2025 International Conference on Advanced Machine Learning and Data Science (AMLDS) | 979-8-3315-1099-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/AMLDS63918.2025.11159340

A Theoretical Probabilistic Framework for Explaining Generative AI

Shahina Begum*, Shaibal Barua, Mobyen Uddin Ahmed, and Mir Riyanul Islam
School of Innovation, Design and Engineering, Mälardalen University, Västerås, Sweden
*Corresponding Author, Email: shahina.begum@mdu.se, Phone: +46 21-10 73 70

***Abstract*—This study uses Generative Artificial Intelligence (gAI) to advance industrial digitization. Although the use of gAI looks promising for industrial digitization, there are significant gaps in current Explainable Artificial Intelligence (XAI) methods, which limit their applicability to such applications. By developing a theoretical framework, the aim is to provide explanations for gAI to improve decision-making processes with actionable insights and explanations for their intended outcomes. The proposed work has an impact on facilitating inspection, monitoring, optimization, and maintenance of industrial equipment and machinery. The theoretical framework proposed in this paper will address this challenge by following a three-step approach: 1) learning prior and posterior from data, 2) feature attribution and counterfactual explanation-based methods, and 3) integrated XAI. While the current study is theoretical, future work will focus on applying the approach to real-world industrial scenarios.**

***Index Terms*—Generative Artificial Intelligence, gAI, Explainable Artificial Intelligence, XAI, Theoretical Framework, Probabilistic Approach.**

I. Introduction

In recent years, Generative Artificial Intelligence (gAI) has been one of the most promising advancements in AI technology, which holds immense potential for revolutionising industrial digitalisation. The gAI refers to algorithms capable of creating new content, such as images, text, or even entire virtual environments, based on patterns learned from existing data. The gAI market is projected to experience significant growth from 2023 to 2030, with nearly 45 billion U.S. dollars in 2023[1]. It is anticipated to increase by almost 20 billion U.S. dollars annually until the end of the decade. This expansion carries substantial business advantages, particularly in streamlining work processes and enhancing productivity. Early adoption of gAI could yield productivity gains of up to 0.6%, a noteworthy contribution considering the global economic scale. However, the realisation of these benefits depends on the prompt integration of this new technology into standard business practices. Currently, there are gAI platforms for businesses, *e.g.*, AWS[2], Google[3], IBM[4] and MainlyAI[5].

Nevertheless, the black-box nature of gAI models poses challenges in understanding the rationale behind prescriptive recommendations, raising concerns about safety and bias issues. This limitation hinders gAI application in industrial settings. Consequently, deploying end-to-end AI solutions tailored to specific industrial use cases becomes challenging. The purpose of this proposed theoretical framework is to improve transparency in gAI systems by incorporating the philosophical concept of Inference to Best Explanation (IBE) [1], [2] into explainable AI (XAI). In this paper, the explanation phenomenon is generated from the base model, providing good control between the model and the generated explanations. By addressing the challenges in XAI and integrating IBE into the framework of XAI, this work will identify the most compelling explanations for gAI systems, facilitating the explanation of AI's outputs to end-users, developers, and other stakeholders. Incorporating IBE into XAI can enhance transparency and accountability in AI systems, fostering greater trust in their decision-making capabilities.

There has been a significant increase in research efforts on XAI, which involves developing secondary (post-hoc) models such as approximation models, derivatives, feature importance measures, and various statistical techniques to explain the inner workings of black box models. Nevertheless, these post-hoc methods can be unreliable and inconsistent [3], [4] and often fail to provide contrastive explanations [5], [6]. Therefore, there are several major gaps in current XAI methods [3]–[7]. From a scientific perspective, IBE can help minimise the identified gaps to provide an explanation for gAI as this method belongs to the form of logical inference suggesting contrastive explanation and offers the most "understandable (loveliest)" explanation covering all the observational data (evidence) [3], [5], [6]. In this approach, the contrastive form of the explanation phenomenon, "Why P rather than Q? where P and Q are two events" [2], [5], is applied to identify the features of explanations that contribute to the degree of understanding they provide, and it is crucial for high-quality prescriptive analytics. The challenge lies in computationally

This study is part of project 1) xBest (Generative AI towards Inference to the Best Explanation), funded by the VR (Vetenskapsrådet - The Swedish Research Council), Diary No. 2024-05613; 2) Trust_Gen_Z, funded by the VINNOVA, Diary No. 2024-01402; 3) TRUSTY (Trustworthy Intelligent System For Remote Digital Tower), financed by the SESAR JU under the EU's Horizon 2022 Research and Innovation programme, Grant Agreement No. 101114838; 4) CPMXai (Cognitive Predictive Maintenance and Quality Assurance using Explainable AI and Machine Learning), funded by the VINNOVA, Diary No. 2021-03679.

[1]Statista forecasts generative AI, https://t.ly/n6vgq

[2]https://aws.amazon.com/ai/generative-ai/

[3]https://cloud.google.com/ai/generative-ai?hl=en

[4]https://www.ibm.com/thought-leadership/institute-business-value/en-us/technology/generative-ai

[5]https://www.mainly.ai/solution

integrating the IBE concept into XAI to provide explanations for gAI. The paper will, in the three steps, 1) develop a probabilistic framework for IBE, 2) implement methods for feature attribution and counterfactual explanations, and 3) create a Hybrid approach for contrastive explanations and evaluation of explanations. These three steps are based only on a theoretical approach, which is a limitation of the current study.

The rest of the paper contains: Section II presents the state of the art for the three steps; Section III discusses the methodological approach used to develop the overall framework; Section IV proposes a probabilistic framework for IBE; Section V discusses methods for feature attribution and counterfactual explanations; Section VI proposes a Hybrid approach for contrastive explanations and evaluation of explanations and Section VII contains discussion and conclusions.

II. State-of-the-art

Existing XAI methods are based on concepts such as functional understanding, explicability, interpretability, transparency, and human-centric XAI [8]–[11]. These concepts involve describing the problem domain, making AI models inspectable, explaining model decisions, ensuring algorithmic behaviour is understandable by humans, and generating explanations with human involvement. However, none of the existing algorithms have fully met all these concepts of XAI. Post-hoc model-agnostic algorithms such as LIME [12], SHAP [13], and DALEX [14] are widely used methods for explaining predictions of complex ML models. However, these methods suffer from consistency issues and face challenges with locality. Locality issues involve identifying appropriate "neighbourhood" data points for generating local surrogate datasets to approximate the AI model.

Traditionally, the most common methods for explaining the decision of deep learning are based on either Gradient-based (Grad-CAM and SmoothGrad) or Layer-Wise Relevance Propagation (LRP) methods [15], [16]. These methods have limitations, including accurate object localisation and reasoning capabilities, can be constrained by the absence of ground truth, and do not provide insight into the underlying reasoning. Again, an advanced deep network architecture called ProtoPNet is proposed, but it cannot always provide accurate explanations due to the semantic gap between latent space and input space similarity [16], [17]. INTERACTION, an XAI model for natural Language inference explanations model, is proposed in [18] that considers predefined Gaussian distribution as priors. However, in the real world, data can come from different distributions. The proposed approach will perform beyond specific data categories and specific types of distribution, which is applicable to gAI for different industrial data sets.

Considering data generation, BayLIME developed by Zhao *et al.* [3], is a Bayesian extension of LIME that tackles inconsistency and locality challenges. However, BayLIME derives priors from application-specific context and the guarantee of "good" priors is dependent on the validation and verification (V&V) tool, which may result in circular arguments regarding reliable V&V tools. Generative modelling techniques, such as Variational Autoencoder (VAE) [19] and Generative Adversarial Networks (GANs), have shown promising results for synthetic data generation and can address the locality problem of model-agnostic methods. However, their potential for XAI has not yet been fully explored for gAI.

Regarding the integration of XAI methods, Statistical Relational Learning (SRL) [20] employs probabilistic graphical models (Bayesian approach) with symbolic reasoning and logic. The proposed approach will address open issues (symbolic representation explanation, probabilistic reasoning, and scaling inference) in XAI by utilizing probabilistic graphical models as the foundation of structural causal models (*i.e.*, counterfactuals) and combining reasoning to deduce explanations. Thus, the proposed approach will advance the modern XAI techniques to more human-like responses to "why" questions and provide understandable (loveliest) explanations.

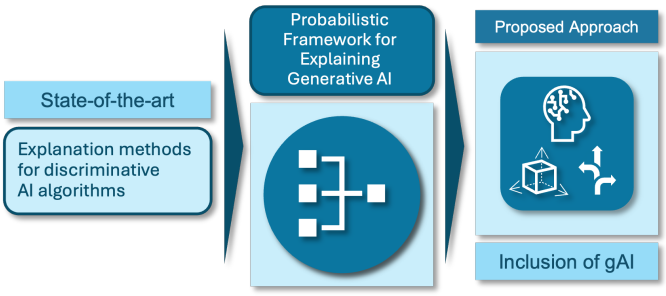


Fig. 1. Expected improvements from the state-of-the-art XAI.

Usually, by design, gAI models are polysemantic in nature, *i.e.*, several concepts are learned from the data and stored in latent space, which need to be investigated to generate explanations for the overall mechanism of the gAI models.

The proposed framework initiative is built upon our prior research experience in XAI, which includes active participation in five international projects (ARTIMATION[6], BrainSafeDrive[7], xAPP[8], Trusty[9], and MONITOR) as well as three national projects (TRUST_GEN_Z, CPMXai and DIGICOGS). As indicated in our previous analysis [10], the bulk of XAI research has centred around neural networks, with a primary emphasis on generating local explanations through post-hoc methods. Based on this observation, we proposed that additional research is needed to develop methods for generating global explanations that do not compromise the performance of the gAI model in its primary task. This requires achieving a good approximation between the base model and the XAI model, which is an essential input in the proposed approach. We anticipate that the proposed causal chain approach will outperform feature importance-based explanations in this regard. As reported in another work [21], the explanation

[6] https://www.artimation.eu
[7] https://brainsafedrive.brainsigns.com/
[8] https://www.es.mdu.se/projects/585-_xApp__Explainable_AI_for_Industrial_Applications
[9] https://research.dblue.it/trusty/

provided was effective in identifying feature importance and selection, with SHAP being the most effective method for this purpose. However, we identified a limitation of SHAP, which is its potential to under-represent the decision-making process. This misalignment with the base model may not meet the expectations of the end-user. Taking these limitations into account, our recent work [22] highlighted the importance of balancing interpretability and accuracy when assessing the quality of explanations. However, a significant challenge lies in maintaining consistency, particularly in achieving accuracy comparable to that of the base predictor while providing consistent explanations in scenarios involving repeated or randomized sampling for surrogate data. In this paper, we propose utilizing the best approximate explanation derived from the posterior distribution to provide sufficient information through a causal chain approach. As Fig. 1 shows, it will provide an improvement over the current XAI methods (which are limited to discriminative AI such as Logistic Regression, Support Vector Machines, Decision Trees *etc.*) by enabling explanations for gAI models.

III. Methodology

"Inference to Best Explanation (IBE)" [1] is a reasoning method that strives to identify the most convincing explanation for a given set of observations. Arguably, IBE is the most appropriate model of explanation suggested in the field of the philosophy of science. This method belongs to the form of logical inference, *i.e.*, given evidence as observations, we infer what would be the most likely cause or explanation for those observations, assuming their truthfulness and prior knowledge. According to Lipton [1], the most understandable explanation is often perceived as the most plausible one, *i.e.*, "loveliness" to determine the likelihood of an explanation. So, IBE can be viewed as "Inference to the Loveliest Explanation".

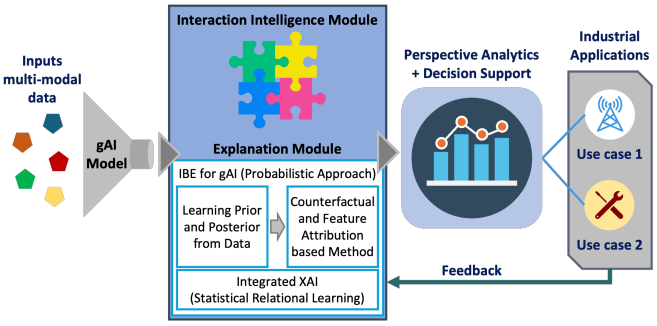


Fig. 2. Overview of the proposed Probabilistic framework to learn the approximate prior and posterior distributions from the observational data for representations of gAIs' decisions for industrial applications.

Drawing inspiration from IBE, the paper aims to develop XAI methods that provide a comprehensive understanding of gAI's decisions. Contrastive forms of explanations that include counterfactuals with the cause and effect of the decisions will be an outcome of this approach. The overall concepts of the work and how it can be used in industrial applications are presented in Fig. 2. It can support prescriptive analytics where explaining gAI outcome is important for informing future measures and decision-making. The IBE for gAI will be based on probabilistic modelling, observational data, and contextual information to provide a better representation and infer the best "understandable (loveliest)" explanations. The approach is discussed in three specific steps:

1) Learning prior and posterior from data,
2) Feature attribution and counterfactual explanations-based methods and
3) Integrated XAI.

The learning model will learn the approximate prior from observational data (Step 1), from which we can estimate the posterior distribution for the Generative Model (GM). GM will provide a better representation and understanding of the observational data. The GM model (Step 2) will be utilized for both feature attribution and counterfactual analysis. Then, in integrated XAI, *i.e.*, the proposed SRL will integrate the methods (Step 3) and provide an XAI model with improved confidence, leading to more understandable (loveliest) explanations. The specific outcome of the work is a *new XAI method for gAI* that will provide consistent and stable explanations using constraint-based numerical techniques and a probabilistic approach from observational data.

IV. Learning Prior and Posterior from Data

This step will focus on generalizing distributions, ensuring convergence, and optimizing computational efficiency for modeling unknown functions. First, it estimates the probabilities of the observation x denoted by $p(x)$. Second, given the decisions y and knowing the prior knowledge of the observation x as probability distributions, a generative model is built that estimates the distribution $p(x|y)$. Generative modelling uses a Bayesian framework by setting prior plausibilities and updating the posterior plausibilities in light of new data. The process can be described as for some dataset, $\mathcal{X} = \{x^{(i)}\}_{i=1}^{N}$, consists of $|\mathcal{X}| = N$ i.i.d. (independent and identically distributed) samples, we assume the data is generated by some random process, involving an unobserved (*i.e.*, unknown to us) probability distribution for some random variable y. The process requires first to know some prior distribution $p_{\theta^*}(y)$ to generate a value $y^{(i)}$ and then second, use some conditional distribution $p_{\theta^*}(x|y)$ to generate the value $x^{(i)}$. Often some fixed parametric family of distribution $p_\theta(y)$ and $p_\theta(x|y)$ are used to define the prior $p_{\theta^*}(y)$ and the likelihood $p_{\theta^*}(x|y)$ and it assumes that their probability density functions (PDF) are differentiable w.r.t. θ and y. In the ML setting, the true parameters θ^* and prior distribution are unknown to us. However, the log-likelihood of the generative model $\log p(x)$ can be evaluated as in (1):

$$\log p_\theta(x) = \log \sum_i p_\theta\left(x|y^{(i)}\right) p\left(y^{(i)}\right) \tag{1}$$

Since the model $p_\theta(x|y)$ is trained with $y = y_{\theta^*}(x)$ from the "Maximum a Posteriori" inference, we can write as in (2):

$$\log p_\theta(x) \approx \log p_{\theta^*}\left(x|y_{\theta^*}(x)\right) p_{\theta^*}\left(y_{\theta^*}(x)\right) \tag{2}$$

TABLE I
EXPECTED QUESTIONS TO BE ANSWERED IN THE PROPOSED APPROACH.

Feature Attribution	Counterfactuals
1. Why did you make that decision? 2. How did you make that decision? 3. What data did you see? 4. How certain are you?	1. Are you sure that it is not something else? 2. What would it take for me to get another decision? 3. How do I correct an error? 4. What happens if I do not take that decision?

which can be evaluated empirically.

Generally, when the parameters are unknown, the posteriors and priors for GMs are assumed to be normally distributed. To provide an explanation with a causal chain, a data-centric full informative prior's elicitation is necessary to better estimate the posterior distribution. To achieve IBE, an XAI model must establish the underlying probability distribution for data of unknown distribution. Here, firstly the probabilistic framework will learn informative priors directly from the training data, enhancing posterior probability updates. Full informative priors provide precise feature information, improving posterior precision and explanations. Poor estimation of priors can lead to biased Bayesian estimates. The framework will define criteria for "good" priors that accurately capture the AI model's behaviour and offer the best approximations. The proposed approach will use Continuous Piecewise Affine mapping to represent complex data distributions, leveraging functions like Rectified Linear Unit (ReLU) to obtain latent space distributions for better density representation.

Secondly, Markov Chain Monte Carlo and Stochastic Gradient Variational Bayes methods will be used to generate samples from the posterior distribution, reducing uncertainty and improving the representation of training data. These methods will transform the XAI model into a probabilistic one to enhance explanation generation.

V. FEATURE ATTRIBUTION AND COUNTERFACTUAL EXPLANATION-BASED METHODS

XAI methods based on feature attribution often rely on combinations of features to explain a decision, which can be unrealistic and result in implausible synthetic instances. This issue arises due to the disregard for the local distribution of features, the density of class labels in the neighbourhood, and causal relationships among input features during data sampling for surrogate models. To address this problem, we propose leveraging the generative modelling approach discussed in the previous section. A principal approach for identifying causal structure in data is to use Structural Causal Models (SCMs). Incorporating knowledge, even partial, of the causal structure of observational data can enhance the understanding of decisions made by the AI model, leading to higher interpretability and more robust explanations. Suppose a random variable $\mathcal{C}$ denotes the cause and $\mathcal{E}$ is the effect. An SCM with graph $\mathcal{C} \rightarrow \mathcal{E}$ can be defined in (3), and (4):

$$\mathcal{C} := \mathcal{N}_{\mathcal{C}} \tag{3}$$

$$\mathcal{E} := f_{\mathcal{E}}(\mathcal{C}, \mathcal{N}_{\mathcal{E}}) \tag{4}$$

where, $\mathcal{N}_{\mathcal{E}}$ and $\mathcal{N}_{\mathcal{C}}$ are the noise associated with $\mathcal{E}$ and $\mathcal{C}$, and $\mathcal{N}_{\mathcal{E}}$ is independent of $\mathcal{N}_{\mathcal{C}}$. We can sample noise values $\mathcal{N}_{\mathcal{E}}$ and $\mathcal{N}_{\mathcal{C}}$ by evaluating $\mathcal{C}$ and $\mathcal{E}$ with the above two equations if we know the function $f_{\mathcal{E}}$ and the noise distributions $P_{\mathcal{N}_{\mathcal{C}}}$ and $P_{\mathcal{N}_{\mathcal{E}}}$. When it comes to explanations consisting of fact and foil, we are specifically interested in understanding the behaviour of the AI model under interventions. This involves inducing a different distribution that deviates from the observational distribution to intentionally change the AI model's behaviour. The two distributions of the AI model become unrelated after the intervention, and we can treat them as two separate models, especially when only certain parts of the data-generating process change due to the intervention. In addition, we modify all noise distributions of an SCM to enable us to effectively address counterfactual questions.

Both feature attribution and counterfactual explanations will leverage generative modelling in the proposed approach. The SRL, which utilizes probabilistic graphical models as a common approach to descriptive modelling and declarative representation, can be used to integrate the two explanation approaches. By combining feature attribution and counterfactual explanations through SRL, we aim to develop a comprehensive XAI model that can address the questions outlined in Table I. Some of the questions related to counterfactuals remain as open challenges that current XAI methods failed to provide a satisfactory answer.

For feature attribution, the generative modelling framework that uses true prior knowledge of the training dataset will build a rational XAI model that allows the "Loveliest" explanations. This task will use the Bayesian approach for GM frameworks, which works in two folds. First, it estimates the probabilities of observation x denoted by $p(x)$. Second, given the decisions y and knowing the prior knowledge of the observation x as probability distributions, a GM is built that estimates the distribution $p(x|y)$.

For generating the explanations, let us consider a classifier f that outputs the prediction $y = f(x)$ for an observation x, where x is a vector consisting of values from m features. The explanation to the prediction with feature attribution is defined by a vector of feature contributions– $[\phi_0, \phi_1, ..., \phi_m]$, where ϕ_0 is the bias term and rest of the values corresponds to the m features denoting the contribution of the features to a particular prediction. Thus, approximation on the confidence of the prediction can be defined in terms of probability with (5):

$$p\left(f\left(x\right)|\, x\right) \approx \sum_{j=0}^{m} \phi_j \tag{5}$$

With the same classifier f, counterfactual explanation to the prediction $y = f(x)$ consists of a set of observations x' such that the prediction by f on x' is different from y, *i.e.*, $f(x') \neq y$, and where the difference between x and x'

is minimal. To identify causal structure in data, we will use SCMs, considering the same generative framework mentioned above.

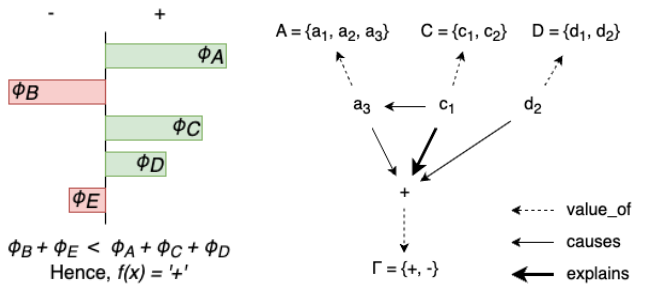
VI. Integrated XAI

A graph-based learning framework will be developed using probabilistic graphical models (*i.e.*, an extension of SRL) to integrate feature attribution and counterfactual methods as a hybrid solution for XAI. Leveraging SRL, we will establish connections between strong associations in feature attribution and causal chains in observational data. As causal models are graph-based, we will design a novel approach to generate graphical representations for feature attribution, allowing us to learn explanations for the outcomes of automated decisions. In statistical learning theory, PAC-Bayes bounds generalize the union bound that allows dealing with both finite and infinite parameters. We aim to bridge the gap between the Bayesian framework and statistical learning theory using a decision-theoretic approach. This theoretical framework for the generalization error bound will provide a high degree of confidence in the explanations given by the XAI model and estimate the underlying uncertainty associated with these explanations.

VII. Outcome, Benefits and Limitations

None of the existing XAI methods applies to gAI to provide comprehensive explanations [12]. By knowing the level of model confidence, the proposed methods increase the faithfulness of explanations to the user. ML methods often lack "control variables" for counterfactual or causal reasoning in inductive reasoning. Counterfactual reasoning is a vital part of human cognition, and it is essential for effective decision-making and problem-solving as it involves considering alternate possibilities or hypothetical scenarios that differ from what has occurred, creating new perspectives on familiar themes and ideas. By learning from interventional data, this approach will provide a better understanding of phenomena and enable better counterfactual representations.

Here, for instance, we can consider a hypothetical use case of binary classification with label $\Gamma = \{`+', `-'\}$ and observations $\mathcal{X}$ containing values from the features A, B, C, D and E. Now, for an observation $x \in \mathcal{X}$, the prediction $f : x \rightarrow `+'$ can be explained with feature attribution, as illustrated in Fig. 3a. From the ϕ values, corresponding answers to the feature attribution from Table I are intended to be answered. For example, if the user asks the classifier model, "How certain are you?", it can be answered with he approximated probability on the prediction $p(f(x)|x)$ from the corresponding feature attribution values ϕ and (5). In addition, using the SCM illustrated in Fig. 3b, the questions intended for the counterfactual-based explanations can be answered, such as "What would it take for me to get another decision?". The SCM is presented with the hierarchy and causation of the concerned features to the prediction associated with the possible values of the highly contributing features to the prediction and the label (Γ). The counterfactuals can be generated from the alternate values of the features as well as the change in predicted label; thus, the query from the user on a different decision can be answered using the SCM.



(a) Feature Attribution (b) Counterfactual

Fig. 3. A hypothetical outcome of the proposed approach for integrated XAI in a binary classification task.

Ultimately, this will allow for learning from past experiences and making decisions that embody the human cognitive process. The proposed unique approach seeks to provide the necessary and sufficient "fact and foil" of AI decisions by incorporating counterfactual reasoning alongside feature attribution, which is a commonly used method for providing possible facts. Doing so will offer a more complete and balanced explanation of AI decision-making than feature attribution alone, as foil information will be provided through the counterfactual approach. Therefore, the main result of this work can be seen as an enabling technology for further rapid innovation of new methods for next-generation gAI-based industrial applications. Incorporating gAI with an explanation will enhance transparency, accountability, and compliance with emerging regulations in AI systems, fostering greater trust in their decision-making capabilities, improving ML fairness and mitigating data bias for industrial applications. It will enhance the level of trust in AI-generated decisions, especially on gAI and prescriptive analytics for different end-users, developers, and other stakeholders.

The framework could provide insights into the future behaviour of machines by quantifying the uncertainty of the AI model for a range of adversarial samples based on historical data, characteristics of data distribution, and perturbation. The knowledge of the capability of gAI, which is crucial for providing actionable measures in explainable decision-making, can be transferable to other domains. So, the next step is to validate the proposed theoretical framework with industrial data. The validation of outcomes is currently restricted to laboratory settings, and the work is ongoing. Once that is done, we may require additional deployment efforts to ensure the overall outcome meets that expectation to be utilized in the real industrial environment.

VIII. Conclusions

"Explanation" is a key research topic in the philosophy of science. The current explanation methods in XAI research lack this connection between the philosophy of science and algorithmic development. As a result, they often

provide insufficient human-understandable explanations and suffer from inconsistency and reliability issues. Trust in AI systems can be increased by providing clear explanations to the users of how the AI arrives at its decisions. One way to achieve this is by using XAI techniques, which support better knowledge representation and can help in understanding the decision-making processes of complex AI models. However, the challenge is to plug in the philosophical concept of explanation that is closer to human-perceived explanations, to the computational/mathematical development of explanations in gAI. The approach will advance the modern XAI techniques to more human-like responses to "why" questions and provide satisfactory explanations that align with human-level comprehension. The proposed approach investigates the use of generative modelling to address the inconsistency and unreliability problems in XAI methods to include it in gAI. Thus, the paper outlines a theoretical framework in 3 steps to achieve "understandable (loveliest)" explanations from data. Here, the significance and scientific novelty are approximations of learning "prior" from observational data, understanding the phenomena and concept based on cognitive process (in counterfactuals), and a unique solution based on SRL approach utilizing the Bayesian framework will provide the understandable (loveliest) explanation.

REFERENCES

[1] P. Lipton, *Inference to the Best Explanation*, 2nd ed. Routledge, 2003.

[2] I. Douven, "Inference to the Best Explanation Made Coherent," *Philos. Sci.*, vol. 66, S424–S435, 1999.

[3] X. Zhao, X. Huang, V. Robu, and D. Flynn, "BayLIME: Bayesian Local Interpretable Model-Agnostic Explanations," in *Proc. Conf. Uncertain. Artif. Intell. (UAI)*, PMLR, 2021, pp. 887–896.

[4] M. Cinquini and R. Guidotti, "Causality-Aware Local Interpretable Model-Agnostic Explanations," in *Explainable Artif. Intell.*, L. Longo, S. Lapuschkin, and C. Seifert, Eds., 2024, pp. 108–124.

[5] P. Lipton, "Contrastive Explanation," *Roy. Inst. Philos. Suppl.*, vol. 27, pp. 247–266, 1990.

[6] T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2019.

[7] C. Rudin *et al.*, "Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges," *Statist. Surv.*, vol. 16, no. none, 2022.

[8] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020.

[9] G. Vilone and L. Longo, *Explainable Artificial Intelligence: A Systematic Review*, arXiv preprint, arXiv:2006.00093v4 [cs.AI], 2020.

[10] M. R. Islam, M. U. Ahmed, S. Barua, and S. Begum, "A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks," *Appl. Sci.*, vol. 12, no. 3, p. 1353, 2022.

[11] G. Schwalbe and B. Finzel, "A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts," *Data Min. Knowl. Disc.*, vol. 38, no. 5, pp. 3043–3101, 2024.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD)*, 2016, pp. 1135–1144.

[13] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Adv. Neural Inf. Process. Syst. (NeurIPS 31)*, 2017, pp. 4768–4777.

[14] H. Baniecki, W. Kretowicz, P. Piątyszek, J. Wiśniewski, and P. Biecek, "DALEX: Responsible Machine Learning with Interactive Explainability and Fairness in Python," *J. Mach. Learn. Res.*, vol. 22, no. 214, pp. 1–7, 2021.

[15] R. R. Selvaraju *et al.*, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, 2020.

[16] C. Chen *et al.*, "This Looks Like That: Deep Learning for Interpretable Image Recognition," in *Systems Adv. Neural Inf. Process. Syst. (NeurIPS 33)*, 2019, pp. 8930–8941.

[17] Q. Huang *et al.*, *Evaluation and Improvement of Interpretability for Self-Explainable Part-Prototype Networks*, arXiv preprint, arXiv:2212.05946v3 [cs.CV], 2023.

[18] J. Yu *et al.*, "INTERACTION: A Generative XAI Framework for Natural Language Inference Explanations," in *Proc. 2022 Int. Joint Conf. Neural Netw. (IJCNN)*, 2022, pp. 1–8.

[19] D. P. Kingma and M. Welling, *Auto-Encoding Variational Bayes*, arXiv preprint, arXiv:1312.6114v11 [stat.ML], 2022.

[20] L. D. Raedt, S. Dumančić, R. Manhaeve, and G. Marra, "From Statistical Relational to Neuro-Symbolic Artificial Intelligence," in *Proc. 29th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2020, pp. 4943–4950.

[21] S. S. Sheuly, M. U. Ahmed, S. Begum, and M. Osbakk, "Explainable Machine Learning to Improve Assembly Line Automation," in *Proc. 4th Int. Conf. Artif. Intell. Ind. (AI4I)*, 2021, pp. 81–85.

[22] M. U. Ahmed, S. Barua, S. Begum, M. R. Islam, and R. O. Weber, "When a CBR in Hand Better than Twins in the Bush," in *Proc. 4th XCBR Workshop, Case-Based Reasoning Explain. Intell. Syst., co-located with ICCBR*, P. Reuss and J. Schönborn, Eds., ser. CEUR Workshop Proceedings, vol. 3389, 2022, pp. 141–152.