WHEN IS EXPLAINABLE AI USEFUL?

Scott Robbins scott.robbins@kit.edu
Karlsruhe Institute of Technology, Department of Philosophy, Faculty of Humanities and Social Sciences, - Karlsruhe, Germany, ORCiD 0000-0002-5338-295X

Article type: Research article

Review process: Double-blind peer review

Topical Collection: Ethics and Normativity of Explainable AI: Explainability as a Social Practice (Guest Editors: Tobias Matzner, Suzana Alpsancar, Martini Philippi, Wessel Reijers)

DOI: 10.59490/jhtr.2025.3.8164
ISSN: 2773-2266
Submitted: 8 May 2025 Revised: 17 September 2025 Accepted: 6 November 2025
Published: 30 December 2025

How to cite (APA): Robbins (2025). When is Explainable AI Useful? *Journal of Human-Technology Relations*, *3*(1), pp.1-13. https://doi.org/10.59490/jhtr.2025.3.8164.

Corresponding author: Scot Robbins

Keywords

Explainable AI; AI Ethics; xAI; Artificial Intelligence

Abstract

In this paper I assess the ethical and epistemic utility of explainable AI algorithms. I first distinguish between different types of outputs that AI can have. The first class of outputs is verifiable (either through a third-party or in virtue of contributing to a win in a game scenario) – that is, there is a way to independently verify the outputs of the model. The second class of outputs is non-verifiable and include outputs like ideals (finding the best of something) and generative AI. While some epistemic value is gained by explaining the outputs of verifiable AI, I argue that explanations created by xAI are unlikely to have any ethical value. Therefore, if there is an ethical problem with the use of opaque AI systems, explainable AI will not be able to help solve it.

1 INTRODUCTION

There is much discussion on the notion of explainability or 'explicability' for AI. It has been suggested as a principle of AI along with the four bioethics principles (Floridi et. al. 2018). Researchers have debated its utility as an ethical principle (see e.g. Robbins 2019, Herzog 2022). It is unclear though, exactly, what kinds of cases the explanations offered by explainable AI would be useful.

In this paper I distinguish between verifiable and non-verifiable AI outputs. In some cases, we can independently verify the outputs via a third party. For example, when an algorithm classifies an image as containing a cat, we can examine the image and verify whether that is the case. In other cases we can verify the outputs because they occur in 'game scenarios'. That is, we can verify that a particular output is good by virtue of its consequences within a context that has a clear win or loss. The output is assessed as correct simply in exactly those cases in which the game is won. The clearest case is that of a game such as chess. An algorithm that suggests moves for a chess game will be good or bad in virtue of its ability to win a game of chess.

For non-verifiable cases, there is, first, a class of outputs that I will call 'ideals'. This class of outputs tries to find the best of something or the most likely. For example, a hiring algorithm might pick the best candidate for a job. This is not something that we can verify independently. Even if the candidate turned out to be a great employee, there may have been a better candidate. Finally, the outputs of generative AI can also not be verified. Although "truth" is important, and assertions made by generative AI must be verified, the goal is creative output that is similar to art and cannot be verified as correct or incorrect.

The point of distinguishing between these cases is to understand when explainable AI would be needed and what it would be needed for. From there, we can assess whether explainable AI could serve that need. I will first argue that there is a fundamental problem regarding the evaluation of an explainable AI method's efficacy. I will then argue that while there is an epistemic purpose that could be served by explainable AI methods (if we could determine that they were effective), there is never an ethical need for the explanations that are possible to provide.

2 EXPLAINABLE AI

It is important to note that explainable AI, as we understand it today, is a result of contemporary machine learning methods that are inherently opaque (Robbins, 2019). Old-fashioned AI was easily able to explain itself. While the output could result from a complex decision-making process that appeared unexplainable, the logic underlying it was explicitly programmed into the algorithm by humans. One could trace the decision tree resulting in the output (Rudin, 2019). While machine learning methods are far more powerful (in some contexts) than old-fashioned AI, their outputs are not the result of encoded human reasoning; rather, through training on many examples, the algorithm is able to (loosely speaking) encode its own reasoning.

This reasoning, however, is not like human reasoning. It is reasoning that looks like weighted nodes that are a part of hidden layers, connected to other layers. One could "open up" the algorithmic model and see these layers, nodes, and weights; however, this would do little to tell us how a particular output was reached. Even the programmers themselves would be lost.

This, of course, can be quite concerning when these algorithms are used in a way that will significantly affect the lives of human beings. Algorithms used to evaluate loan applications, predict crime (Perry, 2013), and evaluate job performance (Ajunwa, 2016) could directly impact the well-being of individual people. It is intuitive that one would demand an explanation for

these outputs. Many have made such demands (AI Principles, 2017; Floridi et al., 2018; The Public Voice, 2018; Vollmer, 2018). Others disagree (Cortese et al., 2023; Robbins, 2019).

Below, I want to highlight some of the proposals made for achieving explainable AI. All types of explanations are not equally useful depending on the context. If explanations are to be useful, they must be fit for purpose.

2.1 EXPLAINABLE AI METHODS

2.1.1 Local vs. Global

Local methods attempt to explain individual outputs of a machine learning system. They answer the question which features of a specific input led to the output in question? Global methods attempt to explain what features the model generally considers (Dwivedi et al., 2023). In the former, the object of consideration is the output. How did we come to this particular output? In the latter, the object of consideration is the model. How does the model generally classify things?

For the purposes of this paper, only local methods will be considered, as our goal here is to understand when an explanation will help us evaluate an output. Users are not concerned with how an application and its model work in general, but why its output is correct. The user has a responsibility for the outputs and their consequences. Knowing how the model generally works may be interesting for those choosing one model over the other, but for considering particular outputs, it would be less useful to know how the model generally works than how it worked this particular time.

If an algorithm classified someone as “very bad” in terms of job performance (with the probable consequence that they will be let go), then it is useful to have a local explanation that tells us how this particular classification was achieved. It is a lot less useful than knowing that the algorithm generally uses things like the number of emails sent and customer satisfaction surveys. Before firing someone, we would want to know which considerations were in fact used.

2.1.2 Ante-hoc vs. Post-hoc

Ante-hoc methods are based on models that are intrinsically explainable (Retzlaff et al., 2024). Decision trees are the most obvious example here. Any output of a decision tree model can be explained by tracing the route taken to achieve the output. For example, a loan approving algorithm based on a decision tree might have if-then statements like: if the applicant has a salary of more than 5000 euros a month, then approve the loan; otherwise, if the applicant has a salary of more than 3000 euros a month and no debt, then approve the loan; otherwise, reject the loan. This simple (but clearly problematic) algorithm would be easily explainable. Any real algorithm designed to approve loans should be much more complicated – and have many more if-then statements. However, any output would be intrinsically explainable. These explanations would all be local explanations – that is, they would explain the particular output.

Post-Hoc explainable AI methods are attempts to use a different black box model to gain an understanding of how the first model reached the output it did. The most popular methods that fall into these categories are Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) and Shapley Additive explanations (SHAP) (Lundberg & Lee, 2017). Both of these methods assign values to input variables, which show how important they are for determining a specific output. LIME and SHAP are themselves machine learning models that explain individual outputs in a model-agnostic way. That is, they can work on any machine learning model – no matter the methodology (Gramegna & Giudici, 2021).

The whole purpose of explainable AI methods is to overcome the problem of models where the explanation is not intrinsic. Therefore, this paper will not be looking at ante-hoc explainability but only Post-Hoc explainability.

3 AI OUTPUTS

For the purposes of understanding when explanations are useful, I will divide the types of AI outputs into verifiable and non-verifiable outputs. There is a relevant difference between evaluating an output that can be independently verified and one that cannot. This is true outside of AI as well. If I ask my daughter if she cleaned her room this morning, and she says 'yes', I can simply go and check if her room is clean or not. Or, after asking her this question many times and getting a truthful answer, I could believe that she is reliable about this and need nothing further for my belief that her room is clean. However, if I ask her how school was today and she says 'great', then I have no way of determining whether it was great or not. I might ask her what made it great – that is, I would like an explanation so that I could better understand her labelling the school day 'great'.

I will further split verifiable AI outputs into those that can be verified by third-party means and those that can be verified in virtue of the consequences of their output. The latter is what I will call 'game scenarios', as the outcome of the game will help determine whether a move was 'correct'.

3.1 VERIFIABLE

3.1.1 Third-Party Verifiability

The explosion of AI is really a result of these types of outputs. When machine learning gave machines the power to classify text, images, and a host of other things, AI took off as a field. Things that couldn't be automated before could now be automated. More importantly, advertisers could make predictions about what ads consumers would click on, making the use of AI highly profitable.[1]

One of the reasons that this was so interesting was that we could say how good these algorithms were. We could claim how well they were able to, for example, classify images that had cats in them. More practically, moles could be classified as cancerous or not (with varying accuracy depending upon skin color). We could extensively test them and say something about their efficacy (Robbins, 2025).

The important thing about these types of classifications (for the purposes of this article) is that they are verifiable. That is, for any particular output, we can verify whether the algorithm was correct or not. If the algorithm classifies an image as having a cat in it, but there is no cat in the image, we can see that it was incorrect. If an algorithm classified a mole as cancerous, but after a biopsy, it is proven to be benign, then we know the algorithm was wrong.

This is not true for all classifications. To take an extreme example, classifying an image as 'beautiful' cannot be verified in any meaningful way. Reasonable people will disagree about what is beautiful.

3.1.2 Game Scenarios

Sometimes outputs by AI can be verifiable in virtue of the end result being so clear. An algorithm recommending a move in Chess or Go, for example, can be somewhat evaluated in virtue of its leading to a win. A surprising move, like the famous move 37 in game two between the then world champion Lee Sedol and DeepMind's AlphaGo algorithm (Silver et al., 2016), was so unusual that it was estimated to have only a 1 in 10,000 chance of being played. That is, humans believed the move, in the moment, to be strange at best, and bad at worst. However,

[1] The reader may not immediately recognize this as a type of classification; however, it is simply an advertisement being categorized as 'will be clicked on' for a particular user, then one can verify how well the algorithm does.

as the game continued, it was clear that this move was genius. The consequences of that move within the game scenario proved the move to be a good one (Metz, 2016, p. 37).

This isn't true in situations outside of games. There often aren't clear win or lose conditions, and when there are, they are often dependent on so many variables that it is impossible to say that any particular variable contributed in a positive or negative way. We often speak loosely of playing 'the game' when it comes to things like job interviews. Many will have advice about what you have to do to succeed in an interview. Of course, even if you receive the job, and therefore 'win' the game, it is difficult to say that, for example, wearing a suit and tie to the interview helped or hurt your chances. Major consultancies have questions[2] like "estimate the total number of gallons of gasoline a typical gas station pumps in the United States on a typical weekday". They expect you to come up with a method of answering that question and to offer an answer. If I asked that question in an interview, anything other than "I have no idea" would ensure that you were not selected. The point of this example is that the outcome depends on people who are outside of your control.

3.2 NON-VERIFIABLE AI OUTPUTS

3.2.1 Ideals

Often, algorithms classify people or things in such a way that we cannot verify the outputs. For example, when a popular AI product classifies a person as being a "good candidate", it is not possible to verify this output as being correct. 'Good' is not objective in this case. Hiring committees argue about which characteristics/achievements/etc should be used to evaluate candidates for a particular job. Different committees can reasonably disagree on the weight and importance of specific considerations. There may be differing opinions on, for example, whether using someone's h-index is a good indicator of future academic performance. There is a growing movement to move away from quantity metrics (like the h-index) towards quality metrics - like having candidates submit a writing sample (Chapman et al., 2019). Maybe some combinations of the two. The point is that these considerations are the core object of the debate - not the outcome. This will be important for understanding whether explainable AI is useful for these types of outputs or not.

One might argue that the output is verifiable in the sense that if the chosen candidate performs well in their job, then the output was correct. However, this ignores a couple of things. First, that candidate may have been a terrible candidate who happened to do very well in their job. Second, there may have been a better candidate in the pool that was not chosen.

3.2.2 Generative AI

Generative AI differs fundamentally from the classification algorithms described above. Its ability to generate coherent prose, images, and even videos doesn't lend itself to the same criteria of "success". The possible criteria for a successful text output are extremely varied: From creating good documentation for my software prototype to making someone fall in love with me. If a text serves my purpose, I can consider it successful. In contrast to "Ideals"-outputs the considerations that lead to generating the text are not relevant as long as it serves its purpose.

Again, one might argue that texts must (and can) be verified. But what is being verified is not the text itself, but the assertions made in it (Robbins & Blundell, 2025). The goal is to prevent the use of text that has factual errors in it. This is not the same as verifying whether the output is "correct".

[2] Stupid questions

4 UTILITY OF EXPLAINABLE AI

We must first acknowledge that the whole point of machine learning is to reach a classification that we struggle to create ourselves. We can't easily explain what features an image must have that depicts a cat. Patterns detected by machine learning algorithms would be inarticulable in human language. Any explanation would thus be a mere effort to translate the actual considerations and logic of the algorithm into human articulable ones. The problem is, however, that we are not able to evaluate such translations as we do not know the source language. If we did, we could translate ourselves. So, for all we know, the explanations could be the result of extremely poor translations. The first subsection will consider the issue of evaluating the efficacy of explainable AI.

Bracketing this problem, it would be good to know when an explanation for a particular output would be useful or not. Importantly, it would be good to know how, exactly, it would be useful. As explanations would provide information on how a particular output was achieved, there is some epistemic utility in such explanations. Taking a real-world example of AI, chess commentators and grandmasters, when analyzing ongoing games, use an AI chess engine to see what the next best moves are. Oftentimes the engine recommends a move and the commentators/grandmasters ask: "What is the idea there?". They can't immediately understand why the algorithm has recommended that particular move. When grandmasters recommend a move, they can give a clear explanation of why the move is a good one (e.g. it threatens a fork, takes space in the center, and discovers an attack on the rook). This helps people learn what to look for in future games. The algorithm fails to do this, leaving us without helpful information to apply to future cases. The issue of epistemic value will be discussed in the second subsection below.

In the last subsection (4.3), the issue of the ethical utility of explainable AI will be discussed. I argue that because the considerations used to make an output should be the object of what we are evaluating, there is no ethical value given by explainable AI. We are simply adding an opaque process on top of an opaque process – without any way to judge its effectiveness.

4.1 EFFECTIVENESS

How do we evaluate whether an explanation is correct? First, I mean 'correct' in terms of whether the considerations used by the model and the weights given to them are represented accurately by the explanation. The point is to have the information needed to evaluate whether the explanation is a good one. For that evaluation to be worth doing, it is important that the explanation accurately represents how the model arrived at the output.

The first difficulty is that the purpose of using machine learning models is to detect patterns that would be inarticulable (and thus codable) by humans. The first interesting use of machine learning was to identify handwritten numbers (LeCun et al., 1989). The point of doing this was that we (humans) are unable to articulate the specific patterns that could be encoded into rules that would detect handwritten numbers. Numbers are written in so many ways. Things become even more complicated when we consider algorithms that identify objects and people in images. The patterns detected by computers will be complex and non-articulable. Any attempt to pick out specific features will be a translation of these complex non-articulable patterns into human language. The 'correctness' of the algorithm providing an explanation depends on the quality of this translation. But as we cannot know the source (the complex pattern detected by the algorithm), we cannot know how close the translation is to approximating it.

For example, if I say the translation of some German is “have a nice day,” you would not be able to assess the quality of my translation unless you knew what the German words, that I was translating from, meant. That is the state we find ourselves in when using xAI (local, post-hoc).

There are also many concerns that the explanations provided by these methods confuse more than they enlighten. One study took participants who had advanced mathematical and machine learning knowledge and asked them to make classifications – one group had SHAP explanations at hand to help, and the other did not. The results showed that the participants were simply confused by the explanations and did not perform any better (Moreira Cunha & Diniz Junqueira Barbosa, 2024).

There is also a concern that with more and more parameters and data, the xAI algorithms significantly lose performance (Wang et al., 2024). Finally, these methods are computationally costly and won’t be worth using (Salih et al., 2025).

4.2 EPISTEMIC UTILITY

It would be quite interesting for researchers to know which features were important in classifying moles as cancerous for an effective AI model. Unexpected features leading to an accurate cancer diagnosis could help researchers better understand skin cancer. That is, there is an epistemic benefit to such explanations.

This is because a perfect alignment between the model's features and the xAI algorithm's explanations is unnecessary. Simply pointing towards a feature that was heretofore unknown to have any connection to skin cancer could, with testing, prove beneficial to future diagnoses. An algorithm with explanatory power could be a “hypothesis generator” – suggesting things to test further.

There would also be epistemic value of explanations in game scenarios. Those interested in these games will find the explanation of particular moves suggested by a powerful algorithm valuable. It will help them learn to play that game better. If a SHAP model were to tell you that a particular move in the game of chess was chosen because it activated your bishop, defended your king, and controlled more of the center, it would help you to learn how to play better chess. If more game scenarios like protein folding are found, then explainable AI could help us better understand how these mechanisms work.

One could think that the epistemic value could also be found for non-verifiable outputs like ideals. A hiring algorithm may find patterns and suggest features that may have been overlooked by a hiring committee. The algorithm, having selected what it determines to be the best candidate, could show which features most contributed to the output.

However, this is a mistake. Adding an algorithm to the hiring committee like this would just be adding one more voice. When a member says what feature they think is important for the evaluation of candidates, a reasonable question is “why do you think that feature is important?” The whole process is a conversation about justifying the criteria. Simply pointing to a criterion is pointless. Further problems arise when you consider that the algorithm, having been trained on past data, will inherently be biased towards features that resulted in people being hired in the past (Robbins, 2023). We have already seen this result in men being selected for management positions simply because they are men (Dastin, 2018).

4.3 ETHICAL UTILITY

To achieve some sort of ethical utility, AI explanations would need to contribute to some ethical value. Values discussed in the literature include: autonomy (Vaassen, 2022), control (Robbins, 2019), safety (Kuznietsov et al., 2024), and trust (Herzog, 2022). A user of an AI system equipped with an explanation for an output could have increased autonomy to make their own decision,

have more control (and therefore responsibility) over the consequences of that output, be able to prevent unsafe outcomes, and trust the system more (which is important if the system works well).

If these values are sufficiently upheld without an explanation, one could argue that the explanation contains no ethical utility. Taking the first class of AI outputs, we can see this quite clearly. When outputs can be independently verified via a third party, then an explanation holds no additional ethical value. An algorithm that classifies an image as containing a cat can be easily verified by looking at the picture. An explanation highlighting the features that lead to the output would add nothing to the user.

If we can already check whether the classification was correct or not, then why would we care about the explanation? For example, if an algorithm identifies a mole as being cancerous, a doctor will most certainly administer a biopsy that would confirm or contradict the algorithm's diagnosis. Then what ethical justification is there for explaining why the algorithm classified that mole as cancerous? (see e.g. Durán & Jongsma, 2021; London, 2019).

Using such an algorithm in practice would, of course, depend on its efficacy. With extremely low efficacy, biopsies would always need to be performed, making the algorithm rather useless. One would hope that such an algorithm would not be used in practice. But at a high efficacy rate there would be a simple way of verifying predictions of cancer - undermining the need for any explanation. Finding out that the mole was classified as cancerous because it had a particular edge pattern would not be very interesting compared to the biopsy verification of the mole being cancerous.

The whole purpose of explainability in this situation would be to give the physician (or the patient) more control over the diagnosis (Robbins 2023). However, all of that control is reached by being able to check the output with a biopsy. The same is true for many other algorithms. Facial recognition algorithms classifying who a particular person is, for example, do not need to provide an explanation if the classification can simply be verified by looking at their faces, checking their ID, etc.

If explanations of outputs in game scenarios were to have any ethical relevance, then there would have to be game-like scenarios that occur outside of games. An algorithm that outputs a terrible move in chess, causing the game to be lost, simply does not have any ethical relevance.

We do often describe certain morally relevant aspects of life as 'a game'. For example, one might say that they are "playing the game" so that they can be promoted. However, this is using this language loosely. Game scenarios occur when the outcome is clear (which this example satisfies), and the contributions to that outcome are well-defined. In a game of chess, the moves are all that contribute to the outcome of a win or a loss. In trying to get promoted, there are so many other variables involved. Knowing that any one action contributed one way or another to getting promoted is difficult.

Ideal classifications provide a more interesting case for explainable AI having ethical value. Without the ability to independently verify the output, it will be difficult for those using the algorithm to have any control over an opaque algorithm. An algorithm output rejecting someone for a loan that doesn't have any accompanying explanation would be extremely difficult for a human to endorse or reject. There wouldn't be enough information to do so.

The explanation would, the idea goes, give the user information about how the algorithm reached its decision, which affords them the control to endorse or reject the decision. A SHAP algorithm (highlighted above), which gives the user the importance of the features of a particular output, could show the user that a feature was heavily used, which should not have been. For example, if the algorithm rejected someone for a loan and the explanation included

the feature “race” being important for the output, then the user would be able to reject the output on ethical grounds.

Of course, the algorithm would most likely not be taking in ‘race’ as an input. We don’t want race to be used at all. The issue is that the algorithm could learn certain proxies for race, like address, name, or some combination of features that we would never know are highly correlated with race. We could determine that some of these are proxies for race and then use an explanation to ensure that this wasn’t being used. However, the entire point of using complex AI models is to discover patterns that we would not be able to figure out ourselves.

Not only will proxies for protected features like race be used, but all kinds of patterns that are simply not understandable to us. We would have to pretend that computers see data and organize it into features the same way a human would. This is not the case. The features picked out by computers will often be features that aren’t articulable to us. A computer doesn’t simply look for a tail and whiskers to determine whether a cat is in a picture. It is so powerful because it is not constrained by human ways of identifying features.

This is the crux of the problem. The entire difficulty in evaluating an ideal output is that the process – or the considerations used – is the object of evaluation. If it were a hiring process without AI, we would look at not only the rubric they used for evaluating potential hires, but the makeup of the committee, and their justifications for not including certain criteria in their evaluations. Using an xAI algorithm to provide an explanation in the AI case would be like using an external committee to come up with what they think was used by the committee, given the inputs and the eventual hire, without ever speaking to the hiring committee. We are placing an opaque process on top of an opaque process and saying that we are increasing our understanding of the output.

The conclusion of all of this is that explainable AI doesn’t offer any ethical value. The explanations offered by explainable AI don’t equip anyone with more autonomy or control. The limits of explainable AI prevent anyone, for example, from ensuring that an algorithm isn’t using some combination of features that, if we understood them, would be unethical to use.

Because the opacity of AI raises real problems, considerable effort is now being devoted to making AI less opaque. Providing explanations is hoped to mitigate these ethical concerns. However, the explanations offered by explainable AI do not align with the explanations that are ethically useful. Consequently, if an opaque AI system poses an ethical problem, explainable AI alone will not be sufficient to resolve it.

5 CONCLUSION

Explainable AI has grown into a major research area, with methods emerging across many application domains. Researchers pursue explainable AI primarily to address ethical concerns that stem from AI’s opacity. The need for an explanation varies with the nature of the AI output—verifiable results often require less justification than non-verifiable ones. While explanations are essential in certain contexts, the ones generated by current explainable AI techniques frequently miss the mark. As this paper demonstrates, some explainable AI outputs have epistemic value—they can highlight features worth further scientific investigation—but they fall short of providing ethically useful explanations. In practice, the straightforward remedy is to avoid black-box models altogether whenever interpretability is a requirement.

Data Access Statement
Not applicable.

Contributor Statement
Scott Robbins is the sole author and prepared the first draft and edited and prepared the final draft.

Use of AI
Not relevant or applicable.

Funding Statement
The authors did not receive any financial support for the work on this article.

Acknowledgments
Thanks to Inga Blundell for her helpful comments and feedback on earlier drafts.

Conflict of Interest
The author reports no conflict of interest.

References
AI Principles. (2017). Future of Life Institute. https://futureoflife.org/ai-principles/

Ajunwa, I. (2016). *Hiring by Algorithm* (SSRN Scholarly Paper ID 2746078). Social Science Research Network. https://papers.ssrn.com/abstract=2746078

Chapman, C. A., Bicca-Marques, J. C., Calvignac-Spencer, S., Fan, P., Fashing, P. J., Gogarten, J., Guo, S., Hemingway, C. A., Leendertz, F., Li, B., Matsuda, I., Hou, R., Serio-Silva, J. C., & Chr. Stenseth, N. (2019). Games academics play and their consequences: How authorship, h-index and journal impact factors are shaping the future of academia. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1916), 20192047. https://doi.org/10.1098/rspb.2019.2047

Cortese, J. F. N. B., Cozman, F. G., Lucca-Silveira, M. P., & Bechara, A. F. (2023). Should explainability be a fifth ethical principle in AI ethics? *AI and Ethics*, *3*(1), 123–134. https://doi.org/10.1007/s43681-022-00152-w

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, *47*(5), 329–335. https://doi.org/10.1136/medethics-2020-106820

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Comput. Surv.*, *55*(9), 194:1-194:33. https://doi.org/10.1145/3561048

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Gramegna, A., & Giudici, P. (2021). SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.752558

Herzog, C. (2022). On the Ethical and Epistemological Utility of Explicable AI in Medicine. *Philosophy & Technology*, *35*(2), 50. https://doi.org/10.1007/s13347-022-00546-y

Kuznietsov, A., Gyevnar, B., Wang, C., Peters, S., & Albrecht, S. V. (2024). Explainable AI for Safe and Trustworthy Autonomous Driving: A Systematic Review. *IEEE Transactions on Intelligent Transportation Systems*, *25*(12), 19342–19364. https://doi.org/10.1109/TITS.2024.3474469

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, *1*(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, *49*(1), 15–21. https://doi.org/10.1002/hast.973

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Metz, C. (2016, March 16). In Two Moves, AlphaGo and Lee Sedol Redefined the Future. *Wired*. https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/

Moreira Cunha, B., & Diniz Junqueira Barbosa, S. (2024). Evaluating the Effectiveness of Visual Representations of SHAP Values Toward Explainable Artificial Intelligence. *Proceedings of the XXIII Brazilian Symposium on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3702038.3702093

Perry, W. L. (2013). *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. Rand Corporation. https://www.rand.org/pubs/research_reports/RR233.html

Retzlaff, C. O., Angerschmid, A., Saranti, A., Schneeberger, D., Röttger, R., Müller, H., & Holzinger, A. (2024). Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. *Cognitive Systems Research*, *86*, 101243. https://doi.org/10.1016/j.cogsys.2024.101243

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Robbins, S. (2019). A Misdirected Principle with a Catch: Explicability for AI. *Minds and Machines*, *29*(4), 495–514. https://doi.org/10.1007/s11023-019-09509-3

Robbins, S. (2023). Recommending Ourselves to Death: Values in the Age of Algorithms. In S. Genovesi, K. Kaesling, & S. Robbins (Eds.), *Recommender Systems: Legal and Ethical Issues* (pp. 147–161). Springer International Publishing. https://doi.org/10.1007/978-3-031-34804-4_8

Robbins, S. (2025). What machines shouldn't do. *AI & SOCIETY*. https://doi.org/10.1007/s00146-024-02169-7

Robbins, S., & Blundell, I. (2025). Losing Our Voice? Generative AI and the Degradation of Human Expression. Minds and Machines, 36(1), 2. https://doi.org/10.1007/s11023-025-09757-6

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

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. *Advanced Intelligent Systems*, *7*(1), 2400304. https://doi.org/10.1002/aisy.202400304

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

The Public Voice. (2018). *AI Universal Guidelines – thepublicvoice.org*. https://thepublicvoice.org/ai-universal-guidelines/

Vaassen, B. (2022). AI, Opacity, and Personal Autonomy. *Philosophy & Technology*, *35*(4), 88. https://doi.org/10.1007/s13347-022-00577-5

Vollmer, N. (2018, September 5). *Recital 71 EU General Data Protection Regulation (EU-GDPR)* [Text]. http://www.privacy-regulation.eu/en/recital-71-GDPR.htm

Wang, H., Liang, Q., Hancock, J. T., & Khoshgoftaar, T. M. (2024). Feature selection strategies: A comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data*, *11*(1), 44. https://doi.org/10.1186/s40537-024-00905-w