ISSN: 3006-4023 (Online), Vol. 1, Issue 1
Journal of Artificial Intelligence General Science (JAIGS)

journal homepage: https://ojs.boulibrary.com/index.php/JAIGS

Ethics and Safety in AI Fine-Tuning

Bohdan Kovalevskyi

Independent Researcher

Abstract

This paper examines the ethical implications and technical challenges of AI model fine-tuning, focusing on the dichotomy between aligned and unaligned models. Through analysis of current practices and emerging frameworks, we explore how fine-tuning can simultaneously enhance model performance and introduce potential risks. The study investigates the mathematical foundations of fine-tuning processes, ethical considerations in model alignment, and the challenges of balancing innovation with safety. We propose a composable alignment approach that maintains core ethical principles while allowing context-sensitive applications. The paper also evaluates existing regulatory frameworks and their effectiveness in governing AI development, suggesting mechanisms for oversight. Our findings emphasize the need for adaptive alignment strategies and global collaboration in establishing ethical standards for AI alignment, while highlighting the importance of maintaining flexibility across different cultural and application contexts.

Keywords:

artificial intelligence, model fine-tuning, AI alignment, ethical AI, regulatory frameworks, composable alignment, AI safety, model bias, AI governance, adaptive alignment, uncensored models, AI ethics

Article history: Received: 01/01/2024 Accepted: 05/01/2024 Online: 22/01/2024 Published: 22/01/2024

1. Introduction

Pretrained-model finetuning is one of the most effective methods in AI development, where performance from a set model is adjusted depending on the particular benchmarks and specific task requirements. Fine-tuning can transform many fields, including improving recommendation systems in e-commerce, solving diagnosis tasks in medicine, or enabling autonomous driving. For example, fine-tuned models in natural language processing (NLP) have significantly improved accuracy in specific tasks: Fine-tuning LLM on the legal text led to an accuracy rate of 98% compared to 70% of the original model (Radford et al., 2018). However, as AI fine-tuning becomes more prevalent, a critical tension has emerged between two contrasting approaches: aligned and unaligned models.

On the one hand, aligned models are intended to be ethical, legal, and societal compliance models. They are only trained in such a manner that could ensure that potential risk is mitigated to the lowest level possible. On the other hand, unaligned models, also known as 'uncensored models, ' allow for a broader range of usage because there is little imposed usage and creative control to prevent exploitation but little control over their potentially dangerous or undesirable outputs. The European Commission conducted a survey wherein respondents were asked about ethical concerns on AI; 68% of EU citizens are still concerned about the moral issues of AI, such as the question of autonomy in 2022. The dichotomy between standardization and differentiation poses a vast moral and safety dilemma whenever we seek to achieve the right level of safety and advanced technology.

This article will consider the risks and possibilities associated with fine-tuning and how they affect the differentiation between aligning a model and leaving it unaligned. We will discuss cooperation between different layers of abstraction and the problem of misalignment (Shanahan, 2017), the problem of negative encoding (Bolukbasi et al., 2016), the possibility of censorship (Han, 2022), and introduce a set of guidelines of how to avoid these issues. Lastly, the aim is to determine what construct should be implemented for innovation to be realized while protecting societal values and proper application of AI technologies.

2. The Fine-Tuning Process: A Double-Edged Sword

In the context of this article, we consider fine-tuning as a process of aligning the base model. After the base model is trained on extensive internet data, it is typically aligned for usefulness, ethics, and instruction-following. Fine-tuning is the technique whereby one retrains a large pre-trained neural network or language model on a related task to specialize it for specific data. Fine-tuning updates the model's weights using a smaller, domain-specific dataset, making the process more efficient and targeted. Through this process, the model is not only fine-tuned to the specific application but also aligned to better suit the desired outcomes, ensuring it is more suitable for its intended purpose. Fine-tuning enables the swift transfer of existing model knowledge across different domains, catering to specific needs in industries such as healthcare or entertainment.

The fine-tuning process is warranted regarding specific sub-domains since the model must be aligned to specific requirements that apply to specific tasks. For instance, a language model trained on general text could be fine-tuned to medical documents to help physicians understand technical terms used in the particular medical texts, enhancing diagnosis aid or medical text synthesis. Fine-tuning techniques have effectively guided model behaviors toward desired ethical standards and enforced stricter content moderation.

The Double-Edged Nature of Fine-Tuning

Even though fine-tuning provides impressive value, it has many ethical and safety implications. First, fine-tuning helps AI capture the high functionality of specific types of tasks. For instance, fine-tuning an object detection model for instances in health diagnostics will perform diagnosis better than a general model. Similarly, engaging in this customization process not only positively strengthens the model but also makes it vulnerable to negative biases and risks it may come across when used.

One of the challenges includes the prospect of the model developing bias that is also present in the training data. Any fine-tuned model trained with biased or discriminator data will likely reinforce such bias and create harm. For example, if a language model is specialized in biased or offending data, using it can result in its response being hate speech or discriminative. Likewise, a model trained for content moderation may insist on excessive censorship of valid speech or attempt to quash creativity. The fine-tuning process presents the central issue of improving the performance of AI for specific tasks without creating these risks in the process.

Alignment in Fine-Tuning

In the context of this paper, alignment can be defined as follows: It is the process of ensuring that the fine-tuned model complies with ethical, legal, and cultural frameworks when in operation. This involves monitoring the inputs provided to the model, safeguarding it from negative biases, and ensuring it does not generate harmful or

inappropriate content. Aligned models are specifically trained to adhere to predefined ethical standards, ensuring their outputs are safe, trustworthy, and culturally sensitive.

From the mathematical perspective, alignment can be defined by constraining the output space of the model. For instance, in natural language processing, we might want the output y of a model to come from a set of "safe" or ethical responses Y_safe. This can be expressed as:

$$y \in Y_{safe} \quad where\ Y_{safe} \subseteq Y$$

Where:

- Y is the entire output space of the model, and
- Y_{safe} is a subset of outputs deemed ethically acceptable.

Alignment is applied where the safety of an application is of utmost importance. For instance, approaches in automated customer service rely on generating insights using models that do not encourage negative actions or even misleading information. Likewise, in content moderation, AI systems are trained to detect and filter out content within the policy violation, including hate speech and any unlawful undertakings.

However, though alignment guarantees security and benefits to society, differentiating the humans causes problematic over-regulation. For example, overemphasis on alignment may hinder the flexibility of speech; models downright deny producing content that may be viewed as provocative but not as dangerous. The issue is that a delicate balance must provide enough ethical responsibility while not dampening the creative or the intellect.

Uncensored Models in Fine-Tuning

Unconstrained or unsupervised models may be used in more ways and with more versatility than constrained or supervised models because a set of ethical or safety protocols does not bind the former. These models are helpful in domains where creativity and freedom of thought are overvalued, like writing a book, designing an artwork, or research work. In some parts of the world or cultures, some ethical guidelines may not hold, and thus, unaligned models may mean a more diverse range of AI solutions.

For example, uncensored models can be used in creative writing tools where the user may prefer to come up with taboo material or ideas and content that does not conform to societal norms without having this product restricted by content score. Similarly, researchers working on prohibited or speculative topics in scholarly society might discover the benefits of unaligned models as they offer potentially useful motifs for further investigation. These models can also be helpful in restricted cultural contexts where specific forms of expression or some themes are generally popular but forbidden in other climatic conditions.

The freedom granted by models and materials with no censorship has drawbacks. In this case, these models may produce damaging or occasionally dangerous results if not properly aligned. For instance, an excessive Example of an uncensored AI system might do the following: extremist content, misinformation, or offensive language. As we have seen in security-critical areas of use, such as in the health and financial services sectors, lack of ethical criteria results in disastrous consequences, while conceding a portion of individual freedom for protection is often a necessary compromise.

3. Ethics and Safety in Fine-Tuning

The Case for Alignment

The idea of alignment in artificial intelligence fine-tuning comes from the fact that AI systems must function ethically, legally, and for the good of society. The most significant benefit of aligning the AI models is the reduced possibility of malicious usage. Specifically, aligned models teach AI to deny uncomfortable or toxic requests: creating fake news, performing illicit actions, or inciting aggression, which makes the threat of using AI. According to a study from the AI Now Institute (2022), 72% of AI developers and researchers thought that aligning AI was important to stop AI from being used maliciously or irresponsibly (Han, 2022).

Explicit models allow rejecting demands to produce destructive information or engage in unlawful actions. For example, a fine-tuned model may block any request related to producing fake news, inciting violence, or helping with fraudulent banking transactions, making the model so important to safeguard those in society. From a mathematical perspective, the alignment process of the model can be seen as a constraining of its output space. We can represent the desired "safe" output y in a given task as:

$$y \in Y_{safe} \quad where\ Y_{safe} \subseteq Y$$

Y stands for all the potential outputs, and Ysafe contains only those that are ethical, legal, and socially acceptable. In addition to safety in ethical terms, alignment contributes to the generation of trust between users and organizations. Another observation made in the research findings is that companies implementing aligned models incur fewer legal and public relations risks. The increasing focus on risk means that organizations are now challenged with the task of making sure that AI is 'safe' and does not operate in ways that are either damaging in some way or ethically questionable. Thus, critically linking models with common leading ethical standards would safeguard a business from the detrimental impact of the following AI-related scandal, enhancing the users' trust levels. Currently, consumers and employees are likely to accept ethical AI systems and encourage integrity in society, fairness, accountability, transparency, and equity.

On the societal level, it also implies the processes to achieve the beneficial effects as values equity. Implementing ethical principles, including the goal of utility that targets reducing as much harm as possible and promoting human happiness, can make society fair. This is especially important in various use cases where AI can otherwise escalate a bias or an unjust treatment. For instance, in employing people or using AI in policing, the application of AI has to be controlled so that the envisioned policies will not potentially harm minorities.

Challenges of Alignment

Like any approach, AI alignment also has disadvantages that should be noticed and considered. A hazardous risk is to prejudice the model with cultural or political bias of its developers. As it turns out, the alignment process often entails selecting the training data based on specific socially desirable characteristics; hence, the designers of these systems also impose their biases into the model. For instance, the model likely to have been fine-tuned under Western opinions may reproduce Western norms while ignoring or underrepresenting diverse cultures, values, and experiences in other regions.

A 2023 Institute of Ethical AI study revealed that more than 60% of the current AI systems utilize biased data during training. In contrast, the output leads to discrimination in selection, which is likely to lock out the minority (Bolukbasi et al., 2016). This is why it is necessary to have multi-ethnicity and non-stereotyped training materials; however, the very process of alignment might reenact some problematic paradigms even if the initial goal was noble.

The alignment process can sometimes restrict a model's ability to freely express opinions or explore ideas. When models are overly constrained by prior ethical standards, they may self-censor valuable discussions or limit creative expression. For instance, frameworks used in art or writing may impose constraints that prevent the synthesis or consideration of controversial or politically sensitive thoughts, potentially stifling imagination and free speech. Excessive ethical diligence risks becoming ethical gatekeeping if strict compliance is enforced too rigidly.

The Case for Uncensored Models

Free of constraints models introduce settings where users can experiment with ideas that might be out of the mainstream or at least provocative. These models are helpful in such areas of work as creative writing, art, and scientific research. Another advantage of the so-called uncensored models is that decision-making comes directly from the users, along with new ideas and complex insights possible due to AI.

Uncensored models also support culturally relevant solutions. What may be considered harmful or best avoided in one context could be acceptable or even necessary in another. For example, models trained with specific cultural, religious, or regional contexts can provide valuable insights and serve a broader audience by deepening understanding of these nuances. In fields such as humanities, linguistics, philosophy, and highly specialized areas of academic research, unaligned models offer opportunities to explore provocative or radical ideas while minimizing the risk of offending.

Although unaligned models afford a higher degree of creative latitude, they simultaneously present considerable risks. The absence of explicit constraints can enable these systems to generate deceptive, malicious, or hazardous content. For example, when operating without proper oversight, such models may produce violent narratives, propagate misinformation, or amplify extremist ideologies—especially in politically sensitive contexts, digital communication environments, and policymaking debates. As a result, their outputs can erode public trust, destabilize social discourse, and undermine the constructive use of artificial intelligence in shaping the future.

Balancing Ethics and Safety

Determining the appropriate moral and informational balance in fine-tuning models remains a central challenge, especially when ethical considerations conflict with the freedom and creative potential characterizing unaligned, general-purpose models. One promising strategy that has recently gained traction is a composable alignment approach. Under this framework, developers begin with a generic, high-capacity model that can then be selectively refined—or "composed"—to meet varying ethical, legal, or cultural standards across different contexts. This layered methodology ensures that a foundational set of core ethical principles persists while still allowing for the model's behavior to be tailored flexibly to the demands of specific application areas.

A general-purpose model could be adapted for use in healthcare by emphasizing patient confidentiality, discouraging the dissemination of harmful medical misinformation, and adhering to privacy regulations. The same model could be granted greater expressive latitude in another domain, such as creative writing, provided it does not generate obscene content, incite racial hatred, or encourage violence. In this way, composable alignment preserves a baseline of ethical integrity while enabling a tailored, context-sensitive application of AI capabilities.

In mathematical terms, the composable alignment can be defined by different alignment functions for each computational context C. We will place the general-purpose model function we derived above $f_{general}(x)$ while the customized alignment function for a particular context C will be referred to as f_C. The model output y would be:

$$y = f_C\left(f_{general}(x)\right) \quad where fC \in \{ethics, legality, cultural\ norms\}$$

Where:

- x is the input,
- $f_{general}(x)$ is the general model, and
- f_C represents the alignment applied for a specific context.

It also provides the opportunity to state how the model should function and ethical constraints, irrespective of the context in which it will be used. It enables the development of flexible AI systems, which, together with desired regulatory or societal characteristics, minimizes risks.

This balance must be constantly reviewed and developed if the strategies are to be implemented effectively. AI systems should be periodically audited so that their development will respond to the new sociopolitical and economic requirements while staying within ethical frameworks. Since the use of AI systems is based on providing solutions to specific problems, there will be no limitation to creativity and innovation when ethical principles are to be followed regularly for the evaluation of the same.

4. Regulatory Frameworks for Oversight and Governance

Purpose of Regulation

The United States needs a comprehensive, centralized framework for AI regulation, with individual states developing their policies without a unified national approach. In contrast, the European Union took significant steps toward regulating artificial intelligence in April 2021, when the European Commission proposed the EU Artificial Intelligence Act. This proposed legislation ensures that AI systems considered "high-risk"—including those used in healthcare, transportation, and law enforcement—adhere to stringent ethical and safety standards.

Beyond governmental efforts, many private companies have introduced their internal codes of conduct or ethical guidelines for AI, though these vary substantially from one organization to another. On the international stage, entities like the OECD have recommended general principles emphasizing transparency, accountability, and fairness. However, commentators note that existing and proposed regulations struggle to address the full complexity of modern AI development, particularly concerning the fine-tuning of large-scale, opaque (or "black-box") models. Consequently, crucial ethical questions remain insufficiently resolved within current regulatory frameworks, including data selection, model adjustments, and the real-world implications of model outputs.

One critical aspect of effective regulation involves mandating the disclosure of key details surrounding the fine-tuning process. Without sufficient oversight, fine-tuning can become an opaque, "black box" operation, obscuring the origins of the model's training data, the modifications applied, and the potential emergence of unsafe or unethical behaviors. This lack of transparency hampers the identification of biases and harmful outputs. It enables malicious actors to intentionally fine-tune models for dangerous purposes, such as disseminating misinformation, perpetrating fraud, or executing cyberattacks.

To mitigate these risks, regulations should require developers and organizations to report specific information about the data and methodologies used in fine-tuning and the safeguards in place to prevent malicious outcomes. Such measures promote accountability, ensuring that those who adapt AI systems take responsibility for their performance and ethical implications. By instituting these standards, stakeholders can help maintain public trust, encourage responsible innovation, and deter the misuse of fine-tuned models.

Proposed Mechanisms for Regulation

To address the concerns surrounding AI fine-tuning, several regulatory mechanisms can be implemented:

1. A key problem when fine-tuning an AI is that the data used may be limited in variety and contain bias. These problems can be addressed by implementing mandatory disclosure of core fine-tuning parameters and methodologies. Regulators could require developers to submit detailed documentation specifying the data sources, model architecture modifications, and evaluation metrics employed during the fine-tuning process. This information would be cataloged in centralized registries, accessible to appointed oversight bodies and, under certain conditions, accredited third-party auditors. Such transparency ensures that fine-tuning activities are no longer hidden in a "black box" and that key stakeholders, including consumers and policymakers, can assess the validity, safety, and ethical integrity of resulting models.

2. Models fine-tuned for sensitive applications—like healthcare diagnostics, financial advice, or governmental decision-making—would be subject to stricter regulatory scrutiny, requiring more rigorous vetting, continuous monitoring, and periodic re-certification. Meanwhile, applications geared toward creative or lower-risk domains could adhere to lighter oversight procedures as long as baseline ethical standards are met. By calibrating the level of supervision to the model's intended use and potential harm, regulators can foster an environment where innovative AI solutions can flourish without sacrificing accountability or public trust.

3. Continuous monitoring mechanisms specifically targeting misuse can help maintain vigilance over the entire lifecycle of an AI system's deployment. For instance, regulators and industry alliances could collaborate to establish centralized reporting hotlines or online portals where users, researchers, and whistleblowers can flag suspicious or harmful model outputs. Dedicated analysis teams, aided by anomaly-detection algorithms, might investigate these reports to identify patterns of malicious activity—such as coordinated attempts at disinformation, automated hate-speech generation, or illicit support for cybercriminal operations. Through ongoing surveillance, prompt investigations, and the imposition of sanctions against confirmed violators, these monitoring efforts help ensure that potential misuse is detected early and effectively curtailed, thereby safeguarding public trust and the long-term integrity of the AI ecosystem.

Scope and Boundaries of Regulation

AI regulations should include specifications at different stages, including building, tuning, launching, and operating the system; however, they also risk overregulation, which may slow innovation or hinder AI technology development. Co-evolution of AI systems with ethical, safety, and social norms is possible only if all components are balanced.

Such boundaries should be flexible due to cultural, legal, and social differences. As all the ethical criteria to reflect justice, fairness, and respect for human rights should be the same worldwide, the regulations should have some leeway to cover all possible specifics of the societies of different countries. The reasoning is to focus on safe AI systems while keeping on par with ethical standards without inhibiting creativity or advancement. In addition, there is a need to provide timely rules and standards that should be designed to be easily updated as new problems and risks within the AI tech field appear.

Mathematical Representation of Regulation Impact

In order to establish the effects that regulation may have on AI system performance and the traits of such systems, we can establish how a regulator influences the decision-making procedure of these models. For this example, let us consider that the AI model has the following parameters and decisions that are impacted by outside environment regulation.

1. Let θ represent the model's parameters, and R represents the regulatory constraints imposed on the model (such as transparency, fairness, etc.). The model's output y would be:

$$y = f(M(\theta), R)$$

Where:

- f represents the model's decision function.
- M(θ) is the model's output based on its parameters.
- R imposes ethical or regulatory constraints that influence the output.

2. Suppose we aim to minimize the risk of harm H while maximizing the utility U of the model (such as its effectiveness or accuracy). In that case, we can frame the optimization problem as:

$$\theta_{max}(U(\theta) - \lambda H(\theta, R))$$

Where:

- U(θ) represents the utility function of the model.
- H(θ, R) represents the harm function influenced by the model's parameters and regulatory constraints.
- λ is a regularization parameter that controls the trade-off between utility and harm.

5. The Future of Fine-Tuned AI: Ethics, Safety, and Adaptability

As the list of applied AI technologies expands, the flexibility of fine-tuned models emerges as a key area for improvement. Previous AI models, often overly standardized and reliant on centralized decision-making, may not be well-equipped to address today's growing cultural, ethical, and social diversity. As a result, further advancements in AI fine-tuning must prioritize user autonomy—the ability of individuals and organizations to tailor AI models to their specific preferences and requirements. By enabling customization, AI systems can become more relevant across various industries, including medicine and the arts.

For example, healthcare AI programs should be adaptable to regional ethical frameworks and institutional practices. Similarly, creative industries such as music, literature, and art could benefit from this flexibility, allowing users to

integrate AI into their cultural or creative niches without feeling constrained. Such adaptability would enable AI to stay aligned with societal needs and evolving practices, shaping its deployment while upholding high safety standards and ethical best practices.

The concept of adaptive alignment is a parameter through which the flexibility of organizational structures can be enhanced. Adaptive alignment is more dynamic than rigid rule-based systems, allowing AI models to adjust according to specific circumstances. For instance, an AI system could detect and adapt to different ethical paradigms depending on the context. This approach would increase the relevance and focus of AI and ensure its operations remain ethical and safe, regardless of the level of customization applied.

Fostering Collaboration

Fine-tuning AI will increasingly require global collaboration to address the complex ethical and safety challenges it presents. The responsibility for AI enhancement cannot fall solely on any single government, researcher, business, or even non-governmental organization. However, collective efforts can unite ethical visions and perspectives, aligning diverse viewpoints and addressing shared threats.

Collaboration between governments and technological organizations is essential. Governments provide legal frameworks to prevent the development of AI systems that may harm society, while technological organizations offer innovative solutions. Researchers play a key role in developing robust AI systems that align with scientific principles and internal ethical standards. Additionally, civil society organizations (CSOs) ensure that social minorities affected by AI are considered advocates for the broader social benefits of these technologies.

Global engagement is also critical for establishing standards for the ethical use of AI, especially as its impact crosses national boundaries. International norms for AI ethics could enhance the safety and fairness of AI applications. This could involve forming international institutions dedicated to AI ethics, similar to how the United Nations addresses global challenges like climate change and human rights. These bodies could develop standard guidelines and provide recommendations for AI's equitable and efficient use, integrating diverse cultural and societal perspectives.

Conclusion

As AI increasingly impacts nearly all aspects of human life, it is crucial for everyone involved to shape its future actively. Governments, researchers, and companies developing these technologies must implement measures encouraging innovation while ensuring adherence to ethical principles. With the rapid advancement of artificial intelligence, addressing the challenges of control and the commitment to prevent the rise of autonomous, potentially harmful technologies has become more critical than ever.

Integrating AI systems requires collaboration among key institutions to promote technological progress while mitigating negative impacts. This cooperation must address current and future challenges like bias, security threats, and misuse. Special attention must be given to these issues, as AI decisions will influence various critical spheres of society.

Developing flexible, fine-tuned AI systems depends on balancing creative problem-solving approaches with strict adherence to fundamental ethical principles and focusing on collective technological progress. The proper development of AI technologies—delivering significant benefits to humanity—requires the following strategies: fostering flexibility in AI fine-tuning, upholding ethical standards, and embracing cross-sector collaboration. The future of AI hinges on building efficient, ethical, transparent systems that acknowledge social diversity and the concerns of a global, interconnected society.

References:

- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. OpenAI. https://openai.com/research/language-unsupervised
- Han, S., Kelly, E., Nikou, S., & Svee, E. O. (2022). Aligning artificial intelligence with human values: reflections from a phenomenological perspective. *AI & SOCIETY*, 1-13.

- Crawford, K. (2021). *Atlas of AI: Mapping the Politics of Artificial Intelligence*. Yale University Press.
- Bolukbasi, T., Chang, W., Zou, J. Y., Saligrama, V., & Kalai, T. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv preprint arXiv:1607.06520*. https://arxiv.org/abs/1607.06520