An open source perspective on AI and alignment with the EU AI Act

Diego Calanzone[1], Andrea Coppari[1], Riccardo Tedoldi[1], Giulia Olivato[2] and Carlo Casonato[2]

[1]Department of Information Engineering and Computer Science, University of Trento
[2]Department Faculty of Law, University of Trento

Abstract

Artificial intelligence systems based on deep learning have increasingly received interest due to their success in complex human tasks. A current trend in deep learning is to study how algorithms learn multiple new abilities as their size and training data increase. "General purpose AI" (GPAI), that is systems that can transfer the acquired knowledge to solve multiple tasks, are candidate to constitute the backbone of many AI algorithms applied in specific fields on industry, e.g. healthcare, customer support, administration. While various research laboratories express safety concerns on GPAI and do not openly share access to their algorithms, others advocate for their "democratization" and an increasing amount of open-source versions is available online. In this study we analyze this phenomenon from two perspectives and try to reconcile them. From one side, research communities support open collaborations, free access to knowledge and resources; on the other, political institutions, involved in the orchestration between the support for innovation and the control of societal impact, aim at preventing violations of fundamental human rights. We particularly focus on the European approach for risk assessment of AI systems. In our opinion, it greatly overlaps with work in ethics and law conducted by AI researchers (e.g. the Stanford Centre for Research on Foundation Models). Specifically we identify some necessary modifications to improve coordination between the two sides, while also discussing viable implementations in the technical field.

Keywords

open source AI, general purpose AI, European AI regulation, AI social impact, technological standards

1. Introduction

Fast-paced progress in deep learning research is currently followed by the proliferation of AI software to generate art [1], music [2] or to follow instructions for textual tasks [3]. Open access to these tools is favored by research organizations and communities (LAION, HuggingFace, EleutherAI), which are receiving increasing attention and funding for collaborative research. The pervasiveness of AI systems in society is unmatched by the progress of lawmakers in assessing their societal implications, alignment with fundamental human rights and safe development.

General-purpose AI. Originally, AI systems based on machine and deep learning were tuned to perform specific tasks ("fixed-purpose systems"[4]) but achieved poor results in others. More recently, research in natural language processing intersected with deep learning revealed the potential of language models, trained in specific tasks such as predicting the next word in a sentence [5][6] or filling the blanks [7], in solving more abstract problems such as processing common sense questions, solving simple maths or identifying patterns from scarce contextual information (*in-context learning*). The ability of these systems to *transfer* knowledge and capabilities to tasks unseen at training time defines them as *general purpose AI*, as also stated in Aguirre et al. [4].

Foundation models. Bommasani et al. [8] analyze the implications of adopting GPAI as the backbone for many specialized systems, defined also as *"foundation models"*. *Transfer learning* is the process of adaption of an already-trained AI to a new task with either further training (*fine-tuning*), few examples (*few-shot learning*) or none (*zero-shot learning*). The majority of such systems has been originally applied to human language (*Language Models*), but applications in computer vision [9] and biology [10] resulted successful as well. Using such models as base for more specialized AI is defined in Bommasani et al. [8] as the process of *homogenization*: it could ease control and development with the risk, however, of propagating bias from design/data to all the downstream applications.

Open source & AI development. To develop software openly means encouraging the study, modification, distribution and re-use of the source code with transparency. Unbounded access to software allows the cooperation of remote developers (also

Macao'23: AISafety-SafeRL 2023 Workshop (IJCAI), August 19–21, 2023, Macao, SAR, China
diego.calanzone@studenti.unitn.it (D. Calanzone); andrea.coppari@studenti.unitn.it (A. Coppari); riccardo.tedoldi@studenti.unitn.it (R. Tedoldi); giulia.olivato@unitn.it (G. Olivato); carlo.casonato@unitn.it (C. Casonato)

organized in communities) on large-scale projects. This principle has been recently transferred to AI research: in communities such as EleutherAI, OpenBioML, LAION, openly accessible research projects are coordinated by community members and AI models, datasets and development tools are publicly released.
Open source also means free access to multimedia information and knowledge: in the field of AI, free courses on public streaming platforms or blogs are made available by academic institutions such as MIT OpenCourseWare, companies such as HuggingFace or organizations such as FastAI. For open AI development, these online collections include documented software frameworks for beginners to start new AI projects.
In the matter of safety and conditions of use, software developers adopt open-source licenses that vary in the level of freedom for multiple dimensions: modification, use, re-distribution, ownership, use case restrictions and downstream licensing. In the field of AI, policies related to these factors of use are heterogeneous among public and private entities: main companies such as OpenAI have defined and enforced policies of use for their models[11] (such as ChatGPT, GPT-4), which can be used by researchers and customers. However, this approach does not reflect the philosophy of open source: no free access to the AI models' code, parameters and data documentation is given, precluding opportunities for the verification of the capabilities of these systems and open discussion on safety measures.

Open research collectives. Tech companies with abundant resources were first in developing increasingly larger language models. With respect to its predecessors GPT-1 and GPT-2, GPT-3 (OpenAI) [7] has not been publicly released: on-demand remote access to use the AI is restricted by OpenAI with paid options. Part of the AI research community discussed whether discretion to access to such powerful algorithms should be left to single private companies, it is arguable that this choice could be also motivated by the non-negligible cost to train these models. Nonetheless, open-source versions of large language models started to emerge among research collectives: the Big Science Project is a global workshop along the lines of CERN or LHC, it promotes joint collaboration on training open large language models applicable in science. Moreover, known academic conferences, such as NeurIPS, expect paper submissions to be accompanied with publicly released code; "grassroot" research collectives such as EleutherAI (AI) or OpenBioML (computational biology), are funded by technical partners advocating for openness, such as HuggingFace or StabilityAI.

The EU AI Act. The European *proposal for a Regulation laying down harmonised rules on artificial intelligence* [12] focuses on the impact of AI systems and it introduces a taxonomy of risks: unacceptable risks AI (enumerated fields of application, e.g. social scoring, exploitation of social groups), high-risk AI (closed list with requirements, e.g. aircraft systems), limited risk AI (open list, optional requirements), minimal risk AI. The proposal adopts a general definition to cover a wide range of AI systems beyond technical specifications. The regulation applies to AI systems (including open source AI) that are put in the European market, so AI developed for the sole purpose of research is not affected by. On 14th June 2023, the European Parliament adopted its negotiating position on the Artificial Intelligence (AI) Act with 499 votes in favour, 28 against and 93 abstentions ahead of talks with EU member states on the final shape of the law.

Council of the EU. On the 6th of December 2022, the Council of the EU adopted the common position on the EU AI Act [13], several changes were introduced to the proposal. The requirements set out for high risk AI systems in the EU AI Act are applicable to GPAI systems as a result of a risk assessment procedure described in a specific implementing act. As originally proposed, high risk AI systems are required to undergo a conformity assessment procedure followed by the emission of a certificate. Requirements for conformity include for instance: a constantly updated technical documentation, a risk management system, an informative of use with reported assessment of robustness and security of the system, record-keeping of the operations of the AI.
Similarly, Bommasani et al. [8] identify the source of harm of AI systems in development and training, where intrinsic bias, (e.g. patterns in data of toxic behavior and associations with social groups), translates into extrinsic harm (e.g. under-representation or misrepresentation, the generation of fake information or hate content). In our opinion, this well aligns with the view of the European proposal, but the fields where the latter applies are too strictly defined, potentially limiting its effectiveness. AI research and industrial applications greatly overlap: large language models developed with data and techniques in research can become market products (e.g. GPT-3, GitHub CoPilot). Consequently, procedures to ensure transparency, safety and quality of systems should take part in the research and development process, as well as they should not hamper them. We believe a potential solution lies in research communities, where a unified collection of datasets and tools is open for testing and improvement. This allows to create a development framework that is widely adopted by researchers and if developed in synergy with political institutions, it could also ensure safety and alignment.

2. The impact of open source AI

In this section we discuss the impacts generated by the use of open source AI systems, for each field we suggest the behaviour to be maintained from the perspective of both users or providers. Our opinion reflects the necessity to develop open, safe, transparent and efficient AI.

2.1. Social impact.

Determining in advance a comprehensive list of social implications of AI is challenging. Pre-trained models describe a recent paradigm shift and demonstrate remarkable capabilities of generalisation beyond training information. Before discussing social impacts, we must comprehend when such models represent some harm. We believe that foundation models could have direct social impact when they are applied in public services or deployed in the market as products. Foundation models have the potential to be dangerous if they have been trained or adapted on questionable data. Especially because such models can inherit biases from toxic information [14]. Disparities in the performance of gender classification [15] and facial recognition systems [16] have been found due to incomplete data. Additionally, it has been proved [17] that in large language models there are exploitable vulnerabilities. Nevertheless, these AI systems already brought tangible benefits in several fields. For instance, they could provide accurate medical diagnoses at a low charge [18]. This could also give people living in areas affected by severe poverty, the opportunity to access medical treatment. Furthermore, these models with generative abilities could provide support to researchers working on discovering new therapies to treat people [19]. In our opinion, foundation models have shown potential for applications in education: an AI can deliver lessons to students in a much more interactive way, depending on the demands of the interlocutor, like a private tutor. Thus, it would be possible to grant everyone a low-cost custom education. Controlling AI development in a risk-based approach, allows to weight potential harms with expected benefits, given a sound and transparent development procedure.

Law enforcement. Training large language models requires considerable amounts of data, which are usually web-scraped. The absence of a data filtering procedure prior to model fitting can lead to legal issues: models put on the market such as GitHub CoPilot (based on GPT-3) have shown to generate licensed software code [20] [21]. Data curation is subject to existing regulations: The General Data Protection Regulation (Regulation (EU) 2016/679) (hereinafter “GDPR") and the California Consumer Privacy Act of 2018 (hereinafter “CCPA”) [22]. Principal issues regard the misuse of private information and ensuring fundamental rights from current law, that is right to ensure anonymity, the right to erasure, the right to be informed (and to know the interlocutor’s nature), the right to object and the right to access. Further critical issues regard liability for the model’s predictions: GPAI can be applied to an unpredictable amount of tasks. Given the consistent advantage in processing large amounts of historical data, model predictions may be overestimated. Moreover, since the mechanisms that drive such systems are in most cases not explainable or difficult to interpret, AI-assisted decisions can be difficult to understand. For instance, applying AI to healthcare and bio-medicine can bring remarkable advantages, although the problem of liability needs to be addressed as these models are not free of bias, and even in the optimal case they can have unacceptable margins of error [23].

Inequities and malicious uses. Producing a dataset without a lack of diversity and biases has been considered an endpoint for a while. Nonetheless, when we deal with a massive quantity of data, it is not straightforward to filter out toxic content [14]. Recently, LAION has released an open-source novel dataset consisting of 5.85 billions of image—text pairs [24], in which images that appear to contain harmful content have been removed. However, they briefly explained that even if they improve the overall safety of the data, there is no guarantee that the generated dataset is completely safe. Evidence showed that models trained on poor/biased data exhibit unexpected behaviours and tend to discriminate against a few marginal social groups, by inherithing stereotypes. Moreover, Weidinger et al. [25], point out that not all social groups may have access to the services provided by these foundation models at the same quality. For instance, research on large language models in English and Chinese languages is extremely active, but it’s quite different for some other languages spoken by a smaller amount of people. If the access to these services predetermined a competitive advantage, it would increase inequality between different social groups. Koenecke et al. [26], found that speech recognition systems perform better with white American English speakers rather than African American English speakers. Moreover, these generative models are able to produce low-cost high-quality content. This may facilitate disinformation campaigns and the dissemination of false information as outlined in Goldstein et al.[27]: users might be over-relying on the AI system, unaware of biased model results. In fact language models can replicate common misconceptions or they can make unfounded statements [28]. Moreover, they could also suggest illegal activities rather than discourage them; recent work in AI safety has focused on adapting generated text with human feedback [29]. Bai et. al. [29],

have conducted some experiments on adapting the model to identify harmful outputs by revising its response without human feedback. Furthermore, a model trained on people's private information may correctly infer sensible data of a person and use it in inappropriate contexts. In this specific case, this might conflict with data privacy, non-discrimination, fairness and other ethical principles [30].

2.2. Economics

Open source AI offers a great opportunity in terms of digitisation for European enterprises. It has been proven during the last decades that open strategies result in efficient normalisation of new technologies. Notably, when speaking of foundation models, whose field of application is very broad, the implementation of open source AI systems would be an important step towards a zero-cost digital innovation of the public sector. Conversely, closed source systems are currently leading all over Europe, establishing a leading minority of tech giants that hampers Small and Medium Enterprises' (SMEs) growth. The following paragraphs describe the economical impact of the use of open-source AI systems.

Digital Innovation. It is probably the greatest achievable benefit from the use of open foundation models. It is estimated that three in five European companies lack digitalisation [31], and that around most of SMEs lag behind in technological innovation because of the high costs. Foundation models can handle a several amount of different tasks, but only the writing skill will be considered for the following example. Most of the jobs have at least one secondary task in which writing is involved, leaving that part to an AI system would make the process much quicker, resulting in higher productivity in terms of produced outputs. The EU AI Act [12], in accordance to the Council of the EU position, defines and labels general purpose AI systems as a separate family of systems, outside the risk-based labeling approach (art. 4a, 4b, 4c). General-purpose systems that should be classified as high-risk systems due to their field of application (art.6, Annex III) are not necessarily open, therefore the exact cost of innovation is not fully predictable. Articles 53, 54 and 55 focus on measures for innovation of tech industries that will develop AI, through regulatory sandboxes and a priority system based on the enterprise scale. Nonetheless, a regulation to encourage the use of open systems applied to the public sector, especially to the fields listed in Annex III, is crucial to ensure faster and stronger innovation, even for non-tech companies.

Mitigating SMEs' burdens. As stated in the previous paragraph, the uptake of foundation models is critical for innovation in terms of costs, but at the current state of the regulation, specifications for general-purpose AI systems are partially defined. The ALLAI organisation published some in-depth studies on the EU AI Act [32]: they point out what could actually hamper the use of such systems. By the Article 8(2), the provider of an AI system is obliged to take into account its intended purpose when adhering to the requirements, but the 'intended purpose', as defined in Article 3(12), is difficult to identify for foundation models due to their intrinsic nature. This way the burden of proving the compliance of those systems with the EU AI Act falls on "downstream users", resulting in stifling innovation for SMEs and micro enterprises that cannot handle the obligations. The solution ALLAI proposes is the definition of 'reasonably foreseeable use', alongside the 'intended purpose', that is "the use of an AI system in a way that is or should be reasonably foreseeable".

Work-force transformation. It is almost certain that the introduction of AI in the European market will have a significant impact on a wide range of fields of occupation, e.g. foundation models applied on specific tasks could replace human workers due to a lower resource cost [25]. Moreover, they should be considered, as economists defined it, a form of *general-purpose technology*, that a new method of producing and inventing that is important enough to have a protracted aggregate impact. Foundations models, thanks to adaptation strategies like fine-tuning and prompting, might be able to solve a considerably large number of problems much more accurately than humans or even tasks that humans cannot perform. As a matter of fact, general-purpose AI systems, are released open-source online at no cost for the majority. Therefore this trend may lead to a significant shift in the labour market due to the fact that those models:

1. perform functions with a zero marginal cost [33];
2. might increase productivity and profit;
3. achieve human-level performance.

Automation will replace and reshape millions of jobs, for instance: worker-less factories are fully automated factories empowered by automated systems [34]. In the mid-term, even knowledge workers like radiologists might be replaced by extremely precise AI systems [35]. To summarise, companies will need a novel work-force that masters the latest emerging skills. Thus, it is not about losing jobs, instead we are introducing a work-force transformation. It has been argued [36] that this process does result in employment increase.

Decentralization of power. As previously stated, some bigger AI providers hold power over smaller ones. In the development of AI systems, power is defined by

the control over data and models. Arguably the European market on AI systems is characterized by the oligopoly of tech giants, which overrule even on computational power. Indeed, open technologies introduce the possibility to compete against those big companies, developing a broader, and less biased, community of experts. Open source AI in general provides better cybersecurity and transparency, but from an economical point of view, it results in an efficient decentralization of power, which in turn leads to a potential improvement in the public sector due to costs reduction and security enhancement.

3. Risks in Foundation Models & the EU AI Act

Several principles for ethical development of AI systems, and specifically on foundation models, have been set out by Bommasani et al. [8]. The object of their paper is to describe risks and opportunities that foundation models bring with them. In this section we present how those highlighted issues are tackled in the EU AI Act [12], and for each of them we discuss our opinion.

3.1. Capabilities

Foundation models exhibited surprising capabilities in language, vision, robotics, reasoning and search, user interaction in lots of downstream tasks [37]. The EU AI Act [12] does not consider general-purpose AI systems, although AI systems are divided following a risk-based approach, Articles 6, 7 establish some classification rules for high-risk systems. In particular Art. 7(2)(a) specifies that whenever a new high-risk area has to be assessed by the Commission, the *intended purpose* of the AI system must be taken in consideration. Article 3(12) defines intended purpose: "Intended purpose means the use for which an AI system is intended by the provider, including the specific context and conditions of use, [...]". By this definition, whatever the capabilities of the system, risk has to be assessed a posteriori by the Commission, based on the declared usage.

Technology standards. Foundation models reach impressive performance across multiple downstream tasks, through internet-size data. Hence, some considerations about data: documentation, access, visualisation, curation and selection. Harmful behavior in AI models find origin in the design process of the model and in data curation processes. Bommasani et al. [8] define the development of an AI system as a sequence of defined stages, ranging from data processing to security assessments and deployment. Such standardised design procedure tackles a variety of issues concerning AI: robustness to adversarial attacks, the generation of harmful content, inconsistent performance depending on the user. The EU AI Act addresses such issues with a series of requirements and procedures: Article 10 concerns procedures for data governance in line with the European law, including the GDPR; Article 15 introduces accuracy and robustness levels to be met for high-risk AI, although strict compliance is not required, according to a recent review by the Council of the EU, as unified metrics are difficult to formalize. To summarize, Chapter 2 of the EU AI Act [12] introduces standards, which have to be harmonised in compliance with Art. 40.

Intrinsic bias & transparency. Due to the ingent amount of information processed by AI systems, contaminated data can introduce critical issues such as social under-representation (e.g. incoherent quality of medical diagnoses for members of a particular minority) or mis-representation (e.g. hate speech or inappropriate associations). In Article 11, the EU AI Act [12] introduces the obligation to write technical documentation, reporting architectural design choices, data pre-processing and task-specific model adaption. Further measures may consist in a more socially inclusive evaluation procedure with metrics developed to specifically measure bias and toxicity. Model Cards (Google [38]) provide with a template completely in line with requirements set out by the EU AI Act in Article 13, and further ensure the interpretability of AI systems. This level of documentation should be mandatory for open source AI systems since "Model cards also disclose the context under which models are intended to be used, details of the performance evaluation procedures"[38], which is crucial information when referring to foundation models.

Legality & liability issues. There are three legality issues that stand out: accountability of AI system providers, output liability, and eventually copyright. Given the broad foundation models' field of application, it is probable that a defection of an AI system might harm people. In order to safeguard the right to an effective remedy and to a fair trial, the provider of the defective AI system should be traceable by the users and should be accountable for the damage caused by his product. Art. 62 of the EU AI Act [12] tackles this accountability requirement implying the obligation for providers of high-risk AI systems to produce reports of serious incidents and malfunctionings. Finally, copyright infringements are considered. They are composed of two major issues, copyright protected data contained in the datasets, and copyright of model output. In the final version of the AI Act there will be a separated section on Generative AI, or models capable of generating content such as images or text. Providers of such models will require to publish summaries of copyrighted data used for training, and to disclose that the content is generated

by AI.

Auditing. AI systems should be periodically tested against possible issues and shifts in data distributions. In order to fulfil these requirements a protocol for periodic testing should be defined, alongside a controlled environment for safe testing. The EU AI Act [12] handles this issue in two different steps. Title V, and in particular Art. 53, defines "AI Regulatory Sandboxes" to safely test high-risk AI systems before entering the market, this is also seen as an innovation enhancement tool, since it gives priority to "small scale" industries. Title VIII defines a post-market monitoring system, in which the Market Surveillance Authority watches over the European market for systems that present new risks, or break the requirements set by the EU AI Act. Wherever a provider establishes a causal link between an AI system and a malfunctioning, the provider is obliged to notify it within a period of 15 days. Notwithstanding the power given to the Market Surveillance Authority (Chapter 3 of EU AI Act), it acts only after notifications, drastically slowing down the process of law enforcement. However Art. 65 states that whenever a provider does not apply corrective actions within a reasonable period, commensurate with the nature of the risk, the *"Market Surveillance Authority shall take all appropriate provisional measures to prohibit or restrict the AI system's being made available on its national market, to withdraw the product from that market or to recall it"*.

4. Changes to the EU AI Act

Most of the changes that we wanted to bring into the regulation have been already proposed by the European Council [13]. Nonetheless, we report ourselves a list of further additions to the EU AI Act to align it to an open-source perspective of AI development. Title 1a of European Council's proposal is a collection of three articles concerning requirements and obligations for General Purpose AI Systems. However, we strongly believe that AI is not a product, but technology to build products. Thus, in Art. 55b we introduce "AI-based general-purpose technology", with the attempt of detaching the concept of product from General Purpose AI, which is a method of producing, see Art. 3(2a). Regarding the addition to ANNEX IV, we noted that data documentation for high-risk AI systems is mandatory only "where relevant" (ANNEX IV(2)(d)), thus we decided to strengthen the requirement for those systems whose use is related to individuals. It is possible to detect the presence of natural persons in the intended purpose of the AI system, because its documentation must include its context and condition of use.

***Art. 3(1a) - Foundation models.** AI systems classified as general-purpose.

***Art. 3(2a) - General-purpose technology.** A new method of producing and inventing that is important enough to have a protracted aggregate impact (e.g. Electricity or Information Technology).

***Art. 55b - Measures to enhance public sector.**

1. General-purpose AI systems classified as high-risk AI systems by compliance with Annex III, shall be considered by the Member States as general-purpose technologies for public sector enhancement.
2. Member States shall undertake the following actions:
   (a) Whenever an AI-based general-purpose technology is chosen by the Member State to enhance innovation within the public sector, priority must be guaranteed to open source solutions in order to reduce innovation costs.
   (b) Exceptions to (a) are those AI systems developed from a closed source, that have proven to achieve better performance in terms of accuracy, transparency or security than any open source solution considered by the Member States, which are obliged to produce documentation explaining the choice of that AI system over the open solutions present in the European market.

***ANNEX IV(2)(dd).** Data requirements listed in ANNEX IV (2)(d) are mandatory for foundation models and whenever natural persons are involved in the intended purpose described in the documentation;

5. Compliance

In this chapter we review proposals to ensure transparency, accountability and fairness in AI systems. We discuss the applicability of software licenses, while underlining potential contributions from the open source AI community in developing sound technology standards.

Licensing. Contractor et al. [39] proposes Intellectual Property (IP) tools for AI software to ensure liability and correctness of use. With respect to patents, licenses protect original creations over conceptual inventions. A license is a legal agreement between an entity (the licensor) and a subject (the licensee), with defined rights

and restrictions. By taking in consideration the 233,601 AI models published on HuggingFace[1], approximately 29% of these are published with a license, including: 14% with the Apache v2.0 License, fully permissive, which allows to use, modify and re-distribute a model also under different licensing terms; 6% with the MIT License, permissive and in line with Apache v2.0; 3% with OpenRAIL-M [39], permissive but it introduces the obligation to propagate use restrictions defined by the licensor or in line with the BigScience Ethical Charter [40]. We denote two main principles in AI licensing: flexibility of use, distribution and modification of the models; guidance in practical applications, in line with ethical principles and intended use. We believe the current licensing approaches are complementary, but their union is not exhaustive: models could share a homogeneous set of guidelines for data processing, training [29] and output generation [41]; the enforcement of restrictions on unethical use is fragmented or excessively delegated to the licensor. Licensor-defined restrictions can result in legally incompatible models that could not be combined in new systems, limiting both research and industrial development [42]. Finally, obligations such as mandatory updates (BigScience RAIL, section IV.7 [43]) are dangerous for task-specific applications, as undesired changes in performance can occur without any power of control for the licensee.
We believe a candidate solution consists in a permissive open AI license with emphasis on traceablity, accountability and limitations solely on use considered against reference codes of ethics such as the EU Charter of Fundamental Rights [44], the AI4People framework from Floridi et al. [41] or the EU Trustworthy AI Ethical Guidelines[30]. This could be achieved with the adoption of documentation standards for data and software [45][38]. The cooperation with political institutions, e.g. in the EU, becomes crucial to achieve effective enforcement, for instance through the legal validation of licenses for AI and standards for the deployment and monitoring of these systems in the public sector.

Standards development. A keyword for safe AI is transparency. In the development of trustworthy AI systems, three types of emerging initiatives from the open source AI community are reported by the BigScience initiative: best practices through guidelines and ethical standards, algorithm transparency with unified evaluation metrics and development toolkits, appropriateness of use through forms of documentation such as AI FactSheets [45] and Model Cards [38]. The HuggingFace platform hosts 233,601 models and 44,309 datasets[1] organized in repositories: each model or data repository can include a documentation, provided by the publisher, on the usage of the model, information on the composition and processing of the training data, analysis on the limitations, intended uses and biases of the model. However, these information largely vary in amount and detail at the discretion of the publisher and the contributors. We believe these elements could be merged in standardized frameworks for development and deployment, with particular emphasis on data acquisition and curation that involve ethical and societal choices. According to Article 40 of the EU AI Act: *"conformity to the requirements for "high-risk AI" or GPAI is ensured through the adoption of harmonised standards or parts thereof published in the Official Journal of the European Union"*. When deemed necessary, the Commission can introduce "common specifications" in addition to such standards, that is directives enforced through bodies of experts. Despite standards or specifications allow to practically apply such provisions, few organisations are in charge for their development. The 2nd of February, 2022, the European Commission announced a new approach to the standardisation system [46], including more active involvement of academic institutions in the process.
A major objective of this article is to promote communication between research organizations and the European Commission: we believe both parts share common concerns in developing trustworthy AI, emerging efforts from researchers can bridge the gap between directives and their practical application without conflicts of interest.

[1]Data updated up to June 18th, 2023

6. Conclusion

Open source AI allows for the collaboration of researchers on a large scale: we believe this setting is ideal to research robustness, reliability and safety in AI models scaling in complexity and capabilities. Openness should underlie the interplay between private research companies, which provide fundamental engineering resources, and researchers in academia and grassroots collectives. Research from lawmakers and AI academics in AI ethics and safety well overlaps as the concerns are shared: improved communication between the two parts could be beneficial to support open scientific research on one hand, while also enforcing safety measures in the process of integration of increasingly complex AI in our society.

References

[1] A. B. e. a. R. Rombach, High-resolution image synthesis with latent diffusion models, 2021.
[2] R. M. e. a. Zalán Borsos, Audiolm: a language modeling approach to audio generation, 2022.

[3] J. W. e. a. L. Ouyang, Training language models to follow instructions with human feedback, 2022.
[4] A. A. e. a. C. I. Gutierrez, A proposal for a definition of general purpose artificial intelligence systems, 2022.
[5] A. R. et al., Language models are unsupervised multitask learners, 2019.
[6] M. C. e. a. J. Devlin, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.
[7] B. M. e. a. T. B. Brown, Language models are few-shot learners, 2020.
[8] R. Bommasani, H. et al., 2021.
[9] L. B. e. a. A. Dosovitskiy, An image is worth 16x16 words: Transformers for image recognition at scale, 2020.
[10] R. E. e. a. J. Jumper, Highly accurate protein structure prediction with alphafold, 2021.
[11] Openai usage policy, 2023. URL: https://openai.com/policies/usage-policies.
[12] EU, Proposal for a regulation of the eu parliament and of the council laying down harmonised rules on ai and amending certain union legislative acts, 2023.
[13] C. of the European Union, Artificial intelligence act: Council calls for promoting safe ai that respects fundamental rights, 2022.
[14] V. U. Prabhu, A. Birhane, Large image datasets: A pyrrhic win for computer vision?, 2020.
[15] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, 2018.
[16] B. F. K. et al., Face recognition performance: Role of demographic information, IEEE Transactions on Information Forensics and Security (2012).
[17] N. C. et al., Extracting training data from large language models, 2020.
[18] L. P. e. a. V. Gulshan, Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, 2016.
[19] C. Jee, The therapists using ai to make therapy better, 2021.
[20] S. Ananya, Lawsuit Raises Copyright Concerns in AI-Generated Work, 2022.
[21] S. e. a. Barke, Grounded copilot: How programmers interact with code-generating models, 2022.
[22] Futurium | European AI Alliance - Trustworthy AI in Practice, 2022.
[23] D. E. Ho, A. Xiang, Affirmative algorithms: The legal grounds for fairness as awareness (2020).
[24] C. S. et al., Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.
[25] L. W. et al., Ethical and social risks of harm from language models, 2021.
[26] A. K. et al., Racial disparities in automated speech recognition, Proceedings of the National Academy of Sciences 117 (2020) 7684–7689.
[27] J. A. G. et al., Generative language models and automated influence operations: Emerging threats and potential mitigations, 2023.
[28] J. H. e. a. S. Lin, Truthfulqa: Measuring how models mimic human falsehoods, 2022.
[29] S. K. e. a. Y. Bai, Constitutional ai: Harmlessness from ai feedback, 2022.
[30] Requirements of Trustworthy AI, EU Commission, 2018.
[31] EIB, Who is prepared for the new digital age? : evidence from the EIB investment survey, Publications Office, 2020.
[32] ALLAI, Aia in-depth (2022) 13–14.
[33] J. Rifkin, The Zero Marginal Cost Society: The Internet of Things, the Collaborative Commons, and the Eclipse of Capitalism, St. Martin's Publishing Group, 2014.
[34] J. Rifkin, Preparing students for" the end of work"., 1997.
[35] The future of Artificial Intelligence and radiology Hunimed, 2022.
[36] T. Taylor, Some journal of economic perspectives articles recommended for classroom use, Journal of Economic Perspectives 33 (2019) 243–48.
[37] Huge "foundation models" are turbo-charging AI progress, 2022.
[38] S. W. e. a. M. Mitchell, Model cards for model reporting, 2018.
[39] D. C. et al., Behavioral use licensing for responsible ai, 2022.
[40] The big science ethical charter, 2022. URL: https://bigscience.huggingface.co/blog/bigscience-ethical-charter0.
[41] J. C. e. a. L. Floridi, Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations, 2018.
[42] R. Stallman, Why programs must not limit the freedom to run them, 2022.
[43] BigScience, Bigscience rail license v1.0, 2022.
[44] Requirements of Trustworthy AI, EU Commission, 2012.
[45] D. P. e. a. J. Richards, A methodology for creating ai factsheets, 2020.
[46] T. E. Commission, New approach to enable global leadership of eu standards promoting values and a resilient, green and digital single market, 2022.