Journal of Posthumanism
2025
Volume: 5, No: 4, pp. 420–432
ISSN: 2634-3576 (Print) | ISSN 2634-3584 (Online)
posthumanism.co.uk

DOI: https://doi.org/10.63332/joph.v5i4.1089

Ethical Alignment in Large Language Models: Interpreting Moral Reasoning in Transformer-Based AI Systems

Awad Alyousef[1], Asem Omari[2], Rund Mahafdah[3]

Abstract

The adoption of Large language models (LLMs) based on transformer networks in contexts of high stakes has much accentuated ethical debates devolving over the alignment of machine values and moral reasoning with those same of the human. This study looks at interpretability of moral-based reasoning in LLMs, in terms of the ability to learn, apply, and justify ethical norms or reasoning. The article also discusses how transformer architectures learn and fix values in the face of normative judgments under a number of interdisciplinary frameworks well at home within psychology, philosophy, and socio-technical perspectives. The article critiques the relevant methodologies: value alignment frameworks, simulation environments, transparency-enhancing tools that, while they can be helpful, can be harmful if not operated carefully, all in an attempt to gauge ethical robustness. It scrutinizes attribute bias detection, fairness interventions, and the limitations of the current moral reasoning in AI-generated outputs. Case studies of healthcare, mental health, and the justice system were provided to show the ethical implications of misalignment. Finally, the paper gives recommendations on how to move towards the development of moral AI systems through inclusive design, explainable decision pathways, and global ethical governance.

Keywords: *Moral Reasoning, Transformer Models, Ethical Alignment, Responsible AI, AI Explainability.*

Introduction

The recent rapid development and eventual widespread use of Large Language Models (LLMs) have elicited several seriously elevated ethical questions regarding the alignment of such models with human values (Boji, 2023). These questions are related to the issues of ethical alignment, controllability, predictability, and the global safety of these models in view of the broad potential for them to reinforce bias in code and inherent need for the propagation of problematized lain perspectives (Boji, 2023), (Hadar-Shoval, 2024). Since LLMs are now slowly integrating into various fields, including the ones of mental health, healthcare, and human decision-making identifiable best practices should be implemented to ensure the alignment of LLMs to human values and to avert the possible disastrous consequences as much as possible for responsible AI development (McLean, 2024), (Baradwaj, 2024). This study expands on the various ethical-execution aspects while hinting at moral reasoning in transformer-based AI models and in-fact is emphasized in the addressing of biases, transparency, and ethically comprehensive game guidelines for their development and implementation. Incorporating LLMs in the sensitive areas needs an evaluation of a kind of ethical footing, as models trained on massive data set can inadvertently take in biases from society and further them by giving outputs that are an

[1] Information Systems Department. College of Computer and Information Sciences. Prince Sultan University, Riyadh, Saudi Arabia, Email: aalyousef@psu.edu.sa, (Corresponding Author).
[2] Computer Information Science, Higher Colleges of Technology Al Ain, United Arab Emirates, Email: aomari@hct.ac.ae
[3] Department of Computer Science. College of Computing and IT, Shaqra University, Shaqra, Saudi Arabia, Email: rundmahafdah@su.edu.sa

archetypal reflection of discriminatory or unfair perspectives (Hadar-Shoval, 2024), (Rajkomar, 2018). This has the potential to create a great deal of apprehension specifically in the domain of mental health, where making biased recommendations or responses might intensify existing disparities and seriously challenge equitable healthcare (Hadar-Shoval, 2024). It is not merely a technical challenge but a profound moral task to ensure an ethical alignment; toward the same, a multidisciplinary approach that combines ethical theories, social sciences, and technical expertise is required (Boji, 2023). For instance, in neuro-oncology, shared decision-making (SDM) is paramount, and here, LLMs will provide a benefit to patient understanding and involvement, but this must be delivered while ensuring ethical matters with respect to patient data protection and informed consent are respected (McLean, 2024). The transformation of Foundation Models (FMs) into real-world clinical settings requires the deployment of the major stewardship and co-design principles for the secure regulation and fair interest representation of all involved parties (Baradwaj, 2024). Focusing on ethical alignment within LLMs is exacerbating the current difficulty within several issues hindering the immediate application of ethical principles. The untransparent nature of many LLMs, often referred to as "black boxes," coupled with lack of external interpretation values and biases embedded in decision-making processes (Hadar-Shoval, 2024). This lack of transparency would only hinder the identification and rectification of problematic viewpoints. Human values, alongside their varying characteristics across cultures and backgrounds, also present difficulty for the designing of LLMs that can be attuned to a plethora of ethical standards (Hadar-Shoval, 2024); Jacoby, 2020). Additionally, the potential for adversarial attacks where deliberately manipulated LLMs are aimed at producing harmful outputs argues for robust safeguards and ongoing monitoring (Boji, 2023). This would mean the ethical challenges posed by LLMs, just like ChatGPT, spotlight the unveiling of comprehensive ethical guidelines and policies for the domain of ramping AI applications.

Literature Review

Methodologies for Assessing Ethical Alignment

For evaluating the ethical alignments of LLMs, techniques emerged unfolding a spectrum of strengths and weaknesses, which in reality provides a good toolkit for researchers and designers ever interested in diving deep into the issue involving ethics. One such technique discussed in the literature uses a value-based framework, Schwartz's Theory of Basic Values (STBV), for measurement of cultural value orientations in LLMs (Hadar-Shoval, 2024). By making a comparison between the value profile of LLMs and human beings in other cultural or societal groups in defining what are the divergences and biases towards another one. Another avenue opening to evaluation is via an autonomous-agent simulation through a Virtual Reality (VR) framework that imitates a real-world setting to explore the interactions between LLMs and automated "digital citizens" (Boji, 2023). This could allow for viewing and fine-tuning of AI behaviors in its own realistic set-up, considering the social, ethical, and theoretical aspects. These methodologies are supportive of examining the realistically differentiable "value-like infrastructure" in the LLMs against human values and raise ethical debates on their counterparts, especially important in, say, mental health sometimes (Hadar-Shoval, 2024).

Moral Reasoning in Transformer-Based AI Systems

Transformer models, the backbone of many contemporary LLMs, have proved an extraordinary tale in natural language processing and generation (Chimenea, 2023), (Mohamed, 2024). However, researchers are currently engaged in exploring the argument that the order of turn-

coats is already occupying a place within these models. Moral reasoning is to engage on the problem of assessing actions, decisions, or outcomes with reference to ethical norms and values, giving reasoned arguments in support of such assessments (Attia, 2023). Can contemporary LLMs participate effectively in moral reasoning while doing it, though? Three categories have thus emerged-viz., having knowledge of ethical concepts and using them in a specific application, being able to provide a coherent justification for one's moral judgments, or would the alternatives be thinking about these drawbacks and biases within their systematics-models and reflect upon such major issues. The survey belongs to the scenario described by Zining Luo et al., suggesting that the LLM-driven clinical reviews tended to have fewer references, offer less insight, and contain lower logical coherence than human-authored reviews (Luo, 2025).

Interpreting and Applying Ethical Concepts

LLMs through exposure to large corpus-based datasets would deal with recognition and classification of ethical concepts such as fairness, justice, and autonomy (Liu, 2024). However, more complex understanding of a concept goes a little further than simply being aware of its face value (Stenseke, 2022). One example can best describe this: an LLM may identify bias in a group without an ability to understand why the bias constitutes an ethical issue or how it could impact people or communities (Rajkomar, 2018). Thus, studying an LLM's knowledge of ethical concepts demands as a constraint his ability to define, explain, and provide examples establishing deeper understanding. The interdisciplinary concerns surrounding AI ethics, as pointed out by Jakob Stenseke, necessitate that cross-disciplinary concepts and practices exercised by AI ethics should surely meet each other-and if in conflict, they should rely on each other to build an environment of understanding and progress rather than argument and destruction of ideas and evidence (Stenseke, 2022). In messy situations, LLMs must balance values with competing interests and then attempt to make a quantitative decision considering consequences. This exercise is quite challenging as ethical dilemmas are usually connected to competing considerations and risks that are hard to evaluate (Salloch, 2024). For instance, in a contact-tracing scenario, the system should be judged on whether individual privacy or public safety should be upheld (Norren, 2022). To assess his ability to apply these principles, we must carefully study the actions the LLM considered, the value priority it gave, and the justifications for its decisions. A discourse connecting the impact of moral intensity on ethical decision-making must be fostered by earnest exploration on very specific ethical dilemmas and with different levels of moral intensity.

Methodology

Addressing Biases in LLMs

One of the most pressing ethical challenges in LLMs is the presence of biases (Hadar-Shoval, 2024). Biases can arise from various sources, including biased training data, biased algorithms, and biased human input. These biases can lead to outputs that reflect discriminatory or unfair perspectives, perpetuating societal inequalities (Hadar-Shoval, 2024). Addressing biases in LLMs requires a multifaceted approach that includes data debiasing techniques, algorithmic fairness interventions, and human oversight (Rajkomar, 2018). As A. Rajkomar et al. point out, historical data capture patterns of health care disparities, and machine-learning models trained on these data may perpetuate these inequities, necessitating proactive design and use of machine-learning systems to advance health equity.

Data Debiasing Techniques and Algorithmic Fairness Interventions

The data debiasing approaches attempt to reduce the level of bias in the training dataset, which is LLMs-abusive (Rajkomar, 2018). The incorporation of determination in the debiasing techniques can involve cases of re-sampling the data to balance the representation of different groups; re-weighting the data to minimize the effect of biased examples; or transforming-despikifying-the data of biased features For instance, if a dataset were to have such a biased representation against certain demographic groups, the data debiasing techniques could be applied to maintain a more even balance for the favored group in the representation of the dataset. This is of particular importance, considering healthcare, where biases in training data can lead to inaccurate or unfair predictions, as discussed in the case studies presented by Algorithmic fairness interventions, which tend to be aimed at changing the algorithms that are used to train LLMs in the interest of fairness. Such changes can encompass the addition of fairness constraints in the training objective; altering the model architecture to mitigate biases; and/or subjecting the model outputs to post-processing in the name of fairness. For instance, fairness constraints could be introduced into the training objective in order to put penalties on the model for making any discriminatory predictions." Such interventions have become necessary in order to mitigate the risks of LLMs perpetuating or exacerbating any existing social bias, especially in sensitive applications, namely criminal justice and healthcare. The integration of trust (ethical), explainable, and Fair (REF-AI) artificial intelligence in medical image analysis serves a stiff need for developing ethical, trustworthy, and transparent AI systems in healthcare, as mentioned by Soheyla Amirian et al. (Amirian, 2024).

Visualization and Interpretation Techniques

Visualization techniques, in fact visualize the dynamic aspects of LLMs, and consequently helps in visualizing internal states and activities. Visualization can include neuron activations, the flow of information through an entire network, or the relationship of different concepts. Understanding the learning, reasoning, and decisions made by LLMs through such visualization works is especially important in high-stakes areas such as healthcare, where having an understanding of the reasoning is essential to curating the model for patient safety and trust in it. Interpretation techniques explain the reasons for such decisions and predictions made by LLMs. Using interpretation techniques could reveal which features or inputs matter most during the model's decision-making process, give justifications for individual predictions, or compare the behavior of the model to that of humans. Overall, decoding LLM decisions would help researchers understand their inner strengths and weaknesses in different real-world applications, including identifying any potential biases or errors in the AI. This is essential for trustworthiness in these AI systems as well as responsible and ethical uses of these systems (Amirian, 2025).

Significance of Results

Algorithmic Fairness, Transparency and Accountability

Algorithmic fairness is yet another ethical aspect involved (Rajkomar, 2018), (Kumar, 2024). The concern here is that the design of such systems should be set to avoid reproduction or aggravation of the already existing societal inequities. There should be scrupulous modelling of the inputs fed into the models, algorithms that process the data and parameters against which to judge how well the models perform. It also includes the series of follow-ups and audits to determine that these models do not yield an output that could be considered discriminatory or unfair. According to Dinesh Kumar and Nidhi Suthar, ethical issues like that of discrimination

and bias are huge threats to this marketing segment and call for applying ethical rules and funding bias detection tools at some point (Kumar, 2024). Transparency and accountability are really considered fundamental values for trust building in such LLMs (Amirian, 2025), (Contini, 2024). It should be understood that the LLMs will not be left in some black-box state of design and use; Their decision processes will need to be made intelligible and explainable, including techniques for visualizing and making interpretable the internal workings of the machines. There must be full accountability chains on the decisions coming out of LLMs; thus, a particular person or organization will carry the responsibility for the outcome of such decisions. Francesco Contini et al. propose that it is the introduction of non-accountable AI in justice systems that marks alterations in actor-network configuration and the distribution of accountability between humans and technology, making it mandatory that judges should exercise control over the outputs generated by these systems.

Human Oversight and Ethical Safeguarding in LLM Deployment

Human monitoring constitutes perhaps the most important foundation for the ethical and responsible use of LLMs. Ideas of human intervention will require a renewed focus as these models become more prevalent in the decision-making processes of high-stakes areas like health diagnostics, financial risk assessment, and judicial systems. Sooner or later, with insufficient human intervention, uncontrolled actions of LLMs can lead to unintended ethical violations such as bias amplification, lack of accountability, or even the infringement of rights of the vulnerable (McLean, 2024). It is critical that LLMs never be put in a position to make decisions of significant consequence for individuals or communities without thorough human examination and contextual validation. Moral reasoning is eminently sensitive and often contextual; thus, it requires very much culturally specific judgments beyond the capability of current machine-learning models. An LLM may produce an apparently moral rationale, but it cannot exercise human empathy, situational acumen, and the struggle to reconcile competing moral perspectives amid uncertainty. There is a conjunction of Atanas et al. stating this shortcoming reiterating that the major actors of public health—and other comparable sensitive areas—need to critically evaluate outputs from AI. Their conclusions stress the need to confirm, contextualize, and explain AI recommendations by means of a human ethical lens, particularly when such implications intersect with human well-being and dignity or even human rights. This means ensuring not only that LLM outputs are technically correct but also that they have been vetted for ethical legitimacy and compliance with social conventions. Human oversight, therefore, plays a major role in building trust into AI systems. When end-users have knowledge that human experts have validated the AI decisions, the legitimacy and transparency of such systems grow. Such oversight mechanisms should be rooted in formal review boards, interdisciplinary ethical committees, and participatory design processes that foreground human values. Ultimately, upholding ethical integrity in AI applications will require a continued presence of human actors—not as passive validators of AI proposals but as co-reasoners who interpret, contest, and steer machine logic toward morally acceptable conclusions.

Discussion

It might be of interest to look at various case studies and applications in terms of ethical challenges and possible solutions concerning LLMs. It gives concrete examples for the issues in question and options for a solution. These example cases would pronounce the implications of ethical alignment in practice and the pressing need to address biases, attain transparency, and lay down thorough ethical guidelines. Entering concrete case studies helps in the realizations of

deeper ethical intricacies pertaining to the creation and usage of LLMs while also suggesting a way forward for ensuring their responsible applications.

LLMs in Mental Health

Possible reformulation with quite a different faith: LLMs-can revolutionize the field of mental health much in that they would offer scalable, personalized, and on-demand support to people unable to take access to traditional therapeutic situations (Hadar-Shoval, 2024; McLean, 2024). Simulating empathy in conversation, processing subtle language, and delivering context-sensitive responses, such models will serve as high-powered tools in augmenting mental health interventions. However, the promises must be tied to many complex ethical issues that will need critical consideration before wide-scale adoption can become safe or responsible. The most challenging risks include the likelihood of biased propagation. These models incorporate the tangible risk of unconsciously reproducing harmful stereotypes or one-sided narratives based on mental health, gender, race, or other identity markers throughout the history, society, or culture that is bringing out a data set for the LLMs. Therefore, these outputs subtle, could pervert any therapeutic value that initiation into this interaction might have presented, harm the well-being of clients, and further promote health inequities. Again, these are not mere hypotheticals; several empirical studies at the time, including those by Hadar-Shoval et al. (2024), show ground evidence of the diversity between moral and cultural values attached to the content LLMs generate versus the lived experiences or ethical norms in the population meant to be targeted. A further matter of grave concern lies in their "black box" nature: i.e. the reason behind the specific response or recommendation is rather difficult to appreciate by clinicians as well as patients who have come into contact with the mental health applications of LLMs. Because of this, it undermines the very principles of informed consent and shared decision-making, both being foundational tenets of ethical mental health practice. Users could be even unaware of how or why such an LLM would give that kind of guidance, which makes the appropriateness, reliability, or even its alignment with professional standards difficult to assess. Rigorous ethical oversight and safety mechanisms will have to guide the deployment of LLMs in mental health. This includes culturally sensitive training protocols, audits and assessments for bias-detection, and the implementation of interpretability tools capable of worldly and understandable descriptions of model behavior. Autonomy would not favor LLMs operating clinically: their outputs would be reviewed and contextu- alized by qualified human professionals to guarantee accuracy and ethical ground. Ultimately, ensuring that LLMs reflect human values and cultural pluralism is not a peripheral concern-it is a prerequisite for their legitimate use in mental health care. Ethical alignment in this domain requires multidisciplinary effort that integrates technical refinement along with philosophical, psychological, and sociocultural perspectives to safeguard the well-being and dignity of all users.

LLMs in Healthcare Decision-Making

By providing access to huge fountains of knowledge on medical information with personalized recommendations (McLean, 2024), LLM can assist healthcare professionals in making complex decisions. Ethical concerns in this area arise, among others. If an LLM has not been validated, it may provide false or misleading information that could potentially harm (Luo, 2025). The same bias that may arise due to the algorithmic nature of LLMs could also disadvantage certain groups in their care by not affording them treatment options which may be more effective or appropriate. Validation and testing of LLMs while ensuring proper review of all recommendations by human experts becomes paramount, especially to mitigate deleterious

effects. This becomes more relevant when looking into neuro-oncology, wherein shared decision-making becomes important, and LLMs can foster patient understanding and engagement, but ethical considerations must be navigated with caution. According to Takanobu Hirosawa and Taro Shimizu, effective development of AI-based clinical reasoning should delineate both the roles of the system and the needs of the user, and all outputs from the system should be rigorously validated against credible medical resources (Hirosawa, 2023).

LLMs in Criminal Justice

And here they are being used for applications such as risk assessment, predictive policing, and sentencing (Rajkomar, 2018), (Contini, 2024) in criminal justice. Controversially, all of them hold to being very susceptible to algorithmic bias and the perpetuation of discrimination. For example, training an LLM with data that have been biased may ultimately lead to the wrong determination of certain demographic groups as the audience and unfair outcomes generally. It is, therefore, imperative to focus on the data, the algorithms, and the evaluation metrics built into their development to be entirely unable to prevent any harm in terms of the application of LLMs in criminal justice and also to formulate strict oversight mechanisms that guarantee accountability and fairness. In line with that, Francesco Contini et al. wrote the actor-network theory framework in highlighting the accountability aspect required for AI systems utilized in hearings, especially when such systems could be employed non-accountably without risking any of the core roles of courts if the judges themselves programmed the system output (Contini, 2024).

Empirical Contributions to the Design of Morally Aligned AI Systems

Recent research has increasingly focused on the ethical alignment of artificial intelligence and the specific implications these hold for areas within artificial intelligence in other domains increasingly around the areas of decision-making, interpretability, and fairness. Al-Omari et al. (2025) elaborated on how AI can be regulated and other ethical complications that arise in higher education, verifying how responsible AI can efficiently manage its deployment via appropriate governance mechanisms to internally address bias and uphold fairness—core themes in aligning LLMs with human values. In the legal sector, where transparency and moral reasoning are almost essential, Hassan et al. (2024) introduced a deep learning-based model for text summarization, which would make complex legal texts interpretable in order to support high-stakes ethical AI decision-making. Jabbar et al. (2024) stated that preprocessing techniques like stemming will improve the NLP accuracy that can accomplish legal classification tasks towards fair and consistent outputs in language models. Ammar et al. (2024) showed that transformer-based models such as BERT and GPT can predict legal judgments in Arabic-forged emphasis about the need for domain-specific fine-tuning to ensure moral coherence and cultural sensitivity in AI outputs. Rehman et al. (2025), on the other hand, dealt with FER systems in terms of evaluation and proposed deep learning-based hybrid approaches to improve the overall accuracy and ethical robustness of these systems-which are very paramount in AI systems that deal with human emotions. On cloud infrastructure, Gaber and Alenezi (2024) also contended serverless computing through FaaS cannot only deliver ethical AI solutions at scale but also counter other challenges, like security and accountability. Heavily linked were also Alyousef and Al-Omari (2024), where by deploying AI in healthcare, ethical concerns transformed into data privacy and algorithmic bias were noted while advocating a more coordinated global regulation with respect to patients' safety and ethical standards. Finally, Semary et al. (2023) employed transformer models such as RoBERTa in sentiment analysis to show how this type of architecture can be

customized for ethically sensitive tasks-in this case, effectively analyzing user-generated content with the least bias possible.

Data Analysis and Interpretation

From a comprehensive perspective, the ethics of LLMs are to be understood not merely by the conventional technical evaluation framework, wherein LLMs are subject to rigorous reflections on their psychological, social, educational, or regulatory consequences. One of the salient dimensions is anthropomorphism in AI: designing systems that seem to have human emotions, intents, or personalities. As Xu et al. (2025) point out, such artificial ecological behavior can leverage the users' cognition and emotional state, steering towards unconscious persuasion, emotional dependence, and detracting critical thinking. Thus, far-reaching ethical issues regarding user autonomy, informed agency, and emotional manipulation in human-AI interactions arise.

AI explanations and public literacy must be put into the spotlight even further by researchers. Explainability allows the public and professionals to see how LLMs reach certain outputs, while AI literacy builds up the ability to be critical and to become involved intelligently. One example is Atenas et al. (2025)'s recommendations for participatory frameworks built upon data justice; i.e., designs that are inclusive to creating and enable educators and learners to contribute defensibly to the development and uses of AI systems. This way, an ethically aligned AI environment is nurtured in which human judgment and democratic principles participate in data-driven domains.

Ethical dimensions of LLMs are even more pressing in professional decision-making contexts, particularly healthcare. For example, Luo et al. (2023) reflect on how moral courage and ethical sensitivity directly impact the integrity of AI-supported decisions. Their observations are essentially persuasive that human traits need to be cultivated to ensure that the socially, professionally ethical, and dignity-aligned AI decisions are made possible by such decisions.

On a larger scale, leadership and company culture create a context for ethical AI outcomes. According to Kim et al. (2024), ethical leadership minimizes the negative psychological impact of AI-induced job insecurity. By nurturing transparency, accountability, and employee well-being, organizations can better synchronize the usage of AI with their more general ethical objectives, such as environmental sustainability and workforce resilience.

Again, as far as high-risk industries such as aviation go, the integration of LLMs in mission-critical systems raises further regulatory, as well as ethical concerns. Azyus et al. (2025) advocate for a rigorous governance framework that combines ethical, security and accountability safeguards to mitigate the risk of the occurrence of catastrophic failures and maintain public trust in automated decision-making pipelines. At the level of policy concerning artificial intelligence capabilities, the piece goes on to state that it is still some way behind the actual reality of things. Kalodanis et al. (2025) have analyzed the condition of preparedness of European healthcare institutions concerning the EU AI Act-and found a considerable gap between AI and the pace of its implementation as well as the associated ethical governance structures. This showed that they should be prompted to make impending recommendations to bridge the gap created by innovation and enforceable ethical standards as LLMs are embedding further and deepening into public services.

Arguably, the findings indicate that the use of LLM outputs has a multidimensional complexity in interpretation and governance. Ethical alignment can never be achieved merely by a technical

fix but requires a systemic integrating psychological, educational, professional, organizational, and regulatory perspectives as is shown in Figure 1. Thus interconnected, these dimensions lay the foundation of a comprehensive framework for the evaluation of the ethical implications of LLMs in the real-world application. For that, it would be better to conceptualize some salient human-centered dimensions based on recent empirical and theoretical explorations that will provide a clearer perspective on what are some multidimensional ethical aspects of LLM deployment. The combined intersection of psychological, educational, professional, organizational, and regulatory will reflect this multidimensionality that translates into how LLMs are perceived, integrated, and regulated across various domains. Table 1 summarizes these multidimensional ethical areas and their corresponding insights, thus indicating the relevance of the holistic evaluation for responsible and trustworthy AI.

Ethical Dimension	Key Insights
Psychological Impact	The AI anthropomorphization fuction shapes autonomy in such a way that the subject ends up being susceptible to manipulation.
Data & AI Literacy	Inclusive data justice frameworks allow critical engagement with AI regarding ethics.
Moral Reasoning in Professions	Such moral courage and sensitivity establish bases for ethical decision-making in AI-assisted professional contexts.
Organizational Ethics	So ethical leadership ensures minimization of adverse effects from the AI-induced job insecurity, together with trust-building.
Regulatory Alignment	Hence, these ethics for the AI shall be important for a high-risk profession like healthcare and aviation.

Table 1: Key Human-Centered Ethical Dimensions in LLM Deployment

Conclusion

The ethical alignment of large language models (LLMs) stands to be a great challenge, especially in relation to the essential sectors like healthcare, mental health, and the justice system, within which artificial intelligence is being integrated. According to this paper, LLMs represent great opportunities for improving decision-making and expanding access to knowledge, yet the quandary manifests itself as serious ethical considerations, including value misalignment, algorithmic bias, opacity of reasoning, and accountability gaps. These are then discussed in this research paper. Even when transformer-based architecture is seen as capable of simulating the process of ethical reasoning, this research indicates that they often do not satisfy deep moral consideration in terms of context awareness and justification harmony. Launching those models into high-stake environments exceeds usability due to biases from data and algorithmic design. Yet, they will require refining and validating with continuing diligence closely related to their assigned fields: fairness interventions, interpretability frameworks, and data debiasing. The human element, the inclusive design, and the culturally alert training protocols are the needs vindicated by the case studies presented. Control and regulation of this nature must restrict the great influence that LLMs threaten to impose on interacting key societal domains by establishing thoroughly defined ethical governance frameworks that lend themselves to psychological, social, and legal concerns. Greater engagement with moral philosophy, interdisciplinary collaboration, and regulatory foresight is pivotal to the symbiosis of the future LLMs that will be deemed

ethical. The thrust of future explorations should be on developing technically sound mitigation approaches and longitudinally assessing their impact on LLMs along the dimensions of cognition, autonomy, and equity. This encompasses a necessary all-inclusive approach toward the morals of trustworthy AI involved in systems designed for this purpose.

Future Directions and Research Gaps

While progress has been made in tackling the ethical challenges of Large Language Models (LLMs), important gaps remain that pose additional challenges for research and innovation. Work should advance toward developing more sophisticated and scalable methods to identify and mitigate algorithmic bias at data and model levels, ensuring that fairness interventions remain robust across various sociocultural contexts. Equally importantly, mechanisms for transparency and explainability of LLMs need to be strengthened. The models today are largely opaque systems or "black boxes," where stakeholders such as users find it difficult to interpret or audit decision-making processes. Future research needs to come up with some interpretability toolkits involving causal reasoning, human-AI interaction paradigms, and modular architectures for clearer tracing of paths in ethical reasoning. The socio-psychological effects of deploying large-language models are far less studied. It will be important to study how these models affect human behavior, cognition, trust, and emotional response to understand better the moral alignment's far-reaching implications. This includes studying the risk of anthropomorphization and user over-reliance and manipulation through deceptive AI-generated content. Global policy guidelines and an exhaustive ethical governance framework are needed for rapid modification to keep pace with advancements in LLM capabilities. Such frameworks should bring affected communities into co-design, ensure AMS is inclusive, and tackle regulatory disparities that cross borders.

The closing of these research gaps will therefore require an interdisciplinary effort comprising specializations in computer science, ethics, psychology, law, and the social sciences collaborating across interconnected boundaries. Only through this integrated collaboration can we ensure that not only the ones that come after will be technologically advanced but also ethically aligned, socially responsible, and globally trusted.

Acknowledgments

The authors would like to acknowledge the support of Prince Sultan University for paying the Article Processing Charges (APC) of this publication. The authors would like to acknowledge the support of Prince Sultan University for their support making this publication successful.

References

Krushnasamy, V. S., Al-Omari, O., Sundaram, A., & others. (2025). LiDAR-based climate change imaging in geoscience using spatio extreme fuzzy gradient model. Remote Sensing in Earth Systems Sciences. https://doi.org/10.1007/s41976-025-00197-5

Nimma, D., Al-Omari, O., Pradhan, R., Ulmas, Z., Krishna, R. V. V., El-Ebiary, T. Y. A. B., & Rao, V. S. (2025). Object detection in real-time video surveillance using attention based transformer-YOLOv8 model. Alexandria Engineering Journal, 118, 482–495. https://doi.org/10.1016/j.aej.2025.01.032

Al-Omari, O., & Al-Omari, T. (2025). Artificial Intelligence and Posthumanism: A Philosophical Inquiry into Consciousness, Ethics, and Human Identity. Journal of Posthumanism, 5(2), 458–469. https://doi.org/10.63332/joph.v5i2.432

Rehman, A., Mujahid, M., Elyassih, A., AlGhofaily, B., & Bahaj, S. A. O. (2025). Comprehensive review and analysis on facial emotion recognition: Performance insights into deep and traditional learning with

current updates and challenges. Tech Science Press. DOI: 10.32604/cmc.2024.058036
Al-Omari, O., Alyousef, A., Fati, S., Shannaq, F., & Omari, A. (2025). Governance and ethical frameworks for AI integration in higher education: Enhancing personalized learning and legal compliance. Journal of Ecohumanism, 4(2), 80–86. DOI: 10.62754/joe.v4i2.5781
Alyousef, A., & Al-Omari, O. (2024). Artificial intelligence in healthcare: Bridging innovation and regulation. Journal of Ecohumanism, 3(8), 10582–10589. DOI: 10.62754/joe.v3i8.5673
Hassan, A. Q. A., Al-onazi, B. B., Maashi, M., Darem, A. A., Abunadi, I., Mahmud, A. (2024). Enhancing extractive text summarization using natural language processing with an optimal deep learning model. AIMS Press, 2024. DOI: 10.3934/math.2024616
Jabbar, A., Iqbal, S., Tamimy, M. I., Rehman, A., Bahaj, S. A., Saba, T. (2024). An analytical analysis of text stemming methodologies in information transformers. arXiv.
Ammar, A., Koubaa, A., Benjdira, B., Nacar, O., Sibaee, S. (2024). Prediction of Arabic legal rulings using large language models. Faculty of Electrical Engineering Banja Luka. DOI: 10.3390/electronics13040764
Gaber, S., & Alenezi, M. (2024). Transforming application development with serverless computing. International Journal of Cloud Applications and Computing, 14(1), 1-12. DOI: 10.4018/IJCAC.365288
Semary, N. A., Ahmed, W., Amin, K., Pławiak, P., & Hammad, M. (2023). Improving sentiment classification using a RoBERTa-based hybrid model. Frontiers Media S.A., December. DOI: 10.3389/fnhum.2023.1292010
Boji, L., Cinelli, M., ulibrk, D., & Delibasic, B. (2023). Cern for ai: a theoretical framework for autonomous simulation-based artificial intelligence testing and alignment. European Journal of Futures Research. https://doi.org/10.1186/s40309-024-00238-0
Hadar-Shoval, D., Asraf, K., Mizrachi, Y., Haber, Y., & Elyoseph, Z. (2024). Assessing the alignment of large language models with human values for mental health integration: cross-sectional study using schwartzs theory of basic values. JMIR Mental Health. https://doi.org/10.2196/55988
McLean, A. L., Wu, Y., McLean, A. L. L., & Hristidis, V. (2024). Large language models as decision aids in neuro-oncology: a review of shared decision-making applications. Journal of Cancer Research and Clinical Oncology. https://doi.org/10.1007/s00432-024-05673-x
Baradwaj, S. S., Gilliland, D., Rincon, J., Hermjakob, H., Yan, Y., Adam, I., Lemaster, G., Wang, D., Watson, K., Bui, A., Wang, W., & Ping, P. (2024). Building an ethical and trustworthy biomedical ai ecosystem for the translational and clinical integration of foundation models. Bioengineering. https://doi.org/10.3390/bioengineering11100984
Rajkomar, A., Hardt, M., Howell, M., Corrado, G. S., & Chin, M. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine. https://doi.org/10.7326/M18-1990
Jacoby, N., Margulis, E., Clayton, M., Hannon, E. E., Honing, H., Iversen, J., Klein, T. R., Mehr, S. A., Pearson, L., Peretz, I., Perlman, M., Polak, R., Ravig2024i, A., Savage, P. E., Steingo, G., Stevens, C., Trainor, L., Trehub, S., Veal, M. E., & Wald-Fuhrmann, M. (2020). Cross-cultural work in music cognition: challenges, insights, and recommendations.. Music Perception. https://doi.org/10.1525/mp.2020.37.3.185
Ghandour, A., Woodford, B., & Abusaimeh, H. (2024). Ethical considerations in the use of chatgpt: an exploration through the lens of five moral dimensions. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3394243
Chimenea, .., Garca-Daz, L., & Antiolo, G. (2023). Exploring the potential of artificial intelligence language models in obstetrics with a focus on fetal medicine: an evaluation of the perplexity ai model. Fetal Diagnosis and Therapy. https://doi.org/10.1159/000535345
Mohamed, Y. A., Mohamed, A. H., Kan2024, A., Bashir, M., Adiel, M. A. E., & Elsadig, M. A. (2024).

Navigating the ethical terrain of ai-generated text tools: a review. IEEE Access. https://doi.org/10.1109/access.2024.3521945

Attia, M., Das, B., Atiyeh, S., & Browne-James, L. (2023). Integrating multicultural competencies in ethical decision-making with immigrant populations. Counseling and Values. https://doi.org/10.1163/2161007x-68010005

P, I. & Mekonnen, N. (2024). Moral intensity and ethical decision-making: a combined importance-performance map analysis for professional accountants. International Journal of Ethics and Systems. https://doi.org/10.1108/ijoes-05-2024-0120

Luo, Z., Qiao, Y., Xu, X., Li, X., Xiao, M., Kang, A., Wang, D., Pang, Y., Xie, X., Xie, S., Luo, D., Ding, X., Liu, Z., Liu, Y., Hu, A., Ren, Y., & Xie, J. (2025). Cross sectional pilot study on clinical review generation using large language models. npj Digital Medicine. https://doi.org/10.1038/s41746-025-01535-z

in, D., health,, Atanas,, Nhs, H. W., Trust, C., Aounallah-Skhiri, H., of, M., Tunis, U., El, T., Tunis,, & INNTA, N. I. 3. (2023). Can chatgpt follow an algorithm for ethical decision-making in public health?. European Journal of Public Health. https://doi.org/10.1093/eurpub/ckad160.1284

Liu, J. (2024). An exploration of the integration of marxs philosophical thought and modern ethical decision-making models for artificial intelligence. Applied Mathematics and Nonlinear Sciences. https://doi.org/10.2478/amns-2024-2615

Stenseke, J. (2022). Interdisciplinary confusion and resolution in the context of moral machines. Science and Engineering Ethics. https://doi.org/10.1007/s11948-022-00378-1

Salloch, S. & Eriksen, A. (2024). What are humans doing in the loop? co-reasoning and practical judgment when using machine learning-driven decision aids.. American Journal of Bioethics. https://doi.org/10.1080/15265161.2024.2353800

Norren, D. E. V. (2022). The ethics of artificial intelligence, unesco and the african ubuntu perspective. Journal of Information, Communication and Ethics in Society. https://doi.org/10.1108/jices-04-2022-0037

Sirgiovanni, E. (2024). Should doctor robot possess moral empathy?. Bioethics. https://doi.org/10.1111/bioe.13345

Amirian, S., Gao, F., Littlefield, N., Hill, J. H., Yates, A. J., Plate, J. F., Pantanowitz, L., Rashidi, H. H., & Tafti, A. P. (2025). State-of-the-art in responsible, explainable, and fair ai for medical image analysis. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3555543

Ingale, S., Shrisunder, N., Gophane, G., & Birajdar, A. (2024). Ascent of artificial intelligence (ai) in pharmacy. International Journal of Technology. https://doi.org/10.52711/2231-3915.2024.00008

Herington, J., Mccradden, M., Creel, K. A., Boellaard, R., Jones, E. C., Jha, A. K., Rahmim, A., Scott, P. J. H., Sunderland, J. J., Wahl, R., Zuehlsdorff, S., & Saboury, B. (2023). Ethical considerations for artificial intelligence in medical imaging: data collection, development, and evaluation. Journal of Nuclear Medicine. https://doi.org/10.2967/jnumed.123.266080

Kumar, D. & Suthar, N. (2024). Ethical and legal challenges of ai in marketing: an exploration of solutions. Journal of Information, Communication and Ethics in Society. https://doi.org/10.1108/jices-05-2023-0068

Contini, F., O2024u, E. A., & Velicogna, M. (2024). Ai accountability in judicial proceedings: an actornetwork approach. Laws. https://doi.org/10.3390/laws13060071

Hirosawa, T. & Shimizu, T. (2023). Enhancing clinical reasoning with chat generative pre-trained transformer: a practical guide. Diagnosis. https://doi.org/10.1515/dx-2023-0116

Xu, Y., Zhao, C., & Cao, W. (2025). Reshaping cognition and emotion: an ethical analysis of ai anthropomorphizations impact on human psychology and manipulation risks. Membrane Technology.

https://doi.org/10.52710/mt.206
Spirnak, J. R. & Antani, S. (2023). The need forartificial intelligence curriculum inmilitary medical education.. Military Medicine. https://doi.org/10.1093/milmed/usad412
Atenas, J., Havemann, L., & Nerantzi, C. (2025). Critical and creative pedagogies for artificial intelligence and data literacy: an epistemic data justice approach for academic practice. Research in Learning Technology. https://doi.org/10.25304/rlt.v32.3296
Luo, Z., Tao, L., Wang, C., Zheng, N., Ma, X., Quan, Y., Zhou, J., Zeng, Z., Chen, L., & Chang, Y. (2023). Correlations between moral courage, moral sensitivity, and ethical decision-making by nurse interns: a cross-sectional study. BMC Nursing. https://doi.org/10.1186/s12912-023-01428-0
Kim, B., Kim, M., & Lee, J. (2024). Code green: ethical leaderships role in reconciling ai-induced job insecurity with pro-environmental behavior in the digital workplace. Humanities and Social Sciences Communications. https://doi.org/10.1057/s41599-024-04139-2
Azyus, A. F., Wijaya, S. K., & Kurniawan, B. (2025). Regulatory, ethical, and security dimensions of ai in aircraft mainte2024ce: a framework for assessing harm. Journal of Ecohumanism. https://doi.org/10.62754/joe.v4i1.6666
Kalodanis, K., Feretzakis, G., Rizomiliotis, P., Verykios, V., Papapavlou, C., Skrekas, A., & Anagnostopoulos, D. (2025). Assessing the readiness of european healthcare institutions for eu ai act compliance.. Studies in Health Technology and Informatics. https://doi.org/10.3233/SHTI250047
Raman, R., Kowalski, R., Achuthan, K., Iyer, A., & Nedungadi, P. (2025). Navigating artificial general intelligence development: societal, technological, ethical, and brain-inspired pathways. Scientific Reports. https://doi.org/10.1038/s41598-025-92190-7.