Est-ce légitime de récolter des pages Web pour entrainer des IA ?

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 18 septembre 2023

https://www.bortzmeyer.org/collecte-pour-l-ia.html

Plusieurs médias ont récemment publié des articles dénonçant GPTBot <https://platform.openai.com/docs/gptbot>, le ramassseur de pages Web de la société OpenAI. Ces articles sont souvent écrits dans un style dramatique, parlant par exemple de « pillage » du Web, ou de « cauchemar ». Je n'ai pas encore d'opinion ferme à ce sujet mais je peux quand même discuter quelques points.

La partie technique d'abord. Le GPTBot est documenté par la société OpenAI <https://platform.openai.com/docs/gptbot>. Il ramasse le contenu des pages Web, ce qui sert ensuite à entrainer un LLM comme celui utilisé par ChatGPT. Étant donné que le ramasseur ne paie pas les auteurs et hébergeurs des sites Web, et que le LLM ainsi entrainé sera utilisé dans des services potentiellement lucratifs, les médias dénoncent cette activité de ramassage comme illégitime. Voici par exemple un article de Télérama du 6 septembre à ce sujet :

Le débat n'est pas nouveau : des protestations avaient déjà été émises <https://www.pixelstech.net/article/1682104779-GitHub-Copilot-may-generate-code-containing-GPL-code> par certains programmeurs à propos du service Copilot de GitHub. Est-ce que Copilot peut s'entrainer sur les programmes stockés sur GitHub ? D'un côté, ces programmes sont en accès ouvert (et la grande majorité sous une licence libre), de l'autre, certaines licences utilisées imposent la réciprocité (c'est notamment le cas de la GPL), alors que les programmes développés avec Copilot, et Copilot lui-même, ne sont pas forcément sous une licence libre.

Notez que, comme documenté par OpenAI, si on n'aime pas le ramassage par le GPTBot, on peut utiliser le traditionnel `robots.txt` (RFC 9309 [1]) et il respectera ce désir. Cela n'empêchera évidemment pas les médias de protester, soit parce qu'ils voudraient être payés, soit parce qu'ils critiquent ce côté « autorisé par défaut ».

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc9309.txt

Alors, est-ce légitime ou pas d'entrainer un LLM avec des données qui sont accessibles publiquement ? Le problème est complexe, plus complexe que la présentation simpliste faite dans les médias. Rappelons d'abord que ce n'est pas parce qu'un contenu est en accès ouvert qu'on peut faire ce qu'on veut avec. Il vient avec une licence (et, s'il n'y en a pas d'explicite, c'est qu'on ne peut rien faire) et celle-ci limite souvent les utilisations. Ainsi, ce blog <https://www.bortzmeyer.org/> est sous licence GFDL et, si vous voulez en réutiliser le contenu, vous pouvez (licence libre) mais vous devez donner les mêmes droits à ceux qui recevront le travail fait à partir de ce blog (comme pour la GPL dans le cas du logiciel). Est-ce qu'un LLM typique respecte ceci ? L'une des difficultés de la question est que le LLM ne copie pas, il ne plagie pas, il produit un contenu original (parfois faux, dans le cas de ChatGPT <https://www.bortzmeyer.org/chatgpt-programmation.html>, mais c'est une autre histoire). Il ne viole donc pas le droit d'auteur. (Comme toujours, la réalité est plus complexe, des cas où Copilot a ressorti tel quel du code GPL ont déjà été cités <https://twitter.com/DocSparse/status/1581461734665367554>.)

Mais est-ce que ce contenu original est dérivé du contenu initial ? Beaucoup d'articles sur la question ont fait comme si la réponse était évidemment oui. Mais cela ne me convainc pas : le mode de fonctionnement d'un LLM fait justement qu'il ne réutilise pas directement son corpus d'entrainement (et cela a de nombreux inconvénients, comme l'impossibilité de sourcer les réponses et l'impossibilité d'expliquer comment il est arrivé à cette conclusion, deux des plus gros problèmes lorsqu'on utilise des IA génératives comme ChatGPT). Il n'est donc pas évident pour moi que la licence du texte original s'applique. Prenons une comparaison : une étudiante en informatique lit, pour apprendre, un certain nombres de livres (certainement sous une licence restrictive) et des programmes (dont certains sont sous une licence libre, et une partie de ceux-ci sous GPL). Elle va ensuite, dans son activité professionnelle, écrire des programmes, peut-être privateurs. Est-ce que c'est illégal vu que sa formation s'est appuyée sur des contenus dont la licence n'était pas compatible avec celle des programmes qu'elle écrit ? Tout créateur (j'ai pris l'exemple d'une programmeuse mais cela serait pareil avec un journaliste ou une écrivaine) a été « entrainé » avec des œuvres diverses.

Dans le cas d'un contenu sous une licence à réciprocité, comme mon blog, notez que le fait que les tribunaux étatsuniens semblent se diriger vers l'idée que le contenu produit par une IA n'est pas <https://www.hollywoodreporter.com/business/business-news/ai-works-not-copyrightable-studios-> "*copyrightable*" fait que la licence sera paradoxalement respectée.

Bon, et, donc, les arguments des médias ne m'ont pas convaincu ? L'un des problèmes est qu'il y a un conflit d'intérêts dans les médias, lorqu'ils écrivent sur des sujets qui concernent leur **business** ; ils sont censés à la fois rendre compte de la question tout en défendant leurs intérêts. Alors, certes, OpenAI est une entreprise à but lucratif, et pas forcément la plus sympathique, mais les médias sont aussi des entreprises commerciales, et il n'y a pas de raison, a priori, de privilégier un requin plutôt qu'un autre (l'IA brasse beaucoup d'argent et attire donc les envies).

Notons aussi que, dans le cas particulier de la francophonie, on verra sans doute les mêmes médias en langue française bloquer GPTBot et se plaindre ensuite que les IA ne traitent pas le français aussi bien que l'anglais…(Pour le logiciel libre, cité plus haut dans le cas de Copilot, un argument équivalent serait qu'exclure le code GPL des données d'entrainement réduirait la qualité du code produit, le code écrit sous les licences libres étant généralement bien meilleur.)

D'autres articles sur le même sujet ? La fondation Wikimedia a analysé la légitimité de ChatGPT <https://meta.wikimedia.org/wiki/Wikilegal/Copyright_Analysis_of_ChatGPT> à récolter leurs données.

Bref, la question me semble plus ouverte que ce qui est présenté unanimement dans les médias français. (Même la FSF considère le problème comme ouvert <https://www.fsf.org/news/publication-of-t

Pour l'instant, je n'ai donc pas ajouté le GPTBot au `robots.txt` (que je n'ai d'ailleurs pas). De toute façon, un examen des journaux ne me montre aucune visite de ce ramasseur, ce qui fait que vous ne retrouverez pas mon style dans les résultats de ChatGPT. Et vous, des avis ? Vous pouvez m'écrire (adresse en bas de cette page) ou bien suivre la discussion sur le fédivers <https://mastodon.gougere.fr/@bortzmeyer/111085936358151236> ou celle sur Twitter <https://twitter.com/bortzmeyer/status/1703735371123634384>.