Issue Brief

Putting Explainable AI to the Test

A Critical Look at AI Evaluation Approaches

Authors
Mina Narayanan
Christian Schoeberl
Tim G. J. Rudner

CSET CENTER for SECURITY and EMERGING TECHNOLOGY

February 2025

Executive Summary

Policymakers frequently invoke explainability and interpretability as key principles that responsible and safe AI systems should uphold. However, it is unclear how evaluations of explainability and interpretability methods are conducted in practice. To examine evaluations of these methods, we conducted a literature review of studies that focus on the explainability and interpretability of recommendation systems—a type of AI system that often uses explanations. Specifically, we analyzed how researchers (1) describe explainability and interpretability and (2) evaluate their explainability and interpretability claims in the context of AI-enabled recommendation systems. We focused on evaluation approaches in the research literature because data on AI developers' evaluation approaches is not always publicly available, and researchers' approaches can guide the types of evaluations that AI developers adopt.

We find that researchers describe explainability and interpretability in variable ways across papers and do not clearly differentiate explainability from interpretability. We also identify five evaluation approaches that researchers adopt—case studies, comparative evaluations, parameter tuning, surveys, and operational evaluations—and observe that research papers strongly favor evaluations of system correctness over evaluations of system effectiveness. These evaluations serve important but distinct purposes. Evaluations of system correctness test whether explainable systems are built according to researcher specifications, and evaluations of system effectiveness test whether explainable systems operate as intended in the real world. If researchers understand and measure explainability or other facets of AI safety differently, policies for implementing or evaluating safe AI systems may not be effective. Although further inquiry is needed to determine whether these results translate to other research areas and the extent to which research practices influence developers, these trends suggest that policymakers would do well to invest in standards for AI safety evaluations and enable a workforce that can assess the efficacy of these evaluations in different contexts.

Table of Contents

Introduction

U.S. policymakers and technologists alike tend to promote several principles that safe and responsible AI systems should uphold, including explainability and interpretability.[1] In the context of AI, explainability and interpretability are sometimes used interchangeably to mean *the ability of a machine learning system to provide human-understandable explanations of its predictions*.* In other words, explainable and interpretable systems can reveal the logic behind their predictions—a contrast to black-box systems whose inner workings are indecipherable. However, the National Institute of Standards and Technology's (NIST, an agency of the U.S. Department of Commerce) differentiation of explainability and interpretability and previous CSET research demonstrate that different groups may ascribe different meanings to these principles.[2] This can lead to confusion and inconsistent operationalization of principles, especially because a generalizable framework for implementing and evaluating AI explainability and interpretability does not exist.

Senator Schumer described explainability as "one of the thorniest and most technically complicated issues we face, but perhaps the most important of all" when introducing his SAFE Innovation framework regarding the AI issues that Congress should prepare legislation for.[3] The subsequent bipartisan AI Insight Forums featured explainability as a topic, and the accompanying road map for congressional action highlighted the importance of making an AI system's components and functions known.[4] In addition, U.S. government agencies and international initiatives have recognized the need to gather evidence that AI systems perform as expected and are safe and trustworthy.[5] The Federal Trade Commission, for example, has reiterated that it will take enforcement action when companies make unsubstantiated claims about the performance and safety of their AI systems.[6]

Yet, the question of *how* developers will demonstrate the explainability and interpretability of their systems remains. Currently, there is no broadly recognized framework for building or evaluating explainable or interpretable AI.[7] To gain more insight into evaluations of explainability and interpretability methods, we examine how researchers evaluate AI-enabled recommendation systems for explainability and interpretability, in addition to how they conceptualize these principles in the first place. More specifically, we ask:

* Machine learning is a branch of AI and computer science that focuses on using data and algorithms to enable AI to imitate the way that humans learn. See https://www.ibm.com/topics/machine-learning.

1. What descriptive approaches do research papers adopt for explainability and interpretability, and do descriptions of explainability meaningfully differ from descriptions of interpretability?

2. What approaches do researchers use to evaluate explainability and interpretability claims?

Given that data on AI developers' evaluation approaches is not always publicly available, we turn to evaluation approaches in the research literature that can guide the types of evaluations that industry adopts.

We focus our analysis on AI-enabled recommendation systems for two reasons. First, governments around the world have enacted regulations that require companies to explain their black-box recommendation systems. Examples include China's 2021 regulation on recommendation algorithms and the European Union's Digital Services Act.[8] As a result, we expect literature on the explainability of recommendation systems to be available for us to analyze. Second, many recommendation systems aim to help users make decisions. To assist with decision making, these systems can provide explanations of their recommendations. However, explanations that are inaccurate or difficult to understand may not help users make informed actions. Given the importance of quality explanations for recommendation systems, we expect evaluations of explainability and interpretability to be relatively well-developed for these systems.[9]

Currently, there is no broadly recognized framework for building or evaluating explainable or interpretable AI.

We study evaluation approaches for explainability and interpretability in the research literature to better understand how these evaluations may be conducted in practice. While previous research has surveyed the evaluation of explainable recommendation systems, we adopt a custom approach to systematically identify and annotate relevant research papers for evaluations of explainable recommendation systems as well as descriptions of explainability and interpretability.[10]

Overall, we find that research papers do not meaningfully differentiate between explainability and interpretability, and that descriptions of the principles use a combination of similar themes. We also find that research papers adopt one or more of five evaluation approaches and observe that *research papers strongly favor evaluations of system correctness over evaluations of system effectiveness*. These

results suggest that explainability and interpretability can convey different meanings to different researchers and that evaluations of these principles may not measure the same variables, achieve the same results, or lead to comparable interpretations.

Given the importance that policymakers have placed on explainable and interpretable AI systems, it will be important to resolve these discrepancies to conduct actionable and informative evaluations. We advise policymakers to invest in standards for AI safety evaluations and establish a talent base that can assess the efficacy of these evaluations to ensure that reported evaluations provide meaningful information.

Background

Recommendation Systems

Recommendation systems are software tools and computational techniques that provide suggestions for items of interest to a user. In their simplest form, such suggestions can be ranked lists of items. These suggestions, or "predictions," reflect which items a recommendation system views as most interesting to the user based on the user's preferences and past behavior.[11] User preferences can be expressed either explicitly (for example, as ratings for products) or implicitly (for example, via user actions).[12] To illustrate, a recommendation system on a video-sharing platform may suggest videos that are similar to content that users have awarded a "thumbs up" to in the past. This system may also base its recommendations on user actions by upranking videos similar to those that users have watched or commented on. Figure 1 illustrates how a user's interactions with an e-commerce website can influence the items that the website's recommendation system suggests to the user.

Figure 1. User Interactions Influence a Recommendation System's Suggested Items

Source: CSET. A user adds running shoes to their cart on an e-commerce website. The website's recommendation system then compares items, including a hat, socks, and a water bottle, that complement the running shoes. The recommendation system determines that the user will be most interested in socks and recommends them to the user, who then adds the socks to their purchase.

Explainability and Interpretability

Explainability and interpretability broadly refer to the ability of a machine learning system to provide human-understandable explanations of its predictions.[13] Methods to achieve explainable and interpretable AI systems allow human operators to ensure that AI systems work as intended and achieve desired outcomes such as fairness and reliability.[14] For example, an employer that uses a recommendation system to select job candidates would want to know why the system recommends particular candidates over others to ensure that the system is not selecting for protected characteristics like race and gender. These methods can enable systems to be audited and help promote accountability when the system has behaved in unexpected ways.[15]

Methods to achieve explainable and interpretable AI systems are important in various settings. In safety-critical settings, such as a doctor using AI-enabled recommendations to assist with diagnosing disease, these methods can help users determine how much to trust a system's outputs when making a consequential decision.[16] Explanations of AI systems may also be warranted in lower stakes settings. Customers can make more informed purchasing decisions if e-commerce websites explain why they recommend particular products.[17] Similarly, users can become more astute consumers of information if social media platforms divulge reasons for prioritizing certain information on users' newsfeeds.[18]

More specifically, these methods improve people's understanding of AI systems in the following ways. First, in addition to validating a system's predictive performance, these methods enable open-ended examination of systems. For example, a developer could check to see how changes to a system's parameters affect not only its predictions but also the way it represents different users or items. Second, they can reveal failure modes of AI systems, which can help developers fix and improve systems.[19]

However, in practice, the usefulness of explainable and interpretable AI systems may depend on a variety of situational factors, including the expertise of the user, the time horizon to make a decision, the complexity of the technology, and the task at hand. For instance, a highly technical explanation from an AI model about how it determines loan eligibility would not benefit most loan seekers. If users cannot understand an explanation, it ceases to be helpful. On the other hand, accessible explanations of the factors that contribute to the extension or denial of a loan could help loan seekers better understand how to adjust their behavior to achieve their financial goals.

Current methods to achieve explainable and interpretable AI models fall into two main categories: intrinsic interpretability methods in which explainability is incorporated into the model design in the first place and post-hoc methods that seek to find explanations for a model's behavior after it has already been trained.[20] Intrinsic interpretability methods involve designing and training AI models to be easier to study or to contain built-in explanations.[21] For example, studying the internals of a model that has been trained with an intrinsic interpretability method can reveal the factors that influenced the model's decisions. In contrast, post-hoc methods seek to elucidate models that have already been trained.[22] For instance, local interpretable model-agnostic explanations, or LIME, is a post-hoc method that fits one model to explain another black box model's individual predictions.[23] Which method is most useful depends on the specific problem domain, which covers aspects such as the problem that the model is trying to solve and the users who are interacting with the system.

Methodology

We used several selection criteria to identify research papers that evaluate AI-enabled recommendation systems for explainability or interpretability from CSET's merged corpus of scholarly literature.[*] We first filtered the research corpus for AI-relevant publications about recommendation tasks. Then we filtered the set of publications to find those with mentions of explainability or interpretability within the title or abstract, along with mentions of some form of evaluation.[†] We used this filtering step to scope the results to a manageable size for annotation and to ensure that the results fell squarely within the research field of explainable AI (although it likely does not capture every research publication that conducts evaluations of explainability or interpretability). After applying the selection criteria, we were left with a total of 100 papers published between 2012 and 2022.[‡] The primary country affiliation associated with these publications was China, followed by the United States. All 100 papers were in English.

Five annotators analyzed the full text of these publications to capture how researchers described explainability and interpretability, and evaluated their explainability and interpretability claims.[§] To improve the quality of annotations and resolve inter-annotator disagreement, we assigned at least two annotators to each publication.[**] After annotators finished reviewing papers, we found that 81 papers actually contained evaluations of a recommendation system's explainability or interpretability. The application areas of recommendation systems within the set of 81 papers spanned entertainment, consumer, health, and education applications. Furthermore, none of the recommendation systems in this set leveraged large language models. This is likely because interest in large language models grew rapidly after 2022, which is the latest publication year in our dataset.

---

[*] CSET's merged corpus of scholarly literature includes Digital Science Dimensions, Clarivate's Web of Science, Microsoft Academic Graph, China National Knowledge Infrastructure, arXiv, and Papers With Code. This corpus has been updated since the time of writing.

[†] See Appendix A for more details on our selection criteria.

[‡] We did not specifically sample for 100 papers. We were left with 100 papers by coincidence after applying the selection criteria.

[§] The decision to annotate full text did not result in significant data loss. Ninety-nine out of the 100 publications had full texts available through either open-source publication aggregators or library access.

[**] See Appendix B for more details on our annotation guidelines.

We reviewed the papers and identified five common evaluation approaches, which we describe in more detail in the findings section. We created these approaches inductively by grouping similar approaches across papers. Certain papers discussed more than one evaluation approach. A few papers used rare approaches that did not fit into our five categories and did not directly relate to evaluating explainability or interpretability; as such, we did not include them in our findings.[*]

[*] For example, a couple of approaches tried to demonstrate the consistency of explanations by measuring the similarity of explanations provided to a user or checking whether the predicted ratings of items aligned with their extracted features. Other approaches calculated the number of recommendations that had accompanying explanations or the number of reasoning paths between users and items. Still other approaches sought to quantify the system's robustness, or its ability to maintain performance in different conditions, by seeing how it performed on datasets that differed in size, quality, or sparsity. None of these approaches were directly related to evaluating the explainability or interpretability of the system itself.

Findings

Explainability Descriptions

Before discussing how researchers evaluated their systems for explainability and interpretability, it is helpful to understand how researchers conceptualized these principles.[*] We investigated the descriptive approaches that papers adopted for explainability and interpretability, and whether descriptions of explainability were meaningfully different from descriptions of interpretability. We found that papers did not meaningfully differentiate between explainability and interpretability, and that the descriptions of these principles were often high-level. Because researchers did not reliably differentiate between explainability and interpretability, we will simply use explainability in the rest of the paper to refer to both explainability and interpretability. However, we acknowledge that explainability may carry a different meaning than interpretability in other settings.

In the 81 papers we reviewed, published between 2012 and 2022, we identified four thematic categories based on how researchers described explainability. Some of the publications used a combination of these descriptive approaches:

- Rely on the use of other principles
- Focus on an AI system's technical implementation
- State that the purpose of explainability is to provide a rationale for recommendations
- Articulate the intended outcomes of explainable systems

Below, we provide examples of each of these uses.

[*] To understand how researchers conceptualized these principles, we paid close attention to text that explicitly mentioned explainability or interpretability and text that summarized the main contribution of the paper. Note that a publication could contain multiple descriptions of explainability or interpretability.

Descriptions That Rely on the Use of Other Principles

Several explainability descriptions referenced principles that are closely related to explainability, such as transparency (see Box 1). To give another example, one description associated explainability with efficiency, effectiveness, and persuasiveness.[24]

> **Box 1. Description That Associates Explainability with Transparency**
>
> "By explainable, we would like our method to be transparent in generating a recommendation and is capable of identifying the key cross features for a prediction."[25]

Descriptions That Focus on an AI System's Technical Implementation

Other descriptions focused on the technical components or functionality of the recommendation system in question. For instance, one description suggested that an explainable recommendation system should use intuitive visual attributes, and another proposed that an explainable system should illustrate the underlying relationship between user preferences and item features.[26]

> **Box 2. Description That Focuses on Technical Mechanisms That Support Explainability**
>
> "Accordingly, our goal is not only to select a set of candidate items for recommendation, but also to provide the corresponding reasoning paths in the graph as interpretable evidence for why a given recommendation is made."[27]

Descriptions That State the Purpose of Explainability Is To Provide a Rationale for Recommendations

Another category of descriptions equates explainability to conveying why or how an item was recommended. These descriptions note that explainability helps users make decisions, understand recommendations, and trust a system (see Box 3). Essentially, they suggest how explainability can be beneficial to users.

Box 3. Description That Links Explainability to Understanding System Decision-Making

"Especially, reasonable explanations are beneficial for user[s] to make better decisions ..."[28]

Descriptions That Articulate the Intended Outcomes of Explainable or Interpretable Systems

We also encountered descriptions that specified the intended outcomes of explainable systems, although they were uncommon in the surveyed literature. These descriptions routinely articulated two elements: who will use the explainable recommendation system and what they will use the explainable recommendation system for. For example, one description posited that explainable systems can help system designers track the behavior of complex recommendation models for debugging.[29]

Box 4. Description That Specifies the Outcomes Of Explainable Systems

"Furthermore, the current approaches do not provide students with actual explanations of the predictions and do not utilise dashboards that provide automatic and intelligent guidance ... such as recommendations that, for instance, guide students towards the learning material or activities that will increase the probability for increased course performance."[30]

Explainability Evaluation Approaches

After studying how researchers conceptualized explainability, we sought to understand what approaches researchers used to evaluate claims about explainability. We identified five common evaluation approaches that appeared in the literature surveyed and named them based on their characteristics. The five approaches are:

- Case studies
- Comparative evaluations
- Parameter tuning
- Surveys
- Operational evaluations

These approaches may involve evaluating whether explainable systems were built according to researcher specifications (what we coin system correctness) or whether explainable systems operate as intended in the real world (what we coin system effectiveness). We summarize each approach below and complement the summaries with examples that are inspired by various approaches in the literature. These examples are edited for simplicity and tailored to the needs of different stakeholders. Table 1 lists the benefits and limitations of each evaluation approach.

Case Study

A case study is an explainability evaluation approach that manually examines entities related to a recommendation system in order to understand how the system works. Entities could be the explanations that the system produces, user or item data provided as inputs to the system, or technical characteristics of the system. Illustrative examples are at the heart of case studies. For instance, mapping a recommendation system's embedded features to customer preferences or comparing items that are referenced in explanations to items that received favorable user reviews in the past would be considered case studies.

Box 5. Case Studies in Practice

- A user experience researcher can examine sentences in explanations for a system's movie recommendations to see if the sentences are relevant and accurately convey user sentiment toward movie genres.
- A marketing analyst can map out system predictions of preference paths for users in a dataset to see if the paths between users and recommended items bear some similarity to users' past purchases.

Comparative Evaluation

Comparative evaluations compare how a system performs on a metric related to explainability relative to baselines (such as other systems), use treatment groups to understand the impact of explanations on users, or demonstrate why a system's design was chosen over alternative designs. They may also conceal different elements of a system to assess each element's contribution to the system's efficacy.* This type of evaluation could express the relevance and personalization of explanations as metrics and compare how different systems perform on these metrics.

Box 6. Comparative Evaluations in Practice

- A software engineer can remove different attention modules, which highlight important features of input data, to understand how they impact a recommendation system's performance.
- A behavioral scientist can compare the behavior of a group of people exposed to explanations to the behavior of a group not exposed to explanations to gain insight into the impact of a recommendation system's explanations.

* This approach is typically coined "ablation." Since ablation studies involve comparing a system to slightly altered versions of itself, we group ablation studies with other comparative evaluations.

Parameter Tuning

Parameter tuning involves tweaking one or more parameters of a recommendation system—usually over a predefined range of values—to show how the parameters affect the system's explanations.* More specifically, this evaluation may entail varying the parameters of a recommendation system, recording their effect on the system's explanations, and deciphering which parameter values result in high-quality explanations.

> Box 7. Parameter Tuning in Practice
>
> A machine learning engineer can tune the parameters of a system to understand how they affect factors related to the system's explanations, such as the presence of redundant sentences in an explanation, and adjust parameters accordingly.

Survey

Surveys ask respondents to evaluate a system through a series of questions. Surveys typically provide outputs of the system such as pre-generated explanations of recommendations and auxiliary information about the recommendation task. Respondents are then asked to judge the explanation quality of the system using these artifacts and record their judgments, which researchers later analyze. For example, a survey could ask respondents to rate the persuasiveness of a hotel recommendation system's explanations to gauge whether real users would be inclined to book a hotel recommended by the system.

> Box 8. Surveys in Practice
>
> A survey analyst can ask respondents to imagine that they are restaurant-goers and determine which explanations would be most helpful when selecting a restaurant.

* Note that parameter settings that are offered with no explanation of how they were chosen and cursory implementation details such as the authors' noting that they performed a search for parameters does not qualify as parameter tuning, according to our definition. We determine that researchers perform parameter tuning only if they include a dedicated section that explains why the choices of parameter values are correct, beyond simply stating that a chosen parameter led to the highest accuracy.

Operational Evaluation

During an operational evaluation, users interact in a typical manner with a recommendation system in a live setting. Users are not necessarily prompted to evaluate the system in these settings. Instead, their interactions are analyzed downstream to infer the system's efficacy. Operational evaluations are typically considered high-quality evaluations of system effectiveness because they occur in the system's actual deployed environment, or a setting that closely approximates the expected real environment. For example, an operational evaluation may involve monitoring how frequently users add items to their cart on an e-commerce website when given explanations for recommendations.

Box 9. Operational Evaluations in Practice

A statistician can run tests on a web browser with real e-commerce users to investigate the effect of automatically generated explanations of recommended phones on user acceptance of recommendations.

Table 1: Benefits and Limitations of Evaluation Approaches

Evaluation Approach	Summary	Benefits	Limitations
Case study	Manual exploration of system components to understand how explanations are generated	Provides a window into how a system works	Does not provide a comprehensive view of system functionality
Comparative evaluation	Compare systems or their elements to assess relative explainability	Helpful for prototyping and debugging systems as well as determining whether a system advances state-of-the-art performance	The baseline for comparison may not be analytically useful
Parameter tuning	Vary one or more parameters to understand their impact on the system's explanations	Illustrates how different parameter values impact system behavior or interact with each other	Does not provide a comprehensive view of system functionality
Survey	Ask respondents to judge explanation quality of a system	Provides insight into how users may perceive a system	The utility of responses depends on the surveyed population
Operational evaluation	User interactions with a system in a live setting are analyzed downstream to gauge effectiveness of explanations	Demonstrates how users may engage with a system in the real world	Resource-intensive

Source: CSET.

Evaluations of System Correctness and Effectiveness

The five evaluation approaches that we identified focus on testing system correctness, system effectiveness, or both. Tests of system correctness seek to ascertain whether an explainable system meets design criteria, whereas tests of system effectiveness seek to determine if an explainable system is useful to a user. Boxes 10 and 11 contain sample questions that each test can address.

Box 10. Questions That an Evaluator May Seek to Answer When Testing System Correctness

- Does this recommendation system accurately model a user's movie preferences?
- Do explanations for learning material recommendations reference students' past performance in school?

Box 11. Questions That an Evaluator May Seek to Answer When Testing System Effectiveness

- Do explanations from a movie recommendation system help a user select a movie they enjoy?
- Do explainable recommendations for learning materials improve students' educational outcomes?

As Table 2 illustrates, case studies and comparative evaluations were the most common evaluation approaches in the literature surveyed, appearing in nearly 88 and 63 percent of the papers we reviewed for this analysis, respectively. Parameter tuning was another relatively popular approach, appearing in nearly 40 percent of the papers. Note that these three evaluation approaches are primarily focused on testing *system correctness—in other words, testing if the explainable system is built correctly or provides correct outputs according to researcher specifications.* For example, evaluators can use case studies to examine the process by which a system generates explanations and ensure that the process is accurate. They can also use parameter

tuning to understand different parameters' impacts on system behavior and use comparative evaluations to benchmark systems' capabilities with respect to a task—both of which provide sanity checks that a system's explainability methods were built correctly. Note that this is distinct from system effectiveness, or whether a system's explanations result in desirable changes in the real world. While comparative evaluations can be used to test both system correctness (benchmarking) and system effectiveness (using treatment groups to understand the impact of explanations on users), their tests of system effectiveness were rare in the surveyed literature.

On the other hand, surveys and operational evaluations were the least common evaluation approaches, appearing in about 19 percent and 4 percent of the reviewed papers, respectively (Table 2). The goal of operational evaluations and surveys is to test *system effectiveness—in other words, determining whether the explainable system helps a user complete a particular task or otherwise achieves an intended effect in the real world*. Unlike system correctness, system effectiveness is concerned with the real-world impact of a system. For example, evaluators can use surveys to gain insight into how users may perceive a system, such as the receptiveness of users to a system's explanations. Similarly, evaluators can use operational evaluations to better understand how users will likely engage with a recommendation system once deployed. The low publication of these evaluation approaches in the surveyed literature may be attributed to their reliance on participants and the related potential resource burden, such as needing researcher hours to obtain institutional review board approval and compensation for participants' time. Researchers may be disinclined to conduct resource-intensive evaluations, especially if other types of evaluation suffice for publication.

Table 2: Case Studies Are the Most Common Evaluation Approach

Evaluation Approach	Count of Papers (Percentage of Total Papers, n=81)
Case study	71 (88%)
Comparative evaluation	51 (63%)
Parameter tuning	32 (40%)
Survey	15 (19%)
Operational evaluation	3 (4%)

Source: CSET. Note that some publications include more than one evaluation approach, so percentages sum to greater than 100%.

Given that the majority of papers contained two or more evaluations, we examined which evaluation approaches co-occurred across papers. We found that the overall counts of approaches across papers did not obscure important interactions between approaches, and that case studies, comparative evaluations, and parameter tuning were often paired together in our publications (see Figure 2). This was expected to some degree, considering that these were the most popular evaluation approaches. Researchers may have implemented several popular evaluation approaches because they were not individually resource-intensive and therefore did not require significant resources to implement together. Researchers may have also been influenced by the evaluation approaches of other researchers and simply decided to implement commonplace approaches.

Figure 2. Case Study, Comparative Evaluation, and Parameter Tuning Are Often Grouped Together in Papers

	Case Study	Comparative	Parameter Tune	Survey	Operational
Case Study	71	42	26	13	1
Comparative	42	51	23	12	3
Parameter Tune	26	23	32	9	2
Survey	13	12	9	15	2
Operational	1	3	2	2	3

Source: CSET. Figure 2 depicts counts of evaluation approach pairings in the dataset.

Policy Considerations

Our findings provide a snapshot of how researchers conceptualized explainability and evaluated their explainability claims for AI-enabled recommendation systems. The trends we observed among explainability descriptions and evaluations suggest that policies for implementing or evaluating explainable AI may not be effective without expert guidance to determine what explainability means in a given context and which evaluations to conduct. Further inquiry is needed to determine if our observations hold for different AI systems and applications, and the extent to which researchers' evaluation approaches influence developers' evaluation approaches. Nevertheless, policymakers can take a meaningful step now by investing in standards for AI safety evaluations and establishing a workforce capable of determining whether these evaluations are adequate.

Explainability Descriptions

Policies that mandate explainable AI systems may not be effective without clear parameters for what explainability means in different contexts. We found that researchers described explainability and interpretability in various ways across papers but did not clearly differentiate between the two principles—a departure from the approach taken by NIST. Policymakers should note that explainability and interpretability are multidimensional concepts that will likely be understood differently by different people. When devising evaluation reporting requirements for explainability or other high-level AI principles, policymakers should not assume that AI developers will operationalize principles in a consistent manner. AI developers may respond to reporting requirements by conflating certain AI principles or operationalizing principles in unintended ways (such as testing the persuasiveness of a system as a means of testing its explainability). Standards for AI safety evaluations can provide clarity on how developers should operationalize principles in a given context.

Evaluations of System Correctness and Effectiveness

Policymakers should allocate resources towards growing a talent base that can assess the efficacy of AI safety evaluations. We found that research papers favored evaluations of system correctness over system effectiveness, which corroborates longstanding criticisms about the lack of user testing in the explainable AI literature.[31] Research papers that neglect effectiveness evaluations can draw an incomplete picture of system capabilities and distort perceptions about the explainability of systems in practice. Trained AI evaluation experts are attuned to these limitations and can suggest ways to build more robust evaluations.

It is not entirely unexpected that evaluations in our dataset were skewed towards system correctness, which may reflect publishing incentives and resource constraints. Explainability evaluations that focus on system correctness provide information about what is happening inside the system and whether the system meets design specifications. On the other hand, effectiveness explainability evaluations reveal whether explanations are meaningful or useful to people interacting with the system in the real world.

Both types of evaluations serve important but slightly different purposes. For example, it may be acceptable for evaluators to only test the correctness of explainable methods of an AI system deployed in a low-stakes setting—in other words, evaluate whether the system's explanations are sound without testing their utility to users. However, by focusing solely on evaluations of system correctness, evaluators ignore the environment in which the system is embedded. This could lead to undesirable consequences in high-risk settings, such as a doctor incorrectly prescribing a course of treatment because they misunderstood an AI system's explanations. In these scenarios, it is important to ensure that explainability methods are customized to users and accurately convey how a system works. Expert guidance on when to use different types of evaluations and what the best practices are for each could help evaluators conduct appropriate evaluations of AI systems in different settings.

Although our findings are derived from a small-scale study, policymakers should keep them in mind when they encounter results from evaluations of explainability and other aspects of safe and trustworthy AI. Without sufficiently detailed instructions for reporting, AI developers may report results from evaluations that are convenient or popular to conduct, rather than evaluations that collect meaningful data about a system's inner workings and effectiveness. Policymakers should take precautionary measures now by growing the technical talent needed to assess the efficacy of AI safety evaluations. Guidance that sets expectations for what a quality evaluation looks like is important, but ultimately technical experts are needed to determine whether certain evaluations are up to par and to tailor evaluation standards to different contexts. We encourage further research to investigate whether our findings are present in other areas of the literature and the extent to which researchers' evaluation approaches guide those of AI developers. More work is urgently needed to understand trends and gaps in AI evaluation science.

Conclusion

Our findings show that policymakers should not assume that explainability evaluations for AI systems are self-explanatory or straightforward. We have demonstrated that in the domain of AI-enabled recommendation systems, where explanations are a cornerstone, researchers have several notions of what explainability means. Furthermore, we found that *research papers strongly favored evaluations of system correctness over evaluations of system effectiveness*. This may be acceptable in certain research settings, but research practices commonly influence real-world applications where effectiveness evaluations are indispensable. Although the knowledge base around AI evaluations continues to evolve rapidly, we advise policymakers to invest in standards for AI safety evaluations and enable a workforce that can assess the efficacy of these evaluations in different contexts.

Authors

Mina Narayanan is a research analyst at the Center for Security and Emerging Technology.

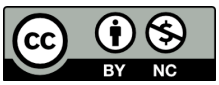
Christian Schoeberl is a data research analyst at the Center for Security and Emerging Technology.

Tim G. J. Rudner is a Non-Resident AI/ML Fellow at the Center for Security and Emerging Technology and a Faculty Fellow at New York University.

Acknowledgments

We thank Heather Frase, Steph Batalis, Mia Hoffmann, Josh Goldstein, Rita Konaev, Emelia Probasco, Helen Toner, Tina Huang, Katherine Quinn, and Matt Mahoney for checking our assumptions, tightening our writing, and providing thoughtful feedback on how to frame our analysis.

We also thank Will McCormack, Xiaojing Ni, Valeria Vera Lagos, and Aanchal Dusija for their persistence and care when annotating papers, and James Dunham for training us to use Snorkel.



Document Identifier: doi: 10.51593/20230003

Appendix A: Selection Criteria

In order to find publications relevant to our research questions, we utilized several selection methods.

CSET AI Classifier

We selected AI-relevant publications leveraging predictions from CSET's arXiv classifier, which learns the definition of AI-relevance from human-labeled examples on arXiv. For more information, see Dunham et al, "Identifying the Development and Application of Artificial Intelligence in Scientific Text," arXiv preprint, arXiv:2002.07143 (2020), https://arxiv.org/abs/2002.07143.

CSET Tasks and Methods

Once we selected AI-relevant publications, we used CSET's tasks and methods extraction model to select only those publications related to recommendation tasks. Tasks and methods are extracted using rule-based and supervised methods from English-language AI-relevant publications in CSET's merged corpus of scholarly literature. This corpus includes publications from Digital Science Dimensions, Clarivate's Web of Science, Microsoft Academic Graph, China National Knowledge Infrastructure, arXiv, and Papers With Code. China National Knowledge Infrastructure is furnished for CSET's use by East View Information Services, Minneapolis, MN, USA. For information on the SciREX model used in extraction, see Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy, "SciREX: A Challenge Dataset for Document-Level Information Extraction," arXiv preprint, arXiv:2005.00512 (2020), https://arxiv.org/abs/2005.00512.

Keywords: Explainability and Interpretability

Given our focus on evaluations of explainability and interpretability, we searched publication titles and abstracts for mentions of "explainability," "interpretability," or close variations thereof. This keyword search failed to capture publications related to explainability or interpretability that do not explicitly use these words in their titles or abstracts. However, we assumed that many researchers advertise their work on explainability using related keywords in their papers' titles or abstracts.

Keywords: Evaluation

Lastly, we wanted to ensure that papers contained some evidence of researchers evaluating the explainability or interpretability of their AI systems. We searched titles and abstracts for mentions of “benchmark,” “baseline,” “evaluation,” “state of the art,” or “state-of-the-art.” This approach was not guaranteed to capture all papers with evaluations of explainability or interpretability but ensured the relevance of those that were captured.

Appendix B: Annotation Guidelines

We provided the full text of each research publication and an annotation guide to a team of five trained annotators. The team consisted of undergraduate students, graduate students, and one full-time employee at CSET, who all had at least some degree of familiarity with the subject matter of papers.

The annotation guide instructed annotators to select spans of text associated with concepts of interest, such as the paper's main research contribution, descriptions of explainability and interpretability, and evaluations of explainability and interpretability. The annotation guide was not overly prescriptive to encourage annotators to capture how researchers conceptualize explainability and interpretability, rather than our preconceived notions of the principles and how researchers might implement them.

Limitations in the user interface of our annotation platform, along with the preprocessing of full text PDFs, presented technical challenges to our analysis. These challenges likely impacted our ability to capture every relevant annotation field for our set of publications. However, we assigned at least two people to select spans of text for each paper and required the CSET employee (the annotator most familiar with the subject matter) to be one of these people. For papers where annotators selected substantially different spans of text for a given field, we considered the CSET employee's annotation to be ground truth. These quality checks helped minimize any negative impact on our analysis.

Endnotes

[1] "Klobuchar, Thune, Commerce Committee Colleagues Introduce Bipartisan AI Bill to Strengthen Accountability and Boost Innovation," U.S. Senator Amy Klobuchar, November 15, 2023, https://www.klobuchar.senate.gov/public/index.cfm/2023/11/klobuchar-thune-commerce-committee-colleagues-introduce-bipartisan-ai-bill-to-strengthen-accountability-and-boost-innovation; "Blumenthal & Hawley Announce Bipartisan Framework on Artificial Intelligence Legislation," U.S. Senator Richard Blumenthal, September 8, 2023, https://www.blumenthal.senate.gov/newsroom/press/release/blumenthal-and-hawley-announce-bipartisan-framework-on-artificial-intelligence-legislation; "Artificial Intelligence Act: MEPs Adopt Landmark Law," News European Parliament, March 13, 2024, https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law; "The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023," 2022 to 2024 Sunak Conservative Government, November 1, 2023, https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023; DigiChina Translation of Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence [《新一代人工智能治理原则——发展负责任的人工智能》], National New Generation Artificial Intelligence Governance Expert Committee, June 17, 2019, https://digichina.stanford.edu/work/translation-chinese-expert-group-offers-governance-principles-for-responsible-ai/; Original CSET Translation of Ethical Norms for New Generation Artificial Intelligence Released [《新一代人工智能伦理规范》发布], PRC Ministry of Science and Technology, October 21, 2021, https://cset.georgetown.edu/publication/ethical-norms-for-new-generation-artificial-intelligence-released/.

[2] Emelia Probasco, Autumn Toney, and Kathleen Curlee, "The Inigo Montoya Problem for Trustworthy AI: The Use of Keywords in Policy and Research" (Center for Security and Emerging Technology, June 2023), https://cset.georgetown.edu/publication/the-inigo-montoya-problem-for-trustworthy-ai/; National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (Washington, DC: Department of Commerce, 2023), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[3] "Event Transcript: Sen. Chuck Schumer Launches SAFE Innovation in the AI Age at CSIS" (Center for Strategic and International Studies, June 21, 2023), https://csis-website-prod.s3.amazonaws.com/s3fs-public/2023-06/230621_Schumer_SAFE_Innovation.pdf?VersionId=jApHm2QrP7nAZvl_B4GJ6s_YjSrfyYBK.

[4] Majority Leader Chuck Schumer, Senator Mike Rounds, Senator Martin Heinrich, and Senator Todd Young, *Driving U.S. Innovation in Artificial Intelligence: A Roadmap for Artificial Intelligence Policy in the United States Senate* (Washington, DC: The Bipartisan Senate AI Working Group, May 2024), https://www.schumer.senate.gov/imo/media/doc/Roadmap_Electronic1.32pm.pdf.

[5] Office of Management and Budget, *Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence* (Washington, DC: Executive Office of the President, March 2024), https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf; *Ensuring Safe, Secure, and Trustworthy AI* (Washington, DC: The White House, July 2023), https://www.whitehouse.gov/wp-

content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf; National Institute of Standards and Technology, *U.S. AI Safety Institute Signs Agreements Regarding AI Safety Research, Testing, and Evaluation with Anthropic and OpenAI* (Washington, DC: Department of Commerce, August 29, 2024), https://www.nist.gov/news-events/news/2024/08/us-ai-safety-institute-signs-agreements-regarding-ai-safety-research; Exec. Order No. 14110, 88 FR 75191 (2023); *Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems* (Hiroshima: Group of 7, 2023), https://www.mofa.go.jp/files/100573473.pdf.

[6] Michael Atleson, "Keep Your AI Claims in Check," *Federal Trade Commission* (blog), February 27, 2023, https://www.ftc.gov/business-guidance/blog/2023/02/keep-your-ai-claims-check.

[7] Haochen Liu, Yiqi Wang, Wenqi Fan et al., "Trustworthy AI: A Computational Perspective," arXiv preprint arXiv:2107.06641 (2021), 28, https://arxiv.org/pdf/2107.06641.pdf.

[8] Matt Sheehan and Sharon Du, "What China's Algorithm Registry Reveals About AI Governance," *Carnegie Endowment for International Peace*, December 9, 2022, https://carnegieendowment.org/2022/12/09/what-china-s-algorithm-registry-reveals-about-ai-governance-pub-88606; UK Parliament, *Online Safety Act 2023* (London: UK Parliament, October 26, 2023), https://bills.parliament.uk/bills/3137; European Commission, *The Digital Services Act Package* (Brussels: European Commission, July 25, 2024), https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package.

[9] Xu Chen, Yongfeng Zhang, and Ji-Rong Wen, "Measuring 'Why' in Recommender Systems: a Comprehensive Survey on the Evaluation of Explainable Recommendation," arXiv preprint arXiv: 2202.06466 (2022), https://arxiv.org/pdf/2202.06466.

[10] Yongfeng Zhang and Xu Chen, "Explainable Recommendation: A Survey and New Perspectives," arXiv preprint arXiv:1804.11192 (2020), https://arxiv.org/pdf/1804.11192; Chen et al., "Measuring 'Why' in Recommender Systems"; Nava Tintarev and Judith Masthoff, "A Survey of Explanations in Recommender Systems" (paper presented at IEEE 23rd International Conference on Data Engineering Workshop, Istanbul, Turkey, April 17–20, 2007), https://ieeexplore.ieee.org/document/4401070; Ingrid Nunes and Dietmar Jannach, "A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems," arXiv preprint arXiv:2006.08672 (2020), https://arxiv.org/pdf/2006.08672; Meike Nauta, Jan Trienes, Shreyasi Pathak et al., "From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI," *ACM Computing Surveys* 55, no. 13s (July 2023): 1-42, https://dl.acm.org/doi/10.1145/3583558.

[11] Francesco Ricci, Lior Rokach, and Bracha Shapira, *Recommender Systems Handbook* (New York City: Springer Publishing Company, November 2015), https://dl.acm.org/doi/10.5555/2857282.

[12] Ricci et al., *Recommender Systems Handbook*.

[13] European Data Protection Supervisor, *TechDispatch #2/2023 – Explainable Artificial Intelligence* (Brussels: European Data Protection Supervisor, November 2023), https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence_en.

[14] Finale Doshi-Velez and Been Kim, "Towards a Rigorous Science of Interpretable Machine Learning," arXiv preprint arXiv:1702.08608 (2017), https://arxiv.org/abs/1702.08608.

[15] Tintarev and Masthoff, "A Survey of Explanations in Recommender Systems."

[16] Tim G. J. Rudner and Helen Toner, "Key Concepts in AI Safety: Interpretability in Machine Learning" (Center for Security and Emerging Technology, March 2021), https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-interpretability-in-machine-learning.

[17] Liu et al., "Trustworthy AI: A Computational Perspective."

[18] Liu et al., "Trustworthy AI: A Computational Perspective."

[19] Rudner and Toner, "Key Concepts in AI Safety: Interpretability in Machine Learning."

[20] Mengnan Du, Ninghao Liu, and Xia Hu, "Techniques for Interpretable Machine Learning," *Communications of the ACM* vol. 63, no. 1 (December 2019): 68–77, https://dl.acm.org/doi/10.1145/3359786.

[21] Du et al., "Techniques for Interpretable Machine Learning."

[22] Du et al., "Techniques for Interpretable Machine Learning."

[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Ernesto Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier" (paper presented at KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, August 13–17, 2016), https://dl.acm.org/doi/10.1145/2939672.2939778.

[24] Weizhi Ma, Min Zhang, Yue Cao et al., "Jointly Learning Explainable Rules for Recommendation with Knowledge Graph" (paper presented at WWW '19: The World Wide Web Conference, San Francisco, California, May 13–17, 2019), 1210–1221, https://dl.acm.org/doi/10.1145/3308558.3313607.

[25] Xiang Wang, Xiangnan He, Fuli Feng et al., "TEM: Tree-enhanced Embedding Model for Explainable Recommendation" (paper presented at WWW '18: Proceedings of the 2018 World Wide Web Conference, Lyon, France, April 23–27, 2018), 1543–1552, https://dl.acm.org/doi/abs/10.1145/3178876.3186066.

[26] Min Hou, Le Wu, Enhong Chen et al., "Explainable Fashion Recommendation: a Semantic Attribute Region Guided Approach," (paper presented at IJCAI '19: Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, August 10–16, 2019), 4681–4688, https://dl.acm.org/doi/10.5555/3367471.3367694; Cong Zou and Zhenzhong Chen, "Joint Latent Factors and Attributes to Discover Interpretable Preferences in Recommendation," *Information Sciences* vol. 505, (December 2019): 498–512, https://www.sciencedirect.com/science/article/abs/pii/S0020025519306760.

[27] Yikun Xian, Zuohui Fu, S. Muthukrishnan et al., "Reinforcement Knowledge Graph Reasoning for Explainable Recommendation" (paper presented at SIGIR '19: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, July 21–25, 2019), 285–294, https://dl.acm.org/doi/10.1145/3331184.3331203.

[28] Zuoxi Yang and Shoubin Dong, "HAGERec: Hierarchical Attention Graph Convolutional Network Incorporating Knowledge Graph for Explainable Recommendation," *Knowledge-Based Systems* vol. 204, (September 2020), https://www.sciencedirect.com/science/article/abs/pii/S0950705120304196.

[29] Juntao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang, "Counterfactual Explainable Recommendation" (paper presented at CIKM '21: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Queensland, Australia, November 1–5, 2021), 1784–1793, https://dl.acm.org/doi/10.1145/3459637.3482420.

[30] Muhammad Afzaal, Jalal Nouri, Aayesha Zia et al., "Explainable AI for Data-Driven Feedback and Intelligent Action Recommendations to Support Students Self-Regulation," *Frontiers in Artificial Intelligence* vol. 4 (November 2021), https://pubmed.ncbi.nlm.nih.gov/34870183/.

[31] Mark T Keane, Eoin M Kenny, Eoin Delaney, and Barry Smyth, "If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques," arXiv preprint arXiv:2103.01035 (2021), https://arxiv.org/abs/2103.01035; Tim Miller, Piers Howe, and Liz Sonenberg, "Explainable AI: Beware of Inmates Running the Asylum," arXiv preprint arXiv:1712.00547 (2017), https://arxiv.org/pdf/1712.00547.