Note d'analyse

Les pratiques *open source* en intelligence artificielle

CNIL.
COMMISSION NATIONALE
INFORMATIQUE & LIBERTÉS

Les pratiques observées dans le domaine de l'IA nécessitent généralement de mobiliser de nombreuses ressources tout au long du processus de développement d'un système, que peu d'acteurs contrôlent intégralement. Un développeur utilisera généralement une bibliothèque en source ouverte (ou OS pour *open source* dans la suite) telle que Tensorflow[1] ou PyTorch[2] afin de créer un nouveau modèle, ou bien il aura recours à la bibliothèque Transformers[3] pour télécharger un modèle pré-entraîné. S'il ne dispose pas de sa propre base de données, il choisira de télécharger des données depuis un site communautaire tel que Kaggle[4], de profiter de la diffusion de bases de données par une équipe universitaire comme le propose l'Université de Californie Irvine sur son site *Machine Learning Repository*[5], ou encore d'utiliser des données diffusées par les services publics, par exemple depuis le catalogue de *datasets* de data.gouv.fr[6]. Il pourra s'assurer depuis les plateformes communautaires comme Github[7] ou HuggingFace[8] que les outils, modèles et données qu'il a téléchargés ont été revus par des tiers, et qu'ils ne présentent pas de défauts critiques. Enfin, afin d'analyser les résultats qu'il obtient, il pourra les comparer à ceux obtenus par d'autre chercheurs et publiés dans une publication scientifique, avant de contribuer à son tour à cette communauté en publiant ses propres travaux.

Ce parcours démontre que le domaine de l'IA repose sur un écosystème fondé sur la diffusion d'outils et de connaissances aux bénéfices nombreux. Les acteurs du secteur ne contribuent toutefois pas tous de la même manière à la communauté, et des rapports de pouvoir se dessinent, tout comme certains risques liés à la confidentialité des données publiées. Cette note propose d'analyser les bénéfices et risques des pratiques de diffusion en OS dans le domaine de l'IA dans l'objectif d'identifier certaines bonnes pratiques, en commençant par une clarification de ce qui ici entendu par l'*open source* en IA.

1 Qu'entend-t-on par l'ouverture d'un modèle d'IA ?

En informatique, la notion d' « *open source* » se confond souvent avec celle de « logiciel libre » bien que le sens soit légèrement différent. Le « logiciel libre » est un logiciel offrant 4 libertés absolues et essentielles à ses utilisateurs :

- liberté d'utilisation du programme ;
- liberté d'étudier le code source du programme ;
- liberté de modifier le programme ;
- liberté de distribuer des copies du programme original ou modifié.

Cette notion permet généralement de mettre en œuvre ces libertés en définissant les conditions d'usage à attacher à l'utilisation d'un logiciel, qui sont généralement incluses dans une « licence ». La fondation Open Source définit ainsi 10 critères à respecter[9] : liberté de redistribution, accès au code source, autorisation de modification, respect de l'intégrité du logiciel original, absence de discriminations envers des personnes ou des secteurs pour la réutilisation, absence de spécificité d'usage ou de restrictions pour différentes raisons, la transmission de la licence lors des redistribution du logiciel, l'absence de spécificité de la licence au produit intégrant le logiciel, l'absence de restriction imposée par la licence sur des logiciels tiers et la licence doit être agnostique de la technologie utilisée pour la distribution du logiciel.

[1] Site web de TensorFlow (en anglais), URL : https://www.tensorflow.org/

[2] Site web de PyTorch (en anglais), URL : https://pytorch.org/

[3] Bibliothèque Transformers, Hugging Face (en anglais), URL : https://huggingface.co/docs/transformers/index

[4] Site web de Kaggle (en anglais), URL : https://www.kaggle.com/

[5] « *Machine Learning Repository* » (en anglais), Université Irvine de Californie, URL : https://archive.ics.uci.edu/

[6] Catalogue des datasets de data.gouv.fr pour le Machine Learning, data.gouv.fr, URL : https://www.data.gouv.fr/fr/pages/donnees_apprentissage-automatique/

[7] Site web de GitHub (en anglais), URL : https://github.com/

[8] Site web de Hugging Face (en anglais), URL : https://huggingface.co/

[9] « *The Open Source Definition (annoted)* » (en anglais), opensource.org, URL : https://opensource.org/definition-annotated/

Le règlement IA européen, tel qu'adopté par le Parlement Européen[10], ne donne pas de définition de l'OS, mais cible certaines catégories de licences au considérant 102 : « Les logiciels et les données, y compris les modèles, publiés dans le cadre d'une licence libre et ouverte grâce à laquelle ils peuvent être partagés librement et qui permet aux utilisateurs de librement consulter, utiliser, modifier et redistribuer ces logiciels et données ou leurs versions modifiées». Le législateur considère que les IA à usage général vérifient un haut niveau de transparence et d'ouverture lorsque « leurs paramètres, y compris les poids, les informations sur l'architecture du modèle et les informations sur l'utilisation du modèle, sont rendus publics ». Il ajoute « La licence devrait également être considérée comme libre et ouverte lorsqu'elle permet aux utilisateurs d'exploiter, de copier, de distribuer, d'étudier, de modifier et d'améliorer les logiciels et les données, y compris les modèles, à condition que le fournisseur initial du modèle soit crédité et que les conditions de distribution identiques ou comparables soient respectées. ».

L'expression fréquemment employée d'« IA ouverte » n'est pas clairement définie, bien que l'Open Source Initiative se soit récemment attelée à cette tâche[11]. L'ouverture dans le domaine de l'IA ne désigne généralement pas la publication du code source lié à l'utilisation ou au développement d'un modèle, bien que cela puisse en faire partie, mais plutôt la publication du modèle et des poids, ou paramètres, qui le constituent.

Concrètement, l' « IA ouverte » vise plusieurs types de pratiques dont une étude menée par Liesenfeld et al., 2023[12] réalise une cartographie pour les modèles de langages (LLM) résumée dans un tableau éclairant[13]. L'étude réalise une classification de plusieurs LLM sur la base de plusieurs critères :

- **La disponibilité des éléments du modèle**, qui peut être associée à la publication du code, des données d'entraînement, des poids du modèle, des données utilisées pour l'apprentissage par renforcement (ou RLHF), des poids correspondants à l'apprentissage par renforcement, ou encore à la licence d'utilisation ;
- **La documentation du modèle**, qui peut être associée à la documentation du code, de l'architecture du modèle, à l'existence d'une publication (ou d'un projet), d'une fiche descriptive du modèle, ou encore d'une fiche descriptive des données ;
- **L'accès au modèle**, qui se matérialise par la publication d'une bibliothèque, ou la production d'une API permettant de faire des requêtes au modèle.

Avec cette grille de lecture, il apparaît qu'un grand nombre de modèles sont qualifiés de modèles « ouverts » alors que leurs conditions d'accès sont fondamentalement différentes, comme l'illustre la très grande diversité de licences d'utilisations que l'on peut rencontrer[14]. Si cette qualification semble parfois relever d'un argument de publicité visant à prouver le caractère éthique du concepteur[15], le non-respect des critères établis par Liesenfeld n'indique pas pour autant un manque de volonté de la part du concepteur. En effet, certains des critères listés semblent particulièrement difficiles à mettre en œuvre, en particulier pour une équipe de recherche ou pour une TPE ou PME. La production d'une API ou d'une interface permettant d'utiliser le modèle par exemple nécessitent des efforts coûteux en raison des coûts d'hébergement du modèle et des tâches de support qu'elles nécessitent. De plus, Liesenfeld relève que certaines pratiques sont rarement mises en œuvre ou de manière incomplète, comme :

- la documentation du modèle ou des données, lorsque le système conçu hérite de la documentation incomplète du modèle sur la base duquel il a été conçu (par paramétrage, ou *fine-tuning*, par exemple) ;
- la publication des données utilisées pour le RLHF en raison du coût qu'à pu représenter leur collecte et leur annotation ;

---

[10] Législation sur l'intelligence artificielle, 13 mars 2024, Parlement Européen, https://www.europarl.europa.eu/RegData/seance_pleniere/textes_adoptes/definitif/2024/03-13/0138/P9_TA(2024)0138_FR.pdf

[11] « *Open Source AI Deep Dive* » (en anglais), opensource.org, URL : https://opensource.org/deepdive/

[12] Liesenfeld, A. & al., 2023, « *Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators* » (en anglais), URL : https://dl.acm.org/doi/10.1145/3571884.3604316

[13] « *Opening up ChatGPT: tracking openness of instruction-tuned LLMs* » (en anglais), URL : https://opening-up-chatgpt.github.io/

[14] « Licenses » (en anglais), opensource.org, URL : https://opensource.org/licenses/

[15] « Le marketing de l'IA « ouverte », 6 septembre 2023, Next, URL : https://www.nextinpact.com/article/72321/le-marketing-ia-ouverte

- la publication d'un article dans un journal scientifique soumis à une revue par les pairs, qui est souvent remplacée par la publication d'un article de blog en pratique.

Pour plus de détails sur les catégorisations des licences et les modèles économiques de l'open source, la note du PEREN « Eclairage sur… Open source et IA : des synergies à repenser »[16] fournit des éléments détaillés.

Dans ce qui suit, les modèles dont le seul accès est fourni via une API ou une interface utilisateur sont exclus du périmètre (comme c'est le cas de GPT3.5 par exemple, qui n'est accessible que via l'outil ChatGPT). En revanche, toutes les pratiques de l'OS sont considérées afin d'en distinguer les intérêts et inconvénients.

Opening up ChatGPT: tracking openness of instruction-tuned LLMs

Liesenfeld, A., Lopez, A. & Dingemanse, M. 2023. "Opening up ChatGPT: Tracking Openness, Transparency, and Accountability in Instruction-Tuned Text Generators." In *CUI '23: Proceedings of the 5th International Conference on Conversational User Interfaces*. July 19-21, Eindhoven. doi: 10.1145/3571884.3604316 (PDF).

There is a growing amount of instruction-tuned text generators billing themselves as 'open source'. How open are they really? ACM paper PDF repo

Project	Availability						Documentation						Access	
(maker, bases, URL)	Open code	LLM data	LLM weights	RL data	RL weights	License	Code	Architecture	Preprint	Paper	Modelcard	Datasheet	Package	API
BLOOMZ	✓	✓	✓	✓	~	~	✓	✓	✓	✓	✓	✓	✗	✓
bigscience-workshop	LLM base: BLOOMZ, mT0			RL base: xP3										§
Mistral 7B-Instruct	~	✗	✓	✗	~	✓	✗	~	~	✗	✗	✗	~	✓
Mistral AI	LLM base: unclear			RL base: unspecified										§
Falcon-40B-instruct	✗	~	✓	~	✗	✓	✗	~	~	✗	~	✗	✗	✗
Technology Innovation Ins…	LLM base: Falcon 40B			RL base: Baize (synthetic)										§
Stable Beluga 2	✗	✗	~	✗	✓	~	✗	~	~	✗	~	✗	✗	~
Stability AI	LLM base: LLaMA2			RL base: Orca-style (synthetic)										§
Stanford Alpaca	✓	✗	~	~	~	✗	~	✓	✗	✗	✗	✗	✗	✗
Stanford University CRFM	LLM base: LLaMA			RL base: Self-Instruct (synthetic)										§
Koala 13B	✓	~	~	~	✗	~	~	~	✗	✗	✗	✗	✗	✗
BAIR	LLM base: LLaMA 13B			RL base: HC3, ShareGPT, alpaca (synt…										§
Falcon-180B-chat	✗	~	~	~	~	✗	✗	~	~	✗	~	✗	✗	✗
Technology Innovation Ins…	LLM base: Falcon 180B			RL base: OpenPlatypus, Ultrachat, Air…										§
Orca 2	✗	✗	~	✗	✓	✗	✗	~	~	✗	~	✗	✗	~
Microsoft Research	LLM base: LLaMA2			RL base: FLAN, Math, undisclosed (sy…										§
LLaMA2 Chat	✗	✗	~	✗	~	✗	✗	~	~	✗	~	✗	✗	~
Facebook Research	LLM base: LLaMA2			RL base: Meta, StackExchange, Anthr…										§
Solar 70B	✗	✗	~	✗	~	✗	✗	✗	✗	✗	~	✗	✗	~
Upstage AI	LLM base: LLaMA2			RL base: Orca-style, Alpaca-style										§
Xwin-LM	✗	✗	~	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	~
Xwin-LM	LLM base: LLaMA2			RL base: unknown										§
ChatGPT	✗	✗	✗	✗	✗	✗	✗	✗	~	✗	✗	✗	✗	✗
OpenAI	LLM base: GPT 3.5			RL base: Instruct-GPT										§

Figure 1 - Extrait du tableau récapitulatif de Liesenfeld avec les modèles les plus utilisés aujourd'hui par la communauté.
***Source :** Liesenfeld, A. & al., « Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators », dl.acm.org ; **DOI :** /10.1145/3571884.3604316*

[16] *Open source et IA : des synergies à repenser ?* (PDF, 654 ko), peren.gouv.fr, URL : https://www.peren.gouv.fr/rapports/2024-04-03_Eclairage%20sur_OpenSource-IAG_FR.pdf

2 Pourquoi ce modèle est-il considéré bénéfique ?

Plusieurs entreprises comme Mistral AI, dans une intervention du fondateur de l'entreprise Arthur Mensch à l'évènement IMAgine Day IA Génératives[17], ou encore Hugging Face[18] ont communiqué sur les bienfaits de l'ouverture des modèles d'IA pour les individus et pour la société dans son ensemble. Plusieurs arguments ressortent.

2.1 Fédérer une communauté autour d'un outil de référence

Pour les entreprises diffusant des modèles d'IA, les bénéfices attendus sont multiples.

Tout d'abord, en diffusant un modèle, son concepteur permet sa réutilisation, mais il bénéficie également des contributions de la communauté. Ses membres peuvent réaliser un audit du modèle, l'améliorer ou proposer de nouvelles fonctionnalités. Ces modifications sont parfois directement intégrées dans de nouveaux produits par le concepteur du modèle.

De plus, en permettant son utilisation, la diffusion du modèle facilite son adoption par la communauté OS et par d'autres entreprises, établissant d'une part le rôle de l'entreprise comme référence dans le domaine, et créant parfois une relation de dépendance aux produits de l'entreprise lorsqu'ils ne sont pas interopérables avec les autres systèmes existants.

Enfin, en diffusant le modèle, en particulier quand celui-ci ouvre la porte à des utilisations inédites, l'attrait pour l'IA et son acceptabilité générale augmentent, ce qui peut avoir un impact sur le marché pour les produits intégrant des modèles d'IA. L'ouverture du modèle permet aussi de démultiplier les usages pertinents et de mieux cibler la stratégie commerciale par exemple.

Les bénéfices de l'OS pour l'entreprise semblent ainsi établis. Toutefois, il faut souligner que cela peut également représenter un coût. Ce coût est, d'une part, opérationnel, en raison du temps et de l'investissement que demandent le maintien des outils diffusés et le support auprès des utilisateurs et contributeurs. La motivation des experts pour contribuer aux biens publics digitaux pouvant s'avérer limitée dans certains cas, comme démontré par Chen et al., 2020[19] dans le cas des contenus publiés sur Wikipédia. D'autre part, il peut être lié à une perte d'opportunité, puisqu'en diffusant un modèle, son concepteur peut alimenter la concurrence.

2.2 Stimuler l'innovation et la productivité

Bien que les acteurs cités plus haut proviennent du secteur privé, les conclusions qu'ils tirent sont partagées par certains acteurs publics. En particulier, une étude de la Commission Européenne[20] sur l'impact du modèle OS des logiciels et du matériel vient en confirmer certaines des conclusions.

D'après cette étude, le modèle OS permettrait aux développeurs d'apprendre les uns des autres en contribuant aux projets en OS (p. 44), de stimuler la croissance et l'emploi en ayant des impacts sur la productivité et la compétitivité des entreprises et au niveau international (p. 175), ou encore d'augmenter le PIB des états membres (p. 202). Toutefois, elle souligne également que la stimulation de l'innovation pourrait se limiter à un bénéfice sur la création de jeunes pousses (p. 175), et manquer d'effets au niveau macro-économique (p. 212). Le PIB n'est pas forcément le meilleur indicateur pour mesurer le gain apporté par l'IA *open source* néanmoins.

En effet, pour Philippe Aghion (*Repenser la croissance économique*, 2016), le système de comptabilité nationale ne parvient pas à intégrer l'impact de la révolution technologique actuelle sur la productivité alors que cet impact est bien réel. L'exemple de la production d'appareil photographiques le démontre particulièrement. Les ventes d'appareils photos diminuant en raison de la capacité des smartphones à réaliser des photographies de qualité

---

[17] « IMAGINE DAY IA Génératives : passé et futur des IA génératives par Arthur MENSCH », 21 juin 2023, youtube.com, URL : https://youtu.be/zX5jGVfQXAs&t=1200

[18] « *Hugging Face CEO tells US House open-source AI is 'extremely aligned' with American interests* » (en anglais), 22 juin 2023, venturebeat.com, URL : https://venturebeat.com/ai/hugging-face-ceo-tells-us-house-open-source-ai-is-extremely-aligned-with-american-interests/?utm_source=substack&utm_medium=email

[19] Chen, Y. & al., 2020, « *Motivating Experts to Contribute to Digital Public Goods: A Personalized Field Experiment on Wikipedia* » (en anglais), URL : https://yanchen.people.si.umich.edu/papers/ExpertIdeas_2020_04.pdf

[20] « *Study about the impact of open source software and hardware on technological independence, competitiveness and innovation in the EU economy* » (en anglais), 2 septembre 2021, ec.europa.eu, URL : https://digital-strategy.ec.europa.eu/en/library/study-about-impact-open-source-software-and-hardware-technological-independence-competitiveness-and

équivalentes, la part de PIB qui leur correspond diminue également. Toutefois, celle correspondant aux smartphones augmente, mais un effort d'interprétation est nécessaire pour réaliser le lien entre les deux. De plus, ce lien est particulièrement difficile à quantifier en raison des multiples usages qu'offrent les smartphones.

Par ailleurs, pour Pierre Veltz (*La société hyperindustrielle*, 2017), l'impact de la révolution technologique a permis de faire mieux avec moins, sans pour autant faire plus. De fait, il y a un gain dans la qualité de la production qui n'est pas mesuré, ou même mesurable. L'apport de Wikipédia par exemple ne sera pas pris en compte dans le calcul de la production alors même que son impact est réel, et ce de la même façon que le seront la diffusion des modèles en OS.

Bien que cette vision globale semble partagée, on peut noter que la diffusion en OS possède certains avantages concrets, tels que :

- De permettre aux étudiants (autodidactes ou encadrés par une institution) d'apprendre et de pratiquer leurs enseignements sur des projets concrets ;
- De stimuler la conception et la publication en OS d'autres outils liés aux modèles diffusés (tels que LM-Eval[21], une bibliothèque permettant de tester la précision et la fiabilité des modèles de langage).
- De favoriser l'harmonisation des pratiques et l'interopérabilité des modèles et systèmes, ce qui permet notamment de les tester plus facilement, comme le démontre l'utilisation des *safetensors* par la plateforme Hugging Face[22] ;
- D'apporter des solutions à certains problèmes que les acteurs privés échouent et ne cherchent pas à résoudre, comme la capacité d'exécuter des LLM sur des ordiphones, ou de réaliser le paramétrage, ou *fine-tuning,* d'un modèle sur un ordinateur portable comme l'indique un document confidentiel de Google qui aurait été diffusé[23].

Le récent rapport de la commission de l'IA missionnée par le Premier Ministre[24] souligne également les bienfaits de l'OS en IA, en soulignant qu'il facilite le développement d'usages bienveillants, y compris les contre-mesures des usages malveillants et permet également d'élargir la base de ses contributeurs et ainsi à les rendre plus sûrs (page 25). Il formule une recommandation claire de « porter une stratégie de soutien à l'écosystème d'IA ouverte au niveau international en soutenant l'utilisation et le développement de systèmes d'IA ouverts et les capacités d'inspection et d'évaluation par des tiers » (recommandation n°4, page 60).

Ainsi, l'effet bénéfique de l'OS sur l'innovation semble avéré pour l'écosystème. Ce modèle pourrait également être bénéfique pour les systèmes eux-mêmes.

2.3 Augmenter la transparence et réduire les biais des systèmes d'IA

En publiant les modèles d'IA en OS, la possibilité est donnée pour explorer leurs capacités, limitations et éventuels défauts. Toutefois, ouvrir les poids du modèle semble ici insuffisant.

L'étude de la Commission Européenne citée plus haut tire des conclusions similaires à ce constat, et ajoute quelques réserves. L'ouverture du modèle serait un moyen de vérifier la présence de biais et d'en reprendre le contrôle (p. 306), cependant l'ouverture est ici insuffisante d'après l'étude, qui souligne que les données utilisées pour la conception du modèle devront être suffisamment fiables (p. 308). En effet, une distinction semble nécessaire entre :

- les systèmes dont seul le modèle est ouvert, permettant ainsi à la communauté de l'utiliser et de vérifier la présence de biais sur des cas concrets ;
- les systèmes dont les poids du modèle et les données d'entraînement sont ouverts, ce qui permet additionnellement de réaliser un audit des données elles-mêmes et de vérifier leur représentativité ;

---

[21] lm-evaluation-harness (en anglais), GitHub, URL : https://github.com/EleutherAI/lm-evaluation-harness

[22] « *Audit shows that safetensors is safe and ready to become the default* » (en anglais), 23 mai 2023, Hugging Face, URL : https://huggingface.co/blog/safetensors-security-audit

[23] « *Google "We Have No Moat, And Neither Does OpenAI"* » (en anglais), 4 mai 2023, semianalysis, URL : https://www.semianalysis.com/p/google-we-have-no-moat-and-neither

[24] *IA : notre ambition pour la France* (PDF, 4,8 Mo), Commission de l'intelligence artificielle, ministère de l'Économie, URL : https://www.economie.gouv.fr/files/files/directions_services/cge/commission-IA.pdf?v=1710339902

- les systèmes dont les poids du modèle et les données d'entraînement sont ouverts, et dont le processus de constitution de la base de données est suffisamment documenté. Cette condition supplémentaire permet alors d'identifier de potentiels biais dans les méthodes de collecte, d'annotation et de pré-traitement des données.

Ainsi, si l'ouverture des modèles semble être bénéfique à la réduction des biais, une gradation semble s'opérer selon le niveau d'ouverture de l'IA en question.

De plus, l'OS offre également certaines garanties en termes de transparence et d'éthique, comme la possibilité pour les individus, pour les pairs et pour le régulateur de vérifier la licéité de l'utilisation des données quand les sources de données utilisées pour la conception sont également ouvertes. Des outils, comme « *Have I Been Trained* ? »[25] qui permet de vérifier si une image est présente dans les jeux d'images Laion, pourraient être développés afin de faciliter l'exploration des bases de données ouvertes. De plus, comme souligné par Piktus et al., 2023[26], l'ouverture des modèles permet d'en vérifier les capacités et défauts, et en particulier d'étudier :

- si le modèle a mémorisé des données personnelles ou protégées lors de son apprentissage ;
- si le modèle possède des comportements qui pourraient être problématiques, comme la génération de contenus haineux ou l'incitation à des comportements dangereux comme ce qui a été observé sur certains LLM ;
- les performances du modèle, et de les comparer à celles annoncées par le concepteur et à celles de modèles similaires afin de sélectionner celui offrant les meilleures garanties en termes de protection de la vie privée.

Par ailleurs, les possibilités offertes par l'OS pour l'audit des modèles et des données pourraient être particulièrement utiles pour les régulateurs. En effet, cet accès n'est pas prévu par défaut par le règlement européen sur l'intelligence artificielle (RIA, actuellement en cours d'adoption[27]), et ne serait possible que lorsque l'audit sur la base des données et de la documentation fournies par le fournisseur est insuffisant (article 63). D'une manière plus générale, la transparence permise par l'OS pourrait apporter une plus grande facilité aux victimes de systèmes défectueux, plus à même de prouver la défaillance du système.

2.4 Faciliter la réutilisation de modèles

Pour finir, la publication en OS des modèles d'IA facilite leur réutilisation, notamment pour des projets pour lesquels les financements n'auraient pas été débloqués pour une conception depuis zéro, comme cela peut être le cas de projets humanitaires, associatifs, éducatifs, ou publics.

En pratique, cela peut être réalisé par le paramétrage, ou *fine-tuning*, d'un modèle de fondation, évitant ainsi le coût environnemental et financier de la conception ; ou encore par le développement de fonctionnalités ou d'interfaces par la communauté (tel que ce plugin permettant d'utiliser Stable Diffusion dans Photoshop[28]).

Les fruits de l'utilisation de ces outils pourraient constituer un bénéfice important pour la société dans son ensemble, d'autant que le coût incrémental de duplication du modèle ou des données est quasiment nul et ne prive pas l'utilisateur initial du bénéfice du système.

3 Pourquoi l'OS en IA pose-t-il question ?

3.1 Les risques pour la concurrence et la fraude

En termes de concurrence, la *Federal Trade Commission* – l'organisme public en charge de la protection des consommateurs aux Etats-Unis – a noté que ce modèle pouvait inciter les entreprises à recourir au modèle

[25] HaveIBeenTrained? (en anglais), URL : https://haveibeentrained.com/

[26] Piktus A. & al., « *The ROOTS Search Tool: Data Transparency for LLMs* » (en anglais), arXiv, URL : https://arxiv.org/pdf/2302.14035.pdf

[27] Proposition de règlement sur l'intelligence artificielle, EUR-Lex, URL : https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:52021PC0206

[28] Auto-Photoshop-StableDiffusion-Plugin (en anglais), GitHub, URL : https://github.com/AbdullahAlfaraj/Auto-Photoshop-StableDiffusion-Plugin

« *open first, close later* »[29] (« ouvrir d'abord, fermer ensuite »). Dans le cas de la diffusion de code ou d'un modèle, ce risque n'est pas aussi grand que pour l'accès à l'IA en tant que service, toutefois l'entreprise pourrait profiter des améliorations sur une version d'un modèle et choisir de ne pas diffuser des versions ultérieures plus performantes.

L'étude de la Commission Européenne précitée mentionne également le risque que les modèles diffusés en OS soient interceptés par des groupes tiers puis monétisés. Ce risque est problématique dans les cas où la licence de réutilisation ne le permettrait pas, ce qui pourrait être particulièrement difficile à prouver en raison de la possibilité pour un réutilisateur de modifier le modèle utilisé, et de la complexité croissante des licences OS comme le souligne la même étude (p. 215).

Enfin, bien que la diffusion en OS possède un coût pour le diffuseur et un avantage pour les contributeurs, mobiliser la communauté pour améliorer les modèles reviendrait à utiliser de la main d'œuvre gratuite pour des fins privées, comme le déplore l'employé de Google dans l'article cité plus haut[30]. En pratique, et notamment pour les modèles d'IA, la diffusion en open source est souvent soutenue par des organisations commerciales ou des fondations qui emploient directement les principaux contributeurs (comme c'est le cas de Mozilla pour Firefox, ou de HuggingFace pour Bloom).

3.2 Les risques règlementaires

Le règlement IA prévoit une exemption pour la diffusion de modèles en OS sauf s'ils sont mis sur le marché ou mis en service en tant que systèmes d'IA à haut risque (paragraphe 12 de l'article 2 du texte) ou s'il s'agit de modèles à usage général. Dans le cas d'un modèle à usage général, le fournisseur devra *a minima* mettre en place une politique visant à respecter le droit d'auteur de l'Union et rédiger et mettre à la disposition du public un résumé suffisamment détaillé du contenu utilisé pour l'entraînement selon un modèle fourni par l'AI Office (article 53). Cette exemption ne concerne toutefois pas les modèles à usage général présentant des risques systémiques auxquels des dispositions spécifiques s'appliquent, prévues à l'article 55 concernant l'évaluation, l'atténuation et la documentation de ces risques.

Cette exemption comporte certains risques liés par exemple à l'utilisation des modèles pour des usages à haut risque qui ne relèvent pas d'une mise sur le marché (comme un usage domestique, ou une utilisation malveillante facilitée par l'accès libre au modèle), ou encore pour le respect des exigences du RIA tout au long de la chaîne de responsabilité.

3.3 Les risques pour la sécurité

Tout d'abord, l'évolution des IA en *open source* est spectaculairement rapide, comme le démontre l'évolution rapide entre les modèles : deux semaines se sont écoulés entre les publications de LLaMA par Meta et d'Alpaca, une version ajustée par *fine-tuning* de LLaMa, par une équipe de Stanford, puis une semaine entre celles d'Alpaca et de Vicuna, une nouvelle version ajustée de LLaMa, par des équipes de l'université de Berkeley[31]. Pour cette raison, les vérifications attendues de la part de la communauté OS pourraient ne pas être réalisées en temps voulu, ce qui implique des risques sur la sécurité de ces systèmes. Certains acteurs admettent par ailleurs frontalement que les délais de développement extrêmement restreints les ont conduits à publier les modèles sans mettre en œuvre les mesures de sécurité adéquates, comme c'est le cas de l'entreprise Adept qui admet publier un LLM sans mesures de contrôle sur les sorties potentiellement toxiques du modèle « *Because this is a raw model release, we have not added further finetuning, postprocessing or sampling strategies to control for toxic outputs.* »[32].

De plus, la mise en OS comporte des risques sur la sécurité. Premièrement, la contribution libre aux poids du modèle introduit une voie d'accès pour des attaquants, cherchant à empoisonner les modèles et à y introduire des portes dérobées. Détecter ces tentatives d'attaque demande une vigilance particulière de la part des

---

[29] « *Generative AI Raises Competition Concerns* » (en anglais), 29 juin 2023, Federal Trade Commission, URL : https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns
Le système d'exploitation Android est une illustration majeure de ce principe : offert initialement comme un projet Open Source afin de conquérir des utilisateurs et des parts de marché, Google limite de plus en plus en les fonctionnalités disponibles ouvertement au profit de composants protégés qu'il fournit directement, gratuitement ou non.

[30] « *Google "We Have No Moat, And Neither Does OpenAI"* », 4 mai 2023, semianalysis, URL : https://www.semianalysis.com/p/google-we-have-no-moat-and-neither

[31] *Ibid.*

[32] « *Releasing Persimmon-8B* », 7 septembre 2023, Adept, https://www.adept.ai/blog/persimmon-8b

diffuseurs de modèles, et il n'est pas certain que les vérifications mises en place leur permettent de les identifier. Deuxièmement, la diffusion en OS présente par nature le risque que les failles du modèle, rendues apparentes, soient exploitées par des attaquants. Ce risque est souvent décrié en raison des améliorations que la communauté OS apporte à la sécurité des systèmes[33], toutefois, comme vu plus haut, la contribution de la communauté OS n'est pas toujours possible. Enfin, les modifications malveillantes des projets en OS comportent également un risque en termes de traçabilité, les contributeurs malveillants pouvant facilement dissimuler leur identité. Comme tout commun numérique, l'efficacité de ce mode de gestion repose en grande partie sur le dynamisme et la qualité des règles de la communauté qui s'organise pour le maintenir et le développer.

Une autre source de risque concerne le détournement des modèles diffusés en OS. Les modèles d'IA sont des outils puissants qui pourraient être utilisés efficacement pour des usages malveillants (comme pour l'envoi de messages de phishing grâce à des LLM ou des campagnes de désinformation alimentées par l'IA, à l'exemple de CounterCloud[34]). Ce risque est particulièrement vraisemblable, comme démontré par Qi et al., 2023[35], en raison de la facilité, accrue par la diffusion en OS, de supprimer ou de contourner les filtres et sécurités ajoutés aux systèmes. La désactivation de ces sécurités peut avoir lieu de manière volontaire (par une attaque spécifique) ou non (par le paramétrage ou fine-tuning du modèle) et notamment rendre les filtres portant sur les sorties des IA génératives caducs (lorsque ceux-ci sont diffusés avec les modèles ce qui n'est d'ailleurs pas toujours le cas). Ce risque est particulièrement mis en avant par les sociétés qui refusent de publier leurs modèles en OS (comme Google et OpenAI) et constitue un défi réglementaire pour le Règlement IA.

3.4 Les risques pour les personnes concernées et leurs droits

Enfin, l'OS pose certaines questions sur la confidentialité des données personnelles utilisées pour l'apprentissage et sur la possibilité pour les personnes d'exercer leurs droits sur les modèles et données diffusés.

Premièrement, alors qu'il a été prouvé qu'il est possible de reconstituer certaines données d'apprentissage à partir des modèles entraînés, quels risques fait porter la publication des modèles en OS sur la confidentialité de ces données ? Comme démontré par Fredrikson et al., 2015[36], il peut être possible de reconstruire un visage ayant servi à entraîner un modèle de reconnaissance faciale par un modèle d'attaque. Toutefois, ce risque peut également advenir par accident, notamment lorsque le modèle est sujet au surapprentissage (ou *overfitting*), pouvant conduire par exemple à la divulgation de données personnelles par la simple utilisation d'un LLM. Dans le cas de la mise en OS des modèles, quantifier l'ampleur et la vraisemblance de ce risque semble particulièrement difficile.

Deuxièmement, comment déterminer si les droits s'appliquent aux modèles ? Les filtres appliqués à la sortie des systèmes d'IA génératifs permettent de diminuer le risque de régurgitation de données personnelles *a posteriori* mais dans le cas de modèles en OS où les filtres sont parfois absents et peuvent être retirés il semble difficile de donner suite aux demandes par ce biais. Il pourrait alors être nécessaire de ré-entraîner le modèle et de mettre à jour la version publiée s'il était avéré que le modèle permet d'extraire des données personnelles concernant une personne ayant exercé son droit d'opposition (sous réserve que ce réentraînement ne porte pas une atteinte disproportionnée à d'autres droits et libertés fondamentaux).

Troisièmement, les demandes d'exercice de droits se propagent-elles dans la communauté OS ? En effet, si une personne exerce ses droits, cela pourrait se répercuter en cascade sur les contributeurs et utilisateurs du modèle. Dans la majorité des cas, seul le diffuseur du modèle sera en capacité d'apporter la modification nécessaire pour donner suite à une demande. Il faudrait alors indiquer ce changement aux utilisateurs et contributeurs et les inciter à mettre à jour leur version locale du modèle.

Finalement, une question subsiste sur l'applicabilité de l'exemption domestique prévue à l'article 2.2.c du RGPD pour les traitements conduits par des particuliers à un projet OS, notamment quand ils participent bénévolement et à titre personnel. Cette exemption pourrait diminuer leur niveau de responsabilité et rendre

---

[33] « *Is open source software a security threat?* » (en anglais), 19 juin 2019, BrightlineIT, URL : https://brightlineit.com/is-open-source-software-a-security-threat/

[34] « *Inside CounterCloud: A Fully Autonomous AI Disinformation System* » (en anglais), The Debrief, 16 août 2023, https://thedebrief.org/countercloud-ai-disinformation/

[35] Qi, X. & al., 2023, « *Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!* » (en anglais), 5 octobre 2023, https://arxiv.org/abs/2310.03693

[36] Fredrikson, M. & al., 2015, « *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures* », URL : https://dl.acm.org/doi/10.1145/2810103.2813677

inapplicable les droits des personnes concernées aux modèles et données téléchargées par les utilisateurs et contributeurs.

Enfin, si de nombreuses utilisations des projets OS ont des visées bénéfiques, il est admis que certaines d'entre elles sont simplement nuisibles, comme l'exemple qui suit.

Dans une publication (dont la probité des résultats a été largement critiquée), Kosinski et al., 2018[37] ont entraîné un modèle d'IA à reconnaître l'orientation sexuelle d'une personne à partir d'une simple photographie de son visage et en utilisant des outils OS comme VGG Face Descriptor[38]. Si le protocole expérimental et les résultats de l'étude ont fait l'objet de nombreuses critiques, la publication démontre néanmoins que de tels systèmes (même inefficaces) peuvent être conçus et des personnes pourraient être persuadées de leur efficacité.

Si les exemples de systèmes nuisibles par nature ne manquent pas, les conséquences négatives de l'utilisation de systèmes d'IA ne s'arrêtent pas ici. Les défaillances, biais, ou conditions d'utilisation de systèmes dont l'utilisation aurait dû être bénéfique, ont souvent été la cause de conséquences graves pour certains individus. La mathématicienne Cathy O'Neil liste plusieurs de ces systèmes dans un livre intitulé *Weapons of Math Destruction*. Parmi cette liste se trouve l'exemple de l'algorithme COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) développé par l'entreprise privée Northpointe Inc et utilisé par certains tribunaux américains afin d'estimer la probabilité de récidive d'un prévenu, dont il a été prouvé qu'il possédait un biais contre les personnes racisées.

4 Quelles pistes peuvent être trouvées ?

Ce qui précède permet de conclure que la mise en OS des modèles bénéficie sans aucun doute à la communauté OS, aux entreprises adeptes de cette pratique et par certains aspects, à la société dans son ensemble, toutefois dans certains cas, des risques ou effets négatifs peuvent persister. Pour anticiper ces cas particuliers et éviter de possibles conséquences négatives, des mesures complémentaires peuvent être apportées afin de garantir la protection des droits des personnes concernées par l'entraînement du modèle publié, ainsi que pour favoriser l'atteinte des bénéfices attendus par la publication.

Ces mesures peuvent être retrouvées dans la fiche focus « *open source* » sur la mobilisation de l'intérêt légitime comme base légale pour le développement d'un modèle d'IA. Cette fiche détaille les conditions selon lesquelles certaines pratiques d'OS bénéfiques peuvent contribuer à la mise en balance des intérêts du responsable de traitements et des personnes concernées. D'une manière plus générale, ces pratiques entrent dans l'évaluation de la nécessité et de la proportionnalité d'un traitement.

Attention

Les réflexions présentées dans l'encadré ci-dessous correspondent à des travaux actuellement conduits par la CNIL et leurs conclusions sont susceptibles d'évoluer. Les résultats de la consultation publique sur les fiches IA lancée en juin 2024, et notamment sur la mobilisation de l'intérêt légitime et sur le statut des modèles, pourront utilement contribuer à ces réflexions.

[37] Wang, Y., & Kosinski, M. 2017, September 7, « Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. » (en anglais), URL : https://doi.org/10.1037/pspa0000098

[38] VGG Face Descriptor (en anglais), URL : https://www.robots.ox.ac.uk/~vgg/software/vgg_face/

4.1 Pour éviter les pratiques contestables en termes de concurrence

Lorsque l'OS est utilisé à dessein par une entreprise afin de conquérir un marché, de bloquer l'interopérabilité, ou d'exploiter les contributions de la communauté, la mise en OS semble servir davantage l'entreprise en question que la communauté. Dans ces cas, plusieurs mesures permettraient de renforcer les bénéfices apportés à la communauté.

Premièrement, la mise en OS des poids du modèle pourrait être accompagnée de la diffusion d'autres éléments afin de réellement contribuer à l'écosystème OS. Parmi ces éléments, reprenant la liste proposée par Liesenfeld, figurent :

- La publication du code ayant permis d'entraîner le modèle et de celui permettant de l'utiliser dans des conditions opérationnelles sécurisées (en incluant par exemple des filtres visant à éviter la production de contenu haineux par un LLM) ;
- La publication des données d'entraînement, lorsque cette publication ne pose pas de risque en termes de propriété intellectuelle et de protection des données personnelles ; *a minima* la publication d'une description détaillée des données d'entraînement utilisées (sources, volume, types, etc.) – cf. *infra* ;
- L'utilisation d'une licence d'utilisation adaptée aux risques liés au modèle (les licences listées par l'Open Source Initiative peuvent servir de référence[39] bien que celles-ci ne permettent pas toujours la restriction des usages dangereux) ;
- La publication d'une documentation suffisamment fournie concernant le code, le modèle (comme les M*odel Cards* introduites par Mitchell et al., 2019[40] par exemple), ou encore d'une fiche descriptive des données (sur le modèle de ce que propose la CNIL dans sa fiche « Tenir compte de la protection des données dans la collecte et la gestion des données »[41] ou sur un autre modèle reconnu).

La publication d'outils favorisant l'accès au modèle, comme une API ou une bibliothèque, plus coûteuse à mettre en place, ne devraient pas être des exigences bien qu'il s'agisse généralement de pratiques bénéfiques. La publication de ces outils est souvent l'occasion pour leurs fournisseurs d'acquérir une position dominante sur le marché en ancrant l'usage de leurs produits dans les habitudes des développeurs. Cette publication pourrait être considérée comme une bonne pratique lorsqu'elle tient compte des enjeux d'interopérabilité et de disponibilité.

La publication pourrait s'accompagner d'un engagement du diffuseur sur sa volonté d'assurer une maintenance des éléments diffusés pour un temps suffisant pour assurer l'intérêt d'une réutilisation pour la communauté. Le diffuseur pourrait également s'engager à diffuser les versions ultérieures du modèle auxquelles la communauté aurait contribué. Enfin, le diffuseur pourrait publier son plan d'intégration des outils diffusés en OS dans ses propres produits afin de fournir à la communauté une visibilité sur la pérennité du projet, et sur les usages qui seront faits de leurs contributions.

4.2 Pour faciliter l'application du cadre règlementaire

Le Règlement IA introduit des exceptions pour l'OS bien que certaines obligations soient tout de même prévues. Cet entre-deux, bien qu'il tende à favoriser le développement de modèles OS, entraîne une difficulté pour les concepteurs de systèmes intégrant des modèles OS à appliquer le règlement.

Afin de garantir l'application du RIA tout au long de la chaîne de développement de systèmes d'IA utilisant des modèles OS, il pourrait être envisagé de responsabiliser davantage les utilisateurs de modèles OS en exigeant d'eux qu'ils puissent démontrer que le modèle respecte un certain niveau d'exigence. Cela pourrait avoir pour effet de forcer les utilisateurs de modèles OS à n'utiliser que les modèles les plus respectueux, et ainsi de conduire les diffuseurs de modèles OS à un plus haut niveau d'exigence pour promouvoir leurs travaux. De plus des travaux avec la communauté OS permettrait d'instaurer un socle de bonnes pratiques et d'identifier celles qui seraient plus contestables. Les travaux de normalisation pourraient également contribuer à cet effort.

4.3 Pour éviter les utilisations détournées ou nuisibles

Les risques liés à l'utilisation des modèles publiés étant de diverses natures, les mesures protectrices suivent également cette tendance.

Premièrement, afin de maîtriser les risques de discriminations liés à la présence de biais du modèle et de garantir que la revue par les pairs permette d'identifier ces risques, il devrait être envisagé de publier les données d'entraînement lors de la diffusion du modèle. Cette publication n'est pas toujours possible en raison

du risque lié à la présence de données personnelles dans le jeu ou d'enjeux de propriété intellectuelle. Dans ce cas, des alternatives pourraient être prévues comme la publication d'un jeu synthétique basé sur les données réelles, la publication d'une documentation intégrant des informations statistiques sur le jeu, ou encore la publication d'un sous-ensemble représentatif du jeu de données dans lequel l'absence de données personnelles a été vérifiée.

Deuxièmement, afin d'éviter les réutilisations détournées, l'utilisation d'une licence restrictive sur les réutilisations à un champ moins risqué semble être une bonne pratique, bien que cela n'empêche pas les réutilisations détournées malveillantes. Afin d'assurer la traçabilité des réutilisations du modèle, il pourrait être envisagé de restreindre l'accès au modèle à la transmission des coordonnées du réutilisateur (bien que cela soit contraire à l'esprit de l'OS, c'est déjà souvent le cas en pratique sur certaines plateformes et pour certains modèles tel que LLama disponible sur HuggingFace[42]). De plus, un tatouage numérique (ou *watermark*) pourrait être apposé au modèle afin de l'identifier lors d'un contrôle ou d'identifier ses sorties dans le cas de l'IA générative, et ainsi de constater un manquement aux termes de la licence de réutilisation (même si cette technique présente des limites)[43].

Afin d'éviter les réutilisations nuisibles, la diffusion même de certains modèles en OS pourrait être questionnée. Toutefois, l'exemple pris plus haut d'un modèle capable de déterminer l'orientation sexuelle d'un individu à partir de sa photo prouve que la diffusion en OS reste utile puisqu'elle aura permis de soumettre les résultats obtenus à une revue par les pairs. Dans ces cas, des mesures de traçabilité poussées pourraient être envisagées afin de conserver une visibilité sur les réutilisations.

Enfin, pour éviter que des modèles non sécurisés soient accessibles trop simplement, il conviendrait d'éviter la mise à disposition d'API si celle-ci n'est pas accompagnée de mesures de protection contre les réutilisations nuisibles. Ces mesures peuvent être techniques ou contractuelles mais suppose le maintien d'un suivi opérationnel. Concernant la mise à disposition de modèles sous forme d'une bibliothèque, des alertes détaillées, claires et directement accessibles sur la responsabilité engagée du réutilisateur semblent *a minima* devoir être incluses avec la bibliothèque.

4.4 Pour sécuriser les données

Tout d'abord, un risque existe pour un modèle d'IA de permette l'accès à des données personnelles utilisées lors de l'entraînement (par régurgitation ou suite à une attaque). Plusieurs pistes semblent possibles pour réduire ce risque.

Afin de limiter le risque de mémorisation de données personnelles, c'est-à-dire que les poids du modèle permettent de retrouver ces données, les mesures de pseudonymisation et d'anonymisation devraient en premier lieu être privilégiées. En effet, elles permettent de garantir l'absence de données identifiantes en aval, mais également de publier la base de données d'entraînement avec un niveau de risque moindre pour les individus.

Il pourrait être envisagé d'exiger pour l'IA générative qu'une analyse soit conduite sur la capacité d'un modèle à régurgiter des données, en réalisant des tests automatisés reposant sur des requêtes ciblant spécifiquement les personnes concernées. Cela prendrait par exemple la forme de requêtes textuelles telles que « Monsieur X habite au … » lorsque cette information existe effectivement dans la base d'apprentissage, ou encore une requête telle que « représente-moi une photo de Monsieur X » pour un système de génération d'images.

Les tests permettant de mesurer le risque d'une fuite de données personnelles suite à une attaque, par inférence d'appartenance ou par inversion du modèle par exemple, est plus complexe. Il pourrait être recommandé de conduire des compétitions du type « *bug bounty* » (également appelées « *red teaming* ») visant à identifier les vulnérabilités du modèle devant ce type d'attaques. Ces compétitions pourraient également permettre de vérifier la sécurité du système, et notamment l'absence de portes dérobées dans le

[39] « Licences, https://opensource.org/licenses/

[40] Mitchell, M. & al., 2019, « *Model Cards for Model Reporting* » (en anglais), URL : https://arxiv.org/pdf/1810.03993

[41] « IA : Tenir compte de la protection des données dans la collecte et la gestion des données », 8 avril 2024, CNIL, URL : https://www.cnil.fr/fr/tenir-compte-de-la-protection-des-donnees-dans-la-collecte-et-la-gestion-des-donnees

[42] « Meta Llama », Hugging Face, URL : https://huggingface.co/meta-llama

[43] « Panorama et perspectives pour les solutions de détection de contenus artificiels [1/2] », 27 octobre 2023, LINC, URL : https://linc.cnil.fr/panorama-et-perspectives-pour-les-solutions-de-detection-de-contenus-artificiels-12

modèle. En tout état de cause, l'existence d'une procédure à tenir en cas de violations de données est une exigence qui pourrait être rappelée aux fournisseurs de modèles OS.

4.5 Pour garantir la transparence et l'exercice des droits

Concernant la possibilité pour les personnes d'exercer leurs droits, celle-ci doit être prévue en amont, en mettant en œuvre une procédure de recueil des demandes lors de la diffusion du modèle, mais également en prévoyant des mesures techniques permettant d'une part de donner suite à une demande (comme des techniques de désapprentissage machine dans certains cas), et de lier juridiquement et d'informer les utilisateurs du modèle d'une demande d'autre part (par une clause ajoutée dans la licence d'utilisation, par des API ou par l'inscription à un canal d'information tenant un suivi des mises à jour du modèle, comme le prescrit la Recommandation de la CNIL sur l'utilisation d'API pour l'exercice des droits[44]). Les utilisateurs des modèles OS n'étant pas les plus à même de donner suite à une demande d'exercice de droits, le diffuseur du modèle devrait mettre en œuvre ce qui est en sa capacité pour que la demande soit prise en compte tout au long de la chaîne d'utilisation du modèle.

[44] « API : les recommandations de la CNIL sur le partage de données », 24 novembre 2023, CNIL, URL : https://www.cnil.fr/fr/api-les-recommandations-de-la-cnil-sur-le-partage-de-donnees