International Journal of Applied Science and Engineering
Citation: *IJASE*: **13**(02): 309-319, December 2025
DOI: 10.30954/2322-0465.2.2025.11

CASE STUDY APPLIED INFORMATICS

Explainable Generative AI: Concepts, Challenges and Future Directions

Saurav Paul and Saptarshi Paul*

Department of Computer Science and Engineering, Assam University, Silchar, India
**Corresponding author: paulsaptarshi@yahoo.co.in*

Received: 16 Sept., 2025 **Revised:** 14 Nov., 2025 **Accepted:** 28 Nov., 2025

ABSTRACT

Large language models (LLMs), diffusion-based image generators, speech synthesizers, code assistants, and multimodal foundation models are examples of generative AI systems that are now deeply integrated in scientific research, creative work, and decision-making. Understanding why these models produce specific outputs has become crucial for safety, accountability, trust, debugging, and regulatory compliance as they continue to acquire autonomy and agency. However, explainability for generative AI is significantly more challenging than explainability for traditional discriminative models, because the generation unfolds over time, the output space is open-ended, and the model internals operate in high-dimensional spaces. Explainable Generative AI (XGAI) is the new interdisciplinary field examined in this paper. The paper defines the scope and fundamental concepts of XGAI, presents a taxonomy of interpretability techniques for generative models, examines technical, social, and regulatory challenges, and outlines future research directions toward transparent and controllable generative systems. This work argues that explainability for generative AI requires coverage of intent, process, uncertainty, and provenance and cannot be reduced to local feature attribution alone.

Keywords: Explainable AI, Generative Models, Interpretability, Large Language Models, Safety Alignment, Model Transparency

Generative AI refers to machine learning systems that synthesize new content including text, code, images, audio, video, molecular structures, or behavior plans, from learned data distributions rather than selecting from predefined labels[5]. Large language models like GPT-style transformers, text-to-image diffusion models, and speech and video generators are now assisting with tasks in education, medicine, law, creative writing, cyber defense, scientific discovery, and governance[6],[7]. These systems increasingly

How to cite this article: Paul, S. and Paul, S. (2025). Explainable Generative AI: Concepts, Challenges and Future Directions. *IJASE.*, **13**(02): 309-319.

Source of Support: None; **Conflict of Interest:** None

act not only as tools but as collaborators and agents, embedding themselves in workflows that carry significant consequences for individuals and organizations. This growing capability creates pressure for explainability, meaning the ability to make a system's behavior understandable to humans in a faithful, useful way. Explainability in generative AI is necessary for at least five interconnected reasons. First, with respect to safety and alignment, researchers and practitioners need to understand why a model produced a harmful, biased, misleading, defamatory, or policy-violating output so that corrections or preventive measures can be applied[8]. Second, for accountability, organizations deploying generative AI will be expected to justify generated advice, decisions, or content to auditors, regulators, courts, and users, as legal frameworks in various regions have already begun to mandate explainability provisions[9]. Third, from a debuggability standpoint, engineers need to trace failure modes such as hallucinations in LLMs, prompt leakage of private data, or copyrighted style cloning in image models[10]. Fourth, regarding user trust, users are more likely to rely on a model's output if they can interrogate its reasoning or content provenance, which directly affects adoption in high-stakes domains such as healthcare and law. Fifth, from a scientific understanding perspective, generative models are currently treated as black boxes despite increasingly agent-like behavior, and making them interpretable is important for both AI safety research and cognitive science[11]. Classical explainable AI (XAI) focused on supervised tasks such as classification and regression, and largely assumed fixed inputs and discrete predictions. Methods such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) attribute importance to input features for a given decision[1],[2]. These methods, while influential, are insufficient for generative models because generators produce sequences or structured artifacts rather than a single label, the model state evolves through iterative sampling such as autoregression or diffusion timesteps, prompts are not the only causal factor since system instructions, safety policies, memory, retrieval-augmented context, and fine-tuning data also shape output, and risk encompasses not only why a model output a specific answer but also what it could plausibly output under nearby conditions[12],[13]. This paper proposes a structured view of Explainable Generative AI. XGAI is treated as the union of techniques, interfaces, and governance mechanisms that help humans answer four questions about any generated output: attribution asking where it came from, intention asking what the model was trying to do, mechanism asking how it was produced, and alternatives asking what else could have been produced and why it was not[14].

Background

Generative Model Families

Modern generative AI mainly relies on four foundational model families. Autoregressive transformers generate output token-by-token, conditioning each next token on previous tokens and internal hidden states, representing the dominant architecture for large language models and code models[5]. Diffusion models learn to iteratively denoise random noise into a coherent sample such as an image, audio, or video through a sequence of refinement steps[15]. Variational Autoencoders (VAEs) learn latent representations and can decode new samples by sampling from the latent space, while GANs (Generative Adversarial Networks) train a generator and discriminator in a minimax game to synthesize realistic data[16]. Although VAEs and GANs remain important in some domains such as compressed latent representations and style

transfer, diffusion and transformer-based models dominate state-of-the-art text, image, and multimodal generation[17],[18].

What Does Explainability Mean?

Explainability has historically split into two broad traditions. Transparency and interpretability aim to make the inside of the system understandable through analysis of neuron circuits, attention heads, steering vectors, and safety classifiers[19]. Post-hoc explanation involves generating an external artifact such as a rationale, a saliency map, or a counterfactual example that helps a human understand why a given decision or output occurred[20]. For generative AI, both traditions are generally necessary. Researchers and practitioners need to interpret model behavior mechanistically to ensure alignment and robustness, while also communicating explanations accessibly to people affected by the output[21].

Why Generative Explainability is Harder

Three properties of generative AI place significant stress on traditional XAI methods. The open-ended output space presents a fundamental challenge because there is no fixed correct label; the model can generate new sentences, new code, or new biological structures, meaning trust must be justified without ground truth in many cases[22]. Long-horizon causal chains are a second source of difficulty because the output at a given time step depends on the entire prefix including the prompt and all previously generated tokens, as well as latent activations, so small changes early in generation can cascade into large differences in the final output. Context mixing further complicates matters because modern LLM pipelines include prompt templates, retrieval-augmented memory, system policies, and tool calls, meaning a single answer may reflect all of these factors and not only the user's original input[23].

A Taxonomy of Explainability Methods for Generative AI

The approaches described in this section are grouped into six major categories. These categories are not mutually exclusive; practical systems often combine multiple techniques.

Prompt and Context-Level Attribution

The goal of prompt and context-level attribution is to explain which parts of the prompt, retrieved context, or system instruction contributed to which parts of the output. This is the most accessible and currently most deployable layer for LLM-based systems. Token-level influence highlighting aligns generated tokens with the most influential prompt or context tokens, often using attention weights, gradient-based attribution, or influence functions adapted to transformers. Attention visualization is attractive but contested because attention weights do not always correspond to true causal importance[24]. In retrieval-augmented generation (RAG), context provenance tracing exposes which retrieved snippet or snippets supported each factual claim, which helps auditors determine whether a hallucinated claim came from poor retrieval, flawed reasoning, or fabrication[25]. Instruction lineage explainability addresses the fact that many LLM responses are shaped by hidden system prompts, safety policies, or role constraints, and exposing which internal instruction blocked or overrode a request is a form of policy explainability useful for safety review and compliance[26]. A key limitation is that attribution at the level of identifying that a sentence came from a particular source document is more tractable than attributing stylistic tone

to a training dataset, and dataset-scale provenance remains an open problem because training corpora are massive and often proprietary[9].

Stepwise Reasoning Traces and Chain-of-Thought Explanations

The goal of stepwise reasoning traces is to reveal the intermediate reasoning, planning, or tool-use steps that led to the final output. In tool-using or planning-oriented agents, the model may decompose a task into subgoals, call tools or external APIs, summarize intermediate results, and then produce a final answer. Surfacing this chain provides humans with a causal story that can be valuable for debugging incorrect answers and for compliance reviews that verify whether the model accessed disallowed tools or risky data[27]. A significant caveat is that generated reasoning traces can themselves represent post-hoc rationalizations rather than faithful accounts of computation. A model can produce a coherent explanation even when the true causal path in its hidden activations was different. Balancing faithfulness, meaning whether the explanation reflects actual internal computation, with helpfulness, meaning whether it is understandable to humans, remains an active research topic[28].

Mechanistic Interpretability for Generative Models

Mechanistic interpretability aims to understand the internal circuits of the model itself by analyzing neurons, attention heads, MLP layers, and residual streams to discover causal roles[19]. Examples include identifying induction heads that copy previous tokens forward in transformer models, detecting neurons that robustly activate on unsafe content such as self-harm instructions or personally identifiable information, and editing internal activations to steer generation through techniques such as activation patching and causal tracing[29]. In diffusion models, researchers probe denoising steps to locate where semantic attributes are introduced and then intervene on those attributes to control generation style[30]. Mechanistic interpretability is promising because it aims at ground-truth faithfulness by attempting to explain what the model is actually doing rather than what it reports doing. However, it is technically intensive, not yet practical for product deployment, and scales poorly as models grow to hundreds of billions of parameters[31].

Concept Activation and Latent Space Probing

The goal of concept activation and latent space probing is to map human-understandable concepts onto internal model features. For vision and diffusion models, linear probes can be trained to detect semantic concepts such as military uniform, medical device, or celebrity face in latent features[32]. If a probe strongly activates on a given concept, it becomes possible to state that the model internally represented that concept at a specific stage of generation. For language models, similar techniques identify directions in embedding space associated with attributes such as sentiment, toxicity, political leaning, gendered phrasing, or programming style. By moving along or suppressing those latent directions, researchers can both explain and control the generated style[33]. This approach is useful for fairness audits that ask whether the model implicitly associates certain jobs with a gendered direction, and for safety purposes asking whether a violent planning concept emerged internally before disallowed instructions were filtered[34].

Counterfactual and Contrastive Explanations

Counterfactual and contrastive explanations address the question of what the model would have generated if some aspect of the input were different. Counterfactual explanations for generative AI can take the form of minimal prompt edits that change a specific part of the output, alternative decoding paths showing which other coherent answers were likely, or style and instruction toggles that reveal which internal constraints suppress certain content[35]. Contrastive explanations are often more satisfying for human users than absolute explanations because they simultaneously explain a behavior and communicate safe interaction boundaries. However, counterfactual search in high-dimensional prompt spaces is computationally expensive, and naive approaches can accidentally reveal how to phrase unsafe requests in ways that would receive a response, which constitutes a security concern[36].

Data Provenance and Training Influence

The goal of data provenance and training influence analysis is to attribute generated content to specific training examples or fine-tuning data. Stakeholders increasingly ask whether a generated image mimics a specific copyrighted artwork, whether a code snippet derives from GPL-licensed code, or whether a medical recommendation quotes a sensitive health record[37]. Influence functions and nearest-neighbor retrieval over training embeddings are being explored to estimate which training samples most influenced a given output, and these approaches can support copyright audits and privacy leak investigations[38]. The central challenge is scale, as frontier models are trained on trillions of tokens, and storing, indexing, and auditing influence at that resolution is highly nontrivial. Legal sensitivities around proprietary datasets add further friction to this area of research[9],[10].

How Do We Evaluate Explanations?

An explanation for generative AI is only useful if it is meaningful to the intended audience and faithful to the model's actual computation. Five desirable properties define a useful explanation in this context. Faithfulness requires that the explanation tracks real causal contributions rather than plausible stories, and mechanistic methods aim most directly at this property[29]. Completeness or coverage requires that the explanation accounts for the main drivers of the output including the prompt, policy, memory, tool calls, and decoding strategy, rather than only one slice of the process[27]. Actionability requires that a developer, auditor, or end user should be able to take some action based on the explanation such as reproducing, correcting, appealing, or refining the request[21]. Cognitive fit requires that the explanation be matched to the audience, as a red-team analyst might need neuron-level traces while an end user might need only a plain-language statement that the answer is based on a specific section of an uploaded document. Robustness requires that the explanation remain stable under small perturbations of the input, because if rephrasing the same question yields a completely different explanation, user trust in the system collapses[39]. A fundamental tension persists between these properties: high-faithfulness, neuron-level causal graphs are not cognitively usable by most people, while highly usable natural-language summaries risk being post-hoc justifications. Generating explanations that are simultaneously faithful and consumable for different stakeholders remains a core unsolved problem[40].

Key Challenges

Hallucination and Invented Evidence

LLMs sometimes produce confident statements with no factual grounding[10]. When asked to justify such a statement, the same model can produce an articulate but fabricated rationale. This creates a recursive hallucination problem in which explanations themselves can hallucinate, making it dangerous to rely solely on self-reported reasoning. Mitigation requires grounding answers in cited sources through RAG provenance, deploying external verification pipelines, and conducting audits that distinguish internal causal factors from post-hoc narratives[25].

Multi-Agent and Tool-Using Systems

Modern generative systems increasingly browse documents, call APIs and code interpreters, write intermediate notes to memory, refine drafts, and then produce a final result[31]. This complexity raises the question of which component should explain the output: the base model, the planner agent that orchestrated tool calls, the retrieval component, or the human who approved a draft. Explainability in this setting becomes workflow explainability rather than simply model explainability, requiring versioned execution traces that record which step produced which part of the final answer and under which policy.

Privacy and Security

Many explanation techniques require surfacing sensitive internal details. Internal instructions or safety prompts may reveal red-team policies when exposed, training provenance analysis may leak personal data, and counterfactual demonstrations can inadvertently reveal how to bypass safety filters[33]. This creates a fundamental paradox in which richer explanations can increase misuse risk. This tension will be central to how future regulations are formulated and enforced[9].

Intellectual Property and Attribution

Image and code generators raise significant intellectual property questions. If an explanation reveals that a generated image strongly resembles a specific artist from a specific training dataset, that finding may support compensation or opt-out mechanisms for the artist[37]. However, it also confirms that the model memorized and can reproduce a particular style, which some vendors consider proprietary. Balancing transparency obligations with legitimate business confidentiality is both politically and economically contested.

Scale and Latency

Producing explanations such as influence traces, provenance metadata, and causal maps can be more computationally expensive than producing the raw output itself. Systems deployed at web scale must generate explanations cheaply, consistently, and in near real time. Meeting these requirements simultaneously remains an open engineering problem[35].

Regulatory Pressure Without Technical Clarity

Governments, standards bodies, and industry consortia are moving toward requirements for meaningful explanation, auditability, or documentation of training data sources. However, what counts as meaningful, sufficient, or reasonable effort is not yet standardized across jurisdictions or technical communities. This gap between legal expectation and technical feasibility is widening as deployment accelerates[9],[37].

Future Research Directions

Native Explainability Architectures

Today, most explanations are added to models after the fact. Future frontier models should be designed to be interpretable from the outset, trained with architectural constraints or auxiliary objectives that yield structured reasoning traces, causal bottlenecks, and disentangled safety controls that are inspectable by default rather than reverse-engineered later[38].

Verified Reasoning Traces

Mechanisms are needed to verify that a model's stated reasoning actually influenced its output. Possible approaches include cryptographic commitments to intermediate states, sandboxed tool calls whose outputs are logged and signed, and causal intervention tests in which removing a step from the trace must change the final answer[28]. Such mechanisms would allow auditors to distinguish honest causal chains from post-hoc rationalizations.

Standards for Provenance and Citation

Every generated claim that appears factual should be linkable to its supporting evidence. For text models this means tracing paragraphs to retrieved sources. For image or code models it could mean associating visual or stylistic features with clusters of training samples from identified licenses. These standards will be essential for copyright enforcement, misinformation control, and scientific reproducibility[39].

Human-Aligned Abstraction Layers

Raw neuron activations are too low-level for practical use by most stakeholders, while English-language rationales are sometimes too high-level and unfaithful. Intermediate representation layers designed for auditors are needed, including task decomposition graphs, safety policy trigger logs, bias and risk scores per subtask, and uncertainty indicators on each part of the generated artifact. These abstractions should be consistent, machine-readable, and comparable across vendors.

6.5 Counterfactual Safety Sandboxes

Organizations deploying generative AI will increasingly want to systematically probe how models respond to large sets of borderline-dangerous prompts before deployment. This approach combines automated red-teaming with counterfactual explainability, producing structured maps of which safety rules fired,

which internal mechanisms were responsible, and where the model remains vulnerable[36]. Such sandboxes could become a standard component of compliance audits and certification processes.

Explainable Alignment Controls

Fine-tuning, reinforcement learning from human feedback (RLHF), direct preference optimization, and safety adapters steer models toward desired behaviors[7]. At present these processes are largely opaque. Transparent alignment reports are needed that document which behaviors were strengthened or suppressed, by which data, and with what expected side effects such as changes in refusal rates, politeness, or hedging behavior. This would make the alignment process itself auditable rather than only the final model.

Multi-Agent Accountability Graphs

As agentic systems chain multiple models, tools, and memory buffers, responsibility graphs are needed to assign provenance to each contributed step. Such graphs would support post-incident forensics by clarifying which module in an autonomous system failed when unsafe content was produced, whether the failure originated in the retrieval filter, the planner, or the base LLM[40].

Socio-Technical Governance and User Experience

Explainability is not only an algorithmic challenge. It is simultaneously a user interface problem asking what explanation a clinician actually needs before acting on an AI-generated diagnosis, an organizational process problem asking who signs off on AI-generated legal language, and a policy problem asking what minimum disclosures are mandatory in regulated sectors[31]. Future work must integrate technical explainability with policy frameworks, legal audit trails, human-in-the-loop review workflows, and sector-specific user experience norms[39].

CONCLUSION

Explainability for generative AI is not simply a matter of computing feature importance for text output. It is a broader requirement to make model behavior inspectable, auditable, and governable across the full life-cycle of generation, encompassing data ingestion, prompting, planning, tool use, decoding, alignment, and safety enforcement. This paper proposed four guiding questions for any generative output, namely attribution, intention, mechanism, and alternatives, and surveyed current approaches including prompt attribution, reasoning traces, mechanistic interpretability, latent concept probing, counterfactual analysis, and data provenance. The analysis demonstrates that useful explanations must be both faithful to the model's causal process and practically usable by humans with varying levels of technical expertise. Eight concrete research and engineering directions were outlined to close that gap, covering native interpretable architectures, verified reasoning traces, provenance standards, human-aligned abstraction layers, counterfactual safety sandboxes, explainable alignment controls, multi-agent accountability graphs, and socio-technical governance. Generative AI will continue to gain autonomy, economic impact, and legal scrutiny in the years ahead. Systems that cannot explain themselves to developers, auditors, regulators, and end users will increasingly be viewed as defective or non-compliant under emerging regulatory frameworks. The future of trustworthy generative AI is therefore inseparable from its explainability.

REFERENCES

1. Lundberg, S.M. and Lee, S.-I. 2017. "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, **30**: 4765-4774.

2. Ribeiro, M.T., Singh, S. and Guestrin, C. 2016. "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, pp. 1135-1144.

3. Miller, T. 2019. "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence,* **267**: 1-38, 2019.

4. Doshi-Velez, F. and Kim, B. 2017. "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608.

5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. 2017. "Attention is all you need," in Advances in Neural Information Processing Systems, **30**: 5998-6008.

6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. *et al.* 2020. "Language models are few-shot learners," in Advances in Neural Information Processing Systems, **33**: 1877-1901.

7. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D. and Christiano, P.F. 2020. "Learning to summarize from human feedback," in Advances in Neural Information Processing Systems, **33**: 3008-3021.

8. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. and Mane, D. 2016. "Concrete problems in AI safety," arXiv preprint arXiv:1606.06565.

9. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A. *et al.* 2021. "Ethical and social risks of harm from language models," arXiv preprint arXiv:2112.04359.

10. Maynez, J., Narayan, S., Bohnet, B. and McDonald, R. 2020. "On faithfulness and factuality in abstractive summarization," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1906-1919.

11. Lipton, Z.C. 2018. "The mythos of model interpretability," *Queue,* **16**(3): 31-57.

12. Montavon, G., Samek, W. and Mueller, K.-R. 2018. "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing,* **73**: 1-15, 2018.

13. Lombrozo, T. 2006. "The structure and function of explanations," *Trends in Cognitive Sciences,* **10**(10): 464-470.

14. Ho, J., Jain, A. and Abbeel, P. 2020. "Denoising diffusion probabilistic models," in Advances in Neural Information Processing Systems, **33**: 6840-6851.

15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. 2020. "Generative adversarial networks," *Communications of the ACM,* **63**(11): 139-144.

16. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. 2022. "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684-10695.

17. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J. 2020. "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research,* **21**(140): 1-67.

18. Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M. and Carter, S. 2020. "Zoom in: An introduction to circuits," Distill, [Online]. Available: https://distill.pub/2020/circuits/

19. Bach, S., Binder, A., Montavon, G., Klauschen, F., Mueller, K.-R. and Samek, W. 2015. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, **10**(7): e0130140.

20. Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K. and Mueller, K.-R. 2019. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer.

21. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuеттler, H., Lewis, M., Yih, W.-T., Rocktaschel, T., Riedel, S. and Kiela, D. 2020. "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems, **33**: 9459-9474.

22. Jain, S. and Wallace, B.C. 2019. "Attention is not explanation," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3995-4005.

23. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V. and Zhou, D. 2022. "Chain-of-thought prompting elicits reasoning in large language models," in Advances in Neural Information Processing Systems, **35**: 24824-24837.

24. Meng, K., Bau, D., Andonian, A. and Belinkov, Y. 2022. "Locating and editing factual associations in GPT," in Advances in Neural Information Processing Systems, **35**: 17359-17372.

25. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y. and Cohen-Or, D. 2023. "Prompt-to-prompt image editing with cross attention control," in Proceedings of the International Conference on Learning Representations.

26. Elhage, N., Nanda, N., Olah, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T. *et al.* 2021. "A mathematical framework for transformer circuits," Transformer Circuits Thread, [Online]. Available: https://transformer-circuits.pub/2021/framework/index.html

27. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F. and Sayres, R. 2018. "Interpretability beyond classification accuracy: A quantitative testing with concept activation vectors (TCAV)," in Proceedings of the 35th International Conference on Machine Learning, pp. 2668-2677.

28. Wachter, S., Mittelstadt, B. and Russell, C. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard Journal of Law and Technology,* **31**(2): 841-887.

29. Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D. and Wallace, E. 2023. "Extracting training data from diffusion models," in Proceedings of the 32nd USENIX Security Symposium.

30. Koh, P.W. and Liang, P. 2017. "Understanding black-box predictions via influence functions," in Proceedings of the 34th International Conference on Machine Learning, **70**: 1885-1894.

31. Bhatt, U., Weller, A. and Moura, J.M.F. 2020. "Evaluating and aggregating feature-based model explanations," in Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, pp. 3016-3022.

32. Doshi-Velez, F. and Kim, B. 2018. "Considerations for evaluation and generalization in interpretable machine learning," in Explainable and Interpretable Models in Computer Vision and Machine Learning. Springer, pp. 3-17.

33. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. *et al.* 2022. "Training language models to follow instructions with human feedback," in Advances in Neural Information Processing Systems, **35**: 27730-27744.

34. Perez, B., Huang, S., Song, F., Shayea, T., Bras, R., Choi, Y. and Ruder, S. 2022. "Red teaming language models with language models," arXiv preprint arXiv:2202.03286.

35. Mitchell, E., Lin, C., Bosselut, A., Finn, C. and Manning, C.D. 2022. "Fast model editing at scale," in Proceedings of the International Conference on Learning Representations.

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

37. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. *et al.* 2020. "Transformers: State-of-the-art natural language processing," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38-45.

38. Ribeiro, M., Wu, T., Guestrin, C. and Singh, S. 2020. "Beyond accuracy: Behavioral testing of NLP models with CheckList," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4902-4912.

39. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E. *et al*. 2021. "On the opportunities and risks of foundation models," arXiv preprint arXiv:2108.07258.

40. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. and Amodei, D. 2020. "Scaling laws for neural language models," arXiv preprint arXiv:2001.08361.