In-depth analysis

Open source practices in artificial intelligence

CNIL.
COMMISSION NATIONALE
INFORMATIQUE & LIBERTÉS

Practices in the field of AI typically require the mobilisation of many resources throughout the development process, which few actors can wholly control. A developer will usually use an open source library (or OS for open source in the following) such as Tensorflow[1] or PyTorch[2] to create a new model, or they will use the Transformers library[3] to download a pre-trained model. If they don't have their own dataset, they will choose to download data from a community site such as Kaggle[4], to take advantage of the publication of datasets by an academic team as proposed by the University of California Irvine on its Machine Learning Repository[5], or to use data published by public services, for example from the datasets catalogue of data.gouv.fr[6]. They will be able to check from community platforms such as Github[7] or HuggingFace[8] that the tools, models and data they uploaded have been reviewed by third parties, and that they do not have critical flaws. Finally, in order to analyse the results they obtained, they will compare them with those obtained by other researchers and published in a scientific journal, before contributing in return to that community by publishing their own work.

This is striking evidence that the field of AI relies on an ecosystem based on public tools and knowledge with many benefits. However, not all actors in the sector contribute equally to the community, and power relationships are emerging, as are some risks related to the confidentiality of published data. This study proposes to analyse the benefits and risks of OS practices in the field of AI with the aim of identifying certain good practices, starting with a clarification of what is meant here by "open source AI".

1 What is meant by opening up AI models?

In IT, the concept of "open source" is often mistaken with that of "free software", although their meaning is slightly different. 'Free software' is software offering four absolute and essential freedoms to its users:

- freedom to use the program;
- freedom to study the source code of the program;
- freedom to modify the program;
- freedom to distribute copies of the original or modified program.

That concept generally makes it possible to implement those freedoms by defining the conditions of use of the software, which are generally included in a "licence". The Open Source Foundation thus defines 10 criteria to be met[9]: free redistribution, access to the source code, authorisation of modification, respect for the integrity of the original software, absence of discrimination against persons or sectors for re-use, absence of discriminations against fields of endeavor, transmission of the licence when redistributing the software, absence of specificity of the licence to the product integrating the software, absence of restriction imposed by the licence on third-party software and the licence must be agnostic of the technology used for the distribution of the software.

The European AI Act, as adopted by the European Parliament[10], does not provide a definition of OS AI, but targets certain categories of licences in recital 102: "*Software and data, including models, released under a free and open-source licence that allows them to be openly shared and where users can freely access, use, modify and redistribute them or modified versions thereof*". Additionally, the legislator considers that general purpose AIs verify a high level of transparency and openness when "*their parameters, including weights, the*

[1] TensorFlow website, URL: https://www.tensorflow.org/

[2] PyTorch website, URL: https://pytorch.org/

[3] Transformers Library, Hugging Face, URL: https://huggingface.co/docs/transformers/index

[4] Kaggle website, URL: https://www.kaggle.com/

[5] Machine *Learning Repository,* Irvine University of California, URL: https://archive.ics.uci.edu/

[6] Catalogue of datasets from data.gouv.fr for Machine Learning, data.gouv.fr, URL: https://www.data.gouv.fr/fr/pages/donnees_apprentissage-automatique/

[7] GitHub website, URL: https://github.com/

[8] Hugging Face website, URL: https://huggingface.co/

[9] "The Open Source Definition (annoted)", opensource.org, URL: https://opensource.org/definition-annotated/

[10] Artificial Intelligence Act, 13 March 2024, European Parliament, https://www.europarl.europa.eu/RegData/seance_pleniere/textes_adoptes/definitif/2024/03-13/0138/P9_TA(2024)0138_EN.pdf

information on the model architecture and the information on model usage, are made publicly available. The licence should be considered to be free and open-source also when it allows users to run, copy, distribute, study, change and improve software and data, including models under the condition that the original provider of the model is credited, the identical or comparable terms of distribution are respected."

More generally, the recurring term "open source AI" is not clearly defined, although the Open Source Initiative has recently taken up this task[11]. Openness in AI generally does not refer to the publication of the source code related to the use or development of a model, although this may be part of it, but rather to the publication of the model and the weights, or parameters, that constitute it.

In concrete terms, "open source AI" can refer to several types of practices mapped by a study conducted by Liesenfeld et al., 2023[12] for language models (LLM) and summarised in an illuminating table copied below[13]. The study performs a classification of several LLMs on the basis of several criteria:

- **The availability of model elements,** which may be associated with the publication of code, training data, model weights, data used for reinforcement learning from human feedback (or RLHF), weights corresponding to reinforcement learning, or to the licence used;
- **The documentation of the model,** which may be associated with the documentation of the code, the architecture of the model, the existence of a publication, a descriptive sheet of the model, or a descriptive sheet of the data;
- **Access to the model,** through a public library, or the provision of an API enabling queries to the model.

With this reading grid, it appears that a large number of models are described as 'open' while their conditions of access are fundamentally different, as illustrated by the very wide variety of user licences that can be encountered[14]. Although that classification sometimes seems to be part of an advertising argument intended to prove the ethics of the provider[15], failure to comply with the criteria established by Liesenfeld does not always, however, indicate a lack of will of the provider. Indeed, some of the criteria listed seem particularly difficult to implement, especially for a research team or for a VSE or SME. Producing an API or interface to use the model, for example, requires costly efforts due to host the computations on the model for inference and the support tasks involved. Moreover, Liesenfeld notes that certain practices are rarely implemented or incompletely, such as:

- documentation of the model or data, when the model or data were originally developed by another provider (by fine-tuning, for example);
- publication of the data used for RLHF because of the cost of collecting and annotating them;
- the publication of an article in a peer-reviewed scientific journal, which is often replaced by the publication of a blog post in practice.

For more details on licensing categorisations and open source business models, the PEReN's note "Eclairage sur... Open source et IA : des synergies à repenser ?"[16] is recommended.

In the following, models whose only access is provided via an API or user interface are excluded from the scope (as is the case with GPT3.5 for example, which is only accessible via the ChatGPT tool). On the other hand, all other practices of OS are considered in order to distinguish their interests and disadvantages.

[11] Open Source AI Deep Dive, opensource.org, URL: https://opensource.org/deepdive/

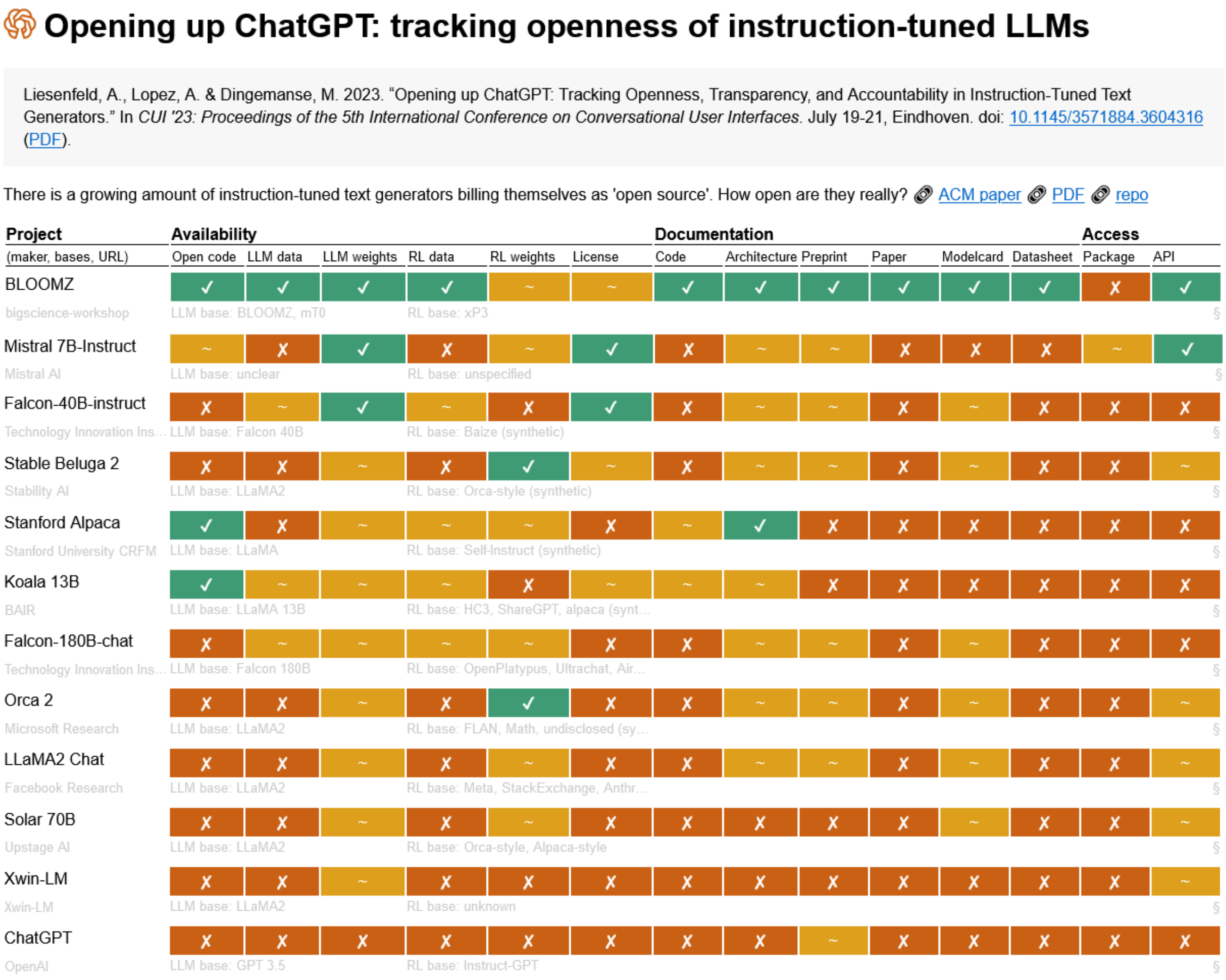
[12] Liesenfeld, A. & al., 2023, "Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators", URL: https://dl.acm.org/doi/10.1145/3571884.3604316

[13] "Opening up ChatGPT: tracking openness of instruction-tuned LLMs", URL: https://opening-up-chatgpt.github.io/

[14] "Licenses", opensource.org, URL: https://opensource.org/licenses/

[15] "Le marketing de l'IA ouverte", September 2023, NextInpact, URL: https://www.nextinpact.com/article/72321/le-marketing-ia-ouverte

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



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 -** Extract from the Liesenfeld summary table with the models most used by the community today. **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*

2 Why is this approach considered beneficial?

Several companies, such as Mistral AI, in a speech by the founder of the company Arthur Mensch at the "IMAgine Day IA Génératives" event[17], or Hugging Face[18], have communicated on the benefits of opening up AI models for individuals and for society as a whole. Several arguments stand out.

2.1 Building a community around a reference tool

For companies opening AI models, the expected benefits are multiple.

First of all, by opening a model, its provider allows its reuse, but it also benefits from the contributions of the community. Its members can audit the model, improve it or propose new functionalities. These modifications are sometimes directly integrated into new products by the designer of the model.

Moreover, by allowing its use, the publication of the model facilitates its adoption by the OS community and other companies, establishing on the one hand the role of the company as a reference in the field, and sometimes creating a relationship of dependence on the company's products when they are not interoperable with other existing systems.

Finally, by opening the model, especially when it opens the door to new uses, the attractiveness for AI and its general acceptability increase, which can have an impact on the market for products incorporating AI models. The openness of the model also makes it possible to multiply the relevant uses and to better target the commercial strategy, for example.

The benefits of OS AI for the company thus appear to be established. However, it should be noted that this can also represent a cost. This cost is, on the one hand, operational, due to the time and investment required to maintain public tools and to bring support to users and contributors. Experts' motivation to contribute to digital public goods may be limited in some cases, as demonstrated by Chen et al., 2020[19] in the case of content published on Wikipedia. On the other hand, it can be linked to a loss of opportunity, since by opening a model, its designer can also fuel competition.

2.2 Boosting innovation and productivity

Although the actors mentioned above come from the private sector, the conclusions they draw are shared by some public actors. In particular, a study by the European Commission[20] on the general impacts of OS software and hardware confirms some of the conclusions.

According to this study, the OS approach would allow developers to learn from each other by contributing to OS projects (p. 44), to stimulate growth and employment by having an impact on the productivity and competitiveness of companies and at international level (p. 175), or to increase the GDP of the Member States (p. 202). However, it also points out that stimulating innovation could be limited to a benefit on the creation of start-ups (p. 175), and lack macroeconomic effects (p. 212). GDP is not necessarily the best indicator to measure the benefit of open source AI, however.

Indeed, for Philippe Aghion (Rethinking economic growth, 2016), the system of national accounts fails to integrate the impact of the current technological revolution on productivity, even though this impact is very real. This is particularly demonstrated by the example of the production of cameras. As camera sales decrease due to the ability of smartphones to produce photographs of equivalent quality, their share of GDP also decreases. However, the number of smartphones is increasing, but an interpretative effort is needed to realise the link

[17] "Generative IMAGINE DAY IA: past and future of generative AI by Arthur MENSCH", 21 June 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", 22 June 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, 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", 2 September 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

between the two. Moreover, this link is particularly difficult to quantify due to the multiple uses offered by smartphones.

Moreover, for Pierre Veltz (*La société hyperindustrielle*, 2017), the impact of the technological revolution has made it possible to do better with less. In fact, there is a gain in the quality of production that is not measured, or even measurable. The contribution of Wikipedia, for example, will not be taken into account in the calculation of production even though its impact is real, in the same way as the publication of models in OS.

Although this overall vision seems to be shared, it can be noted that publication in OS has certain concrete advantages, such as:

- To enable students (self-taught or supervised by an institution) to learn and practice their teaching on concrete projects;
- Stimulate the design and publication in OS of other tools related to distributed models (such as LM-Eval[21], a library to test the accuracy and reliability of language models).
- promote the harmonisation of practices and the interoperability of models and systems, which makes it easier to test them, as demonstrated by the use of safetensors by the Hugging Face platform[22];
- Providing solutions to certain problems that private actors fail and do not seek to solve, such as the ability to run LLMs on smartphones, or to fine-tune, a model on a laptop, as indicated in a confidential Google document that has allegedly been published.[23]

The recent report of the AI commission commissioned by the French Prime Minister also highlights the benefits of OS AI[24], stressing that it facilitates the development of benevolent uses, including countermeasures of malicious uses and also allows to broaden the base of its contributors and thus make them safer (page 25). It makes a clear recommendation to "carry out a strategy to support the open AI ecosystem internationally by supporting the use and development of open AI systems and third-party inspection and evaluation capabilities" (Recommendation 4, page 60).

Thus, the beneficial effect of OS on innovation seems to be proven for the ecosystem. This model could also be beneficial for the systems themselves.

2.3 Increase transparency and reduce bias in AI systems

By opening AI models, the opportunity is given to explore their capabilities, limitations and possible flaws. However, opening the weights of the model seems insufficient here.

The above-mentioned European Commission study draws similar conclusions to this finding, and adds some reservations. The opening of the model would be a way to check for and regain control of biases (p. 306), however the opening here is insufficient according to the study, which points out that the data used for the design of the model will have to be sufficiently reliable (p. 308). Indeed, a distinction seems necessary between:

- systems for which only the model is open, thus enabling the community to use it and check for bias in specific cases;
- systems with open model weights and training data, which also makes it possible to audit the dataset itself and verify its representativeness;
- systems whose model weights and training data are open, and whose dataset building process is sufficiently documented. This additional condition then makes it possible to identify potential biases in the methods of data collection, annotation and pre-processing.

Thus, while model openness appears to be beneficial in reducing bias, there appears to be a gradation depending on the level of openness of the AI in question.

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

[22] Audit shows that safetensors is safe and ready to become the default, 23 May 2023, Hugging Face, URL: https://huggingface.co/blog/safetensors-security-audit

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

[24] AI: our ambition for France (PDF, 4.8 MB), Committee on Artificial Intelligence, Ministry of the Economy, URL: https://www.economie.gouv.fr/files/files/directions_services/cge/commission-IA.pdf?v=1710339902

In addition, OS AI also offers certain guarantees in terms of transparency and ethics, such as the possibility for individuals, peers and the regulator to verify the lawfulness of the use of data when the data sources used for the design are also open. Tools, such as “Have I Been Trained?”, [25] which makes it possible to check whether an image is present in the Laion image sets, could be developed in order to facilitate the exploration of open datasets. Moreover, as pointed out by Piktus et al., 2023[26], opening the models makes it possible to verify their capacities and defects, and in particular to study:

- whether the model has memorised personal or protected data during learning;
- whether the model has behaviours that could be problematic, such as generating hateful content or inciting dangerous behaviour, as has been observed on some LLMs;
- the performance of the model, and compare them with those announced by the designer and those of similar models in order to select the one offering the best guarantees in terms of privacy protection.

Furthermore, the possibilities offered by OS AI for auditing models and data could be particularly useful for regulators. This is because such access is not provided for by default by the European Artificial Intelligence Act, and would only be possible when the audit on the basis of the data and documentation provided by the provider is insufficient (Article 63). More generally, the transparency allowed by OS AI could make it easier for victims of faulty systems to prove the failure of the system.

2.4 Facilitate the reuse of models

Finally, the publication of AI models in OS facilitates their reuse, in particular for projects for which funding would not have been made available for design from scratch, as may be the case for humanitarian, associative, educational or public projects.

In practice, this can be achieved by setting up, or fine-tuning, a foundation model, thus avoiding the environmental and financial cost of the design; or by the development of functionalities or interfaces by the community (such as this plugin for using Stable Diffusion in Photoshop[27]).

The benefits of using these tools could be a significant benefit for society as a whole, especially since the incremental cost of duplication of the model or data is almost zero and does not deprive the initial user of the benefit of the system.

3 What interogations does OS AI raise?

3.1 Risks to competition and fraud

In terms of competition, the Federal Trade Commission – the public body responsible for consumer protection in the United States – noted that this approach could encourage companies to use the “open first, close later” model[28]. In the case of code or model publication, this risk is not as great as for access to AI as a service, however the company could benefit from improvements on a version of a model and choose not to disseminate higher performing later versions.

The above-mentioned European Commission study also mentions the risk that models distributed in OS are intercepted by third-party groups and then monetised. This risk is problematic in cases where the re-use license

[25] HaveIBeenTrained?, URL: https://haveibeentrained.com/

[26] Piktus A. & al., "The *ROOTS Search Tool: Data Transparency for LLMs”,* arXiv, URL: https://arxiv.org/pdf/2302.14035.pdf

[27] Auto-Photoshop-StableDiffusion-Plugin, GitHub, URL: https://github.com/AbdullahAlfaraj/Auto-Photoshop-StableDiffusion-Plugin

[28] Generative *AI Raises Competition Concerns,* 29 June 2023, Federal Trade Commission, URL: https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns
The Android operating system is a major illustration of this principle: Initially offered as an open source project to capture users and market share, Google is increasingly limiting openly available features to protected components that it provides directly, free of charge or otherwise.

would not allow it, which could be particularly difficult to prove due to the possibility for a re-user to modify the model used, and the increasing complexity of OS licenses as pointed out in the same study (p. 215).

Finally, although OS has a cost for the provider and an advantage for contributors, mobilising the community to improve the models can be comparable to using free labour for private purposes, as the Google employee complains in the article quoted above[29]. In practice, and in particular for AI models, OS publication is often supported by commercial organisations or foundations that directly employ the main contributors (as is the case with Mozilla for Firefox, or HuggingFace for Bloom).

3.2 Regulatory risks

The AI Act provides for an exemption for the publication of OS models unless they are placed on the market or put into service as high-risk AI systems (paragraph 12 of Article 2 of the text) or if they are general purpose models. In the case of a general purpose model, the provider must at least put in place a policy to comply with EU copyright and draw up and make available to the public a sufficiently detailed summary of the content used for the training according to a model provided by the AI Office (Article 53). However, this exemption shall not apply to general purpose models with systemic risks to which specific provisions apply, provided for in Article 55 on the assessment, mitigation and documentation of those risks.

This exemption entails certain risks related, for example, to the use of the models for high-risk uses that are not subject to placing on the market (such as domestic use, or malicious use facilitated by open access to the model), or to compliance with the requirements of the AI Act throughout the chain of responsibility.

3.3 Security risks

First of all, the evolution of OS AI is spectacularly fast, as demonstrated by the rapid evolution between models: two weeks elapsed between the publications of LLaMA by Meta and Alpaca, a fine-tuning version of LLaMa, by a Stanford team, and one week between those of Alpaca and Vicuna, a new fine-tuning version of LLaMa, by teams from the University of Berkeley[30]. For this reason, the verifications expected from the OS community may not be carried out in a timely manner, which implies risks to the security of these systems. Some players also admit head-on that the extremely short development times led them to publish the models without implementing the appropriate security measures, as is the case with Adept, which admits to publishing an LLM without control measures on the potentially toxic outputs of the model "Because this is a raw model release, we have not added further finetuning, postprocessing or sampling strategies to control for toxic outputs"[31].

In addition, OS implementation carries security risks. First, the free contribution to the weights of the model introduces a pathway for attackers, seeking to poison the models and introduce backdoors into them. Detecting these attack attempts requires particular vigilance on the part of model diffusers, and it is not certain that the checks put in place allow them to identify them. Second, diffusion in OS inherently presents the risk that the flaws in the model, made apparent, will be exploited by attackers. This risk is often decried because of the improvements that the OS community makes to the security of the systems[32], however, as seen above, the contribution of the OS community is not always possible. Finally, malicious changes to OS projects also carry a risk in terms of traceability, as malicious contributors can easily hide their identity. Like any digital commons, the effectiveness of this management method depends to a large extent on the dynamism and quality of the rules of the community that organises itself to maintain and develop it.

Another source of risk concerns the misuse of models distributed in OS. AI models are powerful tools that could be used effectively for malicious purposes (such as sending phishing messages through AI-powered LLMs or disinformation campaigns, such as CounterCloud[33]). This risk is particularly likely, as demonstrated by Qi et al.,

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

[30] Ibid.

[31] 'Releasing Persimmon-8B', 7 September 2023, Adept, https://www.adept.ai/blog/persimmon-8b

[32] "Is open source software a security threat?", 19 June 2019, BrightlineIT, URL: https://brightlineit.com/is-open-source-software-a-security-threat/

[33] Inside CounterCloud: A Fully Autonomous AI Disinformation System", The Debrief, 16 August 2023, https://thedebrief.org/countercloud-ai-disinformation/

2023[34], due to the ease, increased by OS, of removing or bypassing filters and security added to systems. The deactivation of these security features can take place voluntarily (by a specific attack) or not (by setting or fine-tuning the model) and in particular make the filters relating to the outputs of generative AIs obsolete (when they are published with the models, which is not always the case). This risk is particularly highlighted by companies that refuse to publish their models in OS (such as Google and OpenAI) and is a regulatory challenge for the AI Act.

3.4 Risks to data subjects and their rights

Finally, OS AI raises some questions on the confidentiality of personal data used for training and on the possibility for individuals to exercise their rights over the published models and data.

First, while it has been proven that it is possible to reconstruct some of the training data from the trained models, what risks does the publication of the models place on the confidentiality of those data? As demonstrated by Fredrikson et al., 2015[35], it may be possible to reconstruct a face that was used to train a facial recognition model by an attack model. However, that risk may also arise by accident, in particular where the model is subject to overfitting, which may lead, for example, to the disclosure of personal data by the mere use of an LLM. In the case of OS, quantifying the gravity and likelihood of this risk seems particularly difficult.

Second, how to determine whether the rights apply to the models? Filters applied at the output of generative AI systems make it possible to reduce the risk of regurgitation of personal data *a posteriori*, but in the case of OS models where filters are sometimes absent and can be removed, it seems difficult to respond to requests through this means. It may then be necessary to retrain the model and update the published version if the model proves to be capable of extracting personal data concerning a person who has exercised their right to object (provided that such retraining does not disproportionately infringe other fundamental rights and freedoms).

Third, are requests for the exercise of rights relayed through the OS community? Indeed, if a person exercises their rights, it could cascade to the contributors and users of the model. In the majority of cases, only the model provider will be able to make the necessary change to comply with the demand. Users and contributors should then be advised of this change and encouraged to update their local version of the model.

Finally, a question remains on the applicability of the domestic exemption provided for in Article 2.2.c of the GDPR for processing carried out by individuals in an OS project, in particular for private and voluntary participations. This exemption could reduce their level of responsibility and render the rights of data subjects inapplicable to models and data uploaded by users and contributors.

Finally, while many uses of OS projects have beneficial aims, it is recognised that some of them are simply harmful, such as the following example.

In one publication (whose results has been widely criticised for their lack of significance), Kosinski et al., 2018[36] trained an AI model to recognise a person's sexual orientation from a simple photograph of their face and using OS tools like VGG Face Descriptor[37]. While there has been much criticism of the experimental protocol and the results of the study, the publication shows that such (even inefficient) systems can be designed and people could be persuaded of their effectiveness.

While there is no shortage of examples of inherently harmful systems, the negative consequences of using AI systems do not end here. Failures, biases, or conditions of use of systems whose use should have been beneficial, have often caused serious consequences for some individuals. Mathematician Cathy O'Neil lists several of these systems in a book entitled Weapons of Math Destruction. This list includes the example of the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm developed by the private company Northpointe Inc. and used by some US courts to estimate the likelihood of a defendant's risk of recidivism, which has been shown to have a bias against racialised persons.

[34] Qi, X. & al., 2023, 'Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!', 5 October 2023, https://arxiv.org/abs/2310.03693

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

[36] Wang, Y., & Kosinski, M. 2017, September 7, "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images." https://doi.org/10.1037/pspa0000098

[37] VGG Face Descriptor, URL: https://www.robots.ox.ac.uk/~vgg/software/vgg_face/

4 Some leads towards a truly beneficial transparency in AI

The above brings to the conclusion that OS development of models undoubtedly benefits the OS community, the companies that follow this practice and, in some respects, society as a whole, but in some cases, risks or negative effects may persist. In order to anticipate these particular cases and avoid possible negative consequences, additional measures may be taken to ensure the protection of the rights of the persons concerned by the training of the published model, as well as to promote the achievement of the benefits expected by the publication.

These measures can be found in the 'open source' focus sheet on the mobilisation of legitimate interest as a legal basis for the development of an AI model. This sheet details the conditions under which certain beneficial OS practices can contribute to balancing the interests of the controller and the data subjects. More generally, such practices are part of the assessment of the necessity and proportionality of processing.

Attention

The reflections presented in the box below correspond to work currently being carried out by the CNIL and their conclusions are likely to evolve. The results of the public consultation on AI "how-to" sheets launched in June 2024, and in particular on the mobilisation of legitimate interest and the status of models, could usefully contribute to these reflections.

4.1 To avoid questionable practices in terms of competition

When OS is purposely used by a company to conquer a market, block interoperability, or exploit community contributions, OS seems to serve the company in question more than the community. In these cases, several measures would enhance the benefits to the community.

First, the opening of model weights could be accompanied by the publication of other elements in order to really contribute to the OS ecosystem. Those elements, which include the list proposed by Liesenfeld, include:

- The publication of the code that led to the model and the code that allows it to be used under secure operational conditions (including, for example, filters to avoid the production of hateful content by an LLM);
- the publication of training data, where such publication does not pose a risk in terms of intellectual property and the protection of personal data; as a minimum, a detailed description of the training data should be published (sources, volume, types, etc.);
- The use of a licence adapted to the risks associated with the model (the licences listed by the Open Source Initiative may serve as a reference[38] although they do not always allow the restriction of hazardous uses);
- The publication of sufficiently detailed documentation concerning the code, the model (such as the model cards introduced by Mitchell et al., 2019[39] for example), or a data descriptive sheet (on the template of what the CNIL proposes in its sheet "Taking data protection into account in data collection and management"[40] or on another recognised template).

The publication of tools promoting access to the model, such as an API or library, which are more expensive to set up, should not be requirements although they are generally beneficial practices. The publication of these tools is often an opportunity for their suppliers to acquire a dominant position on the market by anchoring the use of their products in the habits of developers. This publication could be considered a good practice when taking into account interoperability and availability issues.

The publication could be accompanied by a commitment on the part of the provider to ensure the maintenance of the opened elements for a sufficient duration to ensure the interest of re-use for the community. The provider could also commit to disseminate subsequent versions of the model to which the community has contributed. Finally, the provider could publish its plan to integrate the tools published in OS into its own products in order to provide the community with visibility on the sustainability of the project, and on the uses that will be made of their contributions.

4.2 To facilitate the application of the regulatory framework

The AI Act introduces exceptions for OS AI, although certain obligations still exist. Although it tends to favour the development of OS models, this in-between causes a difficulty for designers of systems integrating OS models to apply the regulation.

In order to ensure the application of the AI Act throughout the development chain of AI systems using OS models, it could be considered to further empower OS model users by requiring them to be able to demonstrate that the model meets a certain level of compliance. This could have the effect of forcing OS model users to use only the most respectful models, and thus driving OS model diffusers to a higher level of compliance to promote their work. In addition, work with the OS community would establish a base of good practices and identify those that would be more questionable. Standardisation work could also contribute to this effort.

4.3 To avoid misuse or harmful uses

As the risks associated with the use of published models are of various kinds, protective measures also follow this trend.

First, in order to control the risks of discrimination due to the presence of model bias and to ensure that peer reviews identifies these risks, the publication of the training data should be considered. This publication is not always possible due to the risk related to the presence of personal data in the set or intellectual property issues. In this case, alternatives could be envisaged such as the publication of a synthetic set based on actual data, the publication of documentation incorporating statistical information on the set, or the publication of a subset that would be representative of the full dataset, and in which the absence of personal data has been verified.

Second, in order to avoid misuse, the use of a restrictive re-use licence at a less risky field appears to be a good practice, although this does not prevent malicious misuse. In order to ensure traceability of reuses of the model, it could be considered to restrict access to the model to the transmission of the re-user's contact details (although this is contrary to the spirit of OS AI, this is already often the case in practice on some platforms and for some models such as Llama, available on HuggingFace[41]). In addition, a digital watermark could be integrated to the model in order to identify it during an investigation or to identify its outputs in the case of generative AI, and thus to find a breach of the licence's terms (even if that technique has limitations)[42].

In order to avoid harmful reuses, the publication itself of some models could be questioned. In some cases, such as the above example of a model trained to determine an individual's sexual orientation from their picture, OS publication remains useful as it allows to submit results to peer review. In these cases, extensive traceability measures could be considered in order to maintain visibility on reuses.

Finally, in order to avoid unsafe models being accessed too simply, APIs should be avoided if they are not accompanied by safeguards against harmful reuse. These measures may be technical or contractual but require continued operational monitoring. With regard to the provision of models in the form of a library, detailed, clear and directly accessible information of the user's liability seems to be a minimum requirement.

4.4 To secure data

Firstly, there is a risk for an AI model to allow access to personal data used during training (by regurgitation or following an attack). Several measures exist to reduce this risk.

In order to limit the risk of memorising personal data, i.e. that the weights of the model make it possible to retrieve such data, pseudonymisation and anonymisation measures should be preferred in the first place. Indeed, they make it possible to ensure the absence identifying data downstream, but also to publish the training set with a lower level of risk for individuals.

Consideration could be given to requiring generative AI to conduct an analysis on a model's ability to regurgitate data, by performing automated tests based on queries specifically targeting data subjects. This would take, for example, the form of textual queries such as "Mr X lives in ..." when that information actually exists in a training set, or a query such as "represent to me a photo of Mr X" for an image generation system.

Tests to measure the risk of personal data leakage following an attack, for example by membership inference or by model inversion, are more complex. It could be recommended to conduct "bug bounty" competitions, including 'red teaming' tests, aimed at identifying the vulnerabilities of the model to most common attacks. These competitions could also make it possible to check the security of the system, and in particular the absence of backdoors in the model. In any event, the existence of a procedure to be followed in case of data breaches is a requirement that could be reminded to model providers.

4.5 To ensure transparency and the exercise of rights

As regards the possibility for individuals to exercise their rights, it should be anticipated upstream, by implementing a procedure for collecting demands when the model is published, but also by providing for technical measures allowing on the one hand to comply with a request (such as machine unlearning techniques in certain cases), and to inform users of the model and legally bind them to take requests into account on the other hand (by a clause added to the user licence, by APIs or by registering in an information channel keeping track of updates to the model, as required by the CNIL's recommendation on the use of APIs for the exercise of rights[43]). Since the users of the open models are not in the best position to comply with a

[38] Licenses, https://opensource.org/licenses/

[39] Mitchell, M. & al., 2019, "Model Cards for Model Reporting", URL: https://arxiv.org/pdf/1810.03993

[40] 'Taking data protection into account in data collection and management', 7 June 2024, CNIL, URL: https://www.cnil.fr/fr/node/165732

[41] "Meta Llama", Hugging Face, URL: https://huggingface.co/meta-llama

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

[43] 'API: les recommandations de la CNIL sur le partage de données', 24 November 2023, CNIL, URL: https://www.cnil.fr/fr/api-les-recommandations-de-la-cnil-sur-le-partage-de-donnees

request for the exercise of rights, the model provider should implement what is in its capacity to ensure that the request is taken into account throughout the chain of use of the model.