Exploring the Ethics of Open-Source Artificial Intelligence: Challenges and Considerations

Julian Roddeman   Daan Brugmans   Maarten Berenschot   Mats Robben

Group 1 — January 28, 2024

1 Introduction

This paper examines the ethical dimensions of open-source artificial intelligence (AI). Open-source AI refers to AI systems and algorithms that have freely available source code and learned model parameters, allowing for public use, modification, and distribution. The open-source nature of these models allows users to alter the models through fine-tuning or other methods. Closed-source models, on the other hand, do not allow for this level of freedom; users can only send input and receive output. Open-source AI models may allow users to bypass existing censorship measures aimed at preventing model misuse by altering the model's learned parameters and the creation of alternative uncensored open-source AI models. Open-source AI models, such as Mixture-of-Experts (MoE) 8x7B (Jiang et al., 2024), a pre-trained large language model comparable in performance to GPT-3.5, are gaining popularity. While the performance is undeniably impressive, there are growing concerns about potential abuse, as open-source AI may lack the censorship found in closed-source models, which prevents them from engaging in harmful activities such as providing instructions for illegal drug synthesis or creating dangerous explosives. Unfortunately, uncensoring is not the only risk associated with open-source AI models; other examples include the generation of hate speech, biases in production models that lead to misinformation, and the potential of AI models being used for malicious purposes, in addition to their primary purpose.

It has long been established that, for the field of software development, open-source licences encourage public scrutiny, collaboration, and innovation (Jesiek, 2003). This sentiment is also common among many AI researchers and developers. However, the rapid advancement of large AI models has called this way of thinking into question. The debate over the potential benefits and drawbacks of open-source AI has grown in recent years. According to Widder et al. (2022), using deepfakes to create non-consensual pornographic images is one of the problematic uses of open-source AI. This is the primary distinction between traditional software development and AI, in which the possibility of harmful and unintended use cases is significantly increased.

What specifically increases the risk associated with open-source AI when compared to closed-source alternatives? Open-source AI, by definition, provides unrestricted access to its source code and model parameters, allowing a diverse set of users, including those with malicious intent, to modify and re-purpose these technologies. Closed-source AI systems, while not without drawbacks, are typically controlled by their parent organisations, allowing for more stringent oversight and access restriction, potentially mitigating the risk of misuse. Furthermore, open-source AI's accessibility may accelerate the spread of powerful AI technologies, often outpacing the development of ethical guidelines and regulatory frameworks.

Recognising this changing landscape, global regulatory bodies such as the European Union have begun to recognise the multifaceted risks of AI. They have taken steps such as the AI Act, which focuses on AI's broad scope. Nonetheless, there is a noticeable gap, particularly in terms of open-source AI (Hacker, 2023). Current regulations focus on AI as a whole, often overlooking the unique aspects and challenges of open-source AI. This regulatory oversight is a critical component of this paper. It emphasises the need for extensive research in this field, which motivates this study to delve into the ethical dimensions of open-source AI, leading us to formulate the following research question.

Research Question: What are the ethical implications and responsibilities associated with the development, deployment, and distribution of open-source AI models, and how do they influence the future trajectory of open-source in AI in society?

2 Literature Review

To better understand the current landscape of open-source AI, we look at key contributions in the field. Figure 1 (McIntosh et al., 2023) shows that there is a significant focus on "Ethics / Ethical" in AI research. It reflects a long-standing concern for the moral aspects of regular AI. This highlights the importance of ethical considerations as an ongoing dialogue within the AI discussion.

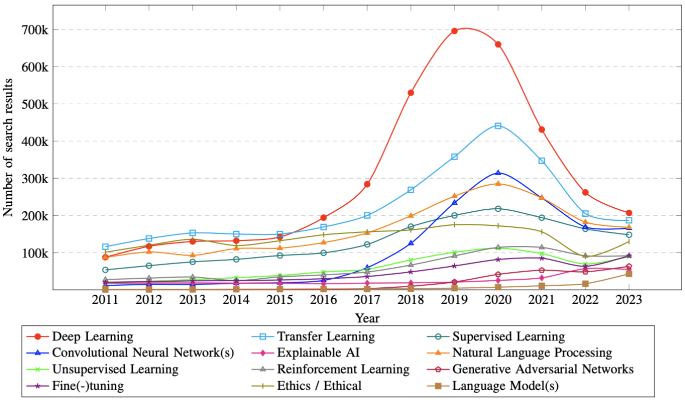


Figure 1: Number of search results on Google Scholar with different keywords by year. Source: (McIntosh et al., 2023)

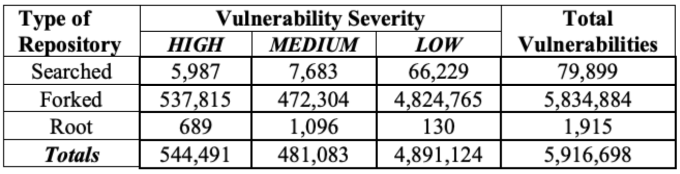
Despite widespread debate about general AI ethics, research into open-source AI ethics remains in its early stages. This literature review discusses relevant research that does exist, and with this paper, we hope to further cross this literature gap.

Work by Hibbard (2008) highlights the importance of open-source methodologies in encouraging ethical and safe advances in AI. He advocates for AI systems that are transparent and understandable in order to increase public trust, as well as highlighting open-source AI's collaborative potential in error detection and risk reduction. To prevent the misuse of open-source AI, Hibbard emphasises the importance of legal and economic reforms that align AI development with human values. The European Union has implemented a version of these AI reforms as per the AI Act. The paper *AI Regulation in Europe: From the AI Act to Future Regulatory Challenges* by Hacker (2023) discusses the AI Act. Hacker critically evaluates the EU's AI Act, recognising its pioneering role in AI regulation, but identifying shortcomings in addressing open-source AI risks such as toxic content (e.g., hate speech, misinformation, and non-consensual pornographic images), environmental concerns (e.g., excessive energy consumption in AI data centres), and hybrid threats (e.g., AI-driven cyberattacks).

Both Hibbard and Hacker emphasise the importance of taking immediate action, such as establishing protocols for regulated access to open-source AI systems. This begs the question of whether regulated open-source AI is still open-source, or if it is partially open-source, and whether this partially open-source type of AI may offer the best of both worlds in comparison to open- and completely closed-source AI.

In their investigation of ethical open-source AI, Widder et al. (2022) identify the risks associated with the open-source community's variable transparency and accountability practices. Their research, which included interviews with contributors to an open-source deepfake project, reveals the widespread belief in technology's neutrality as well as the challenges of managing its application. These findings highlight the complex ethical landscape in open-source AI development. For example, a deepfake tool created with open-source contributions may be intended for entertainment but may also be used to create misleading or harmful content, demonstrating the challenges of ensuring responsible use of such technology.

Acemoglu (2021); Kathikar et al. (2023) delve into the security risks associated with open-source AI, specifically examining the Hugging Face platform, a repository for pre-trained AI models used by over 22,000 organisations, including major companies such as Intel and Microsoft. With 2.6 billion downloads of models hosted by Hugging Face, the paper raises concerns about potential security flaws in these widely used resources, which can have far reaching ethical consequences. The authors carried out a thorough vulnerability assessment, connecting Hugging Face models to their GitHub code bases and analysing over 110,000 models and 29,000 repositories. The study discovered that root repositories have a significantly higher percentage of high-severity vulnerabilities (35.98%) than low-severity vulnerabilities (6.79%). Root repositories refer to the original source code repositories from which other repositories are derived or branched. They often serve as the primary or foundational code base for a project, and changes or updates in these repositories can significantly impact all the derivative or branch repositories (table 1). These findings are concerning, as the security of root repositories specifically is crucial. The significant number of high-severity vulnerabilities in essential AI repositories, such as the Hugging Face Transformers repository, again highlights a major security concern within the foundational infrastructure of open-source AI (Kathikar et al., 2023). These findings show the critical need for improved security measures in open-source AI, aligning with statements made by Hibbard (2008) and Hacker (2023) on the importance of regulation.



Type of Repository	Vulnerability Severity			Total Vulnerabilities
	HIGH	***MEDIUM***	***LOW***	
Searched	5,987	7,683	66,229	79,899
Forked	537,815	472,304	4,824,765	5,834,884
Root	689	1,096	130	1,915
Totals	544,491	481,083	4,891,124	5,916,698

Table 1: Number of Hugging Face models' GitHub repositories security vulnerabilities, grouped by repository type. Source: (Kathikar et al., 2023)

While Kathikar et al. (2023) concentrate on analysing security risks in already published open-source AI models and their hosting platforms, Nikolskaia and Naumov (2020) delve into the ethical considerations faced by publishers during the process of releasing open-source AI models, particularly those with dual-purpose potential that could be used for both peaceful and (unintended) harmful objectives. In line with previous literature, Nikolskaia and Naumov (2020) emphasise the need for ethical and legal guidelines to address the challenges of open-source distribution of dual-use AI technologies. These guidelines can encourage developers to systematically explore and identify any covert, harmful secondary uses of their AI models, resulting in a more responsible and informed approach to open-source AI development and deployment.

The papers discussed in this section all emphasise the need for ethical guidelines and principles in the open-source AI field. The studies emphasise a number of aspects of this need, including addressing dual-use dilemmas and legal responsibilities in AI technology, managing ethical challenges in the open-source community, and ensuring security against potential abuse.

The following sections of this paper will provide an ethical analysis of open-source AI. We hypothesise that this will further align with the overarching theme of the literature discussed: ethical guidelines, principles, and regulations in the field of open-source AI, and that this will have a significant impact on the future trajectory of open-source AI.

3 Ethical Considerations

Before proceeding with our ethical analysis, we first discuss potential issues and considerations of open-source AI. We will build upon the topics discussed here and incorporate them into the analysis.

For the purpose of our discussion and analysis, we emphasise the distinction between the different facets of open-source AI. Most open-source AI models have two main components: the model architecture and the final model parameters. The model architecture is often available as computer code that contains the instructions used during the training and production stages of the final model. The model parameters are the resulting product of the training of a model, such as learned weights and predetermined hyperparameters. For modern state-of-the-art AI models, the training of models is very computationally expensive; training can take weeks or even months, even whilst using clusters of specialised and expansive hardware. The learned weights resulting from such training can consist of millions, if not billions, of numeric values. This effectively transforms the model into a black box, with only the model input, output and architecture that can be inspected and reasoned about.

It is common practice for open-source AI projects to supply both a model's architecture and parameters. The availability of both constitutes the usability of the model. We argue that this availability should be critically assessed.

As the state-of-the-art in AI progresses, the possibilities that AI technology offers expand and intensify. This progress has also come with heightened hardware requirements for the training and use of state-of-the-art models. This means that, as AI gets more powerful, the barrier of entry of AI has increased: it is more difficult for low-resource groups to train and use state-of-the-art AI models. Open-source AI projects weaken or remove the barrier of entry by supplying a complete model architecture and learned parameters. As a consequence, it has become easier for low-resource groups to leverage state-of-the-art AI models. Some of these low-resource groups are those who leverage the availability of powerful state-of-the-art AI to perform tasks that are considered prohibited or malicious. This collection of groups will be referred to as *bad actors*.

Ajder et al. (2019) show examples of how bad actors leverage AI to produce deepfakes. Deepfakes are existing media that are altered to include the likeness of a person. A common example of deepfakes are images where a natural person's face is projected onto another, natural or otherwise, person's face. With current state-of-the-art AI, it is not only possible to realistically project a person's face onto an existing medium, it has also become possible to produce AI-generated media of that person by merely supplying a state-of-the-art AI model with existing media containing that person. Bad actors have used this possibility to create deepfakes of natural persons, without prior consent or approval of the individual depicted, that often include content that is intended to disrupt or disturb or that contains prohibited content. Examples of such deepfakes are AI-generated pornography of natural persons or intellectual property and AI-generated videos of politicians that aim to impersonate (Ajder et al., 2019).

Widder et al. (2022) conducted a set of interviews with leaders of open-source AI deepfake projects. These deepfake projects were set up with certain restrictions that aim for the prevention of prohibitive use of the AI model. However, the project leaders were also aware of users bypassing these restrictions. As a result, the deepfake AI projects can be considered to be *dual-purpose* models. For the purposes of our analysis, we take dual-purpose models to be AI models that are capable of producing outcomes that can be either benign or malicious, without being designed to favour one of these two types of outcomes.

Widder et al. found that some of the interviewed project leaders presented certain arguments that seem to include a notion of *technology neutrality*, a seemingly neutral stance towards an existing or developing technology. These stances of technology neutrality can be interpreted as a way to distance oneself from the responsibility of offering, developing, or maintaining a piece of dual-purpose technology. In the case of Widder et al., project leaders gave various arguments of technology neutrality distancing their work from bad actors. Widder et al. present the arguments presented to them as belonging to one of four (4) categories:

1. *Open Source Licensing as a Frame for Ethics*, where the contents of the open-source licences that define the limitations of end users in regards to the project dictate or outweigh arguments of ethicality. For example, since an open-source licence includes the right to alter and use the project in whatever way, then the project's community would not be responsible for changes made by an end user that result in prohibited material.

2. *Technological Inevitability*, where the assumption of the inevitability of technological progress related to their project justifies current practices. For example, the argument is given that there exist many other similar AI deepfake projects, so attempts at preventing prohibited use would cause end users to switch to a different project instead. Preventative measures are then taken to be useless as a consequence,

3. *Just a Tool?*, where there exists an emphasis on the project being a seemingly neutral tool with no intent to produce prohibited or malicious results and that such results are then only the products of the end user of the project.

4. *Setting and Enforcing Counter-Norms by Denying Support*, where malicious end user behaviour is discouraged by those offering the AI model, but not outright prohibited, the latter requiring the project's community to take measures.

These four types of arguments highlight the issue of technology neutrality as a means to distance oneself from the responsibilities of offering, developing, and maintaining dual-purpose AI models, creating a responsibility void of ethical AI model use.

4 Key Ethical Issues

We present here ethical issues that we consider key to the ethical analysis of open-source AI. These are as follows:

- The removal of barriers of entry that open-source AI provides has allowed for a great increase in accessibility of state-of-the-art models to low-resource groups.

- This accessibility has allowed for an increasing number of bad actors to use state-of-the-art AI models in ways unintended or prohibited by those offering the model, causing the model to be considered as dual-purpose.

- Those who offer, develop, and/or maintain dual-purpose models may attempt to distance themselves from the responsibilities and relevant inherent ethicalities of such models by using arguments of technology neutrality. This creates a void of responsibility for ethical model use.

- There exist security vulnerabilities associated with the offering, developing, and maintaining of specifically open-source AI models, that are not applicable to closed-source AI models.

- Malicious use of dual-purpose AI models may be discouraged by increasing awareness of the ethical consequences that come with malicious use.

- The regulation of open-source AI models may offer a solution in keeping state-of-the-art models available to low-resource groups, while also preventing malicious use.

5 Ethical Analysis

With key ethical issues established, we present our ethical analysis of the topic of open-source AI. We aim to clarify a path that may lead to responsible AI development and application, in line with societal values and positively impacting AI's future role in society. The following sections will critically analyse each ethical aspect, highlighting the importance of striking a balance between technological progress and ethical responsibility.

The removal of barriers of entry has had a great impact on the accessibility, innovation, and social impact of AI. At its core, the removal of barriers has led to an increase in the democratisation of AI: the newest and best models that were once only available to organisations with many resources are now also accessible for individuals and groups with few resources. This aligns with ethical principles by ensuring that the benefits of advanced AI technologies are distributed among all communities.

The ethical implications of this accessibility are considerable. Open-source AI enables marginalised or underrepresented communities to actively participate in and contribute to the advancement of AI technologies. This inclusivity not only promotes diversity within the AI community, but it also helps to reduce the risk of biassed or ethically questionable outcomes by incorporating a broader range of perspectives and experiences into the development process. Furthermore, increased accessibility is consistent with broader societal goals of reducing technological disparities. Open-source AI helps to bridge the technological divide between different segments of society by giving low-resource groups access to state-of-the-art AI models. This is consistent with the ethical imperatives of promoting fairness, justice, and equal opportunity, as it ensures that advancements in AI are not monopolised by a small group.

This accessibility has also come with risks. Making AI open for a varied range of users requires that efforts be made in reaching out to many groups with limited knowledge of AI, addressing issues like digital literacy, training, and support. These risks also introduce a new ethical issue: the possible misuse of an open-source AI system by bad actors. Bad actors may use state-of-the-art AI models to create deepfakes, misinformation, or even harmful surveillance, undermining the model creators' original intentions. The concept of dual-purpose AI models emphasises the value of ethical guidance in open-source development. Developers and maintainers should be held responsible for anticipating potential misuse scenarios and implementing safeguards to prevent unauthorised or harmful applications. To reduce the risk of unintended consequences, striking a balance between openness and responsible use becomes increasingly important. Clear guidelines, terms of use, and ethical frameworks can all be effective deterrents to misuse. Model developers must communicate and enforce ethical standards to ensure that users understand the intended applications and prohibited uses. The ethical responsibility for the model extends beyond its initial release, necessitating ongoing vigilance to adapt to changing risks and challenges. Furthermore, the emergence of dual-purpose applications sparks a broader societal discussion about the ethical application of AI. This discussion should include not only developers and model providers, but also policymakers, regulatory bodies, and the public. Ethical frameworks and guidelines must be adaptable to the dynamic nature of AI technology, requiring ongoing collaboration and adaptation to address emerging ethical challenges. When navigating the ethical implications of dual-purpose AI, it is critical to strike a balance between transparency and accountability. Open-source AI projects should foster a culture of responsible innovation by encouraging collaboration and actively discouraging malicious applications.

Aside from the potential misuse of the open-source AI, there is also a possibility of security vulnerabilities in the models themselves. Developers and maintainers must prioritise the discovery and mitigation of security flaws in their products. Failure to do so may result in unintended consequences, such as malicious actors exploiting weaknesses to gain unauthorised access or misuse of the models, as stated before. Ethical responsibility in this context necessitates open communication from model providers about existing security measures, potential risks, and secure implementation guidelines. The commitment to user safety and data integrity should be central to the ethical framework governing the distribution of open-source AI models. The ethical considerations extend to the development phase, where security flaws may be introduced inadvertently through collaborative efforts. The open and collaborative nature of these projects improves collective intelligence, while increasing the risk of missing potential security threats. Ethical development requires a proactive approach to identifying and addressing vulnerabilities throughout the development lifecycle. Developers should prioritise secure coding practices, conduct regular security audits, and foster a culture of reporting and addressing vulnerabilities in the open-source community. The ethical imperative is to ensure that collaboratively developed AI models are robust and resilient to potential security threats, thereby maintaining user trust and confidence.

Restricting access to learned model parameters might be a viable approach to limit potential misuse, particularly in the context of the rapid growth and anticipated future capabilities of AI models. Requiring ethical evaluations and licences for individuals and institutions seeking to use pretrained models could provide a balanced solution between complete open- and closed-source approaches to AI development. This approach could contribute to responsible AI use and mitigate the ethical challenges associated with unrestrained access to powerful AI technologies.

6 Conclusion and Discussion

The ethical implications of open-source AI are numerous, addressing issues such as privacy concerns, misuse of technologies and the balance between innovation and ethical safeguards. Open-source AI presents both opportunities and challenges, with the potential to democratise access to advanced technologies while also raising concerns about responsible AI use, unintended consequences and use by nefarious groups and individuals. We emphasise the important distinction between publishing model architectures and model parameters. While a model architecture can be inspected and critiqued, model parameters are nearly impossible to interpret. Furthermore, the costs of obtaining model parameters make training them from scratch nearly impossible for most individuals and small businesses. This barrier to entry may enable regulators to limit general open-source access to AI models. This would reduce the likelihood of malicious actors using these models for unintended purposes, but it would also stifle collaboration and innovation in the AI field. The paper suggests imposing ethical evaluations and licensing requirements on individuals and institutions wishing to obtain pre-trained model parameters.

Our study's limitations include the field's evolving nature, as ethical considerations in AI are constantly shaped by technological advancements, societal changes, and regulatory developments. Furthermore, the scope of the literature review may not include every emerging ethical issue in open-source AI, and the analysis provided is based on the state of knowledge as of January 2024.

Future research on the ethical dimensions of open-source AI should delve deeper into specific aspects, such as addressing dual-use dilemmas, spreading awareness of ethical issues and considerations for the use of AI models, and investigating the potential of regulated open-source AI. Investigating the efficacy of regulatory frameworks and ethical guidelines in mitigating the risks associated with open-source AI is critical for shaping responsible practice in the field. Furthermore, research could focus on the impact of open-source AI on marginalised communities, with an eye toward increasing inclusivity and fairness. The role of education and awareness in promoting responsible AI use, as well as the possibility of interdisciplinary collaboration among AI developers, ethicists, and policymakers, should be investigated in order to develop comprehensive solutions.

As open-source AI evolves, ongoing research and dialogue are critical to addressing emerging ethical challenges and guiding responsible AI development and deployment. Our findings add to the ongoing discussion about the ethical implications of open-source AI and serve as a foundation for future research in this rapidly evolving field.

7 Method Acknowledgements

The authors of this paper utilized ChatGPT to enhance the grammatical correctness and structural coherence of the content. While the tool contributed to refining the presentation, it's crucial to note that the views, ideas, and opinions expressed in the paper are entirely the authors' own. ChatGPT was employed as an aid in language refinement.

References

Acemoglu, D. (2021). Harms of ai. Working Paper 29247, National Bureau of Economic Research.

Ajder, H., Patrini, G., Cavalli, F., and Cullen, L. (2019). The state of deepfakes: Landscape, threats, and impact.

Hacker, P. (2023). Ai regulation in europe: From the ai act to future regulatory challenges. *arXiv preprint arXiv:2310.04072*.

Hibbard, B. (2008). Open source ai. *Frontiers in Artificial Intelligence and Applications*, 171:473.

Jesiek, B. (2003). Democratizing software: Open source, the hacker ethic, and beyond. *First Monday*, 8(10).

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2024). Mixtral of experts.

Kathikar, A., Nair, A., Lazarine, B., Sachdeva, A., and Samtani, S. (2023). Assessing the vulnerabilities of the open-source artificial intelligence (ai) landscape: A large-scale analysis of the hugging face platform. In *2023 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6. IEEE.

McIntosh, T. R., Susnjak, T., Liu, T., Watters, P., and Halgamuge, M. N. (2023). From google gemini to openai q*(q-star): A survey of reshaping the generative artificial intelligence (ai) research landscape. *arXiv preprint arXiv:2312.10868*.

Nikolskaia, K. and Naumov, V. (2020). Ethical and legal principles of publishing open source dual-purpose machine learning algorithms. In *2020 International Conference Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS)*, pages 56–58. IEEE.

Widder, D. G., Nafus, D., Dabbish, L., and Herbsleb, J. (2022). Limits and possibilities for "ethical ai" in open source: A study of deepfakes. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2035–2046.