Developing Ethics via Framework Within Machine Learning Systems

(3809 words)

A Research Paper

Presented to The Faculty of the

School of Engineering and Applied Science

University of Virginia, Charlottesville ,Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science in Computer Science

Akira Durham

Spring 2025

On my honor as a University student, I have neither given nor received unauthorized aid

on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Advisor

Sean Murray, Department of Engineering and Society

Introduction

Machine learning has become increasingly integrated into daily life through common technologies, yet biases in artificial intelligence systems remain a significant scientific and cultural issue. Despite the rapid development of machine learning (ML), artificial intelligence (AI), and natural language processing (NLP) systems, bias has continued to be a source of weakness for the effectiveness of these models (Nazer, 2023). While developers focus on improving the capabilities and toolset of these ML systems, not enough attention is placed on combating the rampant growth of bias and the lack of ethical structure inherent in these models, causing the negative effects to be pushed onto the users who are unaware of hidden discrimination (Livingston, 2020). Unfortunately for developers, eliminating bias within models is a slippery task as models heavily depend on large datasets that are expensive to parse for systemic discrimination and implementing ethics checks while targeting malicious prompts is computationally expensive on top of the costs to run the model itself (Ntoutsi et al., 2020). To address the issue of bias, historical and ethical analysis can be used to examine how organizations combat bias in ML systems and to evaluate the effectiveness of existing regulations in reducing systemic discrimination.

Effectively combating bias in machine learning models requires a combination of awareness during dataset curation, ethical ML regulations, and continuous monitoring to reduce systemic discrimination and ensure fairness in ML systems. Data is a large contributor to bias in ML systems as it provides the basis from which these systems learn and develop their patterns from, therefore by addressing systemic issues from the root we can stop unnecessary spreading of representational bias. Ethical regulations play a key role in ensuring model to model equity in reducing discrimination, as well as serving as the penalty for creating malicious or negligent models. Finally, monitoring throughout the development and deployment phase is essential in actively taking steps to reduce bias during model creation as well as being able to follow through with real-time updates if the model produces discriminatory output. By integrating these strategies, ML systems can become more equitable, accountable, and resistant to bias, ultimately improving their reliability and societal impact.

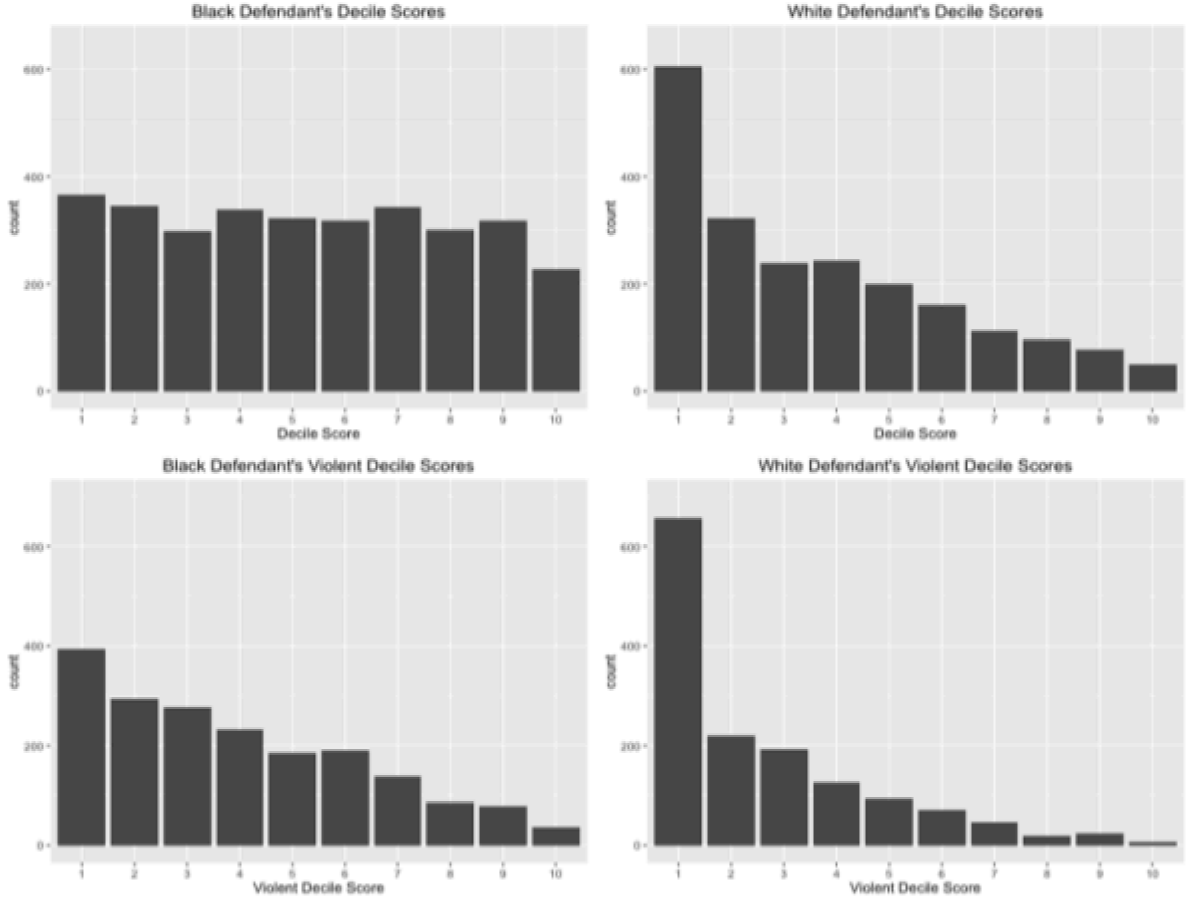
Problem Definition

To address these issues, it is essential to identify existing deficiencies within the current practice of reducing bias in ML and analyze the need for an ethical development framework. Before diving into these topics, a baseline understanding of ML ethics and its importance today must be established. ML ethics can be broadly defined as the ethical domain application of ML, supported by Taylor Grenawalt's description of ML ethics as "moral principles involved in designing, implementing, and deploying machine learning algorithms" (Grenawalt, 2023). While Grenawalt'broad definition allows one to generalize ML ethics, the specific ethics of each model depends on multiple factors such as its impact potential, human oversight levels, as well as domain-specific features. James Moor, a professor from Dartmouth College, describes the importance of ML ethics as a "focus on developing limited explicit ethical agents. Although they would fall short of being full ethical agents, they could help prevent unethical outcomes" (Moor, 2006, p. 21). Moor's realization of the potential uses of ML to prevent human bias is the key point here. Through ethical ML systems, processes can begin to remove interpersonal discrimination. However, that assumes the model's ethics are at least as good as a person's. By introducing this caveat, the importance of creating ethical models becomes apparent as technology continues to evolve (Shadowen, 2019). Current ML ethics implementations stem from a processing stage, before and/or after the model runs, to clean the data as well as validate the output as a way for software engineers to confirm the model is working as expected, which could involve processes that use NLP in large language models to assess prompt danger levels or ensuring a statistically accurate valid output to name a few examples. While the processing stage(s) in ML systems partly addresses deviations from expected input and outputs of the model, these safeguards do not ensure the developers or application of the model are inherently ethical.

Case Studies

Examples of unethical applications of ML are surprisingly common in today's technological landscape. By utilizing historical analysis on these unethical applications, one can understand the underlying causes that lead up to these cases, as well as the harsh effects of using the unethical models, allowing developers to address these mistakes for future models. One landmark case would be ProPublica's analysis of a Florida justice system's racially biased recidivism algorithm, which was designed to calculate the likelihood of a convicted criminal to reoffend. Figure 1 below shows the data discovered by ProPublica during their investigation, with the top graph displaying non-violent crime recidivism prediction scores for white and black defendants respectively, and the bottom shows violent crime recidivism prediction scores for white and black defendants. Within Figure 1, there is a clear low-score skew for white defendants, meaning that white defendants on average receive lower scores compared to black defendants in comparable cases covered by the study. This algorithm was statistically identified as systematically assigning high risk scores to black individuals with no prior history compared to regularly offending white individuals for similar crimes, resulting in 77% higher risk scores on average disparity for non-white defendants in comparable cases (Angwin et al., 2016).

Figure 1. ProPublica COMPAS Recidivism Results



The graph displays the recidivism algorithm scores for non-violent and violent crimes, separated by white and black defendants (Angwin et al., 2016).

A similar binary bias case would be Amazon's 2015 ML resume engine that aimed to reduce manual resume review work, but introduced gender bias by locating female names or women-orientated clubs and reducing the overall score as a result, as the dataset used to train the engine consisted of Amazon's male-dominated staff (Iriondo, 2018). A 2024 article by ACLU of Minnesota details the current bias in facial recognition technology, stating a 33.9% higher error rate for darker individuals in accurate face detection, which in one case lead to a false arrest and court case as the individual was detained solely on facial detection utilized by local police rather than evidence (Fergus, 2024). These three case studies are just the tip of the iceberg in terms of

ML ethics leading to discrimination and bias highlighting the real-world impact of unethical ML systems, as well as demonstrating the need for continuous development towards ethical development and use of these models.

Immediate costs to address these concerns would involve taking down biased models and either re-working the existing model or replacing it with a more ethical counterpart. Considering the similar nature of both models, the time aspect in a hot-fix situation is smaller compared to developing a replacement model, although both cost substantial resources in terms of company down-time on the model use case as well as software engineer time. Other costs for these biases would include lawsuits against proven discriminatory models against the companies that created them, as well as customer or stakeholder engagement with the product and company. Key examples of AI discrimination lawsuits would include Mary Louis' $2.2 million settlement win against a third-party rental scoring algorithm and an ongoing class action lawsuit against Workday, a job management platform, for "systemic discrimination" (Bedayn, 2024; Lyman 2024). While additional development to reduce discrimination potential adds monetary and time costs, the effect of reducing the likelihood of these biased models outweighs the potential for harm against individuals in these discriminated groups.

Literature Review

Current frameworks and committees to oversee these ML systems are usually ad hoc and vary among companies and countries, leading to differing regulations and standards, which are not apparent to the common users (Jobin et al., 2019). Jobin et al. (2019) further contends that while there has been an increase in AI guideline activity in both public and private sectors, there is "significant divergence in the solutions proposed to meet the ethical challenges of AI" (p. 13). Existing frameworks and regulations either address large scale ethical ideas leading to non-actionable goals during development or these guidelines promote fairness in specific use cases. This leads to improvement in one subject, but not the broader field such as FairML reducing successfully improving model fairness but only in classification models (Burgard & Pamplona, 2024). To systematically reduce the potential of discriminatory models, the use of an ethical development framework can be employed to provide a consistent baseline across models as well as ensure minimum standards of care for tech companies. Floridi and Taddeo support this idea in their paper about data: “ethics should be developed from the start … as an overall framework that avoids narrow, ad hoc approaches and addresses the ethical impact and implications” (Floridi & Taddeo, 2016). This ethical framework would involve the creation of the following processes: data cleaning, data bias detection, bias-aware development, and output validity checks. Improving ML ethics requires a stable, consistent development framework that is enforceable, scalable, and adaptable to support the quick progression in the ML field as well as promote accountability and fairness, ensuring these models contribute to a more equitable society.

Research Approach

The previous section consisted of historical analysis via a literature review on current implementation strategies for ethical ML systems, case studies on previous ML models, and reviews of research papers regarding ML ethics. Historical analysis is the primary analysis type in this paper because of its ability to discern patterns over time as well as the fact that much of the work being conducted is through new research, allowing one to compare methodologies to uncover the changing focuses of the ML field. By conducting historical analysis, the trend of inconsistent policies on ML ethical frameworks as well as the injustices they perpetrate become apparent. These inconsistent policies set up the underlying reasons on why an ethical framework is necessary for future developments. The case studies also display the continued use of discriminatory ML models without being checked by the developers or people who created the model, cementing the fact that bias is not a new problem in ML systems. Analyzing the historical trends is key in ML as the field makes fast strides in development, bringing new technologies, frameworks, and abilities for ML models forward, expanding the potential use cases but also expanding the potential for malicious or discriminatory models to exist. Developers can build new models with these past examples in mind, allowing for more informed construction that utilize new technologies to combat the previously found issues.

The sources analyzed in this paper mainly involve research papers in the areas of ML ethics or surveys of multiple research papers providing a consensus and broad overview of the topic. By utilizing journal-published research papers, the authors' analysis was reviewed and generally up-to-date upon their time of publishing, providing current academic understanding of ML ethics. This paper also uses website articles to provide further context for the case studies and their impact on the affected individuals. By combining these different sources, this paper combines academic knowledge with real-world effects, highlighting the results of releasing models without extensive bias checks during the development phase. Future developers can utilize this study to examine what factors played a part in ML ethics, major case studies and their effects, as well as expanding on the topics with more recent case studies and research, continuing the trend of analyzing the development of ML over time.

When conducting the historical analysis, the journal entries and articles were viewed with a virtue and deontological ethical framework, as these ideologies serve to enforce the idea of doing the right thing and creating rules that guide people towards ethical behaviors. Virtue ethics contends that the actions one takes should be the actions a virtuous person would take, and deontological ethics states that one should follow rules that are set in place to guide people along an ethical path. Scholars generally agree with this definition, as discussed by Rosalind Hursthouse and Glen Pettigrove, where virtue ethics "may, initially, be identified as the one that emphasizes the virtues, or moral character, in contrast to the approach that emphasizes duties or rules (deontology)" (Hursthouse & Pettigrove, 2023). A virtue ethics lens enables a reader to view the injustices dealt within each case study, further understanding why one must take action against these continued injustices, as simply standing by and not taking action is violating the virtues one should promote. Deontological ethics supports the creation of a framework or guidelines that would provide developers with steps to follow to minimize bias within their models. By combining these two ethical frameworks, the next step in the overall development of ML systems would be to address injustices from previous models, and create systems that would help to prevent discriminatory models from being made. This next step reflects the stances taken in this paper through case study analysis and ethical framework development, while ensuring marginal groups are not taken advantage of through the use of ML, something that could happen under the utilitarianism ethical framework. Utilitarianism argues that as long as actions are for the greater good of the majority of people, they are justified, creating an inequitable situation for smaller minority groups, directly conflicting with a virtue ethics framework. By integrating historical analysis with virtue and deontological ethics, this paper emphasizes the need for a standardized ethical framework to guide ML development. Recognizing past failures and their consequences enables more responsible innovation, ensuring that future developments focus on fairness and accountability. With this understanding of the analysis, the next step is to explore insights on how to create an effective framework that will push the future of ML in a more equitable direction.

Insights

Bringing the research together highlights the necessity of an ethical ML framework that bridges the gap between rapid innovation and consumer protection against discrimination. The case studies and literature review reveals a common trend that ML systems often perpetuate biases due to flawed training data, insufficient human oversight, or inherent model problems. Ethics are often underprioritized in models, compared to their functionality, creating models without protections against attacks on the model and discriminatory outputs from the model. Ethical considerations must be embedded in ML development from the start this includes planning and integrating ethical factors into standard ML processes and processing stages for data. While a per-model basis solves immediate issues, this does not address the underlying problems of continued biased models that destroy people's lives due to an oversight made by developers and their clients. A consistent framework is necessary to prevent ad hoc solutions that fail to address the systematic issues present in ML development and the world. Implementing a framework directly increases developer accountability by placing the onus on developers rather than poorly managed regulations and out of date guidelines. Without a structured approach, accountability remains ambiguous, allowing companies to shift the blame to external factors such as data availability rather than their internal decision-making that approved the biased model without proper validations.

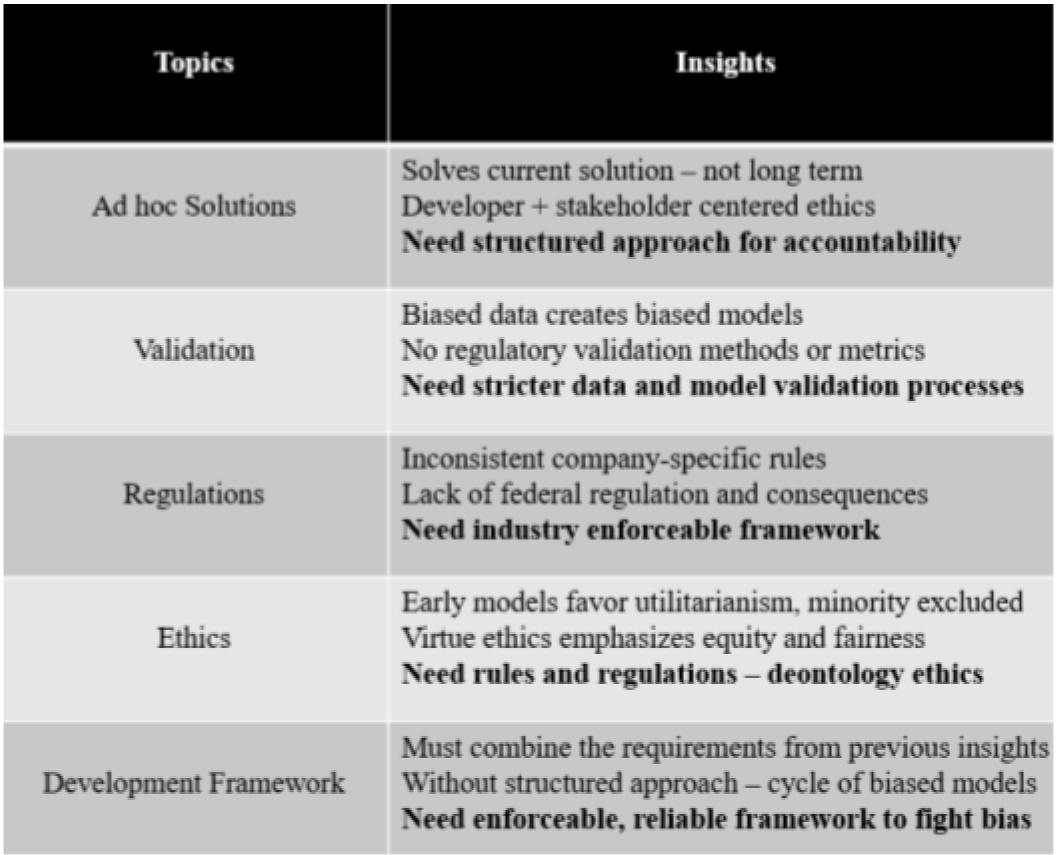
The historical analysis indicates that bias is often created through the data, often through its collection methods involving biases in some form however ignorance does not absolve someone of wrong doing, especially when that model results in false criminal accusations or economic disparities. While the data often creates inherent biases that are difficult to detect, the responsibility to verify the data is on the data collectors, developers, and stakeholders to ensure the data is proper for their use case and is representative. Data should not be a scapegoat for poor management and verification of trends in the data that are not representative of the public or specific data application, meaning that while data can include biases, there should be processes that clean the data of negative biases. Towards this end, companies must implement stricter validation and monitoring processes to hold developers accountable for the potential outcomes of the model to a certain degree. For example if someone jailbroke a model, then that model's output is no longer a valid representation of what the developer worked on, but during normal operations, the model should not continue any discriminatory practices or views. The challenges in keeping developers accountable come from weak and unspecific regulations, resulting in companies focusing on efficiency and development speed over ethics, especially in the technological race in ML that is happening today. However, if developers take a more proactive role, regulations might take on an enforcement role. This would lead to regulations not creating rules but enforcing, creating a loop of regulations' enforcement being proportional to the self-regulation of companies.

Current regulations demonstrate inconsistencies across companies and countries, leaving consumers confused on what exactly is being promised when using one of the countless ML tools, as well as highlighting the need for some centralized regulations to ensure a baseline level of guarantees. Several sources reinforce the idea that there is a lack of regulations, where claims that "developers and deployers of AI systems will operate in an increasing patchwork of state and local laws, underscoring challenges to ensure compliance" and "On May 17, 2024, Colorado enacted the first comprehensive US AI legislation … will go into effect in 2026" (Reem, 2024). Similarly to Reem, a 2025 AI legislation overview article by the Software Improvement Group asserts that "the U.S. lacks a comprehensive AI Act; instead, its strategy revolves around fragmented policies” and “AI’s rapid evolution has outpaced regulatory frameworks … the ISO has published a number of standards to benefit businesses adopting AI, whilst the Organization for Economic Co-operation and Development released a similar series of AI principles”, where both articles indicate a need for centralized, federal regulations as well as consistent frameworks for companies to adhere to (“AI legislation in the US: A 2025 overview”, 2025). Existing guidelines are often on a per company basis, meaning that the rules which dictate the security and bias levels of models are subject to stakeholder whims, putting developers and consumers on the receiving end of bad practices. Wider regulations are often broad and vague, allowing governments to keep a foot in the door without properly creating laws to address bias and discrimination within models as the field continues to evolve quickly. A unified standard for ethical ML development is necessary to create meaningful improvements in reducing discrimination throughout the entire industry and ensuring companies are held liable for their models. The lack of enforcement in current regulations due to their broad nature and company specific rules leads to the perception that ethical safeguards are costly and unnecessary. Companies often reduce funding for ethical development and allocate less money for ensuring strong internal regulations as these rules do not often return a profit and there are no financial penalties for biased models. With an industry-wide framework and increased regulations, companies will be more accountable as a result of these regulations, reinforcing ideas of long-term success and a reduction in model bias.

Historically, ML models favor a utilitarianism approach, where the benefit of the majority supersedes the wrongdoing of the minority in favor of higher overall happiness. This idea can be clearly seen in the Amazon case study because the majority of employees and applicants were men, putting women at a distinct disadvantage even when their experience level was comparable to their male counterparts. By utilizing a virtue ethics approach during this paper, the framework ensures that models are designed with fairness as a core principle, with equity at its core and anti-bias measures built-in. Through this injection of virtue ethics, mistreatment of minority groups no longer becomes acceptable as long as no one investigates or audits the model, the non-discrimination factor will be an integral part of the model to begin with. Alongside virtue ethics, a deontological approach would further exemplify the need for stronger regulations and demand penalties for flagrant violations. Deontology would resolve disputes of company policies versus other policy or regulations by removing the case by case comparisons, and creating a level playing field for these companies to build off from. The failure of utilitarian ethics indicates a need for standardized ethical development practices to continue to decrease the spread of bias in data and models. If fairness and equity are built into ML systems from the start, then companies will not have to constantly update their own policies with new discoveries, but have a steady and secure framework to apply on their new models.

To address these growing concerns, an ethical development framework is necessary to ensure company compliance, consumer safety, and model accuracy. This ethical framework must be integrated with regulation to be enforceable while maintaining the ability to adapt quickly to new technologies and methods as a result of the quickly developing ML industry. The framework must also utilize a combination of legislative and ML domain knowledge, to ensure up-to-date regulations and enable the framework to more dynamically evolve over time through ML domain knowledge from subject matter experts. ML models must be created with built-in bias detection, transparency measures, and human oversight at every stage to hold people accountable while reducing sources of discrimination. Without a structured approach, the cycle of biased models and reactionary band-aid fixes will continue, eroding public trust in ML systems and indicating to stakeholders that they can get away with biased models. A unified framework would provide not only a foundation for equity, but introduce accountability, encouraging innovation without sacrificing the protection of the consumers. These insights are visualized in Figure 2, summarizing the different aspects of the ethical ML framework issue as well as their interactions. As ML continues to shape the world and make technological strides forward, its development must be guided by principles that prioritize long-term societal health over short-term efficiencies.

Figure 2. Insights Visualized



Topics	Insights
Ad hoc Solutions	Solves current solution – not long term Developer + stakeholder centered ethics **Need structured approach for accountability**
Validation	Biased data creates biased models No regulatory validation methods or metrics **Need stricter data and model validation processes**
Regulations	Inconsistent company-specific rules Lack of federal regulation and consequences **Need industry enforceable framework**
Ethics	Early models favor utilitarianism, minority excluded Virtue ethics emphasizes equity and fairness **Need rules and regulations – deontology ethics**
Development Framework	Must combine the requirements from previous insights Without structured approach – cycle of biased models **Need enforceable, reliable framework to fight bias**

This graphic demonstrates the insights visually as they relate to each other (created by author).

Conclusion

Overall, ML ethics needs a structured development framework rather than ad-hoc fixes that vary from company to company, state to state, and country to country in attempts to self-regulate. Through historical analysis and case studies, the need for a universally agreed upon ethical framework to combat the bias found in ML systems becomes obvious. Current methods lack in generalizability across most ML systems, or are too broad of an idea to apply to a majority of the models, thereby necessitating the need for a modular framework to support the entire industry. By creating a framework that applies ethical thinking into ML development, developers can ensure fairer and more equitable models that better represent reality and reduce harmful consequences for users. While improving models themselves is beneficial in removing immediate discriminatory impact on consumers, there are several industry-wide benefits for ML development with the adoption of an agreed upon ethical framework. These benefits include increased accountability over each companies' models, reduced bias in outputs, as well as bringing the ML development community together to solve a key issue in innovative technology. Practical implications include reduced discrimination across all models at the expense of development time and costs. Ethical ML is not a problem that can be solved on an unspecified date in the future, it is a foundational requirement to the continued use of ML which requires action within company guidelines, regulations, and current development practices.

References

AI legislation in the US: A 2025 overview. SIG. (2025, February 18). https://www.softwareimprovementgroup.com/us-ai-legislation-overview/

Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016, May 23). *Machine bias*. Machine bias risk assessments in criminal sentencing. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bedayn, J. (2024, November 21). *Class action lawsuit on AI-related discrimination reaches Final settlement*. AP News. https://apnews.com/article/artificial-intelligence-ai-lawsuit-discrimination-bias-1bc785c24a1b88bd425a8fa367ab2b23#:~:text=A%20federal%20judge%20approved%20a,the%20lawsuit%20alleged%20were%20discriminatory.

Burgard, J. P., & Pamplona, J. V. (2024). FairML: A Julia Package for Fair Classification. *arXiv preprint arXiv:2412.01585*.

Fergus, R. (2024, February 29). *Biased technology: The Automated Discrimination of Facial Recognition*. Biased Technology: The Automated Discrimination of Facial Recognition. https://www.aclu-mn.org/en/news/biased-technology-automated-discrimination-facial-recognition#:~:text=Studies%20show%20that%20facial%20recognition,Facial%20recognition%20automates%20discrimination.

Floridi, L., & Taddeo, M. (2016). What is data ethics?. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160360.

Grenawalt, T. (2023, October 3). Machine learning ethics: Understanding bias and fairness. Vation Ventures Research. https://www.vationventures.com/research-article/machine-learning-ethics-understanding-bias-and-fairness#:~:text=Machine%20learning%20ethics%20refers%20to,%2C%20society%2C%20and%20various%20industries.

Hursthouse, R., & Pettigrove, G. (2022, October 11). Virtue ethics. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/ethics-virtue/

Iriondo, R. (2018, October 11). *Amazon scraps secret AI recruiting engine that showed biases against women*. Amazon Scraps Secret AI Recruiting Engine that Showed Biases Against Women. https://medium.datadriveninvestor.com/amazon-scraps-secret-ai-recruiting-engine-that-showed-biases-against-women-995c505f5c6f

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature machine intelligence*, *1*(9), 389-399.

Livingston, M. (2020). Preventing racial bias in federal ai. Journal of Science Policy & Governance, 16(02).

Lyman, J. (2024, July 24). *Workday facing discrimination lawsuit over AI hiring software*. Pleasanton Weekly. https://www.pleasantonweekly.com/courts/2024/07/18/workday-facing-discrimination-lawsuit-over-ai-hiring-software/

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, *21*(4), 18-21.

Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., ... & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digital Health, 2(6), e0000278.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... & Staab, S. (2020). Bias in data‐driven artificial intelligence systems—An introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1356.

Reem, N. (2024, December 18). Ai Watch: Global regulatory tracker - united states: White & Case LLP. United States | White & Case LLP. https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-united-states

Shadowen, N. (2019). Ethics and Bias in Machine Learning: A Technical Study of What Makes Us “Good”. In: Lee, N. (eds) The Transhumanism Handbook. Springer, Cham. https://doi.org/10.1007/978-3-030-16920-6_12