Human-centric Artificial Intelligence: The road ahead

Prof. Dr. Niklas Kühl
Professor for Information Systems and Human-centric Artificial Intelligence, Universität Bayreuth

kuehl@uni-bayreuth.de

This editorial explores the critical role of Human-centric Artificial Intelligence in marketing, emphasizing the importance of explainable AI, fairness, and human's appropriate reliance on AI systems. It highlights the growing gap between AI's technical advancements and their impact on human behavior, stressing the need for transparent and equitable AI solutions that enable stakeholders to reflect on AI's recommendations critically.

Keywords: › human-centric › Artificial Intelligence › explainable Artificial Intelligence › AI fairness › appropriate reliance

1 Introduction

I could start the extended editorial on this issue of "Artificial Intelligence in market communication and brand management" by stating how AI has rapidly ascended as a transformative technology, reshaping various industries and fundamentally altering how businesses operate (Berg, Raj, & Seamanset, 2023). I could re-iterate how, in marketing, AI's ability to analyze vast amounts of data, predict consumer behavior, and personalize interactions has become an indispensable tool (Kshetri et al., 2023). But, you know that probably already, and I will spare you further details on AI's promise — e.g., how it has a total economic potential of up to 25.6 trillion USD (Chui et al., 2023) — and I will cut to the chase: AI is here to stay.

Today, I want to shed light on one specific lens of AI research that has been neglected in the past and is, in my view, fundamental if we want to "make AI work": the human-centric perspective. In the past, AI research predominantly focused on technical advancements, such as improving algorithms, computational efficiency, and data processing capabilities, often prioritizing performance metrics over human implications (Taylor, O'Dell, & Murphy, 2023). And there was good reason for it: This approach has led to remarkable progress in fields like machine learning (Sarker, 2021) and natural language processing (Ding et al., 2023).

In contrast and complement to this, human-centric AI prioritizes designing, developing, and deploying AI systems that "understand" human needs, enhance human performance and well-being, respect human rights, and align with human values. Why do I believe this perspective to be essential, especially right now? Since OpenAI released ChatGPT in November 2022, the tool has played a pivotal role in democratizing AI by making large language models (LLMs) accessible to a broad audience, sparking widespread interest and fueling the hype around AI's potential—which is, do not get me wrong—in many ways a blessing. Its ability to generate human-like text has helped users across various fields, from creative writing, personalized marketing to customer service. Suddenly, everyone could experience AI for themselves—free of charge and with the most straightforward interface possible. However, as the excitement around AI grows (and this is the "curse"-part), there is also a growing gap in understanding how humans interact with these systems and, more importantly, the long-term implications of AI usage on human behavior, cognition, and us as a society. To put it mildly, we were overrun, and we, as researchers, are only beginning to grapple with the ethical, psychological, and social effects of widespread AI adoption, highlighting the need for more research and careful consideration of how AI is incorporated into our human lives—both privately and professionally.

So, as AI continues integrating more critical functions, a human-centric approach ensures its development and deployment (remain?) aligned with human values and needs. It is vital to stress that human-centric AI is not merely a technological aspiration but a *necessary* approach to ensure that AI serves humanity ethically, transparently, and beneficially. To me, this concept encompasses various dimensions, three of which I would like to shed more detail on: (1) explainable AI, (2) fairness in AI, and (3) human's appropriate reliance on AI systems. Each of these areas addresses fundamental aspects of human-centric AI, aiming to create understandable, equitable, and trustworthy systems. In the remaining editorial, I want to explore these themes and outline future research directions essential for advancing human-centric AI in marketing and beyond.

2 Explainable AI

Explainable AI (XAI) is a cornerstone of human-centric AI. It seeks to make the decision-making processes of AI systems transparent and understandable to human users. The importance of XAI lies in its potential to build trust, facilitate accountability, and enable informed decision-making (Ali et al., 2023). Having explainable models is particularly critical in marketing, where AI-driven decisions can significantly impact consumer experiences and business outcomes (Rai, 2020).

XAI techniques such as feature importance scores, Local Interpretable Model-agnostic Explanations (LIME), and SHAP (SHapley Additive exPlanations) can provide insights into how AI models derive their predictions (Ribeiro, Singh & Guestrin, 2016). These techniques can help marketers understand the rationale behind AI recommendations, e.g., to allow for better strategic decisions and customer interactions. However, while XAI seems to be the panacea for complex and black-box-based AI models like deep neural networks, there is a catch to it—or multiple ones, to be precise. First of all, the typically used post-hoc explainability techniques involve using AI models to interpret and explain the decisions made by other, often more complex, AI models, which can introduce layers of abstraction and uncertainty in understanding the true reasoning behind an AI's output (Retzlaff et al., 2024). This process highlights the challenge of relying on one AI to "demystify" another, potentially compounding the difficulty of achieving clear, transparent, and trustworthy explanations. Thus, we need to be careful when using explanations for high-stake decision-making (Rudin, 2019). In consequence, the effectiveness of XAI is contingent upon two aspects: (1) the fidelity of its explanations (are the explanations truthful to the underlying decision model?) and (2) the human's ability to comprehend and utilize them (are the explanations understandable and actionable by the human?). For instance, simplified explanations can sometimes misrepresent the underlying complexities, leading to a false sense of security or misunderstandings (van der Waa et al., 2021). Moreover, human cognitive limitations pose a challenge in fully grasping AI-generated explanations—which leads to a more fundamental question: Can complex (AI) reasoning be even fully explained in ways that we humans understand (Liao & Varshney, 2021; Riefle et al., 2024)? Can we understand errors if explanations are wrong, even though they sound plausible and convincing (Lakkaraju & Bastani, 2020; Morrison et al., 2024)? Do explanations really foster trustworthiness of humans (Kästner et al., 2021; Weber et al., 2024)? Do we need to personalize explanations for different stakeholders (Langer et al., 2021)? Future research should focus on both understanding and improving the fidelity of explanations and developing metrics to assess and enhance human comprehension. Along those lines, it is also essential to investigate whether XAI genuinely improves decision-making (Senoner, Netland & Feuerriegel et al., 2022; Senoner et al., 2024) or merely provides an illusion of transparency (Fok & Weld, 2023; Schemmer et al., 2022).

3 Fairness in AI

Fair AI is a critical requirement of human-centric AI, ensuring that AI systems operate without bias and do not discriminate against any group or individual. At least, that is how far the theory goes—because AI models' sole purpose is to identify patterns based on "biases" in data; only now, we want to make sure the model does not do so based on sensitive or protected attributes. Unless, it makes sense—for instance, in medicine, differentiating between women and men is an important (although sensitive) aspect (Cirillo et al., 2020). You see, the Fair AI discussion is not trivial. Nonetheless, across all fields, but also in marketing, biased AI algorithms can perpetuate stereotypes, unfairly target specific demographics, and lead to discriminatory practices (Caton & Haas, 2024). But why can AI be biased and (in consequence) discriminative in the first place? AI models learn from data that may contain historical biases, leading them to replicate and even amplify those biases in their outputs. Thus, addressing these issues is vital for creating equitable and inclusive AI systems.

Current research has identified various fairness metrics to measure bias, such as demographic parity, equalized odds, and fairness through unawareness (Friedler, Scheidegger, & Venkatasubramanian, 2021). However, these fairness notions often conflict, necessitating trade-offs that complicate the implementation of fair AI. Within the FATE (Fairness, Accountability, Transparency, and Ethics) community, it is intensively discussed if and how multiple fairness notions can be considered (Bell et al., 2023).

Interestingly, we see a similar rationale like in the XAI discussion: While everyone agrees that fairness is necessary and potentially helpful, the precise implementation for a (real-world) problem is challenging. Who

Abstract

Dieser Artikel beleuchtet die Rolle der humanzentrischen künstlichen Intelligenz im Marketing und betont die Bedeutung von erklärbarer KI, Fairness und das kalibrierte Vertrauen der Menschen in KI-Systeme. Er hebt die wachsende Kluft zwischen den technischen Fortschritten der KI und ihren Auswirkungen auf das menschliche Verhalten hervor und betont den Bedarf an transparenten und nicht-diskriminierenden KI-Lösungen, die es Entscheidungsträgern ermöglichen, die Empfehlungen der KI kritisch zu reflektieren

***Keywords**:* › *humanzentrisch* › *Künstliche Intelligenz* › *erklärbare Künstliche Intelligenz* › *KI Fairness* › *Kalibriertes Vertrauen*

defines what is fair (Deshpande & Sharp, 2022)? What (economical) costs are associated with “making AI fair” (von Zahn, Feuerriegel & Kuehl, 2022)? Furthermore, how can we ensure the different perspectives of system designers, decision-makers, and affected individuals are appropriately considered and balanced (Zhang et al., 2023)?

Future research should explore how to reconcile these conflicting fairness notions and examine the role of explainability in achieving fairness. While XAI can help uncover biases, it may also introduce new biases or obscure deeper issues. Thus, it is crucial to critically assess the interaction between explainability and fairness (Deck, Schoeffer et al., 2024). Researchers should develop context-specific frameworks for fairness assessments and explore how different fairness metrics impact various stakeholders. Additionally, there is a need to investigate how regulatory frameworks and industry standards can support fair AI practices — with the EU AI Act being an important starting point here (Deck, Müller et al., 2024).

4 Appropriate Reliance

The truth is: AI systems make mistakes (Koch, Föhr & Germelmann, 2023) — just like humans do (Culverhouse et al., 2003). The question is: Can humans identify and correct these errors? When humans interact with AI systems to make (better) decisions, past studies observed two specific types of human behavior: Algorithm aversion (Liu et al., 2023) on the one end of the reliance spectrum and automation bias (Goddard, Roudsari, & Wyatt, 2011) on the other. Human’s appropriate reliance on AI systems refers to their ability to depend on AI for decision-making to the extent that is calibrated (Schemmer et al., 2023), i.e., justified by the technology’s capabilities, the psychology of the users, and the organizational context (Goddard, Roudsari, & Wyatt, 2011). Despite its importance, the concept of appropriate reliance is under-researched. Achieving it requires a socio-technical and nuanced understanding of the interplay between AI technology, human psychology, and organizational environments. Humans may over-rely on AI due to overconfidence in its capabilities or their characteristic of being lazy efficient decision-makers. They may under-rely due to skepticism, lack of trust, or overestimating their own abilities. Both, overreliance and underreliance, lead to suboptimal outcomes, as they do not leverage the complementary potential of human-AI teams (Hemmer et al., 2024)—but combining the strengths of each entity is essential for effective teaming (Kühl et al., 2022). Due to the earlier mentioned democratization of AI via tools like ChatGPT, human oversight of AI recommendations in general and their appropriate reliance in specific have been boosted to one of the most crucial aspects in ensuring that AI systems are used effectively (Sterz et al., 2024)—and rightfully so, as human decision-makers legally have complete responsibility for their AI-supported decisions, but the presence of AI support might reduce them to “rubber-stamping” borgs (Fügener et al., 2021; Wagner, 2019).

Future research should aim to optimize appropriate reliance by identifying and understanding factors influencing human-AI decision-making—both technical (like uncertainty and XAI) as well as psychological (like cognitive biases and decision-making heuristics). For instance, studies should examine how different levels of XAI influence (appropriate) reliance (Schemmer et al., 2023), how reliance and team performance are related (Schoeffer, Jakubik et al., 2024), and how effective reliance can be fostered (He, Kuiper & Gadiraju, 2023). Additionally, organizational policies and training programs that promote a balanced reliance on AI should be investigated (Gimpel et al., 2024). These programs should ensure that human judgment complements AI recommendations rather than being overshadowed by them. Finally, the interplay of all mentioned human-centric aspects, XAI, fairness and appropriate reliance remains underexplored (Schoeffer, De-Arteaga et al., 2024).

5 Conclusion

The future of AI in marketing and beyond is undeniably human-centric, focusing on developing technologies that prioritize human welfare, autonomy, and fairness. By addressing critical challenges related to explainability, fairness, and appropriate reliance, researchers can ensure that AI technologies are developed and deployed in a manner that is ethical, transparent, and beneficial to all stakeholders.

To fully harness AI’s potential in market communication and brand management, organizations must understand and prioritize human-centric AI design as a fundamental cornerstone, considering the system’s properties like explainability, appropriate reliance, and alignment with human values. Additionally, addressing bias and fairness proactively within the companry is essential, implementing robust measures that promote inclusivity and ethical AI-driven decision-making for its employees. Finally, encouraging training for effective human-AI collaboration and fostering a balance between AI support and human judgment ensures that AI complements rather than overrides human expertise.

As we move forward, academia, industry, and policymakers need to collaborate and guide AI research and application in ways that enhance its positive impact on humans—while mitigating potential risks. By embedding human-centric principles into AI development, we can build a future where AI not only augments human capabilities but also respects and upholds the values that define our society. A promising field of research lies ahead.

References

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99, 101805.

Bell, A., Bynum, L., Drushchak, N., Zakharchenko, T., Rosenblatt, L., & Stoyanovich, J. (2023). The possibility of fairness: Revisiting the impossibility theorem in practice. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 400-422.

Berg, J. M., Raj, M., & Seamans, R. (2023). Capturing value from artificial intelligence. *Academy of Management Discoveries, 9*(4), 424-428.

Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys, 56*(7), 1-38.

Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L. & Zemmel, R. (2023). *The economic potential of generative AI*. Abruf von https://t1p.de/5u05f.

Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M. J., & Chadha, A. S. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digital Medicine, 3*(1), 1-11.

Culverhouse, P. F., Williams, R., Reguera, B., Herry, V., & González-Gil, S. (2003). Do experts make mistakes? A comparison of human and machine indentification of dinoflagellates. *Marine Ecology Progress Series, 247*, 17-25.

Deck, L., Müller, J.-L., Braun, C., Zipperling, D., & Kühl, N. (2024). Implications of the AI Act for Non-Discrimination Law and Algorithmic Fairness. *Third European Workshop on Algorithmic Fairness (EWAF'24).*

Deck, L., Schoeffer, J., De-Arteaga, M., & Kühl, N. (2024). A Critical Survey on Fairness Benefits of Explainable AI. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1579-1595.

Deshpande, A., & Sharp, H. (2022). Responsible AI Systems: Who are the Stakeholders? *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society,* 227-236.

Ding, Q., Ding, D., Wang, Y., Guan, C., & Ding, B. (2023). Unraveling the landscape of large language models: a systematic review and future perspectives. *Journal of Electronic Business & Digital Economics,* ahead-of-print.

Fok, R., & Weld, D. S. (2023). In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-Advised Decision Making. *ArXiv.* Doi.org/10.48550/arXiv.2305.07722.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM, 64*(4), 136-143.

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI. *Management Information Systems Quarterly, 45*(3), 1527-1556.

Gimpel, H., Gutheil, N., Mayer, V., Bandtel, M., Büttgen, M., Decker, S., Eymann, T., Feulner, S., Kaya, M. F., & Kufner, M., …, Urbach, N. (2024). (Generative) AI Competencies for Future-Proof Graduates: Inspiration for Higher Education Institutions. Hohenheim Discussion Papers in Business, Economics and Social Sciences. Stuttgart: Universität Hohenheim.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121-127.

He, G., Kuiper, L., & Gadiraju, U. (2023). Knowing about knowing: An illusion of human competence can hinder appropriate reliance on AI systems. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1-18.

Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., & Satzger, G. (2024). Complementarity in Human-AI Collaboration: Concept, Sources, and Evidence. *ArXiv* Preprint. Doi.org/10.48550/arXiv.2404.00029.

Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T., & Sterz, S. (2021). On the relation of trust and explainability: Why to engineer for trustworthiness. *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW),* 169-175.

Koch, T., Föhr, J., & Germelmann, C. C. (2023). Who's to blame?: The Effect of Consumers' Role Attributions of Smart Voice-Interaction Technologies during Service Failures. Presentation at Frontiers in Service 2023, Maastricht.

Kshetri, N., Dwivedi, Y. K., Davenport, T. H., & Panteli, N. (2023). Generative artificial intelligence in marketing: Applications, opportunities, challenges, and research agenda. *International Journal of Information Management, 75*(6), 102716.

Kühl, N., Goutier, M., Baier, L., Wolff, C., & Martin, D. (2022). Human vs. supervised machine learning: Who learns patterns faster? *Cognitive Systems Research, 76*(Dec.), 78-92.

Lakkaraju, H., & Bastani, O. (2020). " How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,* 79-85.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence, 296*(July), 103473.

Liao, Q. V., & Varshney, K. R. (2021). Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *ArXiv* Preprint ArXiv:2110.10790.

Liu, M., Tang, X., Xia, S., Zhang, S., Zhu, Y., & Meng, Q. (2023). Algorithm Aversion: Evidence from Ridesharing Drivers. *Management Science.* Doi:10.1287/mnsc.2022.02475.

Morrison, K., Spitzer, P., Turri, V., Feng, M., Kühl, N., & Perer, A. (2024). The Impact of Imperfect XAI on Human-AI Decision-Making. *Proceedings of the ACM on Human-Computer Interaction, 8*(CSCW1), 1-39.

Rai, A. (2020). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science, 48*(1), 137-141.

Retzlaff, C. O., Angerschmid, A., Saranti, A., Schneeberger, D., Roettger, R., Mueller, H., & Holzinger, A. (2024). Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. *Cognitive Systems Research, 86*, 101243.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 1135-1144.

Riefle, L., Hemmer, P., Benz, C., & Vössing, M. (2024). The Role of Cognitive Styles for Explainable AI. Presentation on 32nd European Conference on Information Systems (ECIS 2024), Paphos, Zypern, 13.06.2024 – 19.06.2024.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206-215.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science, 2*(3), 160. Doi.org/10.1007/s42979-021-00592-x.

Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., & Vössing, M. (2022). A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 617-626.

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. *Proceedings of the 28th International Conference on Intelligent User Interfaces,* 410-422.

Schoeffer, J., De-Arteaga, M., & Kuehl, N. (2024). Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. *Proceedings of the CHI Conference on Human Factors in Computing Systems,* 1-18.

Schoeffer, J., Jakubik, J., Vössing, M., Kühl, N., & Satzger, G. (2024). AI Reliance and Decision Quality: Fundamentals, Interdependence, and the Effects of Interventions. *Journal of Artificial Intelligence Research (JAIR).* Preprint. Doi.org/10.48550/arXiv.2304.08804.

Senoner, J., Netland, T., & Feuerriegel, S. (2022). Using explainable artificial intelligence to improve process quality: evidence from semiconductor manufacturing. *Management Science, 68*(8), 5704-5723.

Senoner, J., Schallmoser, S., Kratzwald, B., Feuerriegel, S., & Netland, T. (2024). Explainable AI improves task performance in human-AI collaboration. *ArXiv,* Preprint ArXiv:2406.08271.

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., & Langer, M. (2024). On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. *The 2024 ACM Conference on Fairness, Accountability, and Transparency,* 2495-2507.

Taylor, R. R., O'Dell, B., & Murphy, J. W. (2023). Human-centric AI: philosophical and community-centric considerations. *AI & SOCIETY,* 39, 2417-2424.

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence, 291*, 103404.

von Zahn, M., Feuerriegel, S., & Kuehl, N. (2022). The Cost of Fairness in AI: Evidence from E-Commerce. *Business & Information Systems Engineering, 64*(3), 335-348.

Wagner, B. (2019). Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. *Policy & Internet, 11*(1), 104-122.

Weber, R. O., Johs, A. J., Goel, P., & Silva, J. M. (2024). XAI is in trouble. *AI Magazine 45*(3), 300-316.

Zhang, A., Walker, O., Nguyen, K., Dai, J., Chen, A., & Lee, M. K. (2023). Deliberating with AI: improving decision-making for the future through participatory AI design and stakeholder deliberation. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW1), 1-32.