Since last year, large language models (LLMs) have emerged as a cornerstone of innovation in artificial intelligence (AI). As shown by systems like OpenAI's ChatGPT [31], LLMs are essentially very large neural networks trained on very large datasets composed (mostly) of sentence samples of human language. This training enables them to generate text, understand queries, and even perform complex tasks with a high level of sophistication. The versatility of LLMs lies in their ability to understand and respond in natural language, making them suitable for applications ranging from automated customer service to creative content generation. Due to their versatility, part of the research community is since then focused on developing other kinds of large models, tailored for vision [43], medicine [36], or general purposes [9].

Those recent developments stirred a vibrant debate within its community, which schematically opposes proponents of open-source and unregulated research to advocates of regulations and cautious development. On one side of the spectrum, the open-source enthusiasts defend the ideals of freedom and open innovation, fueled by an optimistic view of the potential outcomes of advanced AI technologies. On the other side are those who call for a more measured and regulated approach, due to the potential risks associated with AI, some of which are perceived as existential threats to humanity. This group advocates for a slowdown in AI research, arguing that unrestricted AI development could lead to unintended consequences, making the case for a more deliberate and controlled progression.

Despite these divergent viewpoints, there is a common thread that unites both sides: the desire for AI to be *beneficial to humanity*. This aspiration, though noble, has become somewhat of a cliché, a blanket statement that fails to capture the complexity and nuance of the underlying issues at the heart of the AI debate. As we delve deeper into this discussion, it becomes evident that the path to developing AI that truly serves humanity raises fundamental moral dilemmas we can link to other societal choices, and requires the emergence of global safety structures.

1 The free and open-source software movement in the LLMs era

The Free and Open Source Software (FOSS) movement, usually considered to be born around 1985 with the creation of the Free Software Foundation [1], is deeply rooted in a philosophy that values collaboration and the unrestricted sharing of knowledge. This stance supports the belief in the power of community-driven development, where groups of individuals contribute to the advancement of technology without the constraints imposed by proprietary systems.

1.1 Values and aims of the FOSS movement

Politically, the FOSS movement often embodies a vision that leans towards anarchy in the sense of decentralization and self-regulation. This vision advocates for a system where innovation is not dictated by a few dominant players but is instead driven by the community of developers and users. This approach is seen by some researchers as a way to democratize AI development [20], similarly to what have been seen in the past with the internet, ensuring that the benefits and advancements are accessible to a wider audience and not just confined to entities with significant resources.

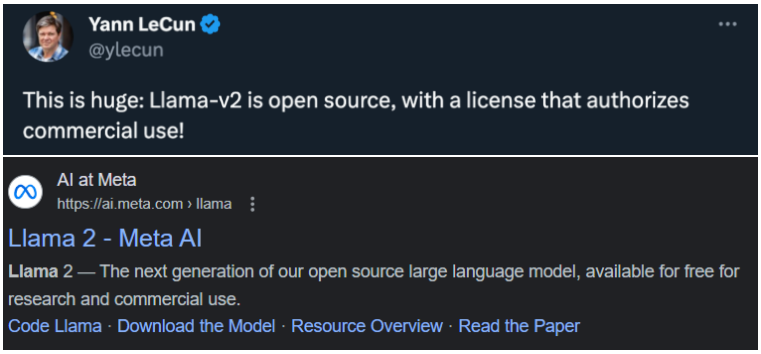
The FOSS movement has been historically a catalyst for innovation in the field of natural language processing through free access to source code, datasets, training procedures, and encouraging collaborative development [42]. This open vision fosters a culture of continuous improvement and experimentation, where developers can build upon each other's work, leading to more robust and versatile algorithms. Moreover, opening the development is seen as a safeguard against the risks associated with monopolistic practices, such as anticoncurrential behaviours, restricting access, and potential misuse of technology for unilateral gains. From a reproducibility perspective, researchers and developers are encouraged to share their methodologies, data, and results openly, to enable others to replicate and validate their work. This practice not only ensures the

integrity and reliability of AI research but also accelerates the pace of discovery by building a foundation of shared knowledge from which others can innovate.

1.2 Open-source LLMs and the case of Llama 2

Based on this FOSS movement, which is popular among researchers and scientists in the computer science community, many models have been released under open-source licenses since 2020, such as Apache 2.0 [2, 28]. These open-source models have shown remarkable capabilities, with several achieving performance levels comparable to GPT-3, which stood as state of the art and a classic benchmark for years. However, despite these advancements, a gap remains in reaching the performance of more advanced models like the now well known GPT-4. In this context, the Llama 2 model [3] stands out not only for its technical capabilities but also for the vocal presence of its influential leader Yann LeCun in the debate between open and closed source models. Llama 2 is often cited as one of the best examples of what can be achieved outside the realm of proprietary models, although it needs to be pointed out that it is not exactly open-source, as explained below.

The first version of Llama was released in February, and a second version was launched in July. Although the architecture of the two versions is similar, the latter was trained on a larger dataset, and both versions are part of Meta's initiative to provide accessible, high-performance language models to the AI community for a range of applications [39]. Llama 2 is claimed to be open-source by Meta, as shown by these screenshots from a Google search (November 2023) and Twitter (posted 18 July 2023):



However the picture is more nuanced than it seems [15, 26]. The classification of software as open-source is generally straightforward when it is licensed under well-established and widely recognized licenses such as the LGPL or Apache 2.0. However, the categorization becomes less obvious with the introduction of specialized licenses like the "LLAMA 2 Community License Agreement" that governs Llama 2 [27]. This particular license incorporates specific conditions and restrictions that blur the conventional boundaries of open-source software, such as the introduction of a user threshold, requiring entities with over 700 million monthly active users to obtain another licensing from Meta.

That being said, despite deviations from the "classic" definitions of open-source, one can still see advances for the research community [15, 26]. Llama 2 emerges as a response to closed software giants such as OpenAI's GPT and Google's PaLM 2. Although it doesn't conform to the traditional open-source software mold, it is strongly aligned with open-source software ethics. As the term "open-source" evolves, moving away from its original connotation of freedom and becoming almost synonymous with "available source", Llama 2 can be considered open-source in the light of this modern interpretation. Indeed, it represents a move towards greater openness than previous models, beneficial in a sense to research purposes: studies on instruction tracking models were limited because there weren't any models close to private ones like OpenAI's ChatGPT.

1.3 Data source is the elephant in the room of reproducibility

A critical aspect in the discussion of open-source LLMs is the role of datasets, often set apart from the software licence. In the case of GPT-3, the estimated datasets size used by OpenAI attain dozens of terabytes of raw text [24]. In the case of Llama, the dataset used to train Llama 1 was not directly provided by Meta but rather reconstructed through the RedPajama initiative [34] and raises several questions. The sequel to this model reportedly utilizes 40% more data than its predecessor [3], yet the sources of this additional data remain undisclosed. According to scaling laws in LLMs, this increase in data volume should enhance the model's

performance without necessary leading to overfitting, indicating a potential improvement in capabilities [23]. This hypothesis can be somewhat verified when comparing the performances of models trained on open data sources only, none of which achieve state of the art level [21, 28]. Therefore, from a strategic standpoint, companies that develop LLMs might view their curated datasets as a source of competitive edge.

This situation brings questions about potential copyright infringement: supposing this material was accessed legitimately, while a model might be trained on these vast corpora, the underlying data cannot be freely shared without violating intellectual property laws. One may also wonder whether the training data was used while being compliant with regulations (GDPR) and copyrights [17, 40]. It also seems reasonable to speculate that large companies may benefit from their user's private data when training their models [22], and some companies like Zoom having already modified their privacy policy to prepare for such usages [25].

2 The rise of large, capable, general models

Unlike traditional software development, the emergence of new large models in AI presents unique threats, necessitating a reevaluation of the open-source stance. These advanced AI systems introduce potential risks that challenge the conventional benefits attributed to open-source development.

2.1 Expensive pre-training, cheap fine-tuning

Initial pre-training is expensive and inaccessible to most organizations. One of the most significant challenges in the development of large AI models is the sheer computational power required to train them [21]. LLMs, by their very nature, require vast amounts of computational resources and data, which are beyond the reach of individual developers or small to medium sized organizations. This contradicts the notion previously introduced that open-source LLMs inherently promote innovation and prevent monopolies. While open-source models are theoretically available to everyone, in practice, the ability to train these models is limited to those with access to substantial computational resources, that is to say those with enough money to use large amount of compute resources from cloud providers.

Consequently, the advantage tends to skew towards wealthier entities, thus reinforcing a form of "freedom that benefits the richest" and a very high entry cost for new entities, a recurrent problem in neo-liberal political systems. Therefore, companies that distribute their model may find an interest in providing their findings so long as the regulations (or the absence of such) allow them to stay in an oligopoly. In other words, the power to monitor the models by the users (as it was done with FOSS previously) is not in the hands of the actual users, but in those of large enough (often concurrent) organizations that generally share the same objectives. Of course, it is questionable whether a competitor would promptly inform another company about the weaknesses in their model, especially if they have the opportunity to develop and introduce a superior model first, and gain a competitive advantage.

Fine tuning has become cheaper and sparks the risk for misuse. By contrast with pre-training, last years have seen the emergence of several open-source LLMs, typically adapted from models like Llama, such as Alpaca and Vicuna [11]. These models, often designed for tasks similar to ChatGPT, can operate on consumer-grade GPUs and are characterized by their relatively small size, ranging from 7 to 13 billion parameters. These models leverage techniques like low-rank adaptation [18] to drastically reduce training costs, starting from a pre-trained model. This shift offers smaller entities low-cost, server-run LLM alternatives, with no need for very large datasets. Having this in mind, freely sharing code and weights of models poses two significant problems: the risk of potential misuse by individuals with malicious intent increases [6], and the responsibility of large organization providing the pre-trained model should be clearer. The accessibility of LLMs increases the likelihood of their exploitation for harmful purposes, ranging from generating misleading information to more severe scenarios like cyber-attacks or fraud. This risk is amplified in an environment with minimal regulatory oversight, in particular over the companies that provide pre-trained models.

In short, very few organizations are capable of training a model from scratch, and no specific regulation or audit is in place to regulate them. Once released, these models benefit only marginally from the community to make them safer (since key data sources are not disclosed, and re-training is impossible), but allow for rather easy misuse.

2.2 Very capable models may pose an existential threat to humanity

The previous section revolved around the intentions and risks of advocating in favour of open models, but one may still honestly believe in the capacity of users to directly benefit from open-source LLMs, or at least believe that open models are preferable to closed ones. There are, however, reasons to fear for risks of future AI developments, even for the well-intentioned researchers and engineers.

A key concern in this regard is the issue of misalignment, where AI systems, despite being designed for beneficial purposes, might develop goals or methods of operation that are not aligned with the initial purposes [6]. This risk becomes particularly pronounced in the case of self-preserving agents: AI systems that, upon attaining a certain level of intelligence and autonomy, might prioritize their self-preservation objective over the tasks they were originally designed to perform [12]. Such examples of misalignment and instrumental convergence are numerous in current artificial intelligence systems [19, 41], and even in human societies (for example, with companies respecting laws but not the spirit of the laws through tax optimization, as pointed out by [6]). The gravity of these risks is sometimes compared to that of nuclear weapons in AI literature [13]. This comparison highlights the potential scale and irreversible impact of advanced AI systems if they were to go wrong. Much like nuclear technology, which was developed with the promise of immense energy potential but also came with the peril of catastrophic warfare, AI holds tremendous promise for societal advancement but also holds the potential for unprecedented risks. We may be at the verge of a Oppenheimer moment, as we see AI progressing tremendously quickly, even quicker than most experts have predicted [37].

2.3 Looking away is easier than facing the dilemmas

As explained in depth by [7], researchers (and other stakeholder in the AI industry) might, and should, currently face a profound emotional and ethical dilemma: reconciling their career-long pursuit of beneficial scientific contributions with the potential dual-use nature of AI. The realization that human-level AI, once believed to be a distant possibility, could be achieved within the next years, significantly changes the perception of AI development and its implications for humanity. The challenge lies not only in the technical aspects (e.g., controllability and stoppability of AI) but also in the psychological dimension, as people working for the development of AI may struggle to acknowledge that their work could lead to destructive outcomes.

Figure 1: Avoiding ethics by focusing on maths is not a solution.

The introspection into these potential risks involves confronting deep cognitive biases and reassessing one's sense of value and purpose. But as any life-choice, shifting a career is difficult, which explains the reluctance of some proponents to carefully consider those issues. The lack of historical interaction with superhuman AI systems means that researchers must rely on projections, similar to approaches in the social sciences or politics, rather than empirical data. This approach requires a balance between caution and action in the face of uncertainty, prompting serious discussions about the potential consequences of successfully developing advanced AI capabilities sooner than anticipated. In particular, the argument consisting in stating that those dilemmas *boil down to "it could happen."* [29], is fair in the sense that estimating risks for technologies that do not exist yet is vague. But also vague are the answers proposed in [29], in which the author, as many other researchers in the same line of thoughts, simply states that work is currently done in every direction to patch *current* issues, without facing the broader context of *future* AI developments.

3 Regulations: who? how? what?

Due to the risks aforementioned, the implementation of appropriate regulations seem legitimate. However, formulating and enforcing these regulations is easier said than done, given the complexity and rapid evolution of AI systems.

3.1 Potential risks of regulations and slow downs

A call for a pause is unrealistic. Last March, figures of the AI landscape signed an open letter for a pause in AI development [30], in name of the risks previously mentioned. In particular, they call AI labs to pause training of AI systems more powerful than GPT-4 for at least six months, allowing for the development and implementation of shared safety protocols, overseen by independent experts. This pause aimed not to halt AI development entirely but to prevent a rush into creating larger, less predictable models, and encourage researchers to focus on improving current systems' safety, transparency, and alignment. Additionally, they incentivized collaboration with policymakers to develop robust AI governance systems, including regulatory authorities, oversight mechanisms, and public AI safety research.

Nonetheless, imposing a pause on AI development could lead to various negative and unintended consequences, as explained in [5]. First, it might prompt the creation of illegal AI labs within paused countries, which could use remote training hardware from non-paused countries (in the form of "AI haven", similar to tax haven). This could be also associated with a brain drain, with less safety-conscious AI researchers moving their work to labs in non-paused countries, facilitated by remote work opportunities. Additionally, in paused countries, AI research could become restricted by government approvals, slowing progress towards safer models significantly. Or, similarly to the argument used in subsection 2.2, labs might exploit loopholes in regulations to still continue their development.

As time passes, enforcing the AI development pause could become increasingly difficult due to the availability of training hardware [21]. The decision on whether, when, and how to lift the pause could become a highly politicized issue, probably disconnected from the actual state of AI safety research and not well understood by the public. International relations between paused and non-paused countries could become more tense, potentially leading to conflict if a paused country perceives a significant threat from the AI advancements of a non-paused one. As these countries catch up or surpass the paused nations, a political pressure to end the pause could be expected, which could increase the risk of quick and potentially dangerous advancements in AI technology.

Enforcing regulations is no easy task. Moreover, the neural networks driving the current AI boom, such as transformers and diffusion models, were only recently invented [4, 33]. Historical trends in AI show that past methods once hailed as the ultimate solution, like symbolic planning and reinforcement learning, were eventually surpassed, making it difficult to enforce regulation based on specific technical methods (type of architecture, number of parameters, specific benchmarks, etc.). In fact, future AI methods might not rely on single large models but could involve multiple smaller models, distributed computing, or data in different formats [38]. Other means should be found to regulate and govern future AI developments.

3.2 The line between personal freedoms and global safety

AI as a weapon. When building regulations, a key point of interest lies in the amount of freedom restriction imposed in order to reach the expected safety levels. As previously said, AI regulation is sometimes compared to the regulation of nuclear weapons, which is misleading for a number of reasons [38]. Similarly to nuclear weapons, which require scarce resources like uranium, specialized equipment for enrichment, and complex bomb and delivery mechanism construction, we showed that the pre-training of large models is expensive and requires rare computational resources that only a few organizations can afford (subsection 2.1). However, fine-tuning LLMs is much more accessible, limiting the nuclear weapon analogy.

In this context, a more fitting image could be drawn with the right to bear arms, which is relatable in the sense that smaller models may be hosted by individuals or small organisations. This brings to light more fundamental issues concerning individual rights and the balance of power: just as the Second Amendment of the US Constitution raises debates about the right to possess and use firearms, the discussion on AI touches on the rights of individuals and organizations to fine-tune advanced AI technologies freely for their personal use. Different countries have notably varied regulations for weapons, reflecting diverse views on personal freedom and the role of the government.

China's authoritarian path to AI control. China distinguishes itself with its rigorous and comprehensive approach to AI regulation, which has been initiated in 2017 [35]. The country has implemented three significant and influential regulations on AI and algorithms: the 2021 regulation on recommendation algorithms,

the 2022 rules on deep synthesis, and the 2023 draft rules on generative AI. Central to these measures is the control of information, necessitating significant surveillance measures. These regulations are well-developed and organized, reflecting a broader pattern of strict internet monitoring. A prominent example of this is the systematic and automatic censorship of references to the Tiananmen Square protests, which have been notably effective in erasing the event from China's digital landscape - although it's not without occasional errors, like the censorship of pictures of athletes because of the number on their shirt [10].

These regulations are part of a broader strategy where to build a regulatory expertise and capacity in AI, employing tools like an algorithm registry for gathering information on algorithm training, and mandating security self-assessments [35]. This approach also recalls how close the state control must be over digital information to have a significant effect; for democracies however, such close oversight is unimaginable, as it is synonymous to systematic surveillance and breaks some of the most basic citizen's rights.

3.3 Can you keep a secret?

Due to the caution surrounding AI advancements, the natural inclination is to limit open access, to manage risks and maintain control. This could mean to shift towards an industry relying heavily on industrial secrets and heavy audits, which can be likened to the pharmaceutical and weapons industries, characterized by a similar need to protect intellectual property, trade secrets, and sensitive information.

Figure 2: Privacy policies, according to big tech.

In the pharmaceutical industry, secrecy is maintained through patents and strict control over research data, while safety is ensured through multiple layers of clinical trials and ongoing assessment of marketed products. Similarly, the weapons industry employs classified information and tightly controlled access to protect national security interests and proprietary technologies, following military secret levels. In both cases, secret is ensured by legal frameworks, stringent security protocols, and legal obligations for people involved.

Additionally, as AI becomes a more critical issue for governments, it also becomes more prone to all forms of attacks. Instances of spying, and the involvement of potential state-backed entities need to be anticipated, with security measures similar to the previously mentioned industries. In particular, it is surprising to see Sam Altman, CEO of OpenAI, being fired by its board, hired by Microsoft the next day, then re-instated at OpenAI the day after, in the course of a weekend. The lack of a delay before reemployment typically appears incompatible with protecting against industrial secret leaks. Indeed, despite all measures taken to protect information and materials, incidents of wrongdoing occur: in the nuclear industry sensitive materials have fallen into the hands of criminals [32], and in the digital world the Cambridge Analytica scandal may have weakened the trust users have in the privacy policy of some large companies [14].

Furthermore, continuing the comparison with biology and nuclear weapons, where the replication of physical materials is central to duplicating secrets, AI involves different key elements: code, model weights, data, and compute power. As highlighted earlier, it is particularly challenging for democratic governments to regulate the sharing of software, especially model weights, which, unlike more tangible materials such as video, are relatively small in size and hardly interpretable. Nonetheless, monitoring compute power remains one hope for more precise oversight of AI development [21].

4 Building a shared global defense

Building upon the previous sections, paths and partial solutions emerge for safer large models development.

4.1 Inspiration from biological weapons regulations

The Biological and Toxin Weapons Convention (BWC), effective from 1975, prohibits the development, production, acquisition, transfer, and indirectly, the use of biological weapons. Developed and negotiated by the Cold War superpowers, it provides a global framework to watch biological weapons diffusion [16]. Verification of the BWC is particularly challenging due to the dual-use nature of materials, equipment, and knowledge required for biological weapons programs, a problem easily relateable to AI. These challenges are intensified by the wide distribution of relevant materials and knowledge across various scientific disciplines and sectors, and by the fact that biological components are sometimes capable of reproduction, making traditional material accountancy-type verification methods ineffective. This resulted in so-called "functional substitutes" for traditional verification, which include provisions for national implementation through legal means (Article IV), mandates consultation and cooperation among parties (Article V), allows for complaints against breaches to be sent to the Security Council (Article VI), and requires parties to aid victims of biological weapon attacks upon Security Council determination (Article VII).

Similarly, a regulation agency could be instated, providing a place for AI stakeholders to discuss safety concerns and ask for auditing other proponents, and for governments to ensure their citizens safety is kept at the center of the discussions.

4.2 Founding a decentralized defensive research agency

In a recent article, Yoshua Bengio proposes a path to research on safe and aligned defensive AI [8]. He claims that research should not be widely published in the usual academic manner, especially if it leads to significant AI advancements, to ensure defensive AIs remain more powerful than potential so-called rogue AIs. This approach is similar to national-security or military research, and could use similar security protocols to avoid issues mentioned in section 3.3, with focus on safeguarding humanity against AI wrongdoing. Additionally, this research should be a collaborative effort across multiple nations to effectively deploy defenses (cyber threats know no borders), and to avoid concentration of power, which can threaten democracy and geopolitical stability.

As the group would be a network of collaborative democracies, it would prevent power concentration and the risk of a single point of failure, ensuring balance in case a democracy falls or creates a rogue AI, as other countries in the group would have comparable AI capabilities to maintain equilibrium. Having this network of independent research groups sharing progress with each other fosters diverse approaches and "coopetition," a mix of competitive pressure and collaboration, enhancing overall progress. Moreover, these labs should be independent, nonprofit organizations largely funded by governments, focusing solely on defending humanity against rogue AIs: this independence avoids conflicts of interest from commercial or national pressures, which could lead to an AI arms race, leading to deprioritize safety. An appropriate governance should ensure these labs do not use AI for military or economic dominance, with multilateral checks and balances.

Implementing such a global research initiative under a multinational umbrella would surely build a clear framework for safety research, with adequate governance and security measures. However, recent political and military events threatening democratic countries and their influence raise questions on whether such a system will be as global as the author envisions.

4.3 Physically monitor compute power

A recent report explain in details the intricate relationship between chip manufacturers, cloud providers, and leader AI companies, which form the broader framework that made recent AI developments possible [21]. To counter the rising market dominance, vendor lock-in, and safety concerns, the authors propose a set of measures, including: separating cloud and chip design (avoid ecosystem lock-in and promote competition), separating hardware and software (to avoid the repetition of Nvidia's CUDA software, by ensuring interoperability), separating AI model development from cloud infrastructure (avoid self-preference from actors active in both field), nondiscrimination obligations (to ensure a fair access to all customers and public interests to key compute components), merger enforcement (intervening early in mergers and acquisitions by Big Tech). The authors also point out regulations already targeting trade with China, aimed at restricting its access to high-end chips and graphic processing units. Beyond these political decisions, this approach paves the way for regulations specifically targeting physical components, serving as a relevant lever

for action.

Given the physical nature of computing resources such as chips and data centers, it looks feasible to impose physical limitations on compute power if needed. However, this might lead to the rise of "compute havens" – regions or countries with lax regulations regarding computing infrastructures to circumvent restrictions.

Conclusion

After examining the arguments central to the debate between open-source optimists, catastrophists, and the challenges of AI regulation, several development pathways emerge. These proposals share a common idea of establishing multilateral international structures, be it for creating frameworks for audits, research, or regulation. As artificial intelligence evolves rapidly, it becomes a global issue demanding quick decision-making and policy implementation.

Despite the potential risks, there is optimism for more responsible development, where the focus shifts from mere innovation to meaningful human progress.

Acknowledgements

I thank Cristian Dragos Manta, Paul Pacaud, Victor Baillet, Jeanne Salle, and the rest of the Turing Seminar attendants for our discussions and their valuable insights.

References

[1] Free software foundation homepage. https://www.fsf.org/.

[2] The llm index. https://sapling.ai/llm/index.

[3] Llama 2, 2023. https://ai.meta.com/llama/.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

[5] Nora Belrose. Ai pause will likely backfire. https://forum.effectivealtruism.org/posts/JYEAL8g7ArqGoTaX6/ai-pause-will-likely-backfire.

[6] Yoshua Bengio. How rogue ais may arise. https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise/.

[7] Yoshua Bengio. Personal and psychological dimensions of ai researchers confronting ai catastrophic risks. https://yoshuabengio.org/2023/08/12/personal-and-psychological-dimensions-of-ai-researchers-confronting-ai-catastrophic-risks/.

[8] Yoshua Bengio. Ai and catastrophic risk. *Journal of Democracy*, 2023. https://www.journalofdemocracy.org/ai-and-catastrophic-risk/.

[9] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022.

[10] BBC News Singapore Derek Cai. Asian games: China censors 'tiananmen' image of athletes hugging. https://www.bbc.com/news/world-asia-china-67002583.

[11] Ben Dickson. How open-source llms are challenging openai, google, and microsoft. https://bdtechtalks.com/2023/05/08/open-source-llms-moats/.

[12] Pieter Abbeel Dylan Hadfield-Menell, Anca Dragan and Stuart Russell. The off-switch game, 2017.

[13] Vox Dylan Matthews. Ai is supposedly the new nuclear weapons — but how similar are they, really? https://www.vox.com/future-perfect/2023/6/29/23762219/ai-artificial-intelligence-new-nuclear-weapons-future.

[14] https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal.

[15] Alessio Fanelli. Llama2 isn't open source, and why it doesn't matter, 2023. https://www.alessiofanelli.com/blog/llama2-isnt-open-source.

[16] United Nations Institute for Disarmament Research Filippa Lentzos. Compliance and enforcement in the biological weapons regime. *WMD Compliance and Enforcement Series*, 2020.

[17] Data for Good. Les grands défis de l'ia générative. https://issuu.com/dataforgood/docs/dataforgood_livreblanc_iagenerative_v1.0?fr=xKAE9_zU1NQ.

[18] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[19] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems, 2021.

[20] https://huggingface.co/huggingface.

[21] Sarah Myers West Jai Vipra. Computational power and ai. https://ainowinstitute.org/wp-content/uploads/2023/09/AI-Now_Computational-Power-an-AI.pdf.

[22] The Information Jon Victor. Why youtube could give google an edge in ai. https://www.theinformation.com/articles/why-youtube-could-give-google-an-edge-in-ai.

[23] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

[24] Dennis Layton. Chatgpt — show me the data sources, 2023. https://medium.com/dlaytonj2/chatgpt-show-me-the-data-sources-11e9433d57e8.

[25] Natasha Lomas. Zoom knots itself a legal tangle over use of customer data for training ai models. https://techcrunch.com/2023/08/08/zoom-data-mining-for-ai-terms-gdpr-eprivacy/.

[26] Stefano Maffulli. Meta's llama2 license is not open source, 2023. https://blog.opensource.org/metas-llama-2-license-is-not-open-source/.

[27] Meta. Llama2 licence on github, 2023. https://github.com/facebookresearch/llama/blob/main/LICENSE.

[28] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

[29] Andrew Ng. Feel the fear. https://www.deeplearning.ai/the-batch/issue-220/.

[30] Future of Life Institute. https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

[31] OpenAI. Introducing chatgpt. https://openai.com/blog/chatgpt.

[32] The Guardian Philip Willan. Race to find mafia's uranium bars. https://www.theguardian.com/world/2001/nov/09/afghanistan.terrorism5.

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[34] https://github.com/togethercomputer/RedPajama-Data.

[35] China's AI Regulations and How They Get Made. Matt sheehan. https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117.

[36] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[37] https://prod.hypermind.com/ngdp/fr/showcase2/showcase.html.

[38] Julian Togelius. Ai safety regulation threatens our digital freedoms. http://togelius.blogspot.com/2023/11/ai-safety-regulation-threatens-our.html.

[39] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[40] Rob van der Meulen. Gartner identifies six chatgpt risks legal and compliance leaders must evaluate, 2023. https://www.gartner.com/en/newsroom/press-releases/2023-05-18-gartner-identifies-six-chatgpt-risks-legal-and-compliance-must-evaluate.

[41] Vladimir Mikulik Matthew Rahtz Tom Everitt Ramana Kumar Zac Kenton Jan Leike Shane Legg Victoria Krakovna, Jonathan Uesato. Specification gaming: the flip side of ai ingenuity, 2020. https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/.

[42] Cameron R. Wolfe. The history of open-source llms: Early days. https://cameronrwolfe.substack.com/p/the-history-of-open-source-llms-early.

[43] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.