Published in IEEE Internet Computing Special Issue, 2024; Link: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10438949

Special Issue Cover Article

Civilizing and Humanizing AI in the Age of Large Language Models

Amit Sheth, *AI Institute, University of South Carolina*

Kaushik Roy, *AI Institute, University of South Carolina*

Hemant Purohit, *George Mason University*

Amitava Das, *AI Institute, University of South Carolina*

Abstract—The advent of large language models (LLMs) has significantly influenced the development of artificial intelligence (AI), expanding its reach beyond researchers to the general public via practical tools. These models demonstrate remarkable capabilities in complex tasks but also present limitations and risks of misuse. For instance, the advances of AI have made it challenging to distinguish AI-generated content from that produced by humans; similarly the adverse outcomes such as hallucinations that these models can produce. In response, regulatory frameworks from the U.S. and the European Union have proposed measures to ensure AI safety, introducing notions such as Constitutional AI to ensure adherence to these guidelines. This special issue introduces and covers the following two interweaving research directions toward AI safety. First, Civilizing AI as balancing AI's ability to generate human-like outputs and mitigate adverse behaviors such as hallucinations and unintended biases, and second, Humanizing AI as aligning AI systems' behavior with human ethics, socio-cultural norms, values, and regulations, ensuring they meet societal expectations similar to those for human roles such as drivers or health professionals.

The rise of LLMs and foundation models has significantly impacted the trajectory of AI development [1, 2]. These advanced models exhibit remarkable capabilities, offering promising insights into both individualized and generalized forms of intelligence [3]. However, they also present challenges - their limitations are becoming increasingly clear, and widespread misuse exists. The sophistication of AI-generated content has reached a point where distinguishing it from human-produced material has become increasingly difficult [4]. Conversely, AI models often produce inaccuracies, unintended biases, or "hallucinations," raising credibility issues [5]. Governments and regulatory bodies worldwide have begun to recognize these issues, prompting the development of regulatory frameworks for AI safety. These frameworks aim to ensure responsible AI usage and penalize misuse. As these regulations take shape, the concepts of *Civilizing AI* - balancing a machine's ability to produce human-like outputs and mitigate its adverse behaviors, and *Humanizing AI* - aligning AI systems with human ethics, socio-cultural norms, policies, regulations, laws, and values, emerge as critical components [6]. In this special issue's cover article, we discuss the current landscape of efforts to civilize and humanize AI, examining various proposed methods such as generative AI frameworks, human-computer-interaction-based approaches, and position papers. If AI were to become an integral part of individual, social, and business activities, it needs to be subjected to similar conditions and governance structures that drive humans and society. We also outline the theme of submissions to this issue and conclude with a discussion on the challenges of civilizing and humanizing AI and the potential paths forward.

Digital Object Identifier 10.1109/XXX.0000.0000000

The Current Landscape

Generative AI-based Frameworks

Researchers have explored different ways to create trustworthy, aligned AI systems. For instance, InstructGPT models leverage Reinforcement Learning from Human Feedback (RLHF) to align model responses with human preferences. This approach uses a reward signal to encourage desired behaviors, effectively guiding the model to produce outputs that humans find more acceptable and useful. The model's training process involves a cycle of generating responses, receiving feedback, and adjusting based on that feedback, thus enhancing its alignment with human values and expectations [7]. Anthropic's Constitutional AI framework takes a different approach by employing a "constitution" to oversee the training process. This constitution serves as a set of guidelines that the AI follows to avoid harmful outputs. The training process involves both supervised and reinforcement learning phases, incorporating feedback from AI systems themselves through Reinforcement Learning from AI Feedback (RLAIF). This approach aims to create AI systems that are inherently safer and more aligned with human values without relying solely on human-labeled data [8], while measured using appropriate metrics.

In more recent work, OpenAI researchers have shown that applying clear, step-by-step rules to evaluate model outputs against safety standards can significantly improve the safety and reliability of generative AI systems (https://shorturl.at/1Dw0A). This approach offers a more efficient alternative to traditional RLHF, which has been essential for fine-tuning models to follow instructions and align with human values. NVIDIA recently open-sourced a tool called NeMo Guardrails, designed to enhance safety in generative AI models (https://shorturl.at/BNqoT). This software enables developers to align LLM-powered applications with use case-specific expertise while ensuring they remain safe. NeMo Guardrails allows developers to establish three types of boundaries: topical guardrails, which keep apps focused on relevant topics; safety guardrails, which ensure accurate and appropriate responses by filtering out unwanted language and referencing credible sources; and security guardrails, which limit app connections to only known safe third-party applications.

Social Science-Inspired Approaches

Social science-inspired approaches integrate concepts from social sciences into AI development. For instance, models like the "helpful, honest, and harmless" model and Schwartz's Theory of Basic Values are used to represent social science dimensions within AI systems. These dimensions are encoded into value vectors, enabling seamless interaction with the vector-based representations used in LLMs [9]. This integration aims to create AI systems that can better understand and align with human values and social norms. Despite their potential, these generative AI frameworks face challenges. Studies have highlighted their susceptibility to hallucinations and adversarial attacks, raising concerns about their reliability and controllability. Moreover, the black-box nature of large neural networks complicates the implementation of effective control mechanisms. Recent advancements, such as mechanistic interpretability work by Anthropic, show promise in addressing these issues [10]. Stanford's Human-Centered AI (HAI) Institute emphasizes the importance of explicitly modeling societal values within AI systems [11]. By reflecting and deliberating on the values to incorporate, this approach aims to align AI systems with a diverse range of social and behavioral concepts. Leveraging advancements in natural language processing (NLP), multiple frameworks strive to represent societal values effectively, influencing and shaping human experiences in a positive manner.

Approaches from the Human-Computer-Interaction Community

Shneiderman proposes a two-dimensional framework called Human-Centered AI (HCAI) that separates human control from computer automation, arguing that high levels of both can be achieved through good design [12]. It challenges the traditional view that increased automation must come at the cost of human control. The author advocates for designing AI systems that are Reliable, Safe, & Trustworthy (RST) by combining high levels of human control with high levels of computer automation. The HCAI framework clarifies when full computer or human control is necessary and warns against the dangers of excessive automation or human control. This framework emphasizes the importance of human-centered design principles that give users appropriate control while providing high levels of automation. Shneiderman argues this approach can dramatically increase human performance while supporting human self-efficacy, mastery, creativity, and responsibility. The author contends that HCAI designs should utilize unique computer capabilities while recognizing and supporting unique human abilities.

Humanistic AI, introduced by Tom Gruber, focuses on using AI to augment and collaborate with human intelligence rather than compete with or replace it. It

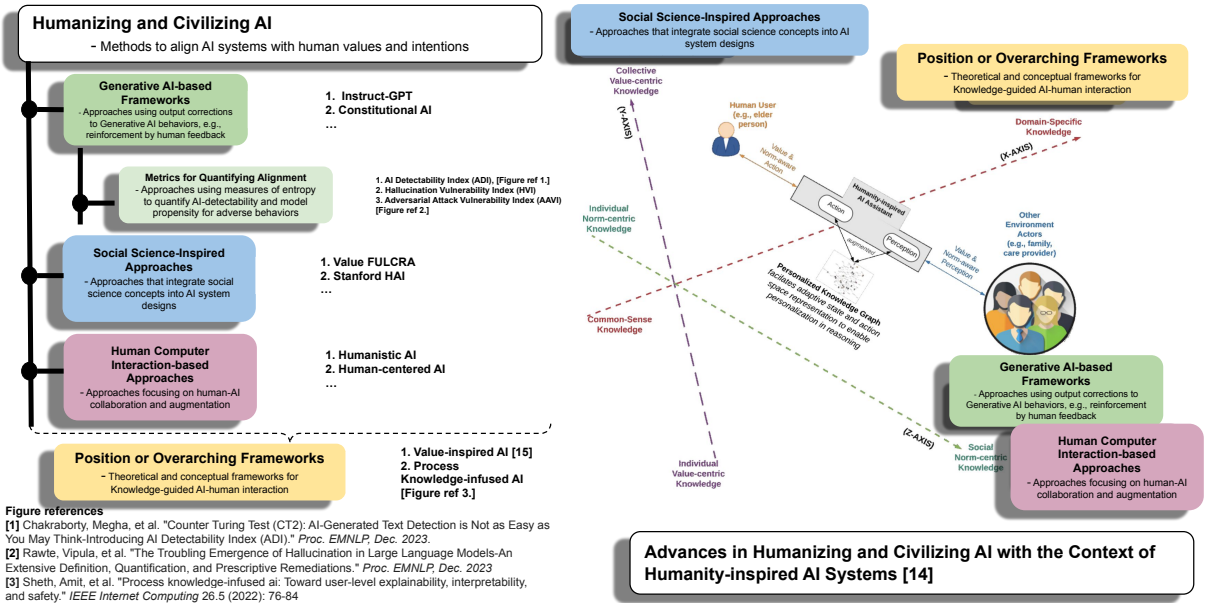

FIGURE 1. Current Landscape of Civilizing and Humanizing-AI: Overview

contrasts this approach with the more traditional view of AI that aims to automate human tasks [13]. Humanistic AI emphasizes applications that enhance human capabilities, such as improving medical diagnoses, aiding in design processes, and developing cognitive enhancement tools. The author also highlights the importance of carefully considering the objectives and consequences of AI systems, particularly in social media, where misguided optimization can lead to negative societal impacts. Gruber advocates for rewriting AI objective functions to prioritize human benefit and provides examples of companies applying Humanistic AI principles in areas like healthcare and assistive technology. The overall message is that adopting a Humanistic AI approach can guide AI development toward more beneficial outcomes for humanity.

Position Frameworks

The concept of Humanity-Inspired AI, as discussed by Purohit et al., focuses on designing mediating AI systems that enhance human-human interaction through considerations inspired by humanity, specifically accounting for ethics, social norms, and values in a mediating AI agent's knowledge and behavior [14]. The authors advocate that AI needs to exhibit socially adaptive behavior by incorporating personalization and an awareness of social context and intentionality. The authors' approach is to leverage knowledge graphs that combine general, common-sense, and domain-specific knowledge with socio-cultural values, norms, and individual cognitive models, as shown in Figure 1. More recently, the Value-inspired AI framework by Sheth et al. provides concrete suggestions on practical designs for encoding society values, norms, and dynamics using symbolic mechanisms within a neurosymbolic framework [15]. The above position frameworks complement a number of prior visions such as Computing for Human Experience (https://shorturl.at/oETmh), which posits for unobtrusive enrichment of human activities, with minimal explicit concern or effort on the humans' part when using intelligent agents or systems, as currently represented by AI.

Figure 1 illustrates an overview of the current landscape.

This Issue

The articles in this issue cover various innovative approaches to civilizing and humanizing AI. **Measuring AI fairness in a continuum maintaining nuances: A Robustness Case Study** proposes a new statistical method for measuring AI fairness on a continuum, allowing for more nuanced observations within groups and focusing on robustness against adversarial attacks. **Towards a Programmable Humanizing AI through Scalable Stance-Directed Architecture** introduces a stance-directed architecture to mitigate toxic content generation in LLMs by fine-tuning them with a focus on core human values and the

common good. **Alignment Studio: Aligning Large Language Models to Particular Contextual Regulations** presents an Alignment Studio architecture that enables application developers to tune models to specific values, social norms, and regulations in particular contexts. Lastly, **AI Design: A Responsible AI Framework for Impact Assessment Reports** proposes AI Design, a semi-automatic framework to generate impact assessment reports for managing risk of AI systems, incorporating stakeholder perspectives and using LLM-based tools to assist in the process.

These approaches collectively advance progress towards civilizing and humanizing AI by aiming to improve AI fairness, reduce harmful content, align AI with specific contextual requirements, and facilitate responsible AI development and assessment.

Challenges for Future Research

The approach towards civilizing and humanizing AI must involve a blend of technologies such as generative AI, HCI-focused conceptual frameworks, and explicit knowledge-based approaches, each with its unique strengths and promises. For example, InstructGPT models and Constitutional AI frameworks focus on aligning AI behavior with human preferences and values through reinforcement learning and constitutions. Meanwhile, social science-inspired and humanity-inspired approaches emphasize the integration of social norms and values into AI systems. Neurosymbolic AI offers concrete ways to implement explicit representations of such values for integration into AI systems. This includes developing knowledge graphs that blend general and domain-specific knowledge with socio-cultural values, handling dynamic societal values, and using NLP to map social science concepts. Representations must facilitate human-like understanding, balancing perspectives and ethical considerations.

Recent progress in the mechanistic interpretability of neural models, the development of comprehensive knowledge graphs for representing aspects related to sociocultural norms and values, and the emergence of neurosymbolic formulations for integrating neural and symbolic representations show promise in addressing these issues [15]. Encouragingly, leaders in the space of LLMs, such as OpenAI and NVIDIA, are also incorporating neurosymbolic mechanisms in their systems to enforce model behaviors consistent with the objectives of humanizing and civilizing AI. Despite advancements across many frontiers, significant challenges remain, which we cover in the next section.

The future of civilizing and humanizing AI lies in the convergence of various approaches and the continuous refinement of existing frameworks. Key areas for future research and development include:

1) Enhanced Interpretability: Developing methods to make AI models more transparent and understandable will be crucial in addressing the black-box nature of deep neural networks. Innovative approaches like mechanistic interpretability and others can help demystify AI decision-making processes.
2) Robustness and Security: Improving the robustness of AI models against hallucinations and adversarial attacks is essential for ensuring their reliability and safety. This involves advancing techniques in adversarial training, anomaly detection, and model validation.
3) Ethical and Societal Considerations: As AI systems become more integrated into daily life, ensuring they align with ethical and societal norms is paramount. This includes ongoing efforts to model societal values and norms accurately and develop frameworks that reflect diverse perspectives, institutional processes, and cultural contexts.
4) Regulatory Compliance: The evolving landscape of AI regulations necessitates the development of AI systems that comply with legal and ethical standards. This involves staying abreast of regulatory changes and incorporating compliance mechanisms into AI design and deployment.
5) Collaborative Efforts: The future of civilizing and humanizing AI will benefit from collaboration across disciplines, including AI research, social sciences, ethics, law, and policy, for instance, to better manage the risk and impact of AI systems on human rights in societal context. Multidisciplinary efforts can foster comprehensive solutions that address the multifaceted challenges of AI development.

The ongoing effort to civilize and humanize AI necessitates a careful balance between technological progress, ethical considerations, and societal values. By integrating diverse approaches and continually refining frameworks, the AI community can aim to create systems that are both intelligent and aligned with human and societal values. While academia has long championed frameworks that enable the practical implementation of these principles, the industry's shift from purely statistical methods to rule-based and guardrails-based neurosymbolic approaches is a significant and promising development. This shift offers encouraging evidence that major companies are taking

concrete steps to develop AI solutions that are not only powerful but also adhere to ethical standards and societal norms. By incorporating society-inspired rules and guardrails, these AI systems are better equipped to navigate complex, real-world scenarios with a focus on safety, reliability, and ethical conduct. This transition marks a move towards more transparent, controllable, and trustworthy AI, reflecting a commitment to responsibly advancing technology for the benefit of humanity.

Acknowledgements

This research is partly funded by NSF awards #2133842, #2335967, and #2210107. The views expressed here are those of the authors, not those of the sponsors.

REFERENCES

1. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.
3. Goertzel, B. (2007). Artificial general intelligence (Vol. 2, p. 1). C. Pennachin (Ed.). New York: Springer. December.
4. Chakraborty, M., Tonmoy, S. T. I., Zaman, S. M., Gautam, S., Kumar, T., Sharma, K., ... & Das, A. Counter Turing Test (CT2): AI-Generated Text Detection is Not as Easy as You May Think-Introducing AI Detectability Index (ADI). In The 2023 Conference on Empirical Methods in Natural Language Processing. December.
5. Rawte, V., Chadha, A., Sheth, A., & Das, A. Tutorial Proposal: Hallucination in Large Language Models. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries (pp. 68-72). May.
6. Smuha, N. A., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., & Yeung, K. (2021). How the EU can achieve legally trustworthy AI: a response to the European Commission's proposal for an Artificial Intelligence Act. Available at SSRN 3899991.
7. Ouyang, Long, et al. "Training language models to follow instructions with human feedback." Advances in Neural Information Processing Systems 35 (2022): 27730-27744.
8. Bai, Yuntao, et al. "Constitutional ai: Harmlessness from ai feedback." arXiv preprint arXiv:2212.08073 (2022).
9. Yao, Jing, et al. "Value FULCRA: Mapping Large Language Models to the Multidimensional Spectrum of Basic Human Values." arXiv preprint arXiv:2311.10766 (2023).
10. Templeton, A. (2024). Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Anthropic, Online at https://shorturl.at/ydYdb.
11. Tuning Our Algorithmic Amplifiers: Encoding Societal Values into Social Media Ais." Stanford HAI, Online at https://rb.gy/q1rrls.
12. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. International Journal of Human-Computer Interaction, 36(6), 495-504.
13. Gruber, T. (2024, June 21). What is humanistic AI? humanistic artificial intelligence - a guiding philosophy. Tom Gruber | Humanistic AI. Online at https://tomgruber.org/humanistic-ai
14. Purohit, Hemant, Valerie L. Shalin, and Amit P. Sheth. "Knowledge graphs to empower humanity-inspired AI systems." IEEE Internet Computing 24.4 (2020): 48-54.
15. Sheth, A., & Roy, K. (2024). Neurosymbolic Value-Inspired Artificial Intelligence (Why, What, and How). IEEE Intelligent Systems, 39(1), 5-11.

Amit Sheth is the NCR Chair & Professor at the University of South Carolina; he founded the AI Institute (AIISC; http://aiisc.ai). He is a Fellow of the IEEE, AAAI, AAAS, ACM, and AIAA, and is a receipient of the IEEE CS Wallace McDowell and IEEE ICSVC Research Innovation awards.

Kaushik Roy is a Ph.D. candidate at AIISC. His research focuses on developing neurosymbolic methods for declarative and process knowledge-infused learning, reasoning, and sequential decision-making, with a particular emphasis on social good applications.

Hemant Purohit is an Associate Professor of Information Sciences and Technology and director of the Humanitarian Informatics Lab at George Mason University. Purohit's research on human-centered computing and human-AI collaboration is highly relevant to this issue.

Amitava Das is a Research Associate Professor at AIISC. His contributions spans challenging topics related to this special issue, including hallucination, human-AI detectability, multimodal misinformation and disinformation.