Grenze International Journal of Engineering and Technology, June Issue

Advancing Frontiers in Explainable Artificial Intelligence (XAI): Bridging Theory and Practice

Vinay Yandrapalli[1] and Shubham Sharma[2]
[1]World Wide Technology
Email: vyandrapalli@gmail.com
[2]National Institute of Technology, Raipur
Email: shubham.willing@gmail.com

***Abstract*— In an era marked by digital transformation, this paper explores the intricate landscape of data governance, focusing on its intersections with artificial intelligence (AI), cybersecurity, ethics, sustainability, and globalization. An extensive literature review reveals the challenges and strategies within each domain. Our research highlights the regulatory implications of the General Data Protection Regulation (GDPR) on AI governance and emphasizes the need for balanced approaches. We also examine the transformative role of AI in data management while addressing associated risks. The study underscores the critical synergy between cybersecurity and data governance and delves into the influence of regulatory frameworks. Furthermore, we advocate for ethical data governance frameworks that foster transparency, accountability, and environmental well-being. Lastly, we propose culturally sensitive global strategies prioritizing human rights and ethical considerations. Our contributions offer insights into GDPR's impact on AI governance, AI's potential in data management, ethical integration, and adaptable global strategies, facilitating more informed, responsible, and effective data governance practices in our increasingly digital and interconnected world.**

***Index Terms*— explainable artificial intelligence (xai), interpretability, model transparency, ethical ai, human-centric ai.**

I. INTRODUCTION

AI has emerged as a revolutionary force in the digital transformation era, driving innovations across industries and reshaping how we live, work, and interact. From healthcare and finance to transportation and communication, AI systems enhance efficiency and productivity and unlock new possibilities. However, as these systems increasingly influence critical aspects of society, the need for transparency, understanding, and trust in AI has never been more pronounced [1]. The growing complexity of AI models, particularly those based on deep learning, has led to superior performance in various tasks at the expense of increased opacity. This opacity, often referred to as the "black box" problem, presents significant challenges: it obscures the decision-making process, making it challenging for users to comprehend, trust, and effectively manage these systems [2]. The crucial requirement for explainability in AI systems is not merely a technical necessity but a societal one, ensuring that AI decisions are just, accountable, and aligned with human values.

XAI has emerged as a critical field, addressing the essential need for transparency and interpretability in

Grenze ID: 01.GIJET.10.2.212_1

increasingly prevalent AI systems. Despite its significance and rapid progress, the field faces persistent challenges [3]. A primary concern is the dilemma of complexity versus explainability. As AI models become more complex to improve accuracy, their explainability often diminishes, making the balance between performance and interpretability a central issue in XAI. The field also has insufficient standardisation, unified methodologies, and benchmarks for creating and evaluating explanations. This makes it harder to compare and understand how well different XAI techniques work. Another significant challenge lies in user diversity; AI system users have diverse backgrounds, expertise, and expectations, necessitating the development of universally meaningful and beneficial explanations.

Moreover, ethical and regulatory considerations are paramount as AI systems integrate into critical domains, necessitating systems to be explainable and comply with strict ethical standards and regulatory requirements [4]. Finally, a critical gap exists between theoretical XAI models and their real-world application; bridging this gap is crucial and involves addressing operational, contextual, and industry-specific challenges to implement XAI effectively in practical scenarios. These challenges highlight the complexity of developing AI systems that are not only advanced and efficient but also transparent, understandable, and ethical, emphasising the ongoing necessity for innovation and research in XAI.

The necessity for advancements in XAI is further illustrated in Figure 1, which contrasts the traditional AI decision-making process with the enhanced clarity provided by XAI frameworks. Figure 2 expands on this by encapsulating the multifarious purposes of XAI, ranging from justification to improvement, and elucidating why these systems must be efficient but transparent and comprehensible.

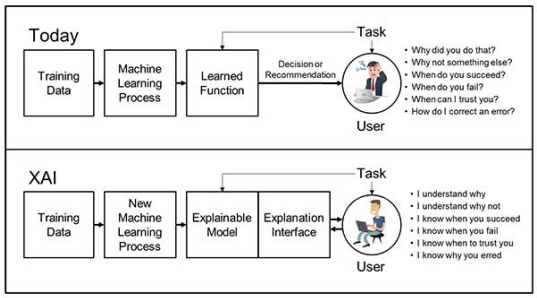

Figure 1. Comparative Workflow of Traditional AI and XAI Systems [5].

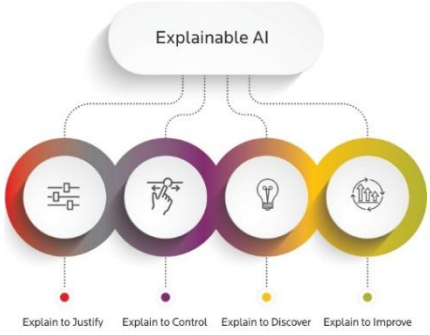

Figure 2. The Four Pillars of Explainable AI Objectives [6].

In light of these challenges and the critical importance of XAI, this paper aims to contribute to the field by achieving the following objectives:

- *Comprehensive Overview*: Provide a comprehensive overview of the concepts, methodologies, and challenges in XAI, synthesising insights from seminal works and recent advancements.
- *Gap Analysis*: Identify and discuss the gaps in current XAI literature, particularly in methodology integration, user-centric design, and empirical validation.
- *Innovative Proposals*: Propose innovative frameworks and models to address the identified gaps, emphasising the need for integrated, dynamic, and user-adaptive explanations.
- *Impact Assessment*: Analyse the impact of XAI in various domains, presenting case studies and applications that demonstrate the real-world effectiveness of explainable AI systems.

- *Future Directions*: Outline future research directions and call for a multi-disciplinary approach to advance the field of XAI, ensuring that AI systems are transparent, understandable, and aligned with ethical standards and human values.

By addressing these objectives, the paper seeks to advance the understanding of XAI, foster the development of more effective and user-friendly explainable systems, and contribute to the responsible and ethical use of AI technologies. The ultimate goal is to ensure that AI systems are robust and efficient but also transparent, understandable, and beneficial for all.

II. Background And Related Work

In the domain of AI, the construct of explainability is central to the capacity of AI systems to articulate the underpinnings of their decisions in a manner that is intelligible to human interlocutors. This facet of AI is concerned with transparently rendering the intricate workings of sophisticated models, thus cultivating the end user's confidence in the technology. While frequently overlapping with interpretability and transparency, explainability clarifies the logic that informs AI decision-making processes [7]. Interpretability is frequently perceived as synonymous with explainability; however, it pertains more precisely to the degree to which the rationale behind an AI's decision-making is accessible. The interpretability aims to decode the technical nuances of a model's decision-making for a diverse user base spanning varying levels of technical acumen [8].

Transparency within AI focuses on unambiguously presenting an AI system's operations and data processing to its stakeholders. This attribute necessitates providing comprehensive information about the system's functional scope, constraints, and underlying logic, facilitating an immediate and clear understanding of how inputs are processed into outputs. The pivotal role of transparency in nurturing trust and advocating for the principled use of AI technologies [9].

Trustworthiness in AI quantifies the confidence that users place in an AI system's ability to perform as expected in an ethical, equitable manner consistent with societal norms and values. The European Commission's High-Level Expert Group on AI (2019) delineates trustworthy AI as adhering to regulatory frameworks and ethical standards, with trustworthiness being emblematic of an AI system's reliability, safety, equity, and accountability, as well as its resilience to external intrusions and operational discrepancies [10].

Although distinct, explainability, interpretability, transparency, and trustworthiness are interwoven and constitute the bedrock of ethically responsible AI system development and deployment. Collectively, they advocate for technologically adept AI while also being ethically oriented and socially advantageous.

The evolution of XAI is inextricably linked with the history of AI. In their early stages, AI systems were simple and naturally interpretable, such as rule-based expert systems where human experts explicitly articulated the reasoning process, creating a transparent decision-making process. Previously, The author acknowledged that even these preliminary systems recognised the imperative for explanations to authenticate and foster trust in AI recommendations [11].

Nevertheless, the rise of machine learning, particularly deep learning, heralded an era where AI began to address more complex tasks with heightened accuracy, giving rise to the 'black box' conundrum—wherein the decision-making process became obscured. The inception of an intensified pursuit for explainability is propelled by the pressing need to comprehend, trust, and manage these potent yet inscrutable systems [12].

The proliferation of deep learning exacerbated the black box issue, as deep neural networks' intricate, non-linear architectures eluded penetrability. This necessitated an imperative for explainability, mainly as AI assumed critical roles in key sectors such as healthcare, finance, and the judiciary. These complex models' interpretability instigated a surge of research into XAI [13].

Additionally, introducing regulatory frameworks, notably the GDPR in Europe, which embodies the 'right to explanation', has significantly propelled research in XAI. An author underscores how such legislation has mandated systems that deliver accurate predictions and elucidate their decision-making processes [14].

At the moment, XAI is an active area of research that is coming up with ways to explain complicated models after the fact, making models that are naturally easy to understand, and creating frameworks for systematic evaluation. According to an author, the field is moving closer to defining benchmarks for explications, incorporating user feedback, and looking at the psychological aspects of explanation reception [15].

As AI becomes increasingly integrated into the fabric of daily life, the prominence of XAI is set to escalate commensurately. The emphasis remains on ensuring fairness, accountability, and nurturing trust and acceptance. XAI's maturation mirrors AI's development, underscoring a field committed to resolving complex quandaries with transparency and comprehensibility and following humanistic principles.

Table 1 encapsulates the foundational principles of XAI, charting the progression and critical contributions within the field. It succinctly defines each principle, catalogues seminal authors, and encapsulates the primary contributions shaping the domain.

TABLE I: OVERVIEW OF CORE CONCEPTS, HISTORICAL EVOLUTION, AND KEY CONTRIBUTIONS IN XAI.

Concept	**Description**	**Key Contributors**	**Contributions/Findings**
Explainability	The ability of AI systems to provide understandable and justifiable reasons for their decisions.	[7]	Proposed a taxonomy for classifying XAI approaches and highlighted the importance of user-centric, ethical, and transparent AI systems.
Interpretability	The degree to which a human can understand the cause of a decision.	[8]	Stressed the need to make the model itself understandable to users.
Transparency	The clarity of AI system mechanisms and data processing.	[9]	Emphasised the importance of transparency for building user trust and facilitating ethical AI use.
Trustworthiness	The confidence in AI systems to perform ethically, reasonably, and in alignment with human values.	[10]	Defined trustworthy AI and outlined the facets of reliability, safety, fairness, and accountability.
Early AI Systems	Simple and inherently interpretable systems with transparent reasoning.	[11]	Recognised the need for explanations in expert systems to validate and trust the system's recommendations.
Black Box Problem	Opaque decision-making processes in complex ML models.	[12], [13]	Highlighted the onset of the quest for explainability due to the opaque nature of ML systems.
Deep Learning and XAI	The challenge of interpretability in multi-layered, nonlinear neural networks.	[13]	Questioned the interpretability of complex models and spurred research into XAI.
Regulatory Push for XAI	The impact of GDPR and the right to explanation.	[14]	Showed how regulations spurred the need for systems that explain their decision-making process.
Current Trends in XAI	Development of post-hoc explanations, interpretable models, and evaluation frameworks.	[15]	Discussed the movement towards standards, user feedback integration, and psychological aspects of explanations.

In the vibrant field of XAI, many methods have been developed to enhance the interpretability and trustworthiness of AI systems. These methods are classified into four principal categories, as illustrated in Table 2. Model-agnostic methods, such as LIME and SHAP, are prized for their flexibility and broad applicability across different AI architectures. Model-specific methods, including saliency maps and CAM, are meticulously designed to align with particular model structures, providing detailed explanations. Post-hoc methods are crucial for complex models that are not inherently interpretable, offering insights after the model's training phase. In contrast, ante-hoc methods focus on creating transparent models from inception, thus obviating the need for post-hoc explanations. This categorisation elucidates the spectrum of XAI techniques, guiding the selection of the most appropriate method based on the model's complexity, the need for explainability, and the intended use case.

TABLE II: CATEGORISATION OF EXPLAINABILITY METHODS IN XAI

Category	Approach	Examples	Key References
Model-Agnostic Methods	Broadly applicable techniques that are independent of the model architecture.	LIME, SHAP	[16]
Model-Specific Methods	Tailored to specific model types, leveraging their internal structures for generating explanations.	Saliency maps, CAM	[17]
Post-Hoc Methods	Applied after-model training to explain decisions, suitable for complex models.	Feature importance, Partial dependence plots	[7]
Ante-Hoc Methods	Building inherently interpretable models from the ground up, prioritising transparency.	Linear models, Decision trees	[18]

III. Challenges In XAI

A Technical Challenges

The endeavour to endow artificial intelligence models with explainability necessitates surmounting technical obstacles attributable to these systems' intricate and heterogeneous nature. These obstacles must be overcome to ensure the practicality and integration of XAI protocols.

- *Complexity of Models:* The increased sophistication of deep learning architectures, with their multi-layered and non-linear configurations, complicates the extraction of intelligible and precise explications. These models' elevated dimensionality and abstractions pose formidable barriers to rendering their operational mechanisms transparent to human stakeholders.
- *Data Integrity and Representation*: The calibre and structure of datasets employed in training AI significantly dictate the viability of furnishing cogent and meaningful elucidations. Data deficiencies, including noise, incompleteness, or intrinsic biases, may culminate in deceptive explanations. Furthermore, portraying data within high-dimensional models exacerbates the difficulty of formulating explications that resonate with user comprehension.
- *Standardisation of Benchmarks:* The absence of universally accepted benchmarks and evaluative norms for gauging explanation quality hinders the comparison of disparate XAI methodologies and obscures their efficacy and constraints. Establishing consensus on such benchmarks is imperative for the progression of the field and the utility and dependability of explanations.
- *Scalability and Computational Efficacy:* The burgeoning size of datasets and the escalating complexity of models demand XAI methods that are both scalable and computationally efficient. The computational expenditure required to generate explanations, especially pertinent to expansive models and voluminous data, may be exorbitant. Refining these methods to optimise performance without compromising explanation integrity is essential.

B. Balancing Act

Navigating the dichotomy between model intricacy and interpretability constitutes a principal challenge within XAI. As models attain heightened accuracy and complexity, their interpretability frequently diminishes, precipitating a conundrum for those in the field.

- *Interpretability versus Performance*: Models characterised by high complexity, such as deep neural networks, are often the benchmark for performance across diverse tasks. However, their elaborate compositions impede interpretability. Conversely, simpler constructs, such as decision trees or linear regressions, afford greater interpretability but may fail to discern more nuanced patterns within data. Balancing interpretability with performance is a critical challenge [8], [19].
- *Intrinsic Interpretability in Design*: One strategy is to design inherently interpretable models to address this equilibrium. This entails the development of models that not only perform effectively but are also intrinsically transparent and comprehensible. Research is directed towards innovative architectures and learning algorithms that harmonise complexity with interpretability [20].
- *Post-hoc Explanatory Techniques*: Another tactic to reconcile this balance involves post-hoc explanatory methods that aim to shed light on complex models after their training. While such methods facilitate the extraction of explanations from pre-existing models, guaranteeing that these explanations authentically represent the model's reasoning process presents a challenge.

C. User-Centric Challenges

The efficacy of XAI extends beyond technical prowess, encompassing the degree to which explanatory outputs resonate with end-user requirements and anticipations. XAI's successful deployment and influence hinge on the aptitude to surmount user-centric impediments.

- *Heterogeneity of User Demographics*: The landscape of AI system users is characterised by a spectrum of expertise, from novices to domain experts, each with distinct objectives and cognitive proficiencies. Crafting bespoke explanations that cater to this gamut of user-profiles necessitates a granular comprehension of their discrete requirements, cognitive processing capacities, and the environments within which decisions are made.
- *Domain-Contingent Explanations*: The pertinence and utility of explanatory content are intrinsically linked to the application domain. An explanation deemed cogent within one sphere may hold no currency in another.

The endeavour to forge explanations adaptable to disparate scenarios and tailored to the contextual exigencies of specific domains presents a formidable challenge.

- *Dichotomy of Trust and Interpretation Accuracy*: The imperative to cultivate trust through explanations while mitigating potential user misinterpretations is pivotal. Explanatory mechanisms that are either excessively arcane or simplistic can precipitate misconceptions and subsequent misapplications of AI systems. Concurrently, simplifications intended to enhance user trust must not detract from the veracity and representational fidelity of the model's logic. Striking an equilibrium between engendering trust and maintaining interpretational verisimilitude and simplicity is a nuanced and critical endeavour.

In the quest to navigate these challenges—technical, balance-focused, and user-centric—the field of XAI aspires to evolve towards models epitomising computational efficacy and a paradigm of transparency, intelligibility, and congruence with human ethical standards. This trajectory mandates not merely the refinement of technical methodologies but an integrative approach that embraces ethical, societal, and pragmatic considerations pertinent to the explicability of AI systems.

IV. Advanced Methodologies In XAI

A. Contemporary Techniques in XAI

XAI methodologies are extensive and diverse, offering many techniques to elucidate AI model decisions. This section critically evaluates prevailing methodologies, incorporating insights from seminal studies and cutting-edge research.

- *Local Interpretable Model-Agnostic Explanations (LIME):* LIME is a widely acclaimed technique that elucidates classifiers' predictions in an interpretable format. It perturbs inputs and observes the resultant variations in outputs to generate locally pertinent explanations, rendering it invaluable for deciphering complex models. Despite its intuitive nature, LIME's locality-specific explanations might not encapsulate the model's overall behaviour [21].
- *SHapley Additive exPlanations (SHAP):* SHAP employs Shapley values from cooperative game theory to ascertain the contribution of each feature to predictions. Offering a unified and coherent measure of feature significance, SHAP stands out for its theoretical rigour. Nevertheless, its computational demands can be intensive, especially for complex models with numerous features [22].
- *Counterfactual Explanations:* Counterfactual explanations provide insights by delineating minimal changes needed to alter model predictions. While these explanations offer actionable insights, crafting pertinent and plausible counterfactuals presents considerable challenges, particularly in high-dimensional spac [23].
- *Feature Importance and Partial Dependence Plots*: These techniques aim to quantify the significance of features and depict the impact of feature variations on predictions. While they offer a macroscopic understanding of feature impacts, they may not adequately capture complex interactions or the nuanced behaviour of models.

B. Emergent Frameworks and Models

In response to existing challenges and identified gaps, the XAI community is incessantly innovating and proposing new frameworks and models to enhance explanations' fidelity, usability, and overall efficacy.

- *Integrated XAI Frameworks*: A promising direction involves formulating integrated frameworks that amalgamate various explanatory methods, providing a comprehensive understanding of model behaviour. These adaptive frameworks are tailored to the model, user, and situational context, delivering bespoke and extensive insights.
- *Interactive and User-Adaptive Explanations*: Innovations are also occurring in developing interactive and user-adaptive explanations. These dynamic systems evolve with user interactions and changing scenarios, engaging users in a dialogic process to refine and deepen explanations based on user queries and feedback.
- *Causal Inference Models in XAI*: A critical and emerging area of research focuses on developing models that furnish causal explanations for AI decisions. These models transcend correlational insights, enabling users to grasp the underlying causal mechanisms that drive predictions, thus providing more profound and more actionable insights.

C. Robust Evaluation Metrics

The evaluation of XAI methodologies is multi-dimensional, encompassing fidelity, comprehensibility, and usability. Developing robust metrics and continuously refining them are paramount to the progression of XAI.

- *Fidelity*: Fidelity pertains to the accuracy and congruence of explanations with the model's actual workings. Metrics might include comparing the predicted outcomes from explanations with those from the model across varied scenarios.
- *Comprehensibility*: This dimension assesses how users understand and process the provided explanations. It can be evaluated through user studies focusing on tasks guided by explanations or surveys gauging perceived clarity and satisfaction.
- *Usability:* Usability examines the practical impact of explanations on user decision-making, trust, and reliance on AI systems. This involves empirical evaluations of how explanations influence user interactions and outcomes with AI systems.
- *Quantitative and Qualitative Assessments*: Both quantitative metrics, such as accuracy, precision, and recall, and qualitative assessments, including user feedback and expert evaluations, are integral to a holistic evaluation of XAI methodologies.

By integrating and refining these metrics, XAI aims to attain technical sophistication, practical value, and accessibility, ensuring that the methodologies developed significantly contribute to the field's advancement and real-world applicability.

V. Advanced Deployments And Case Studies In XAI

The domain of XAI transcends theoretical discourse, finding increasingly prevalent applications across various pivotal sectors. Its implementation showcases XAI's potential in augmenting transparency, fostering trust, and refining decision-making processes across heterogeneous industries. This section delves into XAI's application in essential sectors such as healthcare, finance, and law, scrutinises its impact, and presents pertinent case studies to demonstrate its real-world efficacy.

A. Domain-Specific Deployments

- *Healthcare*: In the medical sphere, XAI is critical in clarifying complex models employed for diagnostics, treatment recommendations, and patient prognosis. It elucidates the underlying factors of AI-driven diagnostics, thereby aiding medical practitioners in comprehending and relying on medical predictions or treatment advisories. This leads to more nuanced clinical decision-making. In predictive health analytics, XAI dissects the variables impacting prognostications, thus facilitating timely and focused healthcare interventions.
- *Finance*: Within the financial industry, XAI finds applications in credit scoring, fraud detection, and algorithmic trading. It makes the factors affecting credit scores transparent, ensuring fairness and clarity in credit assessments. In fraud detection, XAI elucidates the recognition patterns employed by the system, thus enhancing the refinement of these mechanisms and diminishing the incidence of false positives.
- *Law*: XAI assists in legal analysis, contract scrutiny, and predicting judicial outcomes. By explicating the predictors in legal decision-making models, XAI equips legal professionals with a deeper understanding of the AI's reasoning, thereby informing more strategic legal approaches.

B. Impact Analysis

The influence of XAI on decision-making and trust is extensive and varied:

- *Trust and Confidence*: XAI's drive towards greater transparency and comprehensibility enhances trust among users. This is particularly vital in sectors where decisions have profound implications on individual lives, such as healthcare and finance.
- *Regulatory Compliance*: In heavily regulated industries, XAI facilitates adherence to legal and ethical standards by providing the requisite transparency and documentation. This is instrumental in demonstrating compliance with various regulatory frameworks.
- *Empowered Decision-Making*: XAI empowers users to make more informed and nuanced decisions by demystifying the rationale behind AI's recommendations. This is crucial across all sectors where AI's advice must be interpreted and applied judiciously.

C. Illustrative Case Studies

- *Healthcare* - Predictive Analytics for Patient Care: A healthcare institution employs a machine learning model to ascertain patient readmission risks. With XAI integration, the model elucidates the underpinning factors

like medical history or vital signs affecting risk assessments, enabling healthcare professionals to interpret and trust the prognostications, leading to tailored patient management strategies.

- *Finance* - Enhanced Fraud Detection: A financial entity utilises an AI system to detect fraudulent transactions. XAI furnishes insights into the rationale behind flagged transactions, such as anomalous transaction sizes or frequencies. This enables analysts to swiftly validate or dismiss suspicions, augmenting the accuracy and efficiency of fraud detection.
- *Law* - Contract Analysis: A legal firm leverages AI for contract analysis. XAI explicates the AI's analytical logic, including identifying critical clauses, potential issues, and legal justifications. This aids lawyers in rapidly assimilating the focal points and implications, thereby enhancing legal advisory quality.

These case studies underscore XAI's transformative influence across various sectors, substantiating its role in enhancing user interactions with AI and ensuring AI's integration is ethically sound and efficacious.

In summation, the real-world applications of XAI underscore its critical function in bolstering decision-making, nurturing trust, and maintaining transparency across a spectrum of domains. As XAI evolves, its incorporation into broader societal operations is anticipated to escalate, highlighting the ongoing imperative for rigorous research, innovative development, and judicious application of these technologies.

VI. Conclusion

In this rigorous examination of XAI, we have meticulously traversed the field's multifaceted terrain, shedding light on its foundational aspects, diverse methodologies, inherent challenges, and practical applications. We initiated our discourse by delineating the core principles of explainability, interpretability, transparency, and trustworthiness, thus laying a comprehensive groundwork for deeper exploration. Progressing through the evolutionary arc of XAI, we identified and addressed pivotal literature gaps, mainly focusing on explanation integration, empirical validation, and the imperative for user-centric designs. Our analysis dissected and differentiated between various XAI techniques, advocating for novel, integrated, and causally-informed frameworks while underscoring the importance of robust evaluation metrics to ascertain XAI endeavours' fidelity, comprehensibility, and practicality.

Through illustrative case studies in critical sectors like AI, finance, and law, we unveiled the transformative capacity of XAI, demonstrating its significant influence on decision-making, trust enhancement, and operational efficacy. As we look to the future, we must continuously advance this field, embracing interdisciplinary collaboration, empirical rigour, ethical commitment, and a relentless pursuit of innovation and adaptability. The journey towards crafting AI systems that are not only technologically advanced but also ethically sound, transparent, and aligned with human values is ongoing. It is a collective quest that demands our dedication, creativity, and an unyielding commitment to ethical tenets. The future of AI, as we envisage, is not merely about sophistication and power but about ensuring these systems are comprehensible, ethical, and intrinsically beneficial to humanity.

References

[1] S. Makridakis, “The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms,” Futures, vol. 90, pp. 46–60, 2017.

[2] C. Zednik, “Solving the black box problem: A normative framework for explainable artificial intelligence,” Philos Technol, vol. 34, no. 2, pp. 265–288, 2021.

[3] R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold, “A historical perspective of explainable Artificial Intelligence,” Wiley Interdiscip Rev Data Min Knowl Discov, vol. 11, no. 1, p. e1391, 2021.

[4] C. Cath, “Governing artificial intelligence: ethical, legal and technical opportunities and challenges,” Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 376, no. 2133, p. 20180080, 2018.

[5] D. Gunning, E. Vorm, Y. Wang, and M. Turek, “DARPA’s explainable AI (XAI) program: A retrospective,” Authorea Preprints, 2021.

[6] R. ’Verma and P. ’Ainapur, “Demystifying Explainable Artificial Intelligence: Benefits, Use Cases, and Models,” https://www.birlasoft.com/.

[7] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A Survey of Methods for Explaining Black Box Models,” ACM Comput. Surv., vol. 51, no. 5, Aug. 2018, doi: 10.1145/3236009.

[8] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” arXiv preprint arXiv:1702.08608, 2017.

[9] Z. C. Lipton, “The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery.,” Queue, vol. 16, no. 3, pp. 31–57, Jun. 2018, doi: 10.1145/3236386.3241340.

[10] N. A. Smuha, “The EU approach to ethics guidelines for trustworthy artificial intelligence,” Computer Law Review International, vol. 20, no. 4, pp. 97–106, 2019.
[11] W. R. Swartout, “Rule-based expert systems: The mycin experiments of the stanford heuristic programming project: B.G. Buchanan and E.H. Shortliffe, (Addison-Wesley, Reading, MA, 1984); 702 pages, 40.50,” Artif Intell, vol. 26, no. 3, pp. 364–366, 1985, doi: https://doi.org/10.1016/0004-3702(85)90067-0.
[12] J. Burrell, “How the machine ‘thinks’: Understanding opacity in machine learning algorithms,” Big Data Soc, vol. 3, no. 1, p. 2053951715622512, 2016.
[13] Z. C. Lipton, “The Mythos of Model Interpretability,” CoRR, vol. abs/1606.03490, 2016, [Online]. Available: http://arxiv.org/abs/1606.03490
[14] B. Goodman and S. Flaxman, “European Union regulations on algorithmic decision-making and a ‘right to explanation,’” AI Mag, vol. 38, no. 3, pp. 50–57, 2017.
[15] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” Artif Intell, vol. 267, pp. 1–38, 2019, doi: https://doi.org/10.1016/j.artint.2018.07.007.
[16] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in KDD ’16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.
[17] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning Deep Features for Discriminative Localization,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921–2929. doi: 10.1109/CVPR.2016.319.
[18] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” Nat Mach Intell, vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.
[19] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. A. Specter, and L. Kagal, “Explaining Explanations: An Overview of Interpretability of Machine Learning,” 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 80–89, 2018, [Online]. Available: https://api.semanticscholar.org/CorpusID:59600034
[20] E. Rader, K. Cotter, and J. Cho, “Explanations as Mechanisms for Supporting Algorithmic Transparency,” in Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, in CHI ’18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 1–13. doi: 10.1145/3173574.3173677.
[21] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘ Why should i trust you?’ Explaining the predictions of any classifier,” in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
[22] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” Adv Neural Inf Process Syst, vol. 30, 2017.
[23] R. Guidotti, “Counterfactual explanations and how to find them: literature review and benchmarking,” Data Min Knowl Discov, pp. 1–55, 2022.