via la même licence) mais restreint certains usages, notamment ceux jugés illégaux au regard du droit américain ou ceux consistant à entraîner d'autres modèles de langage à l'aide des sorties (*outputs*) de LLAMA 2. La licence du modèle BLOOM[19] autorise quant à elle sa réutilisation, sa modification et sa redistribution mais interdit certains usages considérés non-éthiques par les chercheurs qui l'ont développé. Enfin, la licence du modèle FALCON soumet le déploiement du modèle à l'appréciation de l'entreprise responsable du modèle et à une licence distincte qui n'a pas été rendue publique[20]. Dans ces trois exemples, des usages étant proscrits ou limités, aucune des licences citées ne saurait être qualifiée *d'open source* au sens des définitions de la FSF, de l'OSI et de l'OKF.

Face à cette diversité de situation, et à l'incertitude liée à la portée de la réciprocité des licences *copyleft*, de nouvelles licences *ad hoc* ont émergé. Consciente de ces enjeux, l'OSI mène actuellement des travaux pour aboutir à une définition claire de l'IA *open source*, et qui pourraient mener à la proposition de nouvelles licences types.

[17] Par exemple Mistral, qui a publié les modèles Mistral-7B et Mixtral 8x7B sous licence Apache 2.0. Toutefois, les modèles les plus performants de Mistral sont soumis à des licences plus restrictives.

[18] https://github.com/facebookresearch/llama/blob/main/LICENSE

[19] https://huggingface.co/spaces/bigscience/license

[20] https://huggingface.co/spaces/tiiuae/falcon-180b-license/blob/main/LICENSE.txt

Modèle	Licence	Accès aux poids du modèle	Redistribution et modification libres	Aucune restriction sur les usages commerciaux	Aucune restriction sur les usages non-éthiques	*Open source* au sens de l'OSI
BigScience Bloom	Bloom	✓	✓	✓	✗	✗
Google Gemini	Conditions d'utilisation de l'IA générative Google	✗	✗	✗	✗	✗
Meta Llama2	Llama2	✓	✓	✗	✗	✗
Microsoft Orca2	Microsoft Research	✓	✗	✗	✗	✗
OpenAI GPT-4	Conditions d'utilisation OpenAI	✗	✗	✗	✗	✗
TII Falcon	Falcon	✓	✓	✗	✗	✗
Anthropic Claude	Conditions d'utilisation Anthropic	✗	✗	✗	✗	✗
Mixtral 8x7B	Apache 2.0	✓	✓	✓	✓	✓
Mistral Large	Conditions d'utilisation Mistral	✗	✗	✗	✗	✗

Tableau 1 : Les critères open source selon l'OSI des modèles de fondation récents.

EN QUOI L'*OPEN-SOURCE* INFLUE-T-IL LA DYNAMIQUE DE L'ÉCOSYSTEME DE L'IA ?

Le monde numérique fournit de nombreux exemples d'usages courants de *l'open source*, qui fournissent des enseignements sur les avantages et risques de la présence d'alternatives *open source* ou de chaînons *open source*. Peuvent être cités en exemple : VLC pour la lecture de médias[21], Android et Chromium par Google[22],

[21] VLC permet d'utiliser d'utiliser librement de nombreux codecs indispensables à la lecture de fichiers multimédias, et son code est ré-utilisé par un grand nombre de lecteurs multimédias commerciaux : https://www.videolan.org/vlc/index.fr.html

[22] Pour Android comme pour Chrome, il existe une version épurée du code disponible en open-source : https://source.android.com/ ; https://github.com/chromium/chromium

la librairie de chiffrement[23] OpenSSL, le système d'exploitation Linux[24] ou encore les bibliothèques de code utilisées pour construire des modèles d'intelligence artificielle (PyTorch, scikit-learn, etc.). Aujourd'hui, **la quasi-totalité des acteurs du numérique s'appuient directement ou indirectement sur des ressources *open source*.** Au niveau national, cette tendance est également encouragée par une politique favorable à encore davantage d'ouverture[25].

Le modèle économique des acteurs de l'*open source*

L'omniprésence de *l'open source* ne doit pas cacher la diversité de ses acteurs et la fragilité économique d'un grand nombre d'entre eux. En effet, **les logiciels *open source* sont par essence diffusables librement, donc *a priori* gratuitement, ce qui rend complexe les modèles économiques reposant sur la seule vente du logiciel**. Pour s'assurer une rentabilité durable, un acteur développant des logiciels *open source* doit apporter à son produit une autre forme de valeur ajoutée : soit pour le client (par exemple, un service après-vente), soit pour l'auteur (par exemple l'ouverture de marché pour d'autres produits). Ainsi, trois cas de figure principaux se distinguent, qui ne sont d'ailleurs pas mutuellement exclusifs :

- les projets *open source* n'ayant pas l'ambition d'être rentables. Ce cas recouvre une grande partie de l'*open source*. Les contributeurs sont le plus souvent des associations, des bénévoles, relèvent du mécénat d'entreprises, ou encore des institutions de recherche publique ;
- les projets *open source* dont la justification économique réside dans la valeur ajoutée pour leurs auteurs. Celle-ci peut correspondre par exemple à une stratégie d'adoption pour pénétrer un marché (Android par exemple), ou à permettre de bénéficier du regard extérieur de spécialistes (Signal et son protocole de chiffrement de messages, Netflix et son code pour éprouver la résilience des serveurs, etc.), ou à certaines briques ayant une valeur partagée très forte, telles que le noyau Linux auquel contribuent de nombreux industriels ;
- les projets *open source* qui génèrent des revenus par une la proposition de services additionnels pour les clients. Ceux-ci peuvent correspondre à un support technique (assistance à l'installation et à l'utilisation), au service après-vente (garanties de disponibilité et de corrections de *bugs*), à l'ajout de fonctionnalités (qui peuvent être elles-mêmes ouvertes ou non), etc. Ce modèle de financement n'empêche néanmoins pas d'autres acteurs de se saisir du code ouvert et l'internaliser, et devenir concurrents sans avoir à supporter le coût initial du développement. Ce développement a entraîné l'émergence de licences (non strictement *open source*), qui peuvent exclure la réutilisation pour fourniture de services concurrents[26].

On peut néanmoins souligner que cette diversité de situations ne prémunit de défaillances, qui peuvent avoir des conséquences en cascade, comme l'a illustré la faille Heartbleed, découverte en 2014 et affectant 17 % des serveurs web sécurisés[27]. Cette vulnérabilité, qui concernait la bibliothèque de chiffrement

[23] OpenSSL par exemple est utilisé par une énorme partie des communications sur internet, pour sécuriser le trafic : https://github.com/openssl/openssl

[24] Linux peut être utilisé sur des ordinateurs personnels, mais ce système d'exploitation est également utilisé sur de très nombreux serveurs : https://github.com/torvalds/linux

[25] https://code.gouv.fr/fr/plan-action-logiciels-libres-et-communs-numeriques/

[26] https://lwn.net/Articles/966133/

[27] https://www.usine-digitale.fr/article/les-geants-du-web-tirent-les-lecons-du-fiasco-heartbleed-et-creent-la-core-infrastructure-initiative.N257824

OpenSSL, a illustré la manière dont certains projets *open source*, à l'économie fragile, peuvent pourtant supporter des pans entiers de l'écosystème numérique. Au moment des faits, la bibliothèque n'était maintenue que par quatre développeurs, dont seulement un à temps plein. Des efforts sont faits, notamment au niveau français, pour limiter le problème[28].

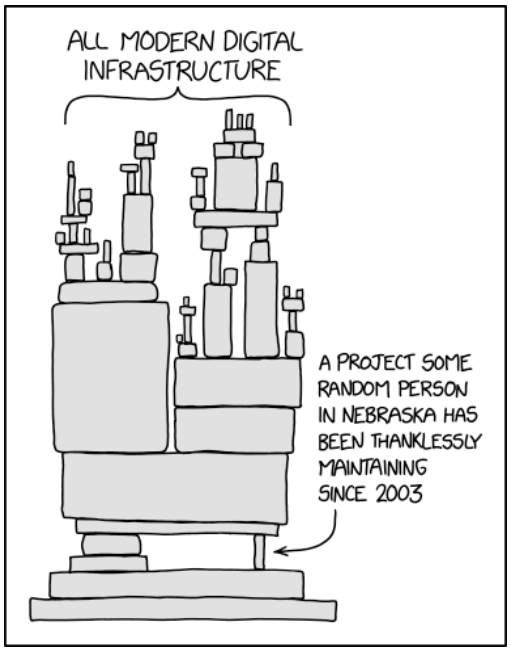

Figure 3 : XKCD n°2347

L'équation économique de l'*open source* dans le domaine de l'IAG

Dans l'ensemble, les différents modèles économiques de *l'open source* se retrouvent également dans le secteur de l'IA. L'État subventionne des acteurs et infrastructures nécessaires au développement des modèles (scikit-learn, Appel à projets Communs Numériques…) afin de dynamiser le secteur, des plateformes font le choix de l'ouverture de briques de code afin de bénéficier d'une aide communautaire (Meta et Google avec PyTorch et Tensorflow respectivement) et certains acteurs publient leurs modèles mais proposent en parallèle des interfaces de programmation (API) prêtes à l'utilisation (Stability.AI et leur modèle Stable Diffusion).

Le recours à l'*open source* dans le champ de l'IAG présente quelques spécificités. Premièrement, **certains grands modèles requièrent une infrastructure puissante et une expertise pour pouvoir être exécutés, ce qui les rend coûteux à l'usage, même en étant fondés sur des licences *open source*, ce qui peut réduire la différence entre modèles ouverts et fermés du point de vue des clients**. Une option est toutefois possible : recourir à des modèles à la fois *open source* et moins volumineux, plus simples à déployer, en faisant de potentiels compromis quant à la performance généralisée, mais qui peut être adéquate ou même meilleure sur des applications spécifiques.
Deuxièmement, la transparence sur les données d'entraînement est un sujet complexe, qui donne aujourd'hui lieu à beaucoup de travaux mais aussi parfois à

[28] https://code.gouv.fr/fr/bluehats/prix-bluehats/

des contentieux. Si certaines entreprises mettant à disposition des modèles propriétaires proposent à leurs clients de prendre à leur charge les contentieux liés à cette problématique[29], cela ne traite pas le problème en amont d'un usage illicite ni ne prémunit contre un éventuel risque réputationnel pour le client. De ce point de vue, les modèles *open source* permettent un contrôle et de déterminer si les données utilisées sont adéquates.
Troisièmement, le recours à certains modèles propriétaires peut emporter la participation à des boucles d'entraînement rétroactif, autrement dit les données des utilisateurs peuvent servir à améliorer le modèle et être utilisées pour fournir des réponses à d'autres clients. De ce point de vue, l'*open source* permet un meilleur contrôle des actions réalisées et de l'absence de fuite de données ou d'expertise de l'utilisateur ou de l'entreprise utilisatrice.

Écosystème de l'*open source* : une gouvernance à appréhender dans sa diversité

Étudier la question de la transparence sous le seul prisme du caractère lisible du code source amène à mettre sur un pied d'égalité d'une part des logiciels développés par de grosses entreprises qui gardent l'entière maîtrise des contributions et d'autre part des logiciels conçus collaborativement par une communauté de bénévoles. Or, i**l existe de très nombreuses formes de gouvernance suivant les projets et logiciels** car tous les projets sont eux aussi différents : par leur objet, mais aussi par leur modèle économique, leur histoire, leur communauté, leurs ambitions, leur positionnement idéologique, etc. Il n'existe pas une gouvernance qui pourrait convenir à tous les projets.

Cependant, cette diversité ne doit pas non plus amener à un relativisme qui gommerait complètement les différences entre gouvernances opaques et transparentes, entre gouvernances verrouillées et ouvertes, entre feuilles de route claires ou non. Ainsi, **certaines modalités de gouvernances peuvent porter des risques distincts**. L'un de ces risques est celui de l'utilisation de l'open-source comme d'un moyen d'augmenter l'adoption d'un produit, pour ensuite exploiter une situation de dépendance induite. Ainsi, certains développeurs de logiciels bénéficiant d'une large adoption tout en étant trop complexe pour être repris en interne par des tiers profitent de leur position pour imposer des conditions drastiques à leurs clients. Le projet Android, emblématique selon Google[30] de l'*open source* est *de facto* sous le contrôle de cette entreprise, qui a pu s'assurer la maîtrise complète de l'écosystème, au point d'avoir été sanctionné pour pratiques anticoncurrentielles par la Commission Européenne[31].

S'agissant du secteur de l'IA, la question de la gouvernance est également centrale, comme le montrent les débats actuels concernant la régulation, la transparence, voire la vocation politique quasi éditoriale de certains modèles de LLM. **La qualité d'*open source* n'apparaissant pas suffisante pour prémunir de dérives, il semble nécessaire de rester particulièrement vigilant à la forme de gouvernance qui peut sous-tendre les projets *open source* dans le domaine de l'IAG.**

[29] https://www.lesechos.fr/tech-medias/intelligence-artificielle/ia-et-copyright-microsoft-sengage-a-payer-les-frais-de-justice-de-ses-clients-1976531

[30] https://techcrunch.com/2011/05/10/andy-rubin-on-androids-openness-light-on-community-heavy-on-open-source/

[31] https://curia.europa.eu/jcms/upload/docs/application/pdf/2022-09/cp220147fr.pdf

ÉCOSYSTÈME DE L'INTELLIGENCE ARTIFICIELLE : VERS QUELLE OUVERTURE ?

La question de l'ouverture des modèles d'IA est cruciale dans le contexte de forte innovation que nous observons, afin d'éviter que le foisonnement actuel ne soit suivi par un mouvement de concentration au profit de quelques acteurs devenus incontournables.

Vers une réelle transparence, au-delà des effets d'annonce ?

Le qualificatif d'*open source* peut être un enjeu de communication pour certains acteurs. En ce sens, le recours aux qualificatifs « *open source* », « ouvert » ou encore « transparent » lorsqu'on traite de modèles d'IAG peut constituer un argument *marketing*. Si *l'open source* désigne comme cela a été vu un dispositif de licences permettant la libre utilisation, redistribution (y compris commerciale) et modification du contenu concerné, la transparence désigne des pratiques très variables de la part des entreprises et beaucoup moins encadrées.

Dans la lignée de son index de comparaison des performances des modèles de langage[32], le Centre pour la Recherche sur les Modèles de Fondation (*Center for Research on Foundation Models*, CRFM) de l'université de Stanford propose un index de transparence des modèles de fondation, le FMTI[33] (*Foundation Model Transparency Index*). Il a cependant été critiqué par certains acteurs, notamment universitaires : certains indicateurs ne seraient pas neutres et la méthode d'agrégation et de pondération semble favoriser les modèles commerciaux. Par exemple, GPT-4 et Bloom obtiennent des scores très proches alors que Bloom est transparent sur les données, l'infrastructure et la méthode d'entraînement, que l'architecture et les poids du modèle sont disponibles gratuitement et que l'évaluation est détaillée sur des jeux de données connus. A l'inverse, GPT-4 n'est transparent sur aucun de ces éléments et les évaluations publiées ne sont pas reproductibles (aucun accès direct au modèle sans filtres en entrée ou en sortie n'est disponible).

Le choix d'indicateurs de transparence dépend de l'objectif poursuivi. Les indicateurs de transparence « en aval » du FMTI répondent aux exigences de certains acteurs économiques : sûreté et conformité du modèle, mises à jour régulières, politiques d'utilisation. En revanche, les indicateurs de transparence « en amont » et quant au modèle semblent mieux correspondre aux attentes du régulateur, de la communauté de la recherche, voire des citoyens : possibilité de consulter les données d'entraînement, reproductibilité, enjeux sociaux et climatiques de la collecte de données et de l'entraînement. Ainsi, **la « transparence » des modèles d'IAG devrait systématiquement être évaluée à l'aune des objectifs poursuivis (utilisateur, régulateur, etc), et ne saurait constituer une unique métrique.**

Vers une plus grande conformité des bases d'entraînement ?

Les jeux de données d'entraînement des modèles d'IAG (grands modèles de langages et modèles de diffusion notamment) contiennent souvent des données protégées par le droit d'auteur ou le RGPD. Selon OpenAI, il serait même impossible de concevoir des modèles d'IAG à l'état de l'art sans données protégées

[32] https://crfm.stanford.edu/helm/latest/
[33] https://crfm.stanford.edu/fmti/

par le droit d'auteur[34]. Bien que les chercheurs et les entreprises à l'origine de ces modèles appliquent des filtres afin de limiter la réutilisation inappropriée de données, ces seuls filtres ne permettent pas de garantir la conformité des jeux d'entraînement[35] et des fuites peuvent même advenir lors de l'utilisation ultérieure des modèles[36].

Il est vrai que garantir la conformité d'ensemble de données est complexe et plus particulièrement pour certaines sources de données, comme le *Common Crawl*[37], une collection de plus de 50 milliards de pages récoltées sur internet régulièrement à l'aide de *crawlers*. Ce jeu de données a peu de chances d'être conforme : les pages web publiques qui sont collectées peuvent contenir des données personnelles ou propriétaires. C'est toutefois parce qu'il est disponible publiquement et visible par chacun qu'il est possible de constater qu'il contient ces données[38], ce qui permet une forme de contrôle et permet au moins théoriquement d'envisager une mise en conformité.

ZOOM SUR… LES *CRAWLERS*

Les *crawlers* sont des robots qui consultent automatiquement un maximum de pages internet accessibles publiquement afin de les indexer (pour les moteurs de recherche) ou de collecter les contenus présents sur ces pages. Il est possible d'indiquer à ces robots via un fichier spécifique (robots.txt) de ne pas collecter certaines pages de données, mais il est nécessaire que les auteurs et éditeurs aient connaissance de ce protocole. Il peut aussi être difficile pour un éditeur de contenus de recenser exhaustivement tous les robots susceptibles de visiter son site et utilisés pour la collecte de données (d'autant que certains robots servent ou ont pu servir à la fois à indexer, donc à favoriser la visibilité du site sur les moteurs de recherche, et à collecter les données sans que l'éditeur ne puisse accepter l'un et refuser l'autre). Face à ces enjeux, le protocole TDMRep[39] vise à communiquer la politique de partage des contenus d'un site web à tous les acteurs qui collectent et traitent automatiquement des données pour l'intelligence artificielle. L'efficacité de ce protocole pour empêcher la collecte massive des données propriétaires par des acteurs de l'IA dépendra en grande partie de son adoption par les éditeurs de contenus.

L'application à venir du règlement européen sur l'IA (*AI Act*) a vocation à mieux encadrer les données d'entraînement de certains systèmes d'IA. L'un des « considérant » prévoit par exemple que les fournisseurs de modèles d'IA à usage général publieront un « résumé détaillé » de leurs données d'entraînement. Cette transparence peut apporter des garanties fortes en matière de conformité. En effet, lorsque :

- les données (ou la méthode utilisée pour les obtenir) sont publiques, il devient facile de vérifier la présence de données protégées ;

[34] https://committees.parliament.uk/writtenevidence/126981/pdf/

[35] https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse

[36] https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html

[37] https://commoncrawl.org/

[38] Le problème se pose également pour les images : des chercheurs de Stanford ont montré que le jeu de données LAION-5B, qui compile plus de 5 milliards d'images accessibles publiquement en ligne, contenait des contenus d'abus sexuels sur mineurs (cf. https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse). À nouveau, cet audit a pu être mené car ce jeu de données est open source.

[39] https://www.w3.org/2022/tdmrep/

- le modèle est public, il est plus facile d'auditer les éventuelles fuites de données[40].

Cependant, pour certains modèles, la publication des données utilisées pourrait présenter des risques (par exemple, les modèles utilisés dans le domaine médical). Aussi, la possibilité d'auditer certains jeux de données ou modèles ne témoigne pas de leur conformité : il est nécessaire que des acteurs conduisent ces audits et publient leurs résultats.

Vers de nouveaux risques liés à l'utilisation de l'IAG *open source* ?

L'ouverture des modèles peut faciliter certains mésusages. En effet, un modèle intégré dans un système d'IA peut bénéficier de filtres en entrée (*input*) et en sortie (*output*) afin de mitiger les utilisations néfastes ou illégales. Lorsque le modèle est publié et accessible à tous, il est possible pour des acteurs malveillants de s'en emparer à des fins détournées en enlevant ces filtres.

Aussi, en permettant à chacun de générer facilement des contenus artificiels en grande quantité, l'ouverture pourrait mener à des risques nouveaux : par exemple, les triches impossibles à prouver lors d'examens, la création de contenus politiquement polarisants et réalistes mais invérifiables même par des experts, ou l'inondation du web par des contenus de faible qualité.

Si ces modèles facilitent certaines pratiques illégales, ils n'en changent pas la nature ni l'appréhension : une atteinte au droit à l'image est punie par la loi que l'auteur ait eu recours à des méthodes algorithmiques ou non. C'est cependant le nombre potentiel d'infractions permises par ces méthodes qui pourrait constituer un risque nouveau.

Les régulations récentes visent à endiguer la génération et la propagation des faux contenus, et à punir leurs auteurs. Le projet de loi pour Sécuriser et Réguler l'Espace Numérique introduit une mention relative à l'emploi de techniques algorithmiques pour générer des montages diffamants dans le code pénal, et punit plus sévèrement les criminels lorsque la diffusion de ces montages se fait via des services de communication au public en ligne. Le Règlement sur l'Intelligence Artificielle oblige quant à lui les fournisseurs de systèmes d'IAG à ajouter une marque détectable par une machine dans les contenus générés (*watermarking*).

CONCLUSION

Les licences peuvent devenir un outil au service de la régulation. De la même façon qu'une licence a été créée pour la réutilisation des informations publiques de l'administration[41], de nouvelles licences pour les modèles sont à l'étude afin de permettre une meilleure adaptation au contexte de l'IAG. L'entraînement des

[40] Les méthodes d'inférence d'appartenance (ou *membership inference*) peuvent être une solution. Explorées par la recherche, elles consistent à évaluer la présence d'un ou plusieurs points de données (documents, images, etc.) dans les données d'entraînement d'un modèle d'IA. Par exemple, le New York Times a tenté d'inférer la présence de ses articles dans les données d'entraînement d'OpenAI (cf. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html). L'efficacité de ces techniques reste néanmoins contestée concernant les modèles génératifs récents (cf. https://arxiv.org/pdf/2402.07841.pdf et https://openaccess.thecvf.com/content/WACV2024/papers/Dubinski_Towards_More_Realistic_Membership_Inference_Attacks_on_Large_Diffusion_Models_WACV_2024_paper.pdf).

[41] https://www.etalab.gouv.fr/wp-content/uploads/2017/04/ETALAB-Licence-Ouverte-v2.0.pdf

modèles d'IAG nécessite de grands volumes de données de qualité. Parmi ces ressources, celles mises à disposition par des acteurs publics, académiques ou subventionnés sont souvent d'une grande valeur pour l'élaboration de modèles d'IAG. Aussi, elles pourraient bénéficier de nouvelles licences, qui pourraient par exemple mentionner des conditions particulières de mise à disposition des poids ou encore interdire les usages illégaux en droit français. De telles contraintes feraient en revanche sortir ces licences de la définition stricte du domaine *open source* tel que défini par l'OSI (*Open Source Initiative*) et l'OKF (*Open Knowledge Foundation*).

Afin d'apprécier le degré d'ouverture des modèles d'IAG de façon souveraine, des travaux autour des métriques de transparence les plus pertinentes pour les acteurs publics et les régulateurs pourraient être nécessaires. Ces métriques devraient être alignées avec les objectifs de la stratégie d'accélération « Intelligence Artificielle », notamment le soutien à l'IA de confiance, en prenant en compte des critères tels que la licence utilisée ou des éléments sur la gouvernance des producteurs d'IAG.

Du point de vue de la dynamique de l'offre et de la réponse aux besoins des utilisateurs il ne faut par ailleurs pas négliger l'intérêt des modèles à la fois ouverts et de petite taille, afin que ceux-ci puissent être ré-utilisés et maintenus à un niveau de performance adéquat, évitant une trop forte dépendance en cas de concentration et de hausse coordonnée des prix. Du point de vue des grands modèles, le maintien d'une recherche de premier plan avec des publications ouvertes contribue également grandement à diminuer les risques évoqués.

Pour ce numéro dédié à l'usage des licences et à la transparence dans le domaine de l'IAG, le PEReN adresse tous ses remerciements à la mission Logiciels libres de la Direction interministérielle du numérique (Dinum) et à la Direction de projets – Intelligence artificielle du Service de l'économie numérique de la Direction générale des entreprises (Bercy) pour leur relecture précieuse, et à la mission Appui au patrimoine immatériel de l'État (APIE) de la Direction des affaires juridiques de Bercy pour nos échanges sur les licences libres.

La collection « Éclairage sur… » du PEReN propose, dans un format didactique, des éléments d'analyse techniques sur des thèmes liés à la régulation des plateformes numériques. Retrouvez l'ensemble des numéros parus à l'adresse www.peren.gouv.fr/publications/

Dépôt légal : Octobre 2022
ISSN (en ligne): 2824-8201

Service à compétence nationale placé sous la tutelle des Ministres de l'économie, du numérique et de la Culture, le Pôle d'expertise de la régulation numérique (PEReN) fournit, aux services de l'État et autorités administratives intervenant dans la régulation des plateformes numériques, une expertise technique dans les domaines du traitement des données, des data sciences et des procédés algorithmiques. Il s'investit également dans des projets de recherche en science des données à caractère exploratoire ou scientifique.

PEReN – 120 rue de Bercy, 75572 Paris Cedex 12 - contact.peren@finances.gouv.fr