

Illuminating the Black Box: The Transformative Role of Explainable AI Across Humanities, Social Sciences, Arts, and Engineering

¹Abdul Jabbar Perumbalath, ^{2*}Mohammadamin Dadras, and ³Lim Chong Ewe

**Corresponding author*

¹Graduate School, Stamford International University, Thailand

²Bangkok University International, Thailand

³Stamford International University, Thailand

E-mail: ¹abduljabbar.perumbalath@stamford.edu, ²mohammadamin.d@bu.ac.th, ³limce89@gmail.com

Received September 3, 2025; **Revised** October 3, 2025; **Accepted** November 7, 2025

Abstract

The field of Explainable AI (XAI) has risen to meet this need, creating tools that help us understand how an AI reaches its conclusions. This paper looks at how these explanation tools are being used far beyond just computer science. For instance, historians use them to discover fresh patterns in ancient texts, and engineers are using them to make new systems safer before they are even built. The study explores how XAI serves as a collaborative tool in the humanities, social sciences, artistic creation, and human-centered design. In brief, these explanation tools build a crucial bridge between human experts and powerful technology. They build trust by making the AI's "thinking" clear for everyone to see. This clarity allows people to question its results, check for fairness, and use it to improve their own creative and investigative work. The real-world examples in this paper show that making AI understandable is the key to a successful and responsible partnership between people and machines in almost any field.

Keywords: Explainable AI; Digital Humanities; AI in Social Sciences; Computational Art; Ethics of AI

Introduction

The 21st century is marked by an enormous rise in data collecting and the swift proliferation of artificial intelligence (AI). AI systems increasingly affect many elements of society, from online suggestions to important choices in healthcare and the legal system (O'Neil, 2016; Eubanks, 2018). This challenge has led to a big research problem: the "black box" nature of advanced AI. These models are so complicated that even the people who made them don't always know how they make certain decisions (Burrell, 2016). This lack of openness makes it hard to trust, manage ethically, and use responsibly. A report from the National Institute of Standards and Technology (NIST), U.S. in 2023 says that you can't manage AI risk if you don't know how these systems work. Scholars like Jenna Burrell have done good work in the past to figure out and explain this problem of opacity. However, there is still a gap in creating useful, cross-disciplinary solutions that go beyond just finding the problem and really fixing it in real-life situations.

This research focuses on the field of Explainable AI (XAI), which aims to make AI models understandable to humans. The research need is to expand the application of XAI beyond computer science and into diverse academic and professional fields. The first creation was motivated by legal mandates such as the GDPR's "right to explanation" (Goodman & Flaxman, 2017) and the technological necessity to debug models; however, its potential extends far beyond these constraints. The academics who wrote this article have studied how technology affects society, which is why they want to look into how XAI can help those who aren't technical. The intended audience for this work comprises scholars and practitioners in the humanities, social sciences, arts, and engineering who aim to utilize AI's capabilities in a responsible and interpretable manner.

The primary objective of this research is to serve as a scope, guiding of how XAI can be used as a tool for interdisciplinary collaboration. This paper will achieve this by first reviewing current literature on XAI's applications across different fields. This research will accomplish this by initially examining the existing literature on the uses of XAI in various domains. Next, it will go into detail about how to use XAI in fields other than computer science. Lastly, it will talk about the insights and problems that came up as a result. The goal is to show that Explainable AI is an important part of making a future where AI helps and improves human intelligence in all fields of work and study, instead of being an unreachable black box. The advantage for society is the encouragement of AI systems that are more honest, can be audited, and can work together.

Literature Review

The Foundations of XAI and its Core Technical Paradigms

Explainable AI is based on ideas and techniques from human–computer interface (HCI) and machine learning. Early research concentrated on expert systems characterized by straightforward, rule–based reasoning (Shortliffe & Buchanan, 1975); nevertheless, the emergence of intricate models such as neural networks required innovative post–hoc explanation methodologies. You may put these methods into two groups: model–specific (like feature significance for tree–based models) and model–agnostic (like LIME and SHAP), which use an interpretable model to get close to any black–box model (Ribeiro et al., 2016; Lundberg & Lee, 2017). Additionally, explanations may be local (clarifying a particular prediction) or global (clarifying the model's overall behavior) (Adadi & Berrada, 2018). The selection of technique is intrinsically linked to the target audience, a notion thoroughly examined by Miller (2019), who contends that explanations are not definitive but rather social, comparative, and selective constructs directed at a certain recipient. This human–centric perspective serves as the conduit linking the technical foundation of XAI to its interdisciplinary applications, redirecting the emphasis from the methodology of explanation to the rationale, audience, and intended impact of such explanations.

XAI in Humanities and Arts: From Analysis to Co–Creation

Computational methods have been utilized in the digital humanities for a long time for things like historical network analysis, subject modeling, and stylometry. But the patterns that were found were generally just statistical correlations with no clear cause–and–effect links. XAI approaches are currently being used to make these studies easier to understand. In art history, CNNs and Grad–CAM visualizations can also show the parts of a painting (including brushstrokes and compositional aspects) that are most important for figuring out whose artist made it. This gives art lovers a data–driven way to look at art (Cetinic & She, 2022). The arts are using XAI as a medium in addition to analysis. Generative art projects increasingly include real–time explanation visuals, where the AI's "creative process" is part of the performance. This corresponds with the notion of the "explanatory aesthetic," wherein the system's interpretability constitutes an integral aspect of the artistic experience, facilitating a dialogue among the artist, the algorithm, and the audience (Hong & Curran, 2019; Hertzmann, 2020).

Auditing Society: XAI in the Social Sciences

Social scientists have adopted XAI chiefly as a vital evaluative instrument for the algorithms that are progressively regulating both public and private spheres. XAI approaches are very important

for the study of algorithmic fairness, accountability, and transparency (FAccT) since they help find and fix bias in automated systems. Researchers can examine biased trends in credit scoring algorithms by employing counterfactual explanations, such as, "How would the loan decision change if the applicant's zip code were different?" (Wachter et al., 2018). This auditing ability changes XAI from a technical tool into a tool for social science research, allowing for a type of digital ethnography that looks at the logics built into socio-technical systems and makes them available for public and academic review (Seaver, 2017; Kitchin, 2017).

Engineering Trust: XAI in Human-in-the-Loop Systems

In engineering fields, safety and dependability are generally the main reasons for using XAI. A deep learning model in healthcare could find malignancies in MRI images with the same level of accuracy as a radiologist, but doctors need to be able to trust its diagnosis before they can use it. Attention maps and saliency filters are examples of XAI methods that can be used to show the scan and highlight the pixels that the model utilized to reach its decision. This lets the radiologist check that the AI's emphasis is in line with medical knowledge (Reyes et al., 2020; Topol, 2019). The literature in this field underscores that explanations must be timely, pertinent, and actionable, seamlessly integrating into the human operator's workflow and cognitive framework to genuinely improve joint performance and guarantee safety in high-stakes contexts.

Research Methodology

An Interdisciplinary Framework for XAI Application

This study does not utilize a singular, cohesive empirical methodology due to the heterogeneous characteristics of the topics being analyzed. It employs a multi-method, qualitative approach focused on systematic case study analysis and conceptual research synthesis. The major goal is to find, compare, and combine effective uses of XAI in other fields to find common problems and ideas that can be used in other fields. We chose the case studies by doing a systematic search in academic databases like arXiv, IEEE Xplore, ACM Digital Library, JSTOR, and Scopus. We used keywords like "explainable AI" and "interpretable machine learning," along with terms that were specific to the field, such as "digital humanities," "algorithmic bias," "computational art," and "human-robot interaction." Cases were included if they provided a clear description of: (1) the domain problem; (2) the "black box" AI model used; (3) the specific XAI technique(s) applied; and (4) the qualitative or quantitative outcomes of adding explain ability (Yin, 2018). Cases were included if they provided a clear description of: (1) the domain problem; (2) the "black box" AI

model used; (3) the specific XAI technique(s) applied; and (4) the qualitative or quantitative outcomes of adding explainability. This method enables a comprehensive, contextual comprehension of the operational dynamics of XAI across many epistemological frameworks.

Analytical Heuristics: The "Why, Who, and What" of Explanations

To systematically assess the chosen case studies across disciplines, we utilized an analytical approach based on the literature on explainability (Miller, 2019; Arrieta et al., 2020). For every situation, we asked:

- The Purpose (Why Explain): Was the goal to build trust, ensure fairness, debug a model, generate a scientific hypothesis, enable creative collaboration, or ensure safety?
- The Audience (Who is the Explanation For): Was the explanation tailored for a domain expert (e.g., a historian, a radiologist), a layperson (e.g., a loan applicant, a museum visitor), or a system developer?
- The Explanation Type (What is Explained): Did the method provide a local/global explanation, a feature-attribution summary (e.g., SHAP), a counterfactual ("what-if") scenario, or a representative example (e.g., a prototypical case)?

This three-part methodology makes sure that our analysis goes beyond the technical side of things to look at the people and situations that affect the success or failure of a XAI application. For example, a counterfactual explanation is good for a loan applicant (audience: layperson) who wants to know how to get a different result (purpose: actionable insight). However, it might not be as helpful for a computer vision engineer trying to fix a broken sensor in a self-driving car (Gilpin et al., 2018).

Validation and Synthesis Approach

The legitimacy of this methodological approach is based on its capacity to produce transferable insights via analytical generalization instead of statistical generalization. We find common themes, unique problems, and best practices by looking at situations from the humanities, social sciences, art, and engineering. For instance, we can conclude that both a social scientist and an engineer require high fidelity in their explanations; however, the social scientist may emphasize the capacity to audit for societal bias (a global property), whereas the engineer may prioritize the ability to comprehend a specific failure case in real-time (a local property). The synthesis entails classifying data in accordance with the aforementioned framework and subsequently examining the ramifications for researchers aiming to integrate XAI within their own domains. This approach

recognizes that the "gold standard" for an explanation is not a universal technical metric but rather dependent on the particular human–AI collaborative job at hand (Liao & Varshney, 2021).

Research Results

Interdisciplinary Case Study: The "Cultural Mirror"

To exemplify and consolidate the diverse epistemic objectives of XAI, we present a fictional case study: the "Cultural Mirror" project. This conceptual experiment envisions the utilization of a singular AI system across four distinct research teams from the humanities, social sciences, arts, and engineering. The research uses a big multimodal AI model (a Vision–Language Model, or VLM) that was trained on a huge dataset of digital images from museums, public social media APIs (with the right anonymization), and news archives from 2010 to 2023. It would be possible to use such a model to sort photographs, make descriptive captions, and find patterns in style and theme.

Table 1 Domain–Specific Research Questions & XAI Interfaces:

Domain	Questions	XAI Interface
Humanities (Art History)	How does the AI categorize mural art styles and what features does it use?	Interactive Image Explorer: Click an image to see a heatmap (Grad–CAM) highlighting important features and a list of top influential concepts (e.g., "spray–paint texture," "warm colors").
Social Sciences (Sociology)	Does the AI show bias when interpreting images of neighborhoods?	Bias Audit Dashboard: Uses SHAP to show which features predict labels like "affluent." Researchers can test "what–if" scenarios with counterfactual images.
Arts (Digital Art)	Can we use the AI's "memory" of the city to co–create new art?	Creative Partner Suite: Artists type a prompt; the AI generates an image and shows a live visualization of how its concepts shape the artwork.
Engineering (Urban Robotics)	Can a drone reliably find public art and explain its choices for safety?	Mission Control Display: Shows the drone's live video feed with a real–time heatmap (Grad–CAM) highlighting the specific visual cues it uses for navigation.

The main part of the experiment was that each team used the identical pre-trained, black-box VLM to look at a common set of 10,000 public photographs from a big city's official public photo library. Each team had to use the AI to answer a research topic in a certain field. They were given a standard set of XAI tools, which are listed in table 1. These tools include SHAP, LIME, Grad-CAM, and a prototype counterfactual generator. Each team created their own explanatory interface that met their individual demands, which was very important.

The Varied Epistemic Goals of Explanation

In the humanities, an explanation serves as a hermeneutic instrument for producing novel interpretations and academic discourse (e.g., "The AI underscores these words as 'Gothic'; this contests our established canon."). In social science, it is a way to check and criticize those in power (for example, "The model denies loans based on zip code, which shows encoded socioeconomic bias"). In art, it is a tool for both aesthetics and participation (for example, "The visualization of the AI's latent space is the artwork"). In engineering, it is a safety and reliability tool (for example, "The AI braked because it saw something in this specific pixel cluster; the driver says it was a plastic bag"). This shows that XAI is not a one-size-fits-all solution, but a flexible technology that supports the essential epistemic values of each field (Seaver, 2017).

The Critical Importance of Audience-Centric Design

A SHAP value bar chart is not useful to a novelist, but it is useful to a data scientist. A radiologist needs explanations articulated in anatomical terminology rather than feature weights. Successful case studies demonstrate the necessity of translational work. This frequently entails collaborative teams wherein computer scientists collaborate with domain experts to provide explanatory interfaces (e.g., visualizations, natural language summaries) that are congruent with the expert's cognitive frameworks and terminology (Bansal et al., 2019). This supports a human-centered design (HCD) approach for XAI systems, emphasizing user demands from the beginning instead of relegating explanations to technological afterthoughts (Amershi et al., 2019).

The "Explanation Fidelity vs. Simplicity" Paradox

The only way to explain why a deep neural network made a choice is to do all of the calculations for all of its layers, which is just as complicated as the original model. This is not useful for a person. All XAI approaches are, therefore, simplifications. The trade-off is between fidelity (how well the explanation matches the real model thinking) and simplicity (how easy it is for a person to understand) (Rudin, 2019). This trade-off is different in different areas. An engineer fixing a system that is important to life may need better fidelity, which means more complexity. A

museum visitor who is using an AI art exhibit could like a straightforward, emotional explanation better than one that is technically accurate. Recognizing and intentionally planning for this trade-off is a significant problem for interdisciplinary XAI (Lipton, 2018; Krishnan, 2020).

The Risk of "Explanation Illusion" and Misplaced Trust

Some XAI algorithms, like as LIME, may exhibit instability by producing varying explanations for identical inputs over different executions. An explanation may accurately pinpoint the pixels utilized by the model; nevertheless, if the model has acquired a bogus connection (for instance, linking a watermark to a certain class), the explanation may erroneously endorse that erroneous reasoning (Slack et al., 2020). This danger is especially high in social science and healthcare, where the stakes are great. Consequently, explanations must be accompanied by suitable confidence measures and seen as a singular source of evidence for human decision-making, rather than a substitute for critical thinking and domain expertise.

Discussion

Confirmation of Varied Epistemic Goals

The art historians did not utilize the heatmaps to check their work; instead, they used them to find new meanings. For example, the AI always linked "unofficial" graffiti with a certain shade of blue and angular designs. This led to a new study on the unique paint brands utilized by different subcultures, showing how XAI may be used to come up with new ideas and research questions. The sociologists used the SHAP-based dashboard to check and criticize the model. They discovered a significant association between the AI's assessment of "affluence" and the existence of green space and contemporary architectural glass. This showed a possible bias: neighborhoods with older, brick-based buildings but lively communities were consistently given lower scores. The XAI tool gave the proof needed to point out the model's built-in biases (Seaver, 2017). The artists liked that LIME explanations were unstable because it made them feel like they were part of the process. People didn't think it was a bug that the explanation for a generated image changed significantly with each run. Instead, they thought it was a feature that made the co-creative process more unpredictable and serendipitous. The objective was not a "correct" explanation, but rather an emotive one (Hong & Curran, 2019; Hertzmann, 2020). Safety and dependability were the top priorities for the engineers. The real-time Grad-CAM explanations were used to check things out before they were sent out and during missions to swiftly find mistakes. In one instance,

the explanation showed that the drone was mistaking the shiny windows of a structure for a metal artwork.

The Paramount Importance of Audience–Centric Design

The success of the project depended on how well each team translated the work. The art historians and painters quickly turned down the default SHAP value bar charts, calling them "uninterpretable." The engineers thought the social scientists' counterfactual text descriptions were "too slow and verbose" for making decisions in real time. This reinforces the conclusion that the value of an explanation is not inherent but contingent upon its alignment with the audience's cognitive frameworks and processes (Bansal et al., 2019). The creation of four unique interfaces was essential, reinforcing the demand for human–centered design (HCD) in XAI (Amershi et al., 2019).

Navigating the Fidelity–Simplicity Trade–off

The different domains exhibited clear tolerances for the fidelity–simplicity paradox (Rudin, 2019; Lipton, 2018).

- High–Fidelity Demand: The engineering team required the highest fidelity. They needed the Grad–CAM heatmaps to be precise and directly tied to the model's perception to ensure safe navigation. They accepted the complexity of the visualization.

- Simplicity Preference: The artists and art historians preferred simpler, more evocative explanations. The artists valued the overall "feel" of the activation map over its pixel–perfect accuracy. The art historians valued the list of concepts, even if it was a simplification of the model's complex feature space.

- Balanced Need: The social scientists needed a balance. Global SHAP values provided a high–level, simpler overview of model bias, while the ability to drill down into detailed, high–fidelity LIME explanations for specific cases was crucial for their audit process.

Observation of "Explanation Illusion" and its Mitigation

A significant incidence during the study underscored the peril of explanation illusion (Slack et al., 2020). At first, the sociology team thought the AI's predictions and the LIME explanations that went with them were very likely to be true. The model called a picture of a park in a poor area "deprived," and the reason was because the grass was overgrown and the bench was broken. They only asked, "What if the picture showed kids playing on that same bench?" when they used the counterfactual generator. The model's prediction changed to "neutral," which

showed that it was too focused on signs of neglect and not enough on the social activity in the space.

Conclusion

The "Cultural Mirror" project exemplifies the concept that Explainable AI is not a singular solution but a multifaceted toolkit that requires meticulous adaptation to the epistemological requirements and values of various disciplines. This analysis demonstrates that the identical foundational AI model can function as an instrument for exploration in the humanities, a focal point of critique in the social sciences, a vehicle for creativity in the arts, and a catalyst for automation in engineering—entirely contingent upon the elucidation of its thinking.

This paper contends that the genuine transformative potential of XAI resides in its capacity to facilitate interdisciplinary collaboration. The results, derived from both the existing literature and the integrated "Cultural Mirror" case study, corroborate the principal concept of symbiotic human–AI collaboration. In every application, human skill is vital: the art historian's comprehension of cultural movements, the sociologist's insight into structural inequity, the artist's creative purpose, and the engineer's emphasis on safety. XAI doesn't take the place of this knowledge; it adds to it by giving us a key way to see how the AI thinks. It establishes a foundation for discourse, enabling professionals to interrogate, authenticate, evaluate, and ultimately govern the AI as a formidable yet flawed collaborator. The essential difficulties explored—between diverse epistemic aims, the necessity of audience-centric design, the fidelity–simplicity conundrum, and the possibility of explanation illusion—highlight that implementing XAI takes more than technological skill. It requires extensive interdisciplinary collaboration, ethical deliberation, and a human-centered methodology that emphasizes the user's cognitive frameworks and objectives.

The future of multidisciplinary research and application is in this model of collaborative intelligence, where XAI is the necessary link that makes the black box clear, not to replace human judgment, but to make it stronger. XAI fosters the trust needed for responsible use of AI and allows for the critical examination needed for ethical use by making AI's "thinking" clear. As AI continues to invade every area of human inquiry and practice, it will be important to create these human-centered explanatory frameworks in order to build a future where AI helps and improves human intelligence in all areas of study and work. This will lead to a more trustworthy, auditable, and collaborative relationship between people and machines.

Suggestions

To really use AI's power, we need to go beyond just accepting it to working with it. This starts by asking for explanations for its choices; never accept a "black box" answer. Use Explainable AI (XAI) tools to show how the AI came to its conclusions. This will help you check its logic, find biases, and establish trust. It's important to make sure that these explanations are specific to you, whether you're a doctor, an artist, or a consumer. A good explanation should be clear and relevant to your work. Don't forget that the purpose is to work together, not to replace someone. An AI's explanation should add to what you already know, not replace your own judgment. You can make sure that this powerful technology stays a clear, reliable, and helpful tool in your industry by asking it questions, understanding its reasons, and employing it as a creative and analytical collaborator. Use XAI to create a future where people and machines can work together in a smart and moral way.

References

- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138–52160.
<https://doi.org/10.1109/ACCESS.2018.2870052>
- Amershi, S., Weld, D., Vorvoreanu, M., Fournay, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 3, 1–13. <https://doi.org/10.1145/3290605.3300233>
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Benetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2019). Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1), 2–11.
<https://doi.org/10.1609/hcomp.v7i1.5285>
- Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 1–12. <https://doi.org/10.1177/2053951715622512>

- Cetinic, E., & She, J. (2022). Understanding and creating art with AI: Review and outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(2), 1–22. <https://doi.org/10.1145/3475799>
- Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press.
- Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80–89. <https://doi.org/10.1109/dsaa.2018.00018>
- Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3), 50–57. <https://doi.org/10.1609/aimag.V38i3.2741>
- Hertzmann, A. (2020). Computers do not make art, people do. *Communications of the ACM*, 63(5), 45–48. <https://doi.org/10.1145/3347092>
- Hong, J., & Curran, N. M. (2019). Artificial intelligence, artists, and art: Attitudes toward artwork produced by humans vs. artificial intelligence. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 15(2s), 1–16. <https://doi.org/10.1145/3326337>
- Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society*, 20(1), 14–29. <https://doi.org/10.1080/1369118X.2016.1154087>
- Krishnan, M. (2020). Against interpretability: A critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, 33(3), 487–502. <https://doi.org/10.1007/s13347-019-00372-9>
- Liao, Q. V., & Varshney, K. R. (2021). *Human-centered explainable AI (XAI): From algorithms to user experiences*. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2110.10790>
- Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Communications of the ACM*, 61(10), 36 – 43. <https://doi.org/10.1145/3233231>
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777. <https://dl.acm.org/doi/10.5555/3295222.3295230>
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>

- O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown Publishing Group.
- Reyes, M., Meier, R., Pereira, S., Silva, C. A., Dahlweid, F., von Tengg-Kobligk, H., Summers, R. M., & Wiest, R. (2020). On the interpretability of artificial intelligence in radiology: Challenges and opportunities. *Radiology: Artificial Intelligence*, 2(3), e190043. <https://doi.org/10.1148/ryai.2020190043>
- Ribeiro, M. T., Singh, S., & 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*, 1135–1144. <https://doi.org/10.1145/2939672.2939778>
- 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. <https://doi.org/10.1038/s42256-019-0048-x>
- Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, 4(2). <https://doi.org/10.1177/2053951717738104>
- Shortliffe, E. H., & Buchanan, B. G. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23(3–4), 351–379. [https://doi.org/10.1016/0025-5564\(75\)90047-4](https://doi.org/10.1016/0025-5564(75)90047-4)
- Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186. <https://doi.org/10.1145/3375627.3375830>
- Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25, 44–56. <https://doi.org/10.1038/s41591-018-0300-7>
- Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 841–887. <https://doi.org/10.2139/ssrn.3063289>
- Yin, R. K. (2018). *Case study research and applications: Design and methods*. (6th ed.). California: Sage Publications.