weizenbaum journal OF THE DIGITAL SOCIETY

WEIZENBAUM JOURNAL OF THE DIGITAL SOCIETY
Volume 4 \ Issue 1 \ w4.1.7 \ 05-27-2024
ISSN 2748-5625 \ DOI 10.34669/WI.WJDS/4.1.7

Information on this journal and its funding can be found on its website:
https://wjds.weizenbaum-institut.de

KEYWORDS

artifical intelligence
human-AI collaboration
explainable artificial intelligence
EU AI act

VOICES FOR THE NETWORKED SOCIETY

Unlocking AI's Potential

Human Collaboration as the Catalyst

Peter Buxmann[1] \ Sara Ellenrieder[*1]

[1] Technical University of Darmstadt
[*] Corresponding author, ellenrieder@is.tu-darmstadt.de

ABSTRACT

Rapid advances in artificial intelligence (AI) have fueled high expectations for the technology's potential to fundamentally transform our economy and society through automation. However, given the inscrutability and, sometimes, susceptibility to error of AI systems, we argue that the focus should shift towards fostering effective human-AI collaboration rather than pursuing automation alone. In this context, system decisions must be made available to decision-makers in an explainable and understandable manner, as further required by the EU's recently passed AI Act. Research shows that there is potential for humans to learn from explainable AI systems and improve their own performance over time. Meanwhile, in addition to enabling humans to benefit from working with AI systems on various everyday tasks, such collaboration ensures the safe and reliable use of AI systems, especially in high-risk areas such as medicine, where human oversight remains paramount.

1 Introduction: Rapid Developments Fuel High Expectations

Rapid advances in artificial intelligence (AI) have not only returned the technology to the public spotlight but also opened up countless new possibilities for implementation in the social, commercial, and industrial domains (Berente et al., 2021; McKinsey Global Institute, 2021). In particular, breakthroughs in generative AI (GenAI) have prompted widespread adoption of the technology, precipitating high expectations for its transformative potential (Dwivedi et al., 2023). For example, after the release of the GPT 3.5-powered GenAI chatbot ChatGPT in November 2022 attracted over one million users within two months, it took OpenAI, the organization behind ChatGPT, only four months to introduce GPT-4, a GenAI chatbot built upon a much more powerful large language model (LLM) (OpenAI, 2023). Shortly thereafter, Google released the PaLM 2 LLM and integrated the technology into its search engine (Roth, 2023). The estimated automation potential of GenAI is impressive, with industry reports suggesting that generative AI can automate 60–70% of daily work. However, the percentage of companies implementing AI in their operations has changed little over the past two years (McKinsey & Company, 2023).

Although companies may not yet be fully keeping pace with the rapid technological developments, research has already demonstrated that AI systems can undertake tasks previously performed by human experts (e.g., Lebovitz et al., 2021; Shen et al., 2019). In some cases, AI systems have even outperformed human experts (e.g., McKinney et al., 2020; Shen et al., 2019), and their use is increasingly being considered in high-risk areas such as medicine and aviation (Lebovitz et al., 2021; Reyes et al., 2020; Rudin, 2019). A notable application of AI is the diagnosis of diseases such as cancer and the classification of tumors from medical images (e.g., Calisto et al., 2021; Lebovitz et al., 2021; McKinney et al., 2020; Pumplun et al., 2023; Silva & Ribeiro, 2011). This raises questions about the possibility of meeting the increasingly high expectations of AI, especially concerning automation potential, or whether these developments should ultimately be characterized as hype.

2 Artificial Intelligence: Challenges Ahead, Opportunities Within Reach

To assess the potential of AI to change how we work, and to understand the barriers to widespread adoption of this technology, it is important to understand the fundamental mechanics of AI. Modern AI applications typically rely on machine learning (ML) algorithms, a subfield of AI that sees algorithms independently recognize patterns in large data sets (Brynjolfsson & Mitchell, 2017; Russell & Norvig, 2021). After training, these ML models can apply learned patterns to new data to make predictions, perform classifications, and either support human decision-making or automatically initiate further actions (Jordan & Mitchell, 2015; Mitchell, 1997). In GenAI, a subarea of ML, models can even generate new data based on learned patterns (Dwivedi et al., 2023; Teubner et al., 2023). Thus, ML models can find solutions independently, potentially generating new knowledge that complements that of human decision-makers (e.g., Fügener et al., 2019; Sturm et al., 2021).

2.1 Unique Characteristics Challenge Expectations

While modern AI systems are powerful and often demonstrate high performance (e.g., McKinney et al., 2020), they are inherently prone to error because they are based on statistical approaches (e.g., Berente et al., 2021; Jordan & Mitchell, 2015). In addition, AI systems can exhibit inconsistent behavior (Schuetz & Venkatesh, 2020) and, in the case of GenAI, produce so-called hallucinations, where systems generate inaccurate information and present it convincingly to the end user (Chui et al., 2022; Dwivedi et al., 2023). It is also critical to be aware that AI systems can be biased due to inherent biases in the training data, which can lead to discrimination in decision-making (Berente et al., 2021; Rai et al., 2019).

A well-known example is the AI system developed by Amazon to help select suitable job applicants. The system did not evaluate candidates in a gender-neutral manner and favored men over women, leading to the project's rapid failure. Contrary to many accusations, AI algorithms are not sexist by design; instead, because Amazon has historically hired primarily men, the system produced output patterns similar to those it learned from the biased data provided (Dastin, 2018). These unique characteristics of AI systems combined with the complexity of modern algorithms to pose challenges for implementation, especially where the complexity results in the output of the systems becoming inscrutable to human decision-makers (Asatiani et al., 2021; Berente et al., 2021). Because the inner mechanisms of AI systems are consequently incomprehensible and the decisions made by AI systems cannot be fully understood by humans, they are often referred to as black boxes (Adadi &

Berrada, 2018; Castelvecchi, 2016; Guidotti et al., 2018; Rudin, 2019). In this way, AI systems differ from traditional information systems, something that must be recognized to understand the challenges of deployment and to more comprehensively and profoundly assess the future impacts of AI.

2.2 Explainable Artificial Intelligence

Promising approaches for enabling safer and more reliable deployment of AI systems appear in the nascent field of research surrounding explainable artificial intelligence (XAI). XAI research aims to make the decision-making processes and outputs of ML systems more understandable to humans. This is accomplished by providing human decision-makers with explanations that help clarify why certain system decisions were made (Arrieta et al., 2020; Meske et al., 2022). Although XAI outputs were initially designed and provided primarily for developers of AI systems, research has recognized the potential for these explanations to positively impact end users too (e.g., Asatiani et al., 2021; Ellenrieder et al., 2023; Gaube et al., 2023).

While AI systems are often admired for their automation potential, XAI research shifts the focus back to humans to enable the successful use of AI systems in the context of collaboration. These efforts are also supported by new regulations. For example, the political consensus reached by the EU in December 2023 manifested in the AI Act, which establishes high requirements for the explainability of AI systems, especially *high-risk AI systems* at the European level (European Commission, 2023). The goal is to empower end users to correctly interpret the output of AI systems and to evaluate the unique system characteristics described above (Panigutti et al., 2023).

3 Conclusion: Shifting From Long-term Automation Goals to Effective Collaboration

Many expectations of AI's potential – some of which may be exaggerated – are focused on the automation of tasks. However, given the current state of technology, effective collaboration between humans and AI systems holds the greatest potential in terms of integrating AI into everyday working life and substantially impacting all functional areas of companies. Here, the AI system and the human work independently to achieve a common goal or task (e.g., McNeese et al., 2018; Siemon, 2022). Products such as Microsoft CoPilot – which is being marketed as an *everyday AI companion* (Microsoft, 2024) – already offer AI support for a significant portion of daily office work. However, it will become increasingly important to enable human oversight of system

decisions and to ensure that humans retain and even improve their own skills through collaboration.

Initial studies have already shown that collaboration with AI systems can have a positive impact on human performance and decision-making, with the explainability of the output being essential in this context (e.g., Abdel-Karim et al., 2023; Gaube et al., 2023). These findings are of particular interest for the deployment of AI systems in high-risk areas, where the human decision-maker will continue to bear responsibility and retain final decision-making authority.

In an experimental study with radiologists, we were able to show that by collaborating with AI-based decision support systems for brain tumor segmentation, radiologists could improve their performance and decision confidence and even learn from the AI systems. Explainable output design not only improved the radiologists' learning outcomes but also prevented incorrect learning in the case of errors made by the AI systems. Notably, some radiologists were able to learn from the mistakes of the AI systems and improve their own performance when the system presented them in an explainable way (Ellenrieder et al., 2023).

In summary, integrating AI into our daily lives and work processes represents a significant milestone in technological evolution. AI's automation capabilities have been highly anticipated. However, the collaboration between humans and AI systems holds the most promise for revolutionizing both the economy and society. Effective collaboration between AI and humans not only utilizes AI's capacity to process and analyze large amounts of data but also allows for human decision-making oversight, particularly in high-risk areas such as medicine. The challenges associated with AI, such as its propensity for errors, inscrutability, and inherent biases, underscore the importance of developing systems that are not only powerful but also explainable and understandable to human users. Hence, the collaborative approach not only addresses several ethical and practical challenges posed by AI but also opens up new avenues for leveraging AI to enhance human capabilities rather than replace them. As regulations like the AI Act come into effect, demanding higher standards for the explainability and reliability of AI systems, we are likely to witness more responsible and beneficial integration of AI across various sectors in the future.

Furthermore, the proposed focus on human-AI collaboration highlights important research needs that must be addressed to advance the field. Understanding and fostering effective human-AI partnerships will require interdisciplinary efforts from fields including computer science, psychology, sociology, and ethics. As a first step, understanding the human decision-making process is becoming increasingly important. This includes investigating human preferences and cognitive biases, which can provide insights into how AI systems can be designed to leverage those preferences and biases to improve future collaborative efforts. The success of AI technology will depend on the seamless integration of AI systems into decision-making processes and applications

through user-friendly interfaces. Microsoft CoPilot is already making progress in integrating AI into daily office tasks, but there remains significant potential to integrate AI systems into the work environment by developing new approaches that encourage effective bidirectional information exchange between humans and AI systems.

Second, in collaborative task completion, one party may predominantly take on certain parts of the task. While humans may become bored or tired of routine tasks over time, AI systems are well-suited for taking on standardized tasks. This demands that future research identify and assess which tasks are best suited for AI systems and which should be performed by humans to optimally achieve common goals by complementing each other's strengths. This is crucial not only for ensuring optimal performance but also for long-term job satisfaction.

Third, the focus on collaborative efforts clearly requires that employees engage with the functionality of AI systems and that companies actively promote AI training for their workforce. Employees and managers must develop an understanding of the AI systems deployed rather than hope for automation. For many types of organizations, building knowledge about AI systems and establishing rigorous long-term knowledge management present significant challenges that require new educational approaches.

Fourth, collaboration raises new questions regarding ethical considerations. Frameworks for the design of AI systems and regulations are needed to determine who bears responsibility for collaborative decisions and to ensure that discrimination by biased AI systems does not affect collaborative decision-making.

Finally, it must be recognized that XAI outputs are insufficiently developed to enable non-experts to fully understand AI systems and their decision-making processes. Interdisciplinary collaboration is required to design XAI for the diverse user groups that will collaborate with AI in the future. This will ensure that companies remain competitive and that society benefits without excluding any segments of the population from this progress.

References

Abdel-Karim, B. M., Pfeuffer, N., Carl, K. V., & Hinz, O. (2023). How AI-based systems can induce reflections: The case of AI-augmented diagnostic work. *MIS Quarterly, 47*(4), 1395–1424.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *An International Journal on Information Fusion, 58*, 82–115.

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2021). Sociotechnical envelopment of artificial intelligence: An approach to organizational deployment of inscrutable artificial intelligence systems. *Journal of the Association for Information Systems, 22*(2), 325–352.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly, 45*(3), 1433–1450.

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science, 358*(6370), 1530–1534.

Calisto, F. M., Santiago, C., Nunes, N., & Nascimento, J. C. (2021). Introduction of human-centric AI assistant to aid radiologists for multimodal breast image classification. *International Journal of Human-Computer Studies, 150*, 102607.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature, 538*(7623), 20–23.

Chui, M., Roberts, R., & Yee, L. (2022, December 22). Generative AI is here: How tools like ChatGPT could change your business. *Quantum Black: AI by McKinsey* https://www.mckinsey.com/capabilities/quantumblac/our-insights/generative-ai-is-here-how-tools-like-chatgpt-could-change-your-business.

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/idUSL2N1VB1FQ/.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, 102642.

Ellenrieder, S., Kallina, E. M., Pumplun, L., Gawlitza, J. F., Ziegelmayer, S., & Buxmann, P. (2023). Promoting learning through explainable artificial intelligence: An experimental study in radiology. *Proceedings of the 44th International Conference on Information Systems.* https://aisel.aisnet.org/icis2023/learnandiscurricula/learnandiscurricula/3/

European Commission. (n.d.). *AI Act*. Shaping Europe's digital future. Retrieved January 11, 2024 from https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2019). Cognitive challenges in human-AI collaboration: Investigating the path towards productive delegation. *Information Systems Research : ISR*. https://repub.eur.nl/pub/115830/ERS-2019-003-LIS_v1.pdf

Gaube, S., Suresh, H., Raue, M., Lermer, E., Koch, T. K., Hudecek, M. F. C., Ackery, A. D., Grover, S. C., Coughlin, J. F., Frey, D., Kitamura, F. C., Ghassemi, M., & Colak, E. (2023). Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Scientific Reports, 13*(1), 1383.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv., 51*(5), 1–42.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260.

Lebovitz, S., Levina, N., & & Lifshitz-Assa, H. (2021). Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. *MIS Quarterly, 45*(3), 1501–1526.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., … Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature, 577*(7788), 89–94.

McKinsey & Company. (2023, August 1). The state of AI in 2023: Generative AI's breakout year. *Quantum Black: AI by McKinsey*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year.

McKinsey Global Institute. (2021, December 8). The state of AI in 2021. *Quantum Black: AI by McKinsey*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/global-survey-the-state-of-ai-in-2021.

McNeese, N. J., Demir, M., Cooke, N. J., & Myers, C. (2018). Teaming with a synthetic teammate: insights into human-autonomy teaming. *Human Factors, 60*(2), 262–273.

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Information Systems Management, 39*(1), 53–63.

Microsoft. (n.d.). *Microsoft Copilot: Your everday AI companion*. Microsoft Copilot: Your Everday AI Companion. Retrieved March 25, 2024, from https://copilot.microsoft.com/.

Mitchell, T. M. (1997). *Machine learning* (Vol. 1). McGraw-Hill.

OpenAI. (2023). *ChatGPT – Release Notes*. Retrieved November 22, 2023, from https://help.openai.com/en/articles/6825453-chatgpt-release-notes.

Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., Scalzo, S., Mazzini, G., Sanchez, I., Soler Garrido, J., & Gomez, E. (2023). The role of explainable AI in the context of the AI Act. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1139–1150.

Pumplun, L., Peters, F., Gawlitza, J. F., & Buxmann, P. (2023). Bringing machine learning systems into clinical practice: A design science approach to explainable machine learning-based clinical decision support systems. *Journal of the Association for Information Systems, 24*(4), 953–979.

Rai, A., Constantinides, P., & Sarker, S. (2019). Next generation digital platforms: Toward human-AI hybrids. *The Mississippi Quarterly, 43*(1), iii–ix.

Reyes, M., Meier, R., Pereira, S., Silva, C. A., Dahlweid, F.-M., von Tengg-Kobligk, H., Summers, R. M., & Wiest, R. (2020). On the interpretability of artificial intelligence in radiology: Challenges and opportunities. *Radiology. Artificial Intelligence, 2*(3), e190043.

Roth, E. (2023, May 10). The nine biggest announcements from Google I/O 2023. *The Verge*. https://www.theverge.com/23718158/google-io-2023-biggest-announcements-ai-pixel-fold-tablet-android-14.

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.

Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach, global edition*. Pearson Deutschland.

Schuetz, S., & Venkatesh, V. (2020). The rise of human machines: How cognitive computing systems challenge assumptions of user-system interaction. *Journal of the Association for Information, 21*(2), 460–482.

Shen, J., Zhang, C. J. P., Jiang, B., Chen, J., Song, J., Liu, Z., He, Z., Wong, S. Y., Fang, P.-H., & Ming, W.-K. (2019). Artificial intelligence versus clinicians in disease diagnosis: Systematic review. *JMIR Medical Informatics, 7*(3), e10010.

Siemon, D. (2022). Elaborating team roles for artificial intelligence-based teammates in human-AI collaboration. *Group Decision and Negotiation, 31*(5), 871–912.

Silva, C., & Ribeiro, B. (2011). Purging false negatives in cancer diagnosis using incremental active learning. *Intelligent Data Engineering and Automated Learning - IDEAL 2011*, 394–402.

Sturm, T., Gerlach, J. P., Pumplun, L., Mesbah, N., Peters, F., Tauchert, C., Nan, N., & Buxmann, P. (2021). Coordinating human and machine learning for effective organizational learning. *MIS Quarterly, 45*(3).

Teubner, T., Flath, C. M., Weinhardt, C., van der Aalst, W., & Hinz, O. (2023). Welcome to the era of ChatGPT et al. *Business & Information Systems Engineering, 65*(2), 95–101.

Date received: April 2024
Date accepted: May 2024