Explainability in Generative AI: An Umbrella Review of Current Techniques, Limitations, and Future Directions

Prabha M. Kumarage[1,*,†], Mirka Saarela[1,†]

[1]Faculty of Information Technology, University of Jyväskylä, P.O.Box 35, FI-40014 Jyväskylä, Finland

Abstract

The rapid development of Generative Artificial Intelligence (GenAI) has introduced a new era of technological innovation by revolutionizing how people interact with information. Explainability is a crucial aspect to ensure transparency, accountability, and trust in these GenAI-driven systems. Interpreting and comprehending the decision-making process of GenAI models is becoming increasingly difficult as they become more complex and widespread over time. This research study aims to undertake a thorough navigation of the current state of explainability in GenAI through an umbrella review to provide an overview by analyzing the existing explainable techniques for GenAI and their limitations. The key limitations in explaining GenAI models include generalization issues, computational inefficiencies, trade-offs between interpretability and model performance, and unknown underlying data. Another significant finding is the absence of a standardized evaluation framework to measure and compare the effectiveness of different explainability techniques. This study highlights the importance of developing well-balanced GenAI-specific explainable techniques to ensure the responsible development of GenAI solutions. In addition, researchers, AI professionals, and policymakers seeking to improve the transparency and explainability of GenAI models can all greatly benefit from the findings.

Keywords

Generative Artificial Intelligence, Explainable Artificial Intelligence, Explainability Techniques, Explainability Challenges

1. Introduction

In the last decade, Generative AI (GenAI) has become among the most revolutionary subfields of AI. It has been rapidly progressing with the evolution and adoption of Large Language Models (LLMs), Generative Adversarial Network (GANs), Variational Autoencoders (VAEs) and other GenAI technologies that have extraordinary capabilities in tasks such as text generation, image creation, music composition, and even the production of codes for programming from training data [1]. Those newly generated synthetic contents in the way that are no longer able to distinguish from human creativity. However, this rapid and broad adoption of these GenAI models creates significant problems with transparency and explainability.

Addressing the explainability issues in complex AI systems has made Explainable Artificial Intelligence (XAI) an increasingly popular research topic. The increasing complexity of AI models makes it harder to understand how they operate internally. Thus, the explainability of these XAI techniques is concerned with providing clear and comprehensible reasons for AI-generated results [2]. Traditional AI systems, referring here to non-GenAI models such as rule-based systems or conventional machine learning approaches, primarily use explicit rules and algorithms to provide classifications or predictions, making it easier to explain. Although, machine learning models like deep neural networks also posed some explainability challenges due to their non-linear and stochastic behavior, GenAI systems introduce additional complexities through open-ended generation, emergent behavior and contest-depended outputs. Thus, the GenAI models have a more opaque black-box nature and make it difficult to

Late-breaking work, Demos and Doctoral Consortium, colocated with the 3rd World Conference on eXplainable Artificial Intelligence: July 09-11, 2025, Istanbul, Turkey

*Corresponding author.

†These authors contributed equally.

prabha.m.kumarage@student.jyu.fi (P. M. Kumarage); mirka.saarela@jyu.fi (M. Saarela)

0009-0002-9146-8647 (P. M. Kumarage); 0000-0002-1559-154X (M. Saarela)

understand the output in addition to the decision-making process behind those models. Lack of explainability causes several issues related to bias, fairness, and trustworthiness of these Gen AI models. In addition, it raises concerns about the use of these modern technologies in high-stakes situations such as autonomous driving, medical diagnostics, and judicial decision-making [3].

Even though explainability has the attention of researchers nowadays, most existing XAI techniques and frameworks are designed for use with the traditional AI models rather than GenAI models. In this paper, the term "traditional AI" refers to non-GenAI systems. As mentioned above, the inner workings behind the traditional AI models are simpler than those of GenAI models. One of the main differences between traditional AI and GenAI is their output. While traditional AI models help with decision making, GenAI models generate novel synthetic data. Therefore, the explainability techniques designed for traditional AI models might not always work well for GenAI models. In addition, the situation is further complicated by the absence of standardized evaluation metrics for explainability in GenAI and there is no way to evaluate the effectiveness of these existing XAI techniques. These gaps must be filled in order to guarantee that GenAI systems are interpretable, reliable, and ethically acceptable.

This research study focuses mainly on exploring current research studies on the topic of explainability of GenAI systems in various domains and critically analyzes current attempts to identify gaps and challenges. For that, the Systematic Literature Review (SLR) methodology has been utilized with an umbrella review of previous literature reviews. Eventually, this research study will provide valuable information for current AI researchers, industry practitioners, and policy makers working toward producing transparent and accountable GenAI systems. Thus, this research study will be a significant contribution to these ongoing conversations by examining the current state of explainability in GenAI systems and identifying directions for future improvement.

2. Explainability Techniques for GenAI

The traditional XAI techniques such as LIME, SHAP, and Grad-CAM are primarily designed to provide comprehensible explanations for AI models used in classification or regression tasks, where local and global approximations can be used to examine how input attributes and output predictions are related [4]. Nevertheless, GenAI systems function fundamentally differently because they are built on top of more complex architectures such as LLMs, GANs, and VAEs. These models frequently use complex internal mechanisms that involve latent variables, multi-step production pipelines, and high-dimensional parameter spaces to produce unstructured outputs like text, images, audio, and video [5]. As a result, novel techniques and considerations that go beyond traditional XAI approaches are needed to explain GenAI outputs.

The nature of the explanation is one significant distinction. Explanations of traditional XAI focus on model choices for fixed inputs by providing insights into the reasoning behind the prediction of a specific label. In contrast, GenAI explainability seeks to address why a specific output was generated, which aspects of the input such as a prompt or context impact particular elements of the output, and how internal representations contributed to the generative process [6]. As a result, techniques specifically designed to address the unique properties of GenAI systems have emerged. For example, concept lens [7] is a fascinating visual analytics framework that offers interpretability by allowing users to explore and analyze semantic manipulations in GANs. It works by identifying latent distributions that corresponds to semantically meaningful changes (e.g., facial expressions, object textures), clustering these directions into concepts, and visually showing how they influence generated images across different regions of the latent space. This concept-based explanation approach goes beyond individual latent vector analysis and expose higher-level semantic behaviors and inconsistencies by helping users comprehend biases, control constraints, and the structure of GenAI models.

Further, many efforts have been made to integrate traditional XAI methods by modifying them to be suitable for GenAI. For example, traditional XAI methods such as SHAP and partial dependence plots, which were originally designed for structured prediction tasks, have been combined with proxy models like LightGBM to interpret feature effects on the output alignment of GenAI models to enhance the explainability [8].

However, these explainable techniques are still mostly in the experimental stage and are frequently developed for specific applications. Therefore, there is currently no unified framework that can generalize across all forms GenAI systems and the majority of existing methods are customized to certain model architectures or modalities. For example, techniques that work well for explaining the latent space of image generators like GANs may not be applicable to transformer-based language models, and vise versa. Thus, there is no widely accepted or standardized categorization of GenAI-specific explainability techniques yet. However, Schneider [6] presents a taxonomy that categorized GenAI explainability techniques based on output properties such as scope, modality, and interactivity, and input and internal properties such as fundamental sources for XAI, required access by XAI methods, model explainers, sample difficulty and dimensions of pre-GenAI. Alongside the taxonomy, the paper outlines a desiderata for effective GenAI systems explainability. This include qualities like verification, descent, personalization and interaction, dynamic flexibility, costs, criteria for alignment, security and unpredictability. These desired properties ensure that explanations are not only understandable, but also trustworthy and adjustable to user needs.

While the terms "XGenAI" [9] and "GenXAI" [6] have been used casually in some literature to describe explainability of GenAI, there is currently no universally acknowledged term for this concept. Although the GenAI field is still developing, these GenAI-specific explanation techniques represent an important step in the direction of increasing the transparency, reliability, and accountability of GenAI systems. Understanding and categorizing these techniques is crucial for both technological advancement and alignment with regularity expectations for responsible GenAI.

3. Research Methodology

An umbrella review was utilized in this study to explore existing review articles on explainability in GenAI systems and this follows the guidelines for Preferred Reporting Items for Systematic Reviews and Meta-Analyzes (PRISMA) [10] to collect the relevant resources for review.

A comprehensive search for literature was carried out in November 2024 through six academic databases, namely Web of Science (WoS), Scopus, IEEE Xplore, ACM Digital Library, ScienceDirect, and Sage to identify relevant resources. The keywords were then carefully chosen to include various aspects of the explainability of GenAI. The boolean search string was formulated as follows to use within the fields of title, abstract, and keywords: (("explainable artificial intelligence" OR "XAI" OR "explainable AI" or "EAI") AND ("generative artificial intelligence" OR "generative AI" OR "GAI")).

This search was restricted based on pre-defined inclusion and exclusion, which required articles to in English, peer-reviewed, and published in journals or conference proceedings between 1 January 2020 and 25 November 2024, to ensure that the most recent and quality advancements were captured maintain strict academic standards [11]. Figure 1 provides the PRISMA flow chart which includes the summary of the literature selection procedure. First, the database search yielded 145 records. After removing duplicates and empirical records, as well as those not related to study content, a total of nine review articles remained for data extraction and synthesis.

Thus umbrella review methodology combines the evidence from multiple systematic reviews to offer a thorough summary of existing research as described by Grant & Booth [12]. The key findings revealed the current status and limitations in this research domain. The data extraction process was conducted systematically using a structured form to maintain consistency among all studies. All articles were assessed using predefined codes and sample questions. Next, we synthesized and analyzed the data using both qualitative and quantitative techniques to enhance the accuracy and reliability of the review. Eventually, we combined all findings the highlight significant advances, limitations, and open challenges in the area of GenAI explainability.

Finally, a reporting bias assessment and a certainty assessment was conducted in to evaluate the overall quality and reliability of the articles that were included with adherence the general recommendations from Da'u & Salim [13], and guidelines proposed by Kitchenham & Charters [14].

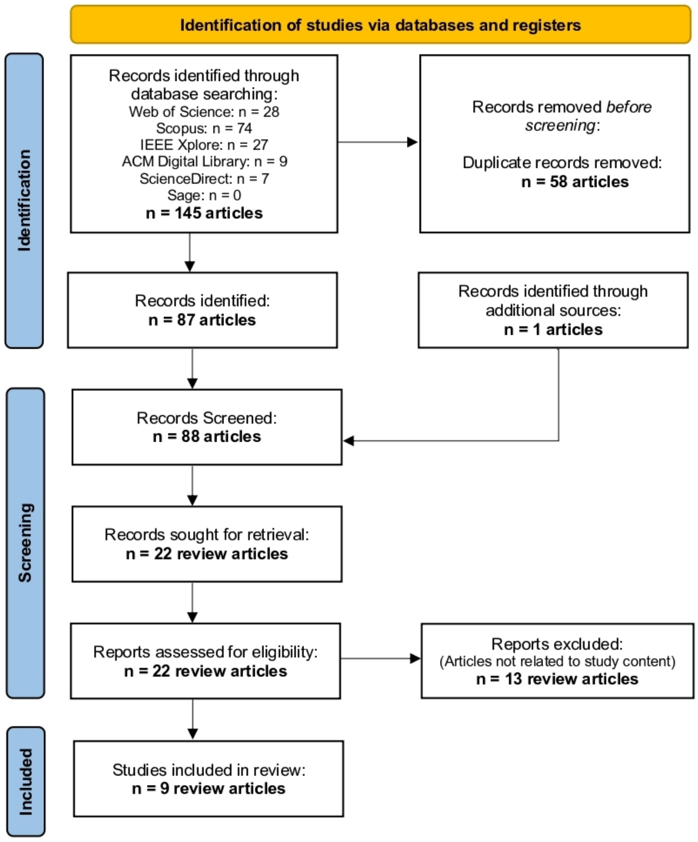

Figure 1: PRISMA flow chart of the study.

4. Results and Analysis

4.1. Temporal Growth and Cross-domain Adoption

The distribution of review articles over time reveals a sharp increase in 2024 with only one article from 2023 and eight out of nine published in 2024. The absence of reviews in previous years indicates that the explainability of GenAI systems are still in their early stages. In addition, this sharp increase in publications in 2024 shows the growing need for explainability of GenAI models which is caused because of both technological advancements and fulfilling the regularity standards.

Next, we explored the application domains covered in the selected review articles to identify where the explainability in GenAI systems have been most actively explored. The healthcare domain has focused on the largest concentration of publications (4 articles) according to our dataset [15, 16, 17, 18]. Then the technology discipline (3 articles) [19, 20, 6] and the next engineering discipline (2 articles) [21, 22] have a significant focus on this research field. Finally, media [15] , legal [15] , finance [16] , education [16], and environmental [16] fields are discovered in at least one article. It is noteworthy that some review articles have focused several domains. These findings suggest that the explainability in GenAI research are widespread in almost all the main domains. However, there are priorities given to the exploration in high-risk sectors. This distribution provides insights about the areas that require further exploration and where GenAI explainability and transparency are now most prominent.

4.2. Explainability Techniques in Reviews Articles

In order to delve deeper into this umbrella review, we concentrated on the explainability strategies that have been examined in these review articles to improve the explainability in GenAI systems.

The post-hoc explainable techniques such as SHAP, Grad-CAM, saliency maps, counterfactual explanations are the most frequently discussed in the review articles [16, 6, 22, 18]. In addition, mechanistic interpretability is another promising post-hoc methods specifically suggested for LLM models [16, 6]. This mechanistic interpretability technique seeks to understand what the model is actually doing by reverse-engineering neural networks, for example, by examining model parameters [16].

Post-hoc techniques closely followed by ante-hoc and human-in-the-loop techniques. For example, Celick & Eltawil [19] recommends developing interpretable latent space representations in VAEs to make it easier to understand how data are generated and manipulated. Furthermore, these articles [16, 6] also state some concept-based learning algorithms specifically for GenAI explainability. These algorithm's explain the predictions of the model in terms of properties and abstractions that humans can understand. On the other hand, the articles [16, 6, 18] have expressed that human-in-the-loop approach bridges the gap between automated decision-making and human interpretation. These techniques are primarily designed for traditional AI systems to provide explanations after model inference, integrating explainability directly into the model design and utilizing human feedback to refine model outputs respectively.

Subsequently, other than the main categories, there are some hybrid techniques and other less common techniques explained in these review articles [20, 21, 6]. Additionally, there was an article that discusses provenance-based techniques [15]. They discuss the usage of Coalition for Content Provenance and Authenticity (C2PA) and Content Authenticity Initiative (CAI) techniques that can use for GenAI explainability. These approaches are basically focuses on embedding meta data to track content provenance. Although these methods are the least discussed, they offer an effective pathway to increase transparency by guaranteeing the traceability of the data used in decision-making.

Overall, the wide variety of methods presented in the review articles suggests that to address the explainability in GenAI systems, the researchers still mostly use the traditional XAI methods and highlight the the lack of GenAI-specific explainability techniques in empirical studies. They indicate that existing empirical work rarely applies or evaluates explainability techniques tailored to GenAI systems. This further emphasizes the need for more targeted research in this area.

4.3. Key Findings, Limitations and Open Challenges

All the review articles together demonstrate the growing need for explainability and transparency of GenAI systems due to their increasing demand across various domains. A common concern highlighted in every article was the back-box nature of these GenAI models [15, 19, 17, 21, 16, 6, 22, 18]. To overcome this black-box issue and enable the adaption of GenAI models more into real-world Scenarios, their underlying mechanisms need to be made explainable. As these review articles outline, although several techniques and frameworks have been suggested and utilized in current literature, there is still a lack of standardized methodologies and evaluation benchmarks specifically designed to improve the explainability of GenAI systems [15, 21, 16, 6, 18]. Some studies claim that the practical applicability of many explanation strategies is limited due to GenAI's complicated nature [21, 20]. Furthermore, the evaluation of explainability methods remains a significant challenge as most of the articles rely on theoretical models rather than validation through evidence [21]. In addition to these technical issues, several studies also discuss the regularity implications of explainability and transparency, particularly on the EU AI Act [15, 19, 17, 16].

Table 1 summarizes the key findings related to limitations and open challenges in explainability in GenAI of the review articles.

Another noteworthy observation from this umbrella review is that most explanations for GenAI are based on existing XAI techniques rather than novel approaches. However, as Longo et al. [16] mentioned, those existing XAI techniques do not generalize well to GenAI models. This limitation arises because of the fundamental differences between traditional AI and GenAI systems. For example, many XAI techniques such as SHAP, LIME are fundamentally based on feature importance scores in structured prediction models. Since GenAI models' outputs are high-dimensional, such as texts, images, or videos, it is challenging to assign clear importance values to specific inputs [6]. Another issue is the lack of deterministic outputs. Explainability for traditional AI models is provided by assuming that

Table 1
An overview of the main findings regarding the limitations in explainability in GenAI

Review Article	Main Findings
Bushey (2023) [15]	Absence of explainability in AI-generated images is a serious challenge to comply with the regularity requirements. Standardized metadata frameworks are required for ensure authenticity and provenance of AI-generated content.
Celik & Eltawil (2024) [19]	Verifying the decision-making process of GenAI is among the main obstacles to adoption. A promising strategy for transparency is the use of visual representations in GenAI models.
Goktas (2024) [17]	Integrating explainability and transparency into GenAI-driven decision-making processes is critically important.
Zarghami et al. (2024) [21]	The proposed taxonomy needs to be refine for specific GenAI models. Research on hybrid XAI techniques for GenAI is limited.
Longo et al. (2024) [16]	Existing XAI methods do not generalize well with GenAI models, hence new ones need to be created.
Schneider (2024) [6]	Current explainable mechanisms have challenges in verifiability, interactivity, security and cost considerations. More interactive and user-controlled explanation mechanisms are required.
Mudabbiruddin et al. (2024) [22]	Existing XAI techniques cannot be fully adopted to GenAI. Strong explainability frameworks for GenAI are required.
Kliestik et al. (2024) [20]	Few frameworks exist for integrating XAI with GenAI.
Abdullakutty et al. (2024) [18]	Standardized evaluation procedures are required to evaluate the explainability and reliability of GenAI models.

there is a fixed mapping between input and output. But GenAI models are stochastic, which means that, for example, the same prompt in an LLM can provide different outputs. This approach complicates the use of current XAI techniques for GenAI systems. Moreover, traditional XAI challenges such as the disagreement problem, conceptual misalignment, and non-faithfulness [23] persist in GenAI settings and are often amplified due to its stochastic outputs and complex data forms. However, due to the lack of established alternative techniques, existing XAI methods are still used. The existing XAI techniques are well-documented and widely used [24]. In addition, although existing XAI techniques may not be entirely generalize and insufficient for GenAI, they can still offer useful insights into model behavior [6]. Therefore, some organizations who are using GenAI in high-stakes domains such as healthcare provide some form of explainability, even if imperfect. This underscores the significant of developing novel explainability techniques specifically for GenAI that can offer more trustworthy and meaningful explanations.

Building on these identified limitations, several open challenges and potential future research directions can be highlight.

One major continuing concern is the lack of transparency in the training data that GenAI systems are relay on. This limits technical explainability, as well as raises concerns about regularity compliance. Thus, future research should also focus on creating technical solutions for data traceability and regulatory frameworks that require GenAI training data to meet basic transparency requirements.

The absence of universally accepted standardized evaluation metrics for explainability in GenAI is another serious concern [15, 21, 18]. It is difficult to compare the utilized explainability techniques as many of the research studies use different benchmarks. Also, this inconsistency in what makes a "good explanation" limits the advancements in the field. Therefore, there is a growing need for standardized benchmarks and protocols to assess the explainability techniques used for GenAI. One possible solution might be defining universal metrics such as usability, faithfulness, and completeness for evaluating explanations. Another trustworthy and considerable solution is to promote the use of human-centered evaluation frameworks, in which the end-users test explanations to make sure they are practically interpretable. Other significant challenge with explainability for GenAI is trade-off between the model performance and explainability. One promising solution could be hybrid approaches that integrate explainability directly into model training. For example, multi-objective optimization techniques can be utilized to balance the accuracy and interpretability of the models [25].

5. Conclusion

This research study evaluated the continued interest in explainability in GenAI by emphasizing the persistent challenges and evolving methodologies in the field. The analysis of selected review articles demonstrated that although there are various techniques have been used and proposed to enhance explainability, still significant limitations remain. Those limitations mainly include the absence of explainability techniques specialized for GenAI and the lack of evaluation standards. The inability to disclose the underlying training data makes it even more difficult to produce meaningful explanations. The findings also highlight the explainability and model performance trade-offs. This indicates the necessity for balanced solutions that maintain accuracy while improving explainability. In addition, this review also identified a significant gap in real-world applicability which suggests that future research should concentrate on creating explainability strategies that are both theoretically sound and practically feasible. Although this research study provides valuable insights, it should be noted that it has some limitations. The umbrella review was limited to the published literature. Therefore, it is possible that relevant insights from industry applications have not been captured. The findings of this research study have practical implications for researchers, AI professionals, and policymakers by providing insights into the current state of explainability in GenAI and its future direction. Explainability in GenAI remains a complex but crucial challenge that needs to be addressed for building trustworthy GenAI solutions.

Acknowledgments

This work was supported by the Academy of Finland (project no. 356314).

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] S. Feuerriegel, J. Hartmann, C. Janiesch, P. Zschech, Generative ai, Business and Information Systems Engineering 66 (2023) 111–126.
[2] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018, Institute of Electrical and Electronics Engineers Inc., 2018, pp. 80–89.
[3] W. Samek, T. Wiegand, K.-R. Müller, Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services 1 (2017) 1–10.
[4] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (xai), IEEE Access 6 (2018) 52138–52160.
[5] H. Amini, M. J. Mia, Y. Saadati, A. Imteaj, S. Nabavirazavi, U. Thakker, M. Z. Hossain, A. A. Fime, S. S. Iyengar, Distributed llms and multimodal large language models: A survey on advances, challenges, and future directions, 2025. arXiv:2503.16585.

[6] J. Schneider, Explainable generative ai (genxai): A survey, conceptualization, and research agenda, Artificial Intelligence Review 57 (2024).
[7] S. Jeong, M. Li, M. Berger, S. Liu, Concept lens: Visually analyzing the consistency of semantic manipulation in gans, in: 2023 IEEE Visualization and Visual Analytics, IEEE, 2023, p. 221–225.
[8] Y. Wang, S. Shen, B. Y. Lim, Reprompt: Automatic prompt editing to refine ai-generative art towards precise expressions, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 2023, p. 1–29.
[9] P. Kumar, Explainable generative ai (xgenai): Enhancing transparency and trust in ai systems, 2024. Accessed: 2025-03-24.
[10] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, D. ... Moher, The prisma 2020 statement: An updated guideline for reporting systematic reviews, International Journal of Surgery 88 (2021) 71.
[11] M. Saarela, T. Kärkkäinen, Can we automate expert-based journal rankings? analysis of the finnish publication indicator, Journal of informetrics 14 (2020) 101008.
[12] M. J. Grant, A. Booth, A typology of reviews: an analysis of 14 review types and associated methodologies, Health Information & Libraries Journal 26 (2009) 91–108.
[13] A. Da'u, N. Salim, Recommendation system based on deep learning methods: a systematic review and new directions, Artificial Intelligence Review 53 (2020) 2709–2748.
[14] B. Kitchenham, S. Charters, Guidelines for Performing Systematic Literature Reviews in Software Engineering, EBSE Technical Report EBSE-2007-01, Software Engineering Group, School of Computer Science and Mathematics, Keele University, Keele, UK, 2007.
[15] J. Bushey, Ai-generated images as an emergent record format, in: 2023 IEEE International Conference on Big Data (BigData), 2023, pp. 2020–2031.
[16] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. D. Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, S. Stumpf, Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions, Information Fusion 106 (2024) 102301.
[17] P. Goktas, Ethics, transparency, and explainability in generative ai decision-making systems: a comprehensive bibliometric study, Journal of Decision Systems 0 (2024) 1–29.
[18] F. Abdullakutty, Y. Akbari, S. Al-Maadeed, A. Bouridane, I. M. Talaat, R. Hamoudi, Histopathology in focus: A review on explainable multi-modal approaches for breast cancer diagnosis, Frontiers in Medicine 11 (2024).
[19] A. Celik, A. M. Eltawil, At the dawn of generative ai era: A tutorial-cum-survey on new frontiers in 6g wireless intelligence, IEEE Open Journal of the Communications Society 5 (2024) 2433–2489.
[20] T. Kliestik, P. Kral, M. Bugaj, P. Durana, Generative artificial intelligence of things systems, multisensory immersive extended reality technologies, and algorithmic big data simulation and modelling tools in digital twin industrial metaverse, Equilibrium Quarterly Journal of Economics and Economic Policy 19 (2024) 429–461.
[21] S. Zarghami, H. Kouchaki, L. Yang, P. M. Rodriguez, Explainable artificial intelligence in generative design for construction, in: Proceedings of the 2024 European Conference on Computing in Construction, volume 5 of *Computing in Construction*, European Council on Computing in Construction, Chania, Greece, 2024.
[22] M. Mudabbiruddin, A. Mosavi, F. Imre, From deep learning to chatgpt for materials design, in: 2024 IEEE 11th International Conference on Computational Cybernetics and Cyber-Medical Systems (ICCC), 2024, pp. 1–8.
[23] B. Barr, N. Fatsi, L. Hancox-Li, P. Richter, D. Proano, The disagreement problem in faithfulness metrics, in: XAI in Action: Past, Present, and Future Applications, 2023.
[24] M. Saarela, V. Podgorelec, Recent applications of explainable ai (xai): A systematic literature review, Applied Sciences 14 (2024).
[25] W. R. Monteiro, G. Reynoso-Meza, A review of the convergence between explainable artificial intelligence and multi-objective optimization, 2022. Preprint.