Aligning the EU AI Act with the open-source trajectory of research

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

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

***Abstract*—Artificial intelligence systems based on deep learning have increasingly become popular due to their success in carrying out various human tasks with high accuracy. Recent research shows multiple skills are acquired by agents with increasing training data and algorithmic parameters. Such general-purpose AI can carry out unseen tasks in a multitude of fields, thus becoming a good candidate for wide application in industry (e.g. GPT-3). While access to these systems has been regulated from developers to prevent misuse, researchers have increasingly advocated for their "democratization", and open-source versions have found wide diffusion in academia and industry. In this study we analyze this phenomenon from two perspectives while reconciling the demands: the academic research community, advocating for freedom of information and open access to resources; political institutions, involved in the orchestration between the support for innovation and the regulation of these technologies to prevent misuse and the violation of fundamental rights. We particularly focus on the European approach for risk assessment of AI systems. We report 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 identified some necessary modifications to improve alignment, finally we discuss the role of licenses and their dual effect on the open-source community.**

***Index Terms*—Open-source, Foundation models, General-purpose AI, Social impact, Risks, Law, Ethics, EU AI Act**

I. Introduction

Lawmakers and tech industries recognize the necessity for regulation on the use of AI systems[1][18]. With increasing training data, more complex AI can leverage a great amount of knowledge to solve multiple different tasks, such as language translation, information retrieval, image classification. These algorithms already find wide spread in online platforms and products, this pace is unmatched from the development of sound regulation: this temporal gap leaves room for potential misuse and misalignment with human values. Recently, Bommasani et al. [11] introduced the term *Foundation models* to define an ecosystem of open source AI that could potentially become the groundwork for AI systems involved in the public.

Foundation models. A set of algorithms (also referred as "models") trained on large scale data that can transfer the acquired knowledge to a multitude of *downstream tasks*. *Transfer learning* defines the process of adaption of an already-trained AI to a downstream task with further training (*fine-tuning*), few examples (*few-shot learning*) or none (*zero-shot learning*). Such algorithms usually are defined in literature as *Language Models*: they're originally trained to predict future or missing words in a sentence, but this can be generalized to image patches, protein sequences etc. With increased in data and computation, such models exhibit new capabilities [17]: solving math world problems, answering commonsense questions, multilingual language understanding and translation. Bommasani et al. [11] discuss the adoption of these general-purpose AIs as base for many AI tools applied in real world problems (homogenization): this could ease control and development, with the risk, however, of propagating bias from design/data to all the downstream applications.

Open research collectives. Tech companies with abundant resources firstly developed increasingly large language models. With respect to its predecessors GPT-1 and GPT-2, the model GPT-3 (developed by OpenAI)[10] 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 one private company, 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[25] 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" diverse research collectives such as EleutherAI (AI) or OpenBioML (computational biology), are backed from technical partners advocating for openness, such as HuggingFace or StabilityAI.

While increasing enthusiasm in research nudges the open approach, the protection of fundamental rights has to be considered. Pre-existing law don't will fit with AI pervading in public administration, healthcare systems and transportation, such norms either can lack of flexibility to allow in this technology, or it leaves undefined how to handle explainability, privacy, accountability and individual autonomy. The cost of developing larger foundation models is prohibitive for most of the research collectives but a few, mostly associated with

"tech giants", which hold the right to choose who can access to these technologies and for which purposes. In parallel with technical literature, political institutions worldwide are developing proposals for regulation of AI and data, in the light of rising concerns about privacy, fairness and fundamental rights.

The European *Proposal for a Regulation laying down harmonised rules on artificial intelligence* [27] focuses on the impact of AI systems rather on the intrinsic nature, it introduces taxonomy of risks: unacceptable risks AI (enumerated fields of application), high-risk AI (closed list, with requirements), limited risk AI (open list, optional requirements), minimal risk AI. With foundation models as general-purpose AI, predicting all the fields of application is unfeasible and tracing extrinsic harmful behavior to intrinsic bias by design is non-trivial. In this study we propose further modifications to the EU AI Act to allow with more flexibility the definition of Foundation Models and its requirements, we moreover investigate the use of intellectual property tools to handle copyright, liability and transparency for the diffusion of open-source AI software.

II. Impact of open source AI

In this section we discuss the impact generated by the use of open source AI systems, in particular for each of them we specify some changes that need to be applied, either to society or to providers. We base our observations on the study of Bommasani et al. [11], as it reflects our ideals: the necessity for openness to develop safe, transparent and efficient AI.

A. *Social impact.*

Predicting the possible social impacts of foundation models and tackling ethical and societal considerations is challenging. Such pre-trained models describe a recent paradigm shift and demonstrate impressive capabilities of generalisation through adaptation. Before the discussion on the social impacts, we must comprehend when those models may become potentially harmful. The ecosystem defined by Bommasani et al. [11] presents a pipeline to achieve a responsible deployment of the AI system in which the main goal is to minimise the potential harms [11]. The first step of the pipeline regarding the creation of the data. We may consider that all the data that empower the training of a foundation model are generated by human-being with a purpose, this may or may not include to be considered a part of the training data-set of the model. The second step of the pipeline concern data curation before the training of the foundation model. This is a challenging step due to the fact that we must ensure the quality of the data collected in the previous step. But most importantly also we must guarantee that the generated data-set is compliant with legal and ethical constraints. The third and fourth steps of the pipeline concern the training of the foundation model on the data-set extracted at step two and the consequent adaptation to the specific task. The model has a direct social impact when at the fifth step of the pipeline becomes a product and is deployed to people. Foundation models may potentially become harmful if they have been trained or adapted to questionable data, due to the fact that they inherit possible flaws from the original data-set. The authors of the Stanford report anticipate a wide range of possible societal consequences according to the various applications. We can already consider the enormous benefits that these AI models could have in many fields. For instance, they could provide accurate medical diagnoses at a low charge. This could also give people living in areas affected by severe poverty the opportunity to access medical treatment. Additionally, these models with generative abilities could provide support to researchers working to discover new therapies to treat people [12]. Further improvements may reach in education. Through these models in the future, we will be able to customise each student's learning plan. In particular, AI models could interact interactively with each student learning through feedback and understand the run-time needs and necessities of each student.Even Though, here we have presented only two scenarios, the spectrum of application of foundation models is awfully wide and the benefits that these models bring are remarkable. However, these models could represent systems that if used without guidelines increase social inequalities, amplify misinformation and may have a wide range of awful social impacts.

Law enforcement. Since the training phase of these models requires huge amounts of data. The common trend may become to accumulate data by scraping from the Web. To illustrate an example of this phenomenon that recently emerged might be GitHub Copilot, which was trained on copyrighted codes [23] [15]. Thus, such a trend appears to be in conflict with major regulations aimed to ensure the protection of data like The General Data Protection Regulation (Regulation (EU) 2016/679) (aka. "GDPR") and the California Consumer Privacy Act of 2018 (aka. "CCPA") [20]. Therefore the principal issues that may arise regarding the misuse of private information and with respect to the use that the data might conflict with the main rights in "CCPA/GDPR" like the right to ensure anonymity, the right to erasure, the right to be informed, the right to object and the right to access. Further critical issue regard who takes responsibility for the model's predictions. The capabilities of these models are impressive. Over several tasks, they perform significantly better than human-being. Since the mechanisms that drive the AI system are in most cases not explainable and difficult to interpret. This entails the problem that humans may not have the skills to take responsibility for the output of the model. These models could have applications in high-risk areas such as healthcare and bio-medicine. Since we know that these models are not free of bias and even in the optimal case may have a certain margin of error [8], who would be liable for these errors?

Inequities and malicious uses. The possible challenging scenarios that emerge from the Stanford report and the DeepMind report by employing these models are an indefinitely large number [11][14]. If a model, for example,

has been trained on improperly filtered data it might inherit some bias. So it might inherit stereotypes and in the case of large language models for instance may tend to discriminate against a few marginal social groups. Moreover, in the DeepMind report, they point out that not all social groups may have access to the services provided by these foundation models at the same quality. This is because in large language models for instance research in English and Chinese is extremely active, so services produced by these models in other languages do not have the same quality. This may increase inequality between different social groups if access to these services produced by the models predetermines a competitive advantage. One example of this phenomenon has been substantiated by a research group. They found that speech recognition systems perform better with white American English speakers rather than African American English speakers [9]. Moreover, these generative models are able to produce low-cost high-quality content. This may facilitate disinformation campaigns and the dissemination of false information. These models could be used to make inferences about people's preferences and to strategically manipulate them. The capabilities of these models could be used to find strategies to violate the law. In addition, users who are unaware of the bias of these models may misuse them or trust them too much. In addition foundation models could be trained on people's private information. This would give these models the ability to make inferences about people's personal information. A model therefore may correctly infer sensitive information about a person and use it in inappropriate contexts. In this specific case, this might conflict with data privacy, non-discrimination, fairness and other ethical guidelines [21].
It is definitely impossible to provide a complete overview of all possible risks associated with these emerging AI models. The main goal still remains to minimize the possible risks and threats associated with and maximize the possible benefits carried by these AI models. Definitely, we should release novel tools to mitigate the risks and encourage the open-source community to responsible innovation.

*B. **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 which the 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. Follow paragraphs describing the economical impact of the use of open AI systems.

Digital Innovation. It is probably the greatest achievable benefit from the use of open foundation models. It is estimated that one in five European companies lacks digitisation, and that around 90% 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. EU AI Act as it is now, does not regulate general-purpose systems, but the Council proposal has given a proposal in which they are defined and labeled 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 the field of application (art.6, Annex III), are not necessarily open, hence the cost of innovation is not predictable, but will probably be high. 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, the updated version of the AI Act is more accurate about it. 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.

Work-force transformation. It is almost certain that the introduction of Al in the European market will result in a complete change in the range of known jobs with novel technologies. Indeed foundation models, and their extended ease of use, are going to replace a significant amount of workers [11][14]. Nevertheless foundation models should be considered, as economists would say, *general-purpose technologies*, just like electricity. Foundations models thanks to adaptation strategies like fine-tuning and prompting, might be able to solve a considerably large number of problems and tasks much more accurately than humans or even tasks that humans can not 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 the functions with a zero marginal cost;
2) might increase productivity and profit;
3) turn out to be effective and are a low-priced solution to replace a significant amount of workers [2].

Automation will replace and reshape millions of jobs, for instance, worker-less factories are right now an example of fully automated factories empowered by automated systems [1]. In the mid-term, even knowledge workers like radiologists might be replaced by extremely precise AI systems [26]. 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 studied [5] that this process does result in

employment increase.

Decentralization of power. As previously said, some bigger AI providers hold the authority over other smaller ones. In the development of AI systems, power is defined by the control over data and models. Undoubtedly the European market on AI systems is characterized by the oligopoly of tech giants, which additionally 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.

C. Environment

We must also consider the environmental impact of these huge models. While the size of the model determines its actual capabilities. In recent years, larger and larger models have been trained. While we get better performance from larger models, on the one hand, we use more and more energy-hungry hardware resources for a long time in order to perform the training of those models. This trend could lead to huge carbon footprint emissions. This leads to exacerbating environmental issues by increasing heat-trapping greenhouse gas levels in the Earth's atmosphere. Definitely, the introduction of pre-train models would greatly reduce the environmental impact. This is because the pre-train model once trained can be re-adapted with reduced environmental cost each time and reused for many tasks. Further considerations concern the data centers used to train these models. In fact, these energy-intensive data centers should consider to reduce environmental impact and the carbon footprint emissions, it is essential to switch to renewable energy [7] [13]. It is also crucial to establish metrics reported in documentation to evaluate the sustainability of the data centres and the associated environmental impact [16]. Moreover, before training a model we might consider conducting a cost-benefit analysis to ensure that the social and environmental benefits counterbalance the costs and hazards.

III. Issues & the EU AI Act

In this section we consider the main principles highlighted by Stanford University's paper on foundation models [11] with respect to the current European law proposal on AI. In particular we interconnect each proposal with its corresponding law in the AI Act tackling it. On 3rd November the Council of EU published an updated version of the Act trying to regulate general-purpose AI systems, such changes will be considered here.

Capabilities. The main goal of the Stanford report is to provide a wider picture of foundations models [11]. The objective of paper is to describe the recent paradigm shift and identify risk and opportunities of foundations models. The authors exhibit the surprisingly capabilities in language, vision, robotics, reasoning and search, user interaction of foundations models in lots of downstream task. They examine also the possible capacity of future models, novel key challenges and possible gaps. The EU AI Act (21st April, 2021) 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 the risk 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. The authors of the Stanford report [11] make some considerations about data documentation, data access, data visualisation, data curation and data selection. Since foundation models learn from data and achieve impressive performance across a multitude of diverse downstream tasks and applications, the authors underline the relevance to provide standards of high-quality data. Nonetheless, issues regarding data privacy and governance controls have shown up. Hence, regulations and guidelines must be draughted in order to ensure prevention of those issues and further violations. Moreover, security vulnerabilities and risks of function creep (unintended use), have been exhibited in the report. An additional open question raises, "How do we access the reliability, robustness and interpretability of the model?". Interpretability is defined as the set of explaining mechanisms that drive the model behaviour. The AI Act already fulfils the technology requirements set by Stanford's report. Article 10 describes how data and data governance are handled by EU. Nevertheless, this European regulation relies on existing Directives and mostly on GDPR (General Data Protection Regulation). Article 15 presents some accuracy and robustness standards, and obligations to follow for high-risk AI systems. In more recent proposals by the European Council, art. 15 has been revised, since proper accuracy metrics are not defined yet. Substantially, Chapter 2 of the AI Act introduces regulations and standards, which have to be harmonised in compliance with Art. 40. For those systems that are difficult to regularize with respect to the aforementioned articles, Art. 41 sets out Common Specifications to be implemented.

Intrinsic bias & transparency. Stanford's paper presents the issue of contaminated data, introducing the duality of under-representation and over-representation. The former

consists in lack of data on minorities, which is a significant concern that hugely affects society. Models applied on those minorities most often produce strongly biased outputs. The AI Act presents some measures to prevent bias in the models by the production of technical documentation (Art. 11) that will provide specific information on used data sets and data selection criterion. Nonetheless, complete prevention is unachievable, thus accuracy and robustness requirements are introduced (Art. 15) in order to minimize the risks. The design procedure for a foundation model includes decisions about the evaluation criterion, the technique to adapt a model to a downstream task, the architecture and the training objective. Model compression amplifies bias, model feedback loops influence social behavior. Although the disclosed information is not enough to guarantee transparency of the models. Google research published a template for documentation under the name of Model Cards [6], which is completely in line with requirements set out by the AI Act in Article 13, and enhances 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", crucial information when referring to foundation models.

Legality & liability issues. The report summarises most relevant legality issues in three points regarding accountability of the AI system providers, output liability, and eventual copyright infringements. Given the broad foundation models' field of application, it is probable that a defection of an AI system might harm people. In order to preserve persons' right to an effective remedy and to a fair trial, the provider of the defected AI system should be traceable by the users, and should be accountable for the damage caused by his product. Art. 62 of the AI Act tackles this accountability requirement implying obligations for the providers of high-risk AI systems consisting in reports of serious incidents and of malfunctioning. Another relevant issue with AI systems is output liability, defined in terms of authorship of the model output. Stanford's paper enlightens the requirements of human oversight and informed interaction with the machine; both are satisfied by the AI Act (Title III, Chapter 2). Finally, copyright infringements are considered. It is composed of two major issues, copyright protected data in the used data sets, and copyright of model output. The former will be handled by the European Data Protection Supervisor in the post-market monitoring system, the latter has not been tackled yet, although possible solutions might discriminate over different model outputs.

Auditing. The paper on foundation models [11] claims that AI systems should be periodically tested against possible issues and shifts in data distribution. 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 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. There is no such thing as a periodic testing protocol, but it has been defined in Title VIII 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 regulations set by the AI Act. Notwithstanding the powers explained in Chapter 3, the market surveillance authority acts only after notifications, drastically slowing down the process of enforcement.

IV. Changes to the EU AI Act

Here follows a list of modifications on articles from the EU AI Act (21st April, 2021) with respect to our proposal. We agree with the changes proposed by the EU Council on 3rd November, thus we will not report their proposals and we will take them for granted. Nonetheless, modifications are on the current version of the AI Act and are not present in any of the Council's submissions. The articles marked with a * are not present in the current version of the regulation, nor in the Council's proposals.

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

***Art. 3(2a) - General-purpose technology.** Technology basis on which other technologies are built.

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

1. General-purpose AI systems classified as high-risk 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 bring innovation on 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 coming from a closed source, that have proven better performance in terms of accuracy, transparency or security than any open source solution. Member States shall produce documentation in which they motivate 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;

V. The role of licensing

Contractor et al. [18] investigated the use of licenses for responsible, open source AI based on Intellectual Property (IP) rights. Relevant IP tools are copyright licenses, which protect original creations, and patents, which define royalties and credits for conceptual inventions. We believe the latter does not reflect the philosophy of academic AI research: knowledge about algorithms for intelligence should be open and free to humanity. A license is a legal agreement between an entity (the licensor) and a subject (the licensee), with defined rights and restrictions. We argue that introducing the latter reveals to be problematic: firstly, it is not possible to identify all the possible cases of misuse beforehand; secondly, restrictions can generate an undesired effect: according to R. Stallman [24], software with specific use-cases become difficult to integrate and compose, which contrasts the concept of openness potentially pushes developers towards private solutions. Furthermore, we argue some points in the RAIL framework of licenses may result critical: mandatory product updates (BigScience RAIL, section IV.7 [19]), may give excessive power of influence to the developer since changes propagate to all the downstream applications, including potential degradation in performance or new vulnerabilities/bias. While licenses represent a way to regulate the circulation of open source AI, we suggest a shift in purpose: free of use restriction, this tool should emphasize obligations for transparency, it should clearly the interplay of roles with related responsibilities and it should encourage ethical use and distribution.
We derive the work of [18] and refine a list of essential components for an open source AI license :

- Definitions and roles:
  I. AI system defined as a composition of the software (source code and possible executable files), the configuration (referred as "hyperparameters" from technical literature), the learned parameters (referred as "pre-trained weights" when provided). In machine learning and deep learning, the latter encodes learned patterns about the training data, both semantic and informative, also in case of bias.
  II. The provider, or licensor, defined as the person, business or organization openly publishing an AI system either self-developed or re-distributed.
  III. The user, or licensee, defined as the person, business or organization that uses in any context the system released by the provider.

- Liabilities: a section defining the legal liabilities for both the provider and the user. While [19] retains the user fully responsible for any output generated by the system, extrinsic harm is not traceable back to the problematic design choices of developers. The latter may be consider accountable for causing harm through fallacies in the design of the model and/or the training data, with exception for any modification applied by the user. The licensee may be responsible for any use beyond the contexts mentioned in the documentation, and any further application in conflict with the law and ethical guidelines. While such partitioning may result excessively neat, one line shared with technical literature [11] is the necessity for transparency and informed use: negligence may be considered as purposeful harm.
- Ethical guidelines: Floridi et al. [4] analyzed background literature in ethics for AI and proposed 5 fundamental principles in an ethical framework: beneficence, non-maleficence, autonomy, justice, explicability. Integrating this work could provide with a more robust ethical baseline in licenses, similarly to [18].
- Copyright and intellectual property. The presence of copyrighted data in large, web-scraped, training data-sets posed large, open source AI systems as the center of discussion for intellectual property violation. Machine-generated content may be considered as the transformation of input data, which does not threat protected work[22]. However, increasingly capable generative models such as DALL-E 2 or Stable Diffusion can reproduce partially or completely training data such as artworks, which could potentially violate the copyright of original work.

VI. Conclusion

Policymakers should examine and assess all the possible strategic factors that may lead to broadly desirable technological progress. Most importantly, political parties should think carefully regarding how to promote an open collaboration between research collectives. Substantial strategic consideration in policymaking still has the potential to guide openness to long-term fair outcomes and to maximize the possible achievable benefits. Specifically, openness may speed up the production of the AI systems compliant with the current regulations, while providing wider engagement between researchers [3]. Wherever we achieve a responsible development and deployment the advantages of open-source AI outweighs the disadvantages.

References

[1] Jeremy Rifkin. *Preparing Students for" The End of Work"*. 1997.

[2] Jeremy Rifkin. *The zero marginal cost society: The internet of things, the collaborative commons, and the eclipse of capitalism*. St. Martin's Press, 2014.

[3] Nick Bostrom. "Strategic implications of openness in AI development". In: *Global policy* 8.2 (2017), pp. 135–148.

[4] Josh Cowls et al. Luciano Floridi. *AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*. 2018. DOI: 10.1007/s11023-018-9482-5. URL: https://link.springer.com/article/10.1007/s11023-018-9482-5.

[5] Timothy Taylor. "Some Journal of Economic Perspectives Articles Recommended for Classroom Use". In: *Journal of Economic Perspectives* 33.3 (Aug. 2019), pp. 243–48. DOI: 10.1257/jep.33.3.243. URL: https://www.aeaweb.org/articles?id=10.1257/jep.33.3.243.

[6] Andrew Zaldivar et al., eds. *Model Cards for Model Reporting*. 2019. URL: https://dl.acm.org/citation.cfm?id=3287596.

[7] Karen Hao. *Training a single AI model can emit as much carbon as five cars in their lifetimes*. Dec. 2020. URL: https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/.

[8] Daniel E. Ho and Alice Xiang. "Affirmative Algorithms: The Legal Grounds for Fairness as Awareness". In: (2020). DOI: 10.48550/ARXIV.2012.14285. URL: https://arxiv.org/abs/2012.14285.

[9] Allison Koenecke et al. "Racial disparities in automated speech recognition". In: *Proceedings of the National Academy of Sciences* 117.14 (2020), pp. 7684–7689. DOI: 10.1073/pnas.1915768117. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1915768117. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1915768117.

[10] Benjamin Mann et al. Tom B. Brown. *Language Models are Few-Shot Learners*. 2020. DOI: 10.48550/ARXIV.2005.14165. URL: https://arxiv.org/abs/2005.14165.

[11] Rishi Bommasani et al. *On the Opportunities and Risks of Foundation Models*. 2021. DOI: 10.48550/ARXIV.2108.07258. URL: https://arxiv.org/abs/2108.07258.

[12] Charlotte Jee. *The therapists using AI to make therapy better*. Dec. 2021. URL: https://www.technologyreview.com/2021/12/06/1041345/ai-nlp-mental-health-better-therapists-psychology-cbt/.

[13] David Patterson et al. *Carbon Emissions and Large Neural Network Training*. 2021. DOI: 10.48550/ARXIV.2104.10350. URL: https://arxiv.org/abs/2104.10350.

[14] Laura Weidinger et al. *Ethical and social risks of harm from Language Models*. 2021. DOI: 10.48550/ARXIV.2112.04359. URL: https://arxiv.org/abs/2112.04359.

[15] Shraddha Barke, Michael B. James, and Nadia Polikarpova. *Grounded Copilot: How Programmers Interact with Code-Generating Models*. 2022. DOI: 10.48550/ARXIV.2206.15000. URL: https://arxiv.org/abs/2206.15000.

[16] Google. *Google 2022 environmental report*. 2022. URL: https://www.gstatic.com/gumdrop/sustainability/google-2022-environmental-report.pdf.

[17] R. Bommasani et al. J. Wei Y. Tay. *Emergent Abilities of Large Language Models*. 2022. URL: https://arxiv.org/abs/2206.07682.

[18] Danish Contractor et al. *Behavioral Use Licensing for Responsible AI*. URL: https://facctconference.org/static/pdfs_2022/facct22-63.pdf.

[19] BigScience. *BigScience RAIL License v1.0*. URL: https://huggingface.co/spaces/bigscience/license.

[20] *Futurium — European AI Alliance - Trustworthy AI in Practice*. https://futurium.ec.europa.eu/en/european-ai-alliance/best-practices?language=en&check_logged_in=1.

[21] *Requirements of Trustworthy AI - FUTURIUM - European Commission*. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html.

[22] S.M.Hedrick. *"Think," Therefore I Create: Claiming Copyright in the Outputs of Algorithms*. URL: https://jipel.law.nyu.edu/vol-8-no-2-1-hedrick/.

[23] Ananya Singh. *Lawsuit Raises Copyright Concerns in AI-Generated Work*. https://theswaddle.com/lawsuit-raises-copyright-concerns-in-ai-generated-work/.

[24] Richard Stallman. *Why programs must not limit the freedom to run them*. URL: https://www.gnu.org/philosophy/programs-must-not-limit-freedom-to-run.html.

[25] *The Big Science Community*. URL: https://bigscience.huggingface.co.

[26] *The future of Artificial Intelligence and radiology - Hunimed — hunimed.eu*. https://www.hunimed.eu/news/the-future-of-artificial-intelligence-and-radiology/.

[27] The European Union. *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.