# The NEW Ethics of Artificial Intelligence

Ehaan Sair

## Abstract

As artificial intelligence systems become increasingly integrated into decision-making processes in society, the ethical reasoning of these models needs closer examination. This paper investigates how different AI models are able to handle certain ethical scenarios differently based on three key factors: training data, algorithmic architecture, and human intention. These factors are essential for understanding why ethical outcomes vary so widely across systems that may seem similar at first glance.

First, the training data that shapes an AI model fundamentally determines its moral responses. Diverse or biased datasets can influence how a model interprets ethical dilemmas, reflecting the cultural, historical, and systemic values present in the data. Second, the architecture of AI systems, from rule-based frameworks to deep learning models, affects how information is processed and how moral decisions are made. The complexity and structure of these systems can influence the depth, flexibility, and adaptability of their ethical reasoning. Finally, human intention plays an important role in guiding AI development, from the values engineers embed into systems to the goals they prioritize. This human oversight can steer, or even distort, the ethical performance of AI.

Together, these elements show that AI's moral reasoning is not natural or universal, but instead very constructed and highly variable. Understanding these influences is crucial not only for improving current systems but also for designing future AI that can function with ethical integrity across diverse contexts. This research emphasizes that creating ethical AI is less about enforcing singular rules and more about intentional and representative design.

## Introduction

As artificial intelligence becomes more integrated into fields like healthcare, law enforcement, and social media, the ethical decisions made by these systems carry increasing importance. From diagnosing medical conditions to determining parole eligibility or moderating harmful content, AI is frequently used in situations that require moral judgment. Yet, unlike humans, AI does not have an inherent sense of right or wrong. Its ethical decision-making is constructed, shaped by the data it receives, the design of its architecture, and the human priorities built into its code. This raises an important question: How do various artificial intelligence models handle ethical scenarios differently based on their training and data input? To what variables is this difference attributable, and how can this disparity be understood or mediated?

To address these questions, three foundational aspects are explored: the role of training data in shaping ethical behavior, the influence of algorithmic architecture on moral reasoning, and the impact of human intention in guiding AI development. Together, these factors show that

an AI's ethical framework is not fixed or universal, but rather a dynamic result of its construction - one that mirrors the values, limitations, and decisions of its creators.

## "Training Data"

The ethical capabilities of artificial intelligence are not innate, but carefully created through training data. Just as human moral frameworks are shaped by cultural experiences and education, AI systems develop their understanding of ethics based on the sources used in their training. The composition of training datasets significantly affects how AI models deal with complex ethical scenarios, presenting a nuanced landscape of technological moral reasoning.

One of the most striking discoveries is how different training datasets lead to very different ethical frameworks. When AI systems are trained on a variety of sources, they exhibit notably different approaches to making moral and ethical decisions. For instance, models trained primarily on Western academic philosophical texts tend to focus on individual rights and utilitarian thinking, while those using Eastern philosophical perspectives lean towards collective benefit and harmony. This was concluded in research by Marc-Etienne Brunet and colleagues, which analyzed word embeddings and discovered systemic biases in how AI interprets gender roles and leadership qualities based on historical textual sources (Brunet et al., 2019).

The diversity of training sources also plays a large role in an AI's moral reasoning capabilities. Datasets that are multilingual and geographically diverse have proven to be essential for creating a more nuanced ethical understanding. A study by Tao et al. (2024) investigated how large language models trained on culturally specific datasets approached ethical dilemmas. The research showed that models trained on Western sources emphasized individualism, while those trained on non-Western cultural content were more likely to prioritize collective benefit, showing that geographic and cultural training diversity improved ethical reasoning across different moral contexts. These consistent findings further reveal how AI systems trained on multiple cultural perspectives show an improved ability to recognize and navigate complex ethical dilemmas. These findings also suggest that ethical reasoning in artificial intelligence is not a fixed characteristic but a dynamic capability shaped by the depth and variety of its training data.

Historical biases also present a significant challenge in AI ethical development. Training data from different historical periods can inadvertently maintain outdated and potentially harmful perspectives. For instance medical AI systems have shown significant biases when trained on pre-1990s data, especially in areas like mental health, where historical stigmas and limited understanding were incorporated into the model's decision-making process (Farhud & Zokaei, 2021). This reveals a critical challenge in AI development: creating systems that can learn from historical information without reproducing historical prejudices.

The evolution of training data over time offers another interesting dimension to ethical AI development. Comparing models trained on older versus newer data sources reveals substantial differences in ethical reasoning. Moreover, recent training data shows more advanced approaches to complex issues such as digital privacy, social justice, and racial equality compared to models trained on older data, highlighting the dynamic nature of ethical understanding in artificial intelligence.

However, these findings stress a fundamental reality about artificial intelligence: ethical reasoning is not a predefined trait, but a very carefully crafted capability. A model's training data forms the basis of an AI's moral framework, similar to how human experiences and education shape our ethical perspectives. This research reveals that creating a truly ethical AI is not just about implementing a single, universal set of rules, but about gathering a diverse and representative array of training sources.

Yet, this approach has many challenges. The process of selecting and balancing training data is extremely complex and requires careful consideration. Researchers must find a balance between representing diverse perspectives and avoiding the perpetuation of harmful biases. It has become a form of ethical curation aimed at developing A.I. systems that can recognize and handle moral complexities effectively while increasing their sophistication.

## "Algorithmic Decision-Making"

While training data provides the foundation for an AI's ethical framework, the way models process and apply this information is determined by their algorithmic structures. The mechanisms behind AI decision-making play a crucial role in shaping ethical outcomes, often intensifying biases or introducing new types of moral reasoning that go beyond their initial training data. This research investigates how algorithmic designs, reinforcement learning strategies, and model interpretability affect the ethical behavior of AI systems.

One of the most critical aspects of ethical AI decision-making is the inherent nature of its algorithmic structure. Different models process moral dilemmas in fundamentally different ways based on whether they use rule-based logic, statistical probability, or deep learning networks. Rule-based systems, which explicitly encode ethical principles, often struggle with complex or ambiguous scenarios that need contextual judgment. On the other hand, deep learning models create ethical responses based on patterns found within their training data, sometimes leading to unpredictable or emergent behaviors. Research by Askell et al. (2021) demonstrated that transformer-based models, due to their self-attention mechanisms, had more consistent moral reasoning across tasks than earlier approaches through the alignment of human values through preference-based training. This highlights how transformer-based models exhibit significantly different ethical decision-making patterns compared to earlier machine-learning approaches, reinforcing the idea that model architecture directly affects moral reasoning.

Reinforcement learning introduces another aspect to AI ethics by shaping behavior through feedback loops. Systems trained with reinforcement learning develop ethical responses based on predefined reward systems, which can both improve and distort ethical decision-making. For example, studies on reinforcement learning in self-driving cars demonstrate how these models prioritize safety, efficiency, or legal compliance based on how reward functions are designed. A study by Askell et al. (2021) showed that AI models trained with reinforcement learning, especially those guided by profit-based goals, were more likely to overlook fairness in favor of utility. This revealed how reward function design can influence whether ethical or exploitative behavior is reinforced. These findings show that ethical AI is not only about training data but also about how decision-making rewards and penalties shape an AI's evolving ethical framework.

Another major challenge in ethical AI development is the issue of interpretability. Unlike human moral reasoning, which can often be explained through introspection, AI decision-making processes are still largely opaque, creating what is known as the "black box" problem. This lack

of transparency raises concerns about accountability, especially in high-stakes ethical decisions like medical diagnosis, criminal sentencing, and financial lending. Research by Blodgett et al. (2020) emphasized the importance of transparency and interpretability in AI systems, arguing that unclear models obscure moral reasoning and limit accountability. The study found that models with clear, explainable processes, such as decision trees or attention-based architectures, allow for better human oversight and correction of ethical inconsistencies. Efforts to create explainable AI (XAI) are crucial in ensuring that ethical decision-making remains understandable and justifiable.

Bias amplification also becomes a key ethical issue when looking at algorithmic decision-making. Even when AI models are trained on carefully selected datasets, their algorithms can strengthen and even magnify existing biases through learning cycles. For example, predictive policing models trained on historical crime data have been shown to disproportionately target minority communities, not because of explicit bias in the data but due to the way algorithms weigh and extrapolate patterns. A study by Udupa et al. (2022) found that content recommendation systems trained on user engagement data often created feedback loops that amplified extreme viewpoints and biased narratives. These models unintentionally reinforced societal inequalities by favoring content that matched existing patterns, worsening ethical blind spots. These findings highlight that ethical AI development must address not only biases in training data but also the structural tendencies of algorithms to reinforce systemic issues.

Furthermore, emergent behavior in AI models adds another layer of ethical complexity. As AI systems become more sophisticated, they display behaviors that were not explicitly programmed or predicted. In general, research on large language models has found that when trained on diverse ethical frameworks, they sometimes develop new moral reasoning approaches that blend multiple perspectives. While this suggests a form of adaptive moral intelligence, it also raises concerns about unpredictability. Should AI be allowed to independently develop ethical principles, or must it strictly follow human-defined moral frameworks? The debate over AI autonomy versus human control remains a central concern in algorithmic decision-making.

Ultimately, the ethical behavior of AI systems is not just determined by their training data but by the complex interactions between data, algorithmic processing, and feedback mechanisms. The structure of decision-making models directly affects how AI systems understand and apply ethical principles, sometimes reinforcing biases and at other times creating new moral reasoning paradigms. As AI continues to advance, the challenge lies in creating algorithms that balance efficiency with fairness, transparency with complexity, and autonomy with human oversight. Developing ethical AI cannot rely solely on data curation; it must also consider the underlying logic that drives AI decision-making.

## “Developer Intent”

The ethical aspects of artificial intelligence are not determined only by training data or model architecture, but also by the intentions and values of the developers behind them. Just as a teacher’s worldview affects their curriculum or a filmmaker’s perspective shapes their storytelling, AI developers bring their own assumptions, goals, and cultural perspectives into the

systems they create. Developer intent, both conscious and subconscious, plays an essential role in shaping how AI systems make moral decisions.

One of the most critical ways this influence appears is through goal-setting during model development. Every AI system is designed to optimize something, whether it is accuracy, efficiency, safety, or user engagement. However, what a system focuses on often reflects the values of its creators. AI used in online platforms, for example, is often built to maximize user attention and revenue through click-through rates or engagement scores. This profit-driven goal may unintentionally overlook ethical issues such as misinformation, harmful content, or user well-being. These design decisions are not random, however, they directly illustrate the commercial priorities of the developers or the institutions funding them. Research from Bender et al. (2021) argued that AI systems designed for engagement metrics often favor polarizing or emotionally charged content. Their analysis showed that priorities such as profit or engagement heavily influenced how ethical considerations were included or ignored in final model behavior and found that models trained for engagement optimization often made decisions that increased polarization and decreased content diversity.

Another key area where developer intent plays a foundational role is in the ethical alignment and constraint of AI systems. Techniques like reinforcement learning from human feedback (RLHF), rule-based alignment, and bias filtering are all implemented with the goal of making AI outputs more ethical, but the effectiveness and success of these techniques depends entirely on what the developers define as “ethical.” As seen in a study by Tao et al. (2024) found that models trained with input from Western annotators often showed preferences for liberal democratic values such as freedom of speech and individual rights, while those trained under different ideological perspectives may have a stronger tendency toward collective order or respect to authority. These differences and variations are not necessarily flaws, but they show how deeply embedded human values are within AI frameworks, even when developers seek objectivity.

In addition, transparency around developer decision-making is a serious limitation in ethical AI development. Many leading AI companies do not disclose how models are trained, who selects the data, or what ethical principles are prioritized in system design. This lack of transparency makes it difficult to hold developers accountable for the ethical consequences of their products. In high-risk areas such as predictive policing, loan approval, and medical diagnostics, this opacity can lead to devastating real-world outcomes, especially when developers overlook how their own implicit biases or institutional pressures may influence model behavior.

Furthermore, even when developers attempt to “de-bias” AI systems, their own interpretations of fairness, justice, and harm will influence the solutions they choose. Designing ethical AI often requires creating balance between competing moral ideals: should AI focus on reducing harm or supporting user autonomy? Should it uphold universal values or adapt to local cultural norms? These are not simply technical questions but are ethical decisions made by people. While some developers may consult ethicists or diverse communities during the design process, the final decisions are ultimately shaped by those with the power and resources to implement them.

Therefore, developer intent must be seen as a key factor of AI ethics, not just as a secondary consideration. Even the most advanced systems are not moral agents on their own, but are tools crafted in the image of their creators, complete with their creators' blind spots, limitations, and ideologies. Just as training data shapes an AI's perspective, developer choices influence how that perspective is understood, prioritized, and acted upon. To build ethical AI it is not enough to compile diverse datasets or refine technical models. We must also demand transparency, accountability, and ethical responsibility from those leading current AI development.

## Conclusion

The ethical development of Artificial Intelligence is an ongoing challenge, shaped by its training data, programmed decision-making structures, and real-world applications. This research has shown that AI's ethical reasoning is neither essential or universal, but is instead a direct reflection of the data it is trained on. The sources used to create AI models, from diverse cultural perspectives or historical archives, significantly impacts the system's moral decision-making. Additionally, these frameworks must be designed properly so AI can handle moral complexities without bias and ethical failures.

In today's world, AI systems are being used in major fields, from healthcare and finance to law enforcement and governance. With AI systems becoming increasingly powerful, ethical AI development is even more critical. Without controls or biases, AI decision-making can reinforce existing societal inequalities or introduce new ethical issues. This necessitates training methods to continuously change, there to be transparency in AI design, and the application of wide ethical oversight.

Although there has been significant progress, there are still many issues. Careful selection of training data is needed in order to balance different points of view without continuing previous prejudices. Furthermore, achieving ethical AI can not be accomplished overnight; it is a continuous process of refinement. With technology making AI more independent, the collaboration of technologists, ethicists, policymakers, and the public will be crucial to ensure that AI will be utilized as a tool for good rather than a producer of harm.

Overall, creating ethical AI is a collective effort. The training and constriction of AI systems today will set the course of human-machine interactions in the future, and by understanding the complexities of ethical AI and working toward its transparency and fairness, we can use the full potential of AI while also safeguarding fundamental human values.

## Works Cited

Askell, Amanda, et al. "A General Language Assistant as a Laboratory for Alignment." *Arxiv.org*, Cornell University, 1 Dec. 2021, arxiv.org/abs/2112.00861. Accessed 26 Jan. 2025.

Bender, Emily M., et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *ACM Digital Library*, Association for Computing Machinery, 1 Mar. 2021, dl.acm.org/doi/10.1145/3442188.3445922. Accessed 10 Nov. 2024.

Blodgett, Su Lin, et al. "Language (Technology) is Power: A Critical Survey of 'Bias' in NLP." *ACL Anthology*, July 2020, aclanthology.org/2020.acl-main.485/. Accessed 9 Dec. 2024.

Brunet, Marc-Etienne, et al. "Understanding the Origins of Bias and Word Embeddings." *Proceeding of Machine Learning Research*, MLResearchPress, 2019, proceedings.mlr.press/v97/brunet19a/brunet19a.pdf. Accessed 22 Nov. 2024.

Farhud, Dariush D., and Shaghayegh Zokaei. "Ethical Issues of Artificial Intelligence in Medicine and Healthcare." *PubMed Central*, National Library of Medicine, Nov. 2021, pmc.ncbi.nlm.nih.gov/articles/PMC8826344/. Accessed 3 Feb. 2025.

Tao, Yan, et al. "Cultural bias and cultural alignment of large language models." *Oxford Academic PNAS Nexus*, Oxford UP, National Academy of Sciences, 17 Sept. 2024, academic.oup.com/pnasnexus/article/3/9/pgae346/7756548. Accessed 18 Nov. 2024.

Udupa, Sahana, et al. "Ethical Scaling for Content Moderation: Extreme Speech and the (In)Significance of Artificial Intelligence." *Harvard Kennedy School Shorenstein Center on Media, Politics and Public Policy*, Harvard Kennedy School, 9 June 2022, shorensteincenter.org/ethical-scaling-content-moderation-extreme-speech-insignificance-artificial-intelligence/. Accessed 10 Nov. 2024.