A MULTIDISCIPLINARY DESIGN AND EVALUATION FRAMEWORK FOR EXPLAINABLE AI SYSTEMS

A Dissertation

by

SINA MOHSENI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Xia Hu
Co-Chair of Committee,	Eric Ragan
Committee Members,	Frank Shipman
	Shuiwang Ji
Head of Department,	Scott Schaefer

December 2020

Major Subject: Computer Science

ABSTRACT

Nowadays, algorithms analyze user data and affect the decision-making process for millions of people on matters like employment, insurance and loan rates, and even criminal justice. However, these algorithms that serve critical roles in many industries have their own biases that can result in discrimination and unfair decision-making. Explainable Artificial Intelligence (XAI) systems can be a solution to predictable and accountable AI by explaining AI decision-making processes for end users and therefore increase user awareness and prevent bias and discrimination. The broad spectrum of research on XAI, including designing interpretable models, explainable user interfaces, and human-subject studies of XAI systems are sought in different disciplines such as machine learning, human-computer interactions (HCI), and visual analytics. The mismatch in objectives for the scholars to define, design, and evaluate the concept of XAI may slow down the overall advances of end-to-end XAI systems. My research aims to converge knowledge behind design and evaluation of XAI systems between multiple disciplines to further support key benefits of algorithmic transparency and interpretability. To this end, I propose a comprehensive design and evaluation framework for XAI systems with step-by-step guidelines to pair different design goals with their evaluation methods for iterative system design cycles in multidisciplinary teams.

This dissertation presents a comprehensive XAI design and evaluation framework to provide guidance for different design goals and evaluation approaches in XAI systems. After a thorough review of XAI research in the fields of machine learning, visualization, and HCI, I present a categorization of XAI design goals and evaluation methods and show a mapping between design goals for different XAI user groups and their evaluation methods. From my findings, I present a design and evaluation framework for XAI systems (*Objective 1*) to address the relation between different system design needs. The framework provides recommendations for different goals and ready-to-use tables of evaluation methods for XAI systems. The importance of this framework is in providing guidance for researchers on different aspects of XAI system design in multidisciplinary team efforts. Then, I demonstrate and validate the proposed framework (*Objective 2*) through one

end-to-end XAI system case study and two examples by analysis of previous XAI systems in terms of our framework.

I present two contributions to my XAI design and evaluation framework to improve evaluation methods for XAI system. First, I investigate temporal patterns of user trust and reliance in XAI systems (*Objective 3*). My study results show that model explanations not only affected user final trust but also shape how user trust evolves over time; indicating the importance of of user behavior for evaluating XAI systems. Lastly, I propose an open-sourced human-attention evaluation baseline for direct evaluation of saliency map explanations (*Objective 4*). I demonstrate my human-attention benchmark's utility for quantitative evaluation of model explanations by comparing it with single-layer feature masks baseline. My experiments also show the advantage of my evaluation baseline by revealing different user biases in the subjective rating evaluation of model saliency explanations.

DEDICATION

To my wife, who has always supported my Ph.D. life.

To my parents, to whom I owe an eternal debt.

ACKNOWLEDGMENTS

I want to express my gratitude to my advisor, Dr. Eric Ragan, for advising and supporting me in my Ph.D. journey. Thank you for carefully reviewing my research products and allowing me to think independently. I appreciate how much my research life was tailored to my goals, something that few graduate student can experience.

I would like to thank my committee chair, Dr. Xia Hu, who patiently guided me through my graduation. I would not have been able to do this without his support and expertise. Special thanks to my committee members Dr. Shuiwang Ji and Dr. Frank Shipman, as well as many colleagues like Rhema Linder and Nic Lupfer who have mentored and supported me over my years at Texas A&M University.

I also would like to thank my colleagues in University of Florida's INDIE Lab and Texas A&M University's DATA Lab for their constructive criticism and helpful suggestions regarding my research. I want to thank Mahsan Nourani of the CISE department at the University of Florida for accepting the interviews for their XAI system presented in Chapter 4.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Dr. Xia Hu (Chair) of the TAMU Department of Computer Science and Engineering and Dr. Eric Ragan (Co-Chair, research advisor) of the Department of Computer & Information Science & Engineering (CISE) at the University of Florida and Dr. Frank Shipman and Dr. Shuiwang Ji of the TAMU Department of Computer Science and Engineering.

The case study presented in Chapters 4 and 5 is a team effort including project PIs Dr. Xia Hu, Dr. Shuiwang Ji, and Dr. Eric Ragan and a group of graduate students. The design and implementation of machine learning models and training data used to produce user study results were conducted by a group of graduate students including Fan Yang, Mengnan Du, Shiva Pentyala, and Yi Liu of the Department of Computer Science and Engineering. Nic Lupfer of the Department of Computer Science and Engineering was in part involved in the development of the user interface for the case study.

Jeremy Block of the CISE department at the University of Florida helped in interface development and data cleaning of the evaluation benchmark presented in Chapter 6. All other work conducted for this dissertation was completed by the student independently.

Funding Sources

My graduate study was supported by the NSF 1565725 award and the DARPA XAI program under N66001-17-2-4031.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION*

Impressive applications of Artificial Intelligence (AI) and data mining have become prevalent in our time. Tech giants like Google, Facebook, and Amazon have collected and analyzed enough personal data through smartphones, personal assistant devices, and social media that can model individuals better than other people. Recent negative interference of social media bots in political elections [4] were yet another sign of how susceptible our lives are to the power of artificial intelligence and big data [5]. In these circumstances, despite tech giants and the thirst for more advanced systems, others suggest holding off on fully unleashing AI for critical applications until they can be better understood by those who will rely on them. The demand for predictable and accountable AI grows as tasks with higher sensitivity and social impact are more commonly entrusted to AI services. Hence, algorithm transparency is an essential factor in holding organizations responsible and accountable for their products, services, and communication of information.

Explainable Artificial Intelligence (XAI) systems are a possible solution towards accountable AI, making it possible by explaining AI decision-making processes and logic for end users [6]. Specifically, explainable algorithms can enable control and oversight in case of adverse or unwanted effects, such as biased decision-making or social discrimination. An XAI system can be defined as a self-explanatory intelligent system that describes the reasoning behind its decisions and predictions. The AI explanations could benefit users in many ways such as enabling appropriate trust and reliance as well as enabling ethical and fairness analysis of machine learning models and their decision-making process.

While the increasing impact of advanced black-box machine learning systems in the big-data era has attracted much attention from different communities, interpretability of intelligent systems has also been studied in numerous contexts [7, 8]. The study of personalized agents, recommendation systems, and critical decision-making tasks (e.g., medical analysis and powergrid control)

* Parts of the material in this chapter are reprint or adapted from [3]. Mohseni et al. "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems" accepted for publication at accepted for publication in ACM Transactions on Interactive Intelligent Systems. Reproduced with permission.

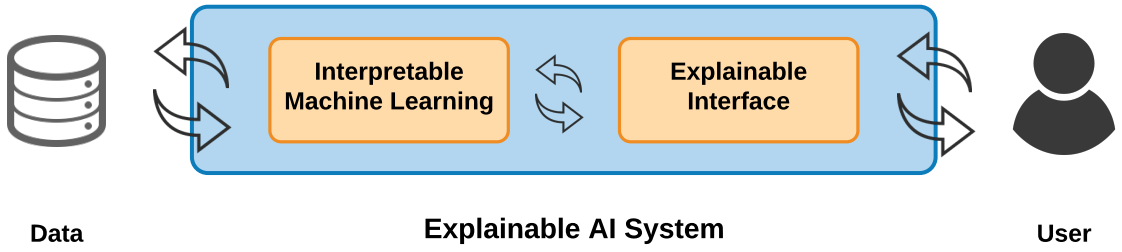


Figure 1.1: XAI system components and interactions. The user interacts with the explainable interface to send queries to the interpretable algorithm and receive model output and explanations. The interpretable model interacts with the data to generate explanations for user queries. Reprinted from Mohseni et al. [3].

has added to the importance of machine-learning explanation and AI transparency for end-users. For instance, as a step towards this goal, the legal right to explanations has been established in the European Union General Data Protection Regulation (GDPR) commission. While the current state of regulations is mainly focused on user data protection and privacy, it is expected to cover more algorithmic transparency and explanations requirements from AI systems [9].

Addressing such a broad array of definitions and expectations for XAI requires multidisciplinary research efforts, as existing communities have different requirements and often have drastically different priorities and areas of specialization. For instance, research in the domain of machine learning seeks to design new interpretable models and explain black-box models with ad-hoc explainers. Along the same line but with different approaches, researchers in visual analytics design and study tools and methods for data and domain experts to visualize complex black-box models and study interactions to manipulate machine learning models. In contrast, research in human-computer interaction (HCI) focuses on end-user needs such as user trust and understanding of machine generated explanations. Psychology research also studies the fundamentals of human understanding, interpretability, and the structure of explanations.

1.1 Problem Statement

An XAI system includes multiple components which directly affect the system design process, see Figure 1.1. There exist diverse sets of design goals, evaluation methods, and research back-

ground for each of the XAI components. For example, numerical analytic methods are employed in machine learning fields to evaluate computational interpretability, while human interpretability and human-subjects evaluations are more commonly the primary goals in HCI and visualization communities. In this regard, although there seems to be a mismatch in specific objectives for designing and evaluating explainability and interpretability, a convergence in goals is beneficial for achieving the full potential of XAI. Additionally, looking at the broad spectrum of research on XAI, even though different aspects of XAI research are following the general goals of AI interpretability, it is evident that scholars from different disciplines have different goals in mind. However, to the best of my knowledge, there exists no multidisciplinary XAI system framework to unify the efforts from multiple disciplines in building XAI systems. A multidisciplinary framework for end-to-end XAI system design and evaluation can identify the relation between diverse design goals to enhance system design process. Further, a unified framework can reveal potential design and evaluation gaps between XAI system requirements and final outcomes. For example, the importance of a unified XAI framework is very higher in multidisciplinary teams focused on critical applications of XAI aiming to leverage psychology-grounded theories (i.e., design requirements) for designing interpretable machine learning techniques and presenting with explanation interfaces (i.e., system outcomes).

1.2 Contributions

My main research contribution is proposing a multidisciplinary design and evaluation framework for XAI systems, followed by a case study and a series of evaluation studies to demonstrate, validate, and improve the proposed framework. The following briefly introduces my four contributions (C1-C4) in this dissertation.

1.2.1 C1: A Design and Evaluation Framework for Explainable AI Systems

I propose a multidisciplinary framework to share knowledge and experiences of XAI design and evaluation methods across multiple fields. I first present a categorization and mapping of XAI design goals and evaluation methods with a thorough review of related literature (over 200

papers) across the fields of machine learning, visualization, and HCI. From the findings, I develop a framework with step-by-step design guidelines paired with evaluation methods to close the iterative design and evaluation loops in multidisciplinary teams. The impetus for this framework is the desire to organize and relate the diverse set of existing design guidelines and evaluation methods in a unified model. The framework is intended to give guidance on what evaluation measures are appropriate to use at which design stage of the XAI system design. Further, I provide summarized ready-to-use tables of evaluation metrics for different goals in XAI system design steps.

1.2.2 C2: Case Study and Examples for XAI Framework

I present a case study of a collaborative design and development effort for an XAI system to showcase a practical example of using the framework. In the scenario of this case study, a multidisciplinary team of researchers designed a XAI system for fake news detection for non-expert (not AI experts or news analysts) daily newsreaders. I review system design steps and evaluation outcomes for the effects of interpretable fake news detector on users' overall experience and performance in detecting fake news. Also, specific to the domain of fake news detection, I aim to examine whether model explanations can help users to avoid overtrusting the fake news detector when explanations are nonsensical to users. Study results revealed the challenges rising from the inherent difference between models' feature learning (word-level features in this case) and human understanding of news and information. Overall, I observed that users' interaction with the AI and XAI assistants affected their performance, mental model, and trust. However, model explanations in these studies did not improve task performance or increase user trust and mental model. Quantitative results and qualitative participants' feedback indicate that explanations helped users' to build an appropriate mental model of intelligent assistants and adjust their trust accordingly given the limitations of the models.

In addition to the XAI system design case study, I analyze two existing XAI systems to demonstrate the descriptive functionality of the framework to describe design process workflow (between-layers) and design and evaluation choices (within each layer). In the first example, I analyze Nourani et al.'s [1] paper in which authors present an XAI system to support AI novice users

tasked with activity recognition in a series of videos. For the second example, I review and analyze Hamidi-Haines et al.'s [2] paper that authors present the "interactive naming" interface that allows the end-user to explore and manually cluster model activation maps to create meaningful groups of "visual concepts". Both analyses are aiming to find insights from their work and intended to suggest future design iterations. I conducted interviews with the first author of these papers for reviewing their design step and main considerations during the process including interactions between machine learning designers and interface designers in the team.

1.2.3 C3: User Trust Dynamics in Explainable AI

I contribute to the proposed framework by elaborating on XAI evaluation methods with a study to demonstrate the importance of dynamics of user behavior with XAI systems. Studying dynamics of user trust is particularly important to understand temporal patterns of user behavior and improve system design and evaluations accordingly. The recurring measurements of user trust in complex systems (e.g., AI-based systems) is invaluable to understand the dynamics of user behavior and complement the limitations of static measurements. I investigate the effects of interpretability on user behavior and trust and their evolution over time in a human-XAI collaborative setup for fake news detection. Specifically, I study the role of explanations in human-AI collaboration by studying dynamics of user trust in the fake news detection case study.

My analysis of results show model explanations affect on how user trust morphs over time during their interactions with the XAI intelligent agent. This was in addition to the case study findings that indicated users working with the same intelligent system can perceive the system competence differently depending on how the model and its decision making is explained. Recurring measurements of user reliance revealed whether model explanations are persuasive (resulting in an increase of user overtrust) or implausible (resulting in a decrease of user trust) to the user. However, my findings suggest the dynamics of self-reported subjective performance measures were not aligned with the objective behavioral measures. This could be an indicator of possible lead or lag in reflections of trust between my two measurements of trust. This latency between users' exposure to the system, adapting their behavior, and coming to their conclusions have also been reported in

previous research, see [10].

1.2.4 C4: A Human-Attention Benchmark

Lastly, I contribute to the proposed framework by proposing a human-attention baseline to quantitatively evaluate model saliency explanations. A limitation of human subject studies to evaluation machine learning explanations is that user feedback tend to be more costly, imprecise, and subjective to the task. My publicly available human-grounded benchmark enables fast, replicable, and objective execution of evaluation experiments for saliency explanations. To foster the interest of the machine learning community, I demonstrate the benchmark's utility for quantitative evaluation of model explanations and compare it with the single-layer feature mask ground-truth and human judgment rating evaluations.

In a series of experiments, I study the relationships and trade-offs between two different human-grounded evaluation approaches (i.e., binary annotation mask and human subjective feedback) to present the efficiency of the proposed human-attention baseline. The study results indicated the significant difference between threshold-agnostic evaluation with a human-attention baseline as compared to previous methods with binary ground-truth mask. My experiments also revealed user biases in their subjective rating when exposed to different visual appearance and error types of saliency explanations. I conclude that human-attention baseline is the most accurate ground-truth for direct evaluation (i.e. feature- level) of model saliency explanations when compared to binary segmentation mask and human subjective review.

`https://github.com/SinaMohseni/ML-Interpretability-Evaluation-Benchmark`

2. BACKGROUND *

I review the XAI research background, terminologies, and literature related to XAI systems from a broad and multidisciplinary perspective. Then, I present design techniques, evaluation measures, and XAI-related surveys from three fields of HCI, visual analytics, and machine learning. The review of literature in this chapter is relatively light and intended to provide the necessary context for XAI framework in Chapter 3. I leave the in-depth review and categorization of XAI design and evaluation techniques to Chapter 3 as a part of the proposed framework. In the end, I present my survey methodology used for in-depth literature review and identification of XAI design goals and evaluation methods.

2.1 AI and Explanations

Nowadays, algorithms analyze user data and affect decision-making processes for millions of people on matters like employment, insurance rates, loan rates, and even criminal justice [11]. However, these algorithms that serve critical roles in many industries have their own disadvantages that can result in discrimination [12, 13], and unfair decision-making [5]. For instance, recently, news feed and targeted advertising algorithms in social media have attracted much attention for aggravating the lack of information diversity in social media [14]. A significant part of the trouble could be because algorithmic decision-making systems—unlike recommender systems—do not allow their users to choose between the recommended items, but instead, present the most relevant content or option themselves. To address this, Heer [15] suggests the use of shared representations of tasks that are augmented with both machine learning models and user knowledge to reduce the negative effects of immature AI autonomous systems. They present case studies of interactive systems that integrate proactive computational support into interactive systems. Bellotti and Edwards [16] argue that intelligent context-aware systems should not act on our behalf. They suggest

* Parts of the material in this chapter are reprint or adapted from [3]. Mohseni et al. "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems" accepted for publication at accepted for publication in ACM Transactions on Interactive Intelligent Systems. Reproduced with permission.

user control over the system as a principle to support the accountability of a system and its users. Transparency can provide essential information for decision-making that is hidden to the end-users and causes blind faith [17]. The key benefits of algorithmic transparency and interpretability include: user awareness [18]; bias and discrimination detection [19, 13]; interpretable behavior of intelligent systems [20]; and accountability for users [21]. Furthermore, considering the growing body of examples of discrimination and other legal aspects of algorithmic decision making, researchers are demanding and investigating transparency and accountability of AI under the law to mitigate adverse effects of algorithmic decision making [22, 23, 24].

2.1.1 Auditing Inexplicable AI

Researchers audit algorithms to study bias and discrimination in algorithmic decision making [25] and study the users' awareness of the effects of these algorithms [26]. *Auditing* of algorithms is a mechanism for investigating algorithms' functionality to detect bias and other unwanted algorithm behaviors without the need to know about its specific design details. Auditing methods focus on problematic effects on the results of algorithmic decision-making systems. To audit an algorithm, researchers feed new inputs to the algorithm and review system output and behavior. Researchers generate new data and user accounts with the help of scripts, bots [12], and crowdsourcing [27] to emulate real data and real users in the auditing process. For bias detection among multiple algorithms, cross-platform auditing can detect if an algorithm behaves differently from another algorithm. A recent example of cross-platform auditing is a work by Eslami et al. [28], in which they analyzed user reviews in three hotel booking websites to study user awareness of bias in online rating algorithms. These examples demonstrate that auditing is a valuable yet time-intensive process that could not be scaled easily to large numbers of algorithms. This calls for new research for more effective solutions toward algorithmic transparency.

2.1.2 Explainable AI

Along with the methods mentioned above for supporting transparency, machine learning explanations have also become a common approach to achieve transparency in many applications

such as social media, e-commerce, and data-driven management of human workers [29, 30, 31]. The XAI system, as illustrated in Figure 1.1, is able to generate explanations and describe the reasoning behind machine-learning decisions and predictions. Machine-learning explanations enable users to understand how the data is processed. They aim to bring awareness to possible bias and system malfunctions. For example, to measure user perception of justice in intelligent decision making, Binns et al. [32] studied explanations in systems for everyday tasks such as determining car insurance rates and loan application approvals. Their results highlight the importance of machine learning explanations in users' comprehension and trust in algorithmic decision-making systems. In a similar work studying knowledge of social media algorithms, Radar et al. [33] ran a crowdsourced study to see how different types of explanations affect users' beliefs on news feed algorithmic transparency in a social media platform. In their study, they measured users' awareness, correctness, and accountability to evaluate algorithmic transparency. They found that all explanations caused users to become more aware of the system's behavior. Stumpf et al. [34] designed experiments to investigate meaningful explanations and interactions to hold users accountable by machine learning algorithms. They show explanations as a potential method for supporting richer human-computer collaboration to share intelligence.

The recent advancements and trends for explainable AI research demand a wide range of goals for algorithmic transparency which calls for research across varied application areas. To this end, my review encourages a cross-discipline perspective of intelligibility and transparency goals.

2.1.3 Explainable AI Terminology

To familiarize the readers with common XAI concepts and terminologies that are repeatedly referenced in this review, the following four subsections summarize high-level characterizations of XAI explanations. Many related surveys (e.g., [35, 36]) and reports (e.g., [37, 38]) also provide useful compilations of terminology and concepts in comprehensive reports. For instance, Abdul et al. [39] present a citation graph from diverse domains related to explanations, including intelligible intelligent systems, context-aware systems, and software learnability. Later, Arrieta et al. [40] present a thorough review of XAI concepts and taxonomies and arrives at the concept of

Responsible AI as a manifold of multiple AI principles including model fairness, explainability, and privacy. Similarly, the concept of *Safe AI* has been reviewed by Amodei et al. [41], which is an interest in safety-critical intelligent applications such as autonomous vehicles [42]. Table 2.1 presents descriptions for 14 common terms related to this survey's topic and organizes their relation to *Intelligible Systems* and *Transparent AI* topics. I consider *Transparent AI* systems as the AI-based class of *Intelligible Systems*. Therefore, properties and goals previously established for *Intelligible Systems* are ideally transferable to *Transparent AI* systems. However, challenges and limitations for achieving transparency in complex machine learning algorithms raise issues (e.g., ensuring the fairness of an algorithm) that were not necessarily problematic in intelligible rule-based systems but now require closer attention from research communities.

The descriptions presented in Table 2.1 are meant to be an introduction to these terms, though exact definitions and interpretations can depend on usage context and research discipline. Consequently, researchers from different disciplines often use these terms interchangeably, disregarding differences in meaning [35]. Perhaps the two generic terms of *black-box model* and *transparent model* are in the center of XAI terminology ambiguity. The *black-box* term refers to complex machine learning models that are not human-interpretable [43] as opposed to *transparent models* which are simple enough to be human-interpretable [40]. I find it more accurate and consistent to separate the transparency of an XAI system (as described in Figure 1.1) from the interpretability of its internal machine learning models. Specifically, Table 2.1 shows that *Transparent AI* could be achieved by either *interpretable AI* or *Explainable AI* approaches. Other examples of terminology ambiguity include the terms *interpretability* and *explainability* that are often used as synonyms in the field of machine learning. For example, the phrase "interpretable machine learning technique" can refer to techniques for generating ad-hoc explanations for non-interpretable models such as Deep Neural Network (DNN) [44]. Another example is the occasional case of using the terms *transparent system* and *explainable system* interchangeably in HCI research [45], while others clarify that explainability is not equivalent to transparency because it does not require knowing the flow of the bits in the AI decision-making process [22].

Table 2.1: Table of common terminologies related to Intelligible Systems and Transparent AI. Higher-level main concepts are shown in gray, while related terms for the main concepts are listed below and categorized as a desired outcome, property, or practical approach. Explainable AI is one particular practical approach for intelligible systems to enable improve transparency. Note that definitions and interpretations can vary across the literature, and this table is meant to serve as a quick reference. Reprinted from Mohseni et al. [3].

Concept	Category	Description
Intelligible System	Main Concept	A system that is understandable and predictable for its users though transparency and explanations [39, 16, 36].
Understandability (Intelligibility)	Desired Properties	Intelligible systems support user understanding of system's underlying functions [40, 46].
Predictability		Intelligibility supports building a mental model of the system that enables user to predict system behavior [36].
Trustworthiness	Desired Outcomes	Enabling positive user attitude toward the system that emerges from knowledge, experience, and emotion [47, 48].
Reliability		Supporting user trust to rely and follow systems advice for higher performance [47, 48].
Safety		Improving safety by reducing user unintended misuse due to misperception and unawareness [42].
Transparent AI	Main Concept	An AI-based system that provides information about its decision-making processes [43, 37].
Interpretable AI	Practical Approaches	Inherently human-interpretable models due to their low complexity of machine learning models [44].
Explainable AI		Supporting user understanding of complex models by providing explanations for predictions [49].
Interpretability	Desired Properties	The ability to support user understanding and comprehension of the model decision making process and predictions [43, 40].
Exaplainability		The ability to explain underlying model and its reasoning with accurate and user comprehensible explanations [43, 40].
Accountable AI	Desired Outcomes	Allowing for auditing and documentation to hold organizations accountable for their AI-based products and services [50, 22].
Fair AI		Enabling ethical and fairness analysis of models and data used in decision-making processes [50, 40].

2.2 Human Factors in Explainable AI

Research goals for XAI systems in the field of HCI are to improve the end-user experience, reliance, and ultimately task performance with the help of intelligent systems. The main targeted user group in HCI research are AI novices who use AI products in daily life but have no (or very little) expertise in machine learning systems. These include end-users of intelligent applications like personalized agents (e.g., home assistant devices), social media, and e-commerce websites. In most smart systems, machine learning algorithms serve as internal functions and APIs in a more extensive application. In these cases, XAI systems are expected to respond directly to their end-users with a human-understandable explanation of their predictions or suggestions. In this regard, creating an abstract and yet accurate representation of complicated machine learning explanations for novice end-users is a challenge for XAI explanation design.

User studies for human subject experiments are common methods in evaluating XAI systems. This line of research explores different XAI designs and studies the effects of different machine learning explanation types and complexity on end-users. In the rest of this section, I review HCI papers studying different aspects of XAI systems with end-users.

2.2.1 Explanations and User Trust

Trust is an essential factor in human-AI collaboration to maximize task performance. Users justified trust could boost the collaboration by avoiding erroneous model predictions. Measuring user trust in AI-based systems is a particularly sensitive task due to complex interactions between various human factors. Previous research studied how the variety of these factors such as user pre-knowledge [51], the system's stated performance [52], user perception of system performance [53], the expectation of AI performance [54], and user experience with interfaces [55] could affect user trust. Further, Eiband et al. [56] present a study in which placebic explanations (randomly generated explanations) in a food recommendation system improved user trust in the intelligent agent compared to the no-explanation baseline.

In studying the effects of model explanations in user trust, multiple studies show that trans-

parency helps users to see the strength and weaknesses of the intelligent agent and adjust their trust accordingly. For example, Nourani et al. [53] show users perceive a significantly lower level of accuracy when seeing model explanations that do not align with their reasoning. When comparing the effects of model performance and transparency on trust, studies in different domains have shown the model's performance is more effective on user reliance compared to its explanations. For instance, Wang et al. [57] study shows that the effect of transparency on adjusting user trust is less than the effect of the agent's success rate in a human-robot collaboration setup. They observed a moderate correlation between the robot's success rate and user trust on the robot. However, when comparing effects of robot ability and explanations on human-robot interactions, they did not observe significant effect from transparency on users' trust and compliance on high-ability robot. On the importance of user pre-knowledge and biases, Yin et al. [52] show that the effects of system stated performance on users' trust is significantly higher than the effects of user observed performance during the study. However, it still remains unclear whether user trust (with the help of transparency) could improve human-AI collaboration in complex tasks and scenarios.

2.2.2 Explanations and User Mental Model

Multiple studies on transparent AI explore design choices for building accurate mental model of algorithms and adjust end-users' reliance on AI systems. For instance, Kocielnik et al. [54] investigate accuracy indicator, example-based explanations, and user control as design choices to improve human-AI collaboration. Their findings indicate that users' perception of control had a significant positive effect on user trust. In the evaluation of XAI interfaces, Poursabzi et al. [58] present a comprehensive evaluation study for users' mental model (via user prediction task) and trust (via user agreement with AI) in interpretable models. Their results indicate the positive effect of interpretability on participants' mental model, however, they did not observe improvement on user trust. On the other hand, Papenmeier et al. [59] present a case study in which users could potentially lose trust in AI when exposed to low fidelity explanations. This is an indicator of the effects of transparency on user mental model and appropriate reliance of users on algorithms. Another example is the effect of AI system's updates on users' mental model. Bansal et al. [60]

present a case in which updates to increase AI's predictive performance may infact, hurt human-AI collaborative performance. Their study on an AI-advised decision-making setup shows that a better user mental model improves overall team performance, however, this break when users see behavioral changes after AI system updates.

2.2.3 Explanations and Task Performance

Studying task performance in human-AI collaboration is an essential topic as more intelligent systems are integrated in our day to day interactions. Since every intelligent system has its own limitations, a successful partnership could be built with users developing insights into intelligent system's strengths and weaknesses. In a recent paper, Ray et al. [61] run human subject studies to examine the effects of different types of explanations on user satisfaction and performance. Their results indicate a positive correlation between user satisfaction and task performance. Also, they found that correct explanations at the time of model failure help were the most effective to improve task performance. This indicates that the complex nature of machine learning algorithms requires users to build a mental model of the intelligent system. On studying users' mental model, Bansal et al. [62] measured attributes of AI systems that help users to build a better mental model and hence boost human-AI team performance. In a low-dimensional setup, they show positive effects of parsimony and non-stochasticity of AI error boundaries on the human mental model. However, their findings are based on low dimensional task may not be generalizable on more complex tasks such as image and text classification. For instance, Lai and Tan [63] run a series of studies to study model explanation types on deception detection task. Their results show that model explanations do not have a significant effect on end-users task performance.

2.3 Visual Analytics to Enable Transparency

Visual analytics and data visualization fields study methods and tools for expert users, including data scientists and domain experts who use machine learning for analysis, decision-making, or research in different domains. Additionally, in recent years, there has been an increase in visual analytics tools for machine learning experts who design and tune machine learning algorithms for

different domains and applications. In the following, I divide the review of visual analytics tools to enable interpreting machine learning models in two parts for data experts and machine learning experts.

2.3.1 Visual Analytics for Data Scientists

Data experts often use interactive data analysis tools, recommender systems, or visual analytics systems that combine interactive interfaces and algorithms. Examples of visual analytics exists in different applications such as cybersecurity [64, 65], medical [66, 67], text analysis [68, 69], and satellite image analysis [70]. Data experts can benefit from machine explanations to inspect uncertainty and investigate algorithms prediction accountability. For example, machine-learning explanations help data experts to find problems with training-bias in supervised machine learning models. Therefore, the main challenge for data-analysis and decision-support systems is to increase model transparency and user awareness with visualization and interaction techniques [71]. Visual analytics approaches can help data experts tune machine learning parameters for their specific data in an interactive visual fashion. Visualizing details and explanations of machine learning output may result in a better understanding of the machine algorithms' behavior [68]. Lastly, visual analytic systems have been used to aid fair data-driven decision making by quantifying and visualizing different notions of fairness for diagnosis and bias mitigation [72].

Similar to evaluations with AI novices, evaluating analytics tools for data knowledgeable users and domain experts often involves human subjects. However, many interpretable analytics tools are designed for data and machine learning experts. Visual analytics expert evaluations enter when controlled experiments fail due to high cognitive tasks [73]. In practice, it can be challenging to gain access or take the time of large numbers of experts for evaluation, which often makes it difficult to evaluate with controlled studies.

2.3.2 Visual Analytics for Machine Learning Experts

Machine learning researchers and engineers use visual analytics tools to visualize model architecture and training process to verify model performance and robustness [74, 75]. A line of

visual analytics tools present interactive visualization of model internals. For example, Kahng et al. [76] present a tool for visualizing instance-level and subset-level of neuron activation that is designed for machine learning engineers. In another work, Wang et al. [77] presented DNN Genealogy, an interactive visualization tool that offers a visual summary of DNN representations. Similarly, Hohman et al. [78] present an interactive system that scalably summarizes and visualizes what features a DNN model has learned and how those features interact in instance predictions. Their visual analytic system presents activation aggregation to discover important neurons and neuron-influence aggregation to identify interactions between important neurons. LSTMVis [79] and RNNVis [80] are also tools to interpret Recurrent Neural Network (RNN) models for natural language processing tasks.

Another critical role of visual analytics for machine learning experts is to visualize model training processes [81]. An example of a visual analytics tool for diagnosing the training process of a deep generative model is DGMTracker [74], which helps experts understand the training process by visually representing training dynamics. An evaluation of DGMTracker was conducted in two case studies with experts to validate the efficiency of these tools in supporting understanding of the training process and diagnosing a failed training process.

2.4 Interpretabilty for Machine Learning Algorithms

Machine learning experts are scientists and engineers who design interpretable machine learning algorithms, as well as other machine learning algorithms. Here, I first briefly review different types of interpretability techniques and then go over evaluation techniques for model explanations in more extend.

Interpretation methods to explain predictions of DNNs and other black-box models could generally be grouped into four categories [82]. The first category is the back-propagation based methods, which calculate the gradient or variants of gradients of a model prediction in terms of the model input [83]. The features in the input with large gradient values would have more significant contribution to the model prediction. The second category is perturbation based methods in which the key idea is to perturb the input sample and the features with more contributions once

perturbed would cause higher changes in the model prediction [84, 85]. Another approach is the local approximation of deep models to explain each prediction. Although the whole model behavior is highly intricate, the local behavior around an input instance could be approximated and well explained. Local model behavior for an input instance could be either approximated using a linear model (such as sparse linear model [86]), or an interpretable non-linear model (such as if-then rules [87]), depending on the property and the need for explanations. The last category is decomposition-based methods [88]. Note that the former three categories are mainly based on heuristics or approximations and thus generate explanations that might not be faithful to the original model. In contrast, decomposition techniques could be more faithful in reflecting the decision making process of the original deep model. In a recent paper, Du et al. [89] present a technique for recurrent neural networks to decompose predictions into additive contribution of each input word by modeling the information flow process from the input text to the model output.

Model explanations can be evaluated with computational methods rather than human-subject reviews to validate the explanations' trustworthiness. Computational evaluation methods are common in the field of machine learning and focus on measuring the correctness and completeness of the explanations in terms of mirroring what the model has learned. In the following subsections, I review two evaluation approaches for measuring *trustworthiness of model explanations* with and without ground-truth and inspecting *fidelity of the interpretability technique* with computational methods.

2.4.1 Explanations Trustworthiness

I review two approaches for evaluation of model explanations with and without ground truth. The two groups show a trade-off in objectivity of the evaluation methods.

An objective way to quantify the correctness of model explanation is to examine it against a ground truth baseline. Ground truth is often defined by human annotation of representative features (i.e., feature masks) and provide a baseline for quantitative evaluation of explanations quality. Examples include annotations of the target class (e.g., objects in image, sentences in text) to create "binary mask" in natural datasets [90], and synthesized datasets [91], that represent specific

features associated with the target class. Different similarity metrics, such as Intersection over Union (IoU) (also called Jaccard index) and mean Average Precision (mAP), are used to quantify the quality of model saliency explanations or bounding boxes compared to the ground truth. For instance, Li et al. [92] used IoU, between the model saliency map from a Convolutional Neural Network (CNN) and the ground truth binary mask from the validation set, to measure their quality as a weakly-supervised semantic segmentation task. In another work, Du et al. [93] calculate the mAP between the bounding boxes of an objects' saliency mask and the ground truth bounding boxes to evaluate their interpretability technique as an object localization task. Similarly, in the text domain, direct comparison of model attention explanations with human annotated sentences, e.g., evidence supporting the target label [94], provides an explanation quality score [95]. However, the relationship between the evaluation of machine learning explanations and the auxiliary tasks, such as binary object localization and semantic segmentation, is not clear yet.

Another common approach for evaluating machine learning explanations is the direct review of model explanations with end-users for their subjective feedback. Multiple papers have reported measurements of users' understanding of explanations as a proxy for human interpretability of explanations [96, 58]. Others have measured user-reported trust as a proxy for explanation goodness. For example, Papenmeier et al. [59] studied the effects of explanation meaningfulness and ad-hoc explainer fidelity on user reliance. Both studies show that model accuracy and explanation fidelity impact users' trust in the model and conclude that providing nonsensical explanations (i.e., those that do not align with users' expectations) may harm users' reported trust and observed reliance on the system. With a crowdsourced evaluation approach, Schmidt and Biessmann [97] present quantitative measures for system interpretability and human trust. They propose that analyzing user interaction time can serve as a proxy for users' understanding of the explanation and level of trust. They recommend that model explanations need to enhance the information transfer rate to users, help users establish an intuitive understanding of system performance and perform well independent from the user task. Taking a different perspective, Schneider et al. [98] inspected the effects of deceptive model explanations in a user study. Their findings indicate that explanations

that are unfaithful to the black-box model can fool users in accepting wrong predictions. Following a similar goal, Lakkaragu et al. [99] present an approach to generate misleading explanations and a case study with law and criminal justice domain experts. Their study results found that misleading explanations were able to significantly increase users' trust. Conclusively, various research efforts have shown the limitations of human judgment for robust evaluation of machine learning explanations.

2.4.2 Fidelity of the Interpretability Techniques

Research shows different approaches to examine the fidelity of interpretability techniques to the black-box model. A basic method to evaluate the ad-hoc explainer's fidelity is to examine it in comparison to an inherently interpretable model. For example, Ribeiro et al. [86] compared explanations generated by their ad-hoc explainer to explanations from an interpretable model. They created gold-standard explanations directly from the interpretable models (sparse logistic regression and decision trees) and used these for comparisons in their study. A downside of this approach is that the evaluation is limited to generating a gold standard by an interpretable model. In some cases, comparing a new explanation technique with existing state-of-the-art explanation techniques is a way to verify explanation quality [100, 101, 102]. For instance, Ross et al. [103] designed a comprehensive set of empirical evaluations and compared their explanations' consistency, features, and computational cost with the LIME technique [86].

To present a comprehensive evaluation setup, Samek et al. [104] and Hooker et al. [105] proposed a framework and benchmark for evaluating different aspects of saliency explanations for image data that quantify the importance of pixels with respect to the classifier prediction. They compared multiple saliency explanation technique for image data (e.g., sensitivity-based [106], deconvolution [107], and back-propagation [108]) and investigated the correlation between saliency map quality and network performance on different image datasets under input perturbation. On the contrary, Kindermans et al. [109] show interpretability techniques have inconsistencies on simple image transformations, hence their saliency maps can be misleading. They define an input invariance property for reliability of explanations from saliency methods. To extend a similar idea,

Adebayo et al. [110] propose three tests as sanity checks to measure correctness and completeness of interpretability techniques for tasks that are sensitive to either data or model.

2.5 Related Surveys and Guidelines

In recent years, there have been surveys and position papers suggesting research directions and highlighting challenges in interpretable machine learning research [111, 43, 112]. Although my literature review is limited to computer science literature, here I summarize several of the most relevant peer-reviewed surveys related to the topic of XAI across active disciplines including *Social Science*. While all surveys, models, and guidelines in this section add value to the XAI research, to the best of my knowledge, there is no existing comprehensive survey and framework for evaluation methods of explainable machine learning systems.

2.5.1 Social Science Surveys

Research in the social sciences is particularly important for XAI systems to understand how people generate, communicate, and understand explanations by taking into account each others' thinking, cognitive biases, and social expectations in the process of explaining. Hoffman, Mueller, and Klein reviewed the key concepts of explanations for intelligent systems in a series of essays to identify how people formulate and accept explanations, ways to generate self-explanations, and identified purposes and patterns for causal reasoning [113, 114, 115]. They lastly focus on DNNs to examine their theoretical and empirical findings on a machine learning algorithm [116]. In other work, they presented a conceptual model of the process of explaining in the XAI context [48]. Their framework includes specific steps and measures for the goodness of explanations, user satisfaction and understanding of explanations, users' trust and reliance on XAI systems, effects of curiosity on the search for explanations, and human-XAI system performance.

Miller [117] suggests a close collaboration between machine learning researchers in the space of XAI with social science would further refine the explainability of AI for people. He discusses how understanding and replicating how people generate, select, and present explanations could improve human-XAI interactions. For instance, Miller reviews how people generate and select

explanations that are involved with cognitive biases and social expectations. Other papers reviewing social science aspects of XAI systems include studies on the role of algorithmic transparency and explanation in lawful AI [22] and of fair and accountable algorithmic decision-making processes [50].

2.5.2 Human Computer Interactions Surveys

Many HCI surveys discuss the limitations and challenges in AI transparency [118] and interactive machine learning [119]. Others suggest a set of theoretical and design principles to support intelligibility of intelligent system and accountability of human users (e.g., [120, 16]). In a recent survey, Abdul et al. [39] presented a thorough literature analysis to find XAI-related topics and relationships among these topics. They used visualization of keywords, topic models, and citation networks to present a holistic view of research efforts in a wide range of XAI related domains; from privacy and fairness to intelligent agents and context-aware systems. In another work, Wang et al. [49] explored theoretical underpinnings of human decision-making and proposed a conceptual framework for building human-centered decision-theory-driven XAI systems. Their framework helps to choose better explanations to present, backed by reasoning theories, and human cognitive biases. Focused on XAI interface design, Eiband et al. [45] present a stage-based participatory process for integration of transparency in existing intelligent systems using explanations. Another design framework is XAID from Zhu et al. [121], which presents a human-centered approach for facilitating game designers to co-create with machine learning techniques. Their study investigates the usability of XAI algorithms in terms of how well they support game designers.

2.5.3 Visual Analytics Surveys

XAI-related surveys in the visualization domain follow visual analytics goals such as understanding and interacting with machine learning systems in different visual analytics applications [122, 123]. Choo and Liu [124] reviewed challenges and opportunities for Visual Analytics for explainable deep learning design. In a recent paper, Hohman et al. [125] provide an excellent review and categorization of visual analytics tools for deep learning applications. They cover

various data and visualization techniques that are being used in deep visual analytics applications. Also, Spinner et al. [126] proposed a XAI pipeline which maps the XAI process to an iterative workflow in three stages: model understanding, diagnosis, and refinement. To operationalize their framework, they designed explAIner, a visual analytics system for interactive and interpretable machine learning that instantiates all steps of their pipeline.

2.5.4 Machine Learning Surveys

Du et al. [82] present a survey and categorization of interpretability methods for black-box models. They review explanation techniques for DNNs in four groups of (1) back-propagation based methods (2) perturbation based methods (3) local approximation of deep models and (4) decomposition-based methods. Looking at a broader spectrum, Guidotti et al. [127] present a comprehensive review and categorization of machine learning interpretability techniques by their explanation method and type of black box system. Also, Montavon et al. [128] focus on interpretability techniques for DNN models. On CNN models, Zhang et al. [129] reviews research on interpretability techniques in six directions including visualization of CNN representations, diagnosing techniques for CNNs, approaches for transforming CNN representations into interpretable graphs, building explainable models, and semantic-level learning based on model interpretability. In another work, Gilpin et al. [130] reviews interpretability techniques in machine learning algorithms and categorizes evaluation approaches to bridge the gap between machine learning and HCI communities.

3. XAI DESIGN AND EVALUATION FRAMEWORK*

3.1 Introduction

One if the contributions of my research is to organize findings and share knowledge between disciplines to further enhance the XAI research. Reviewing the broad spectrum of XAI research indicates that scholars from different disciplines pursue different objectives and aspects of XAI systems to achieve the general goals of accomplishing explainability of AI. The diverse objectives between disciplines results in different design goals and evaluation measures for machine learning models and interface design (see Figure 1.1) of the XAI system. Therefore, a holistic and more actionable vantage will require consideration of interests from the different research communities (as identified from the literature review in Section 2.5) to help promote interdisciplinary progress in the XAI research.

This section presents my categorization of XAI system goals and evaluation measures (Section 3.1.2) drawn from my systematic review of literature in the fields of machine learning, HCI, and data visualization. The categorization is concluded by a design and evaluation framework (Section 3.1.3) to present the relationship between the goals and measure in a multidisciplinary XAI system design process. The framework presents step-by-step guidance for iterative design and evaluation loops in multidisciplinary teams with summarized ready-to-use evaluation methods for different goals for each design step.

3.1.1 Survey Method

I conducted a survey of the existing research literature to capture and organize the breadth of designs and goals for XAI evaluation. I used a structured and iterative methodology to find XAI-relevant research and categorize the evaluation methods presented in research articles (summarized in Figure 3.1). In an iterative paper selection process, I started by selecting existing work

* Parts of the material in this chapter are reprint or adapted from [3]. Mohseni et al. "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems" accepted for publication at accepted for publication in ACM Transactions on Interactive Intelligent Systems. Reproduced with permission.

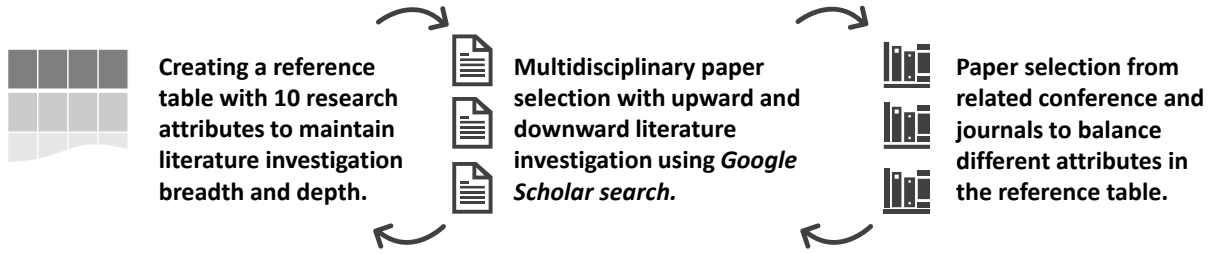


Figure 3.1: A diagram summarizing my iterative and multi-pass literature selection and review process to achieve desired literature investigation breadth and depth. I started with 40 papers to create the reference table. Then I added 80 papers by upward and downward literature investigation to improve review breath and depth. Finally, I added another 80 papers from related conferences proceedings and journals to balance the reference table. Reprinted from Mohseni et al. [3].

from top computer science conferences and journals across the fields of HCI, visualization, and machine learning. However, since XAI is a quite fast growing topic, I also wanted to include *arXiv* preprints and useful discussions in workshop papers. I started with 40 papers related to XAI topics across three research fields including but not limited to research on interpretable machine learning techniques, deep learning visualization, interactive model visualization, machine explanations in intelligent agents and context-aware systems, explainable user interfaces, explanatory debugging, and algorithmic transparency and fairness.

Then I used selective coding to identify 10 main research attributes in those papers. The main attributes I identified include: research discipline (social science, HCI, visualization, or machine learning), paper type (interface design, algorithm design, or evaluation paper), application domain (machine learning interpretability, algorithmic fairness, recommendation systems, transparency of intelligent systems, intelligent interactive systems and agents, explainable intelligent systems and agents, human explanations, or human trust), machine learning model (e.g., deep learning, decision trees, SVM), data modality (image, text, tabular data), explanation type (e.g., graphical, textual, data visualization), design goal (e.g., model debugging, user reliance, bias mitigation), evaluation type (e.g., qualitative, computational, quantitative with human-subjects), targeted user (AI novices, data experts, AI experts), and evaluation measure (e.g., user trust, task performance, user mental model).

In the second round of collecting XAI literature, I conducted an upward and downward literature investigation using the *Google Scholar* search engine to add 80 more papers to the reference table. I narrowed down the search by XAI related topics and keywords including but not limited to: interpretability, explainability, intelligibility, transparency, algorithmic decision-making, fairness, trust, mental model, and debugging in machine learning and intelligent systems. With this information, I performed axial coding to organize the literature and started discussions on my proposed design and evaluation categorization.

Finally, to maintain reasonable literature coverage and balance the number of papers for each of the categories of design goals and evaluation measures, I added another 80 papers to the reference table. The conferences from which I selected XAI related paper were including but not limited to: CHI, IUI, HCOMP, SIGDIAL, UbiComp, CHI EA, AIES, VIS, ICWSM, IJCAI, KDD, AAAI, CVPR, and NeurIPS conferences. The journals included: Trends in cognitive science, Transactions on Cognitive and Developmental Systems, Cognition Journal, Transactions on Interactive Intelligent Systems, International Journal of Human-Computer Studies, Transactions on Visualization and Computer Graphics, and Transactions on Neural Networks and Learning Systems journals.

Following a review of over 200 papers, my categorization of XAI design goals and evaluation methods is supported by references from papers preforming design or evaluation of XAI systems. The reference table is available online to the research community to provide further insight beyond the discussions in this document. Table 3.1 shows a digest of my surveyed papers that contains 40 papers with both design and evaluation of XAI system. Later in the Sections 3.2, 3.3, and 3.4, I provide a series of tables to organize different evaluation methods from research papers with example references for each.

3.1.2 Categorization of XAI Design Goals and Evaluation Methods

While an ideal XAI system should be able to answer all user queries and meet all XAI concept goals [6], individual research efforts focus on designing and studying XAI systems with respect to specific interpretability goals and specific users. Evaluating the explanations can demonstrate and

https://github.com/SinaMohseni/Awesome-XAI-Evaluation

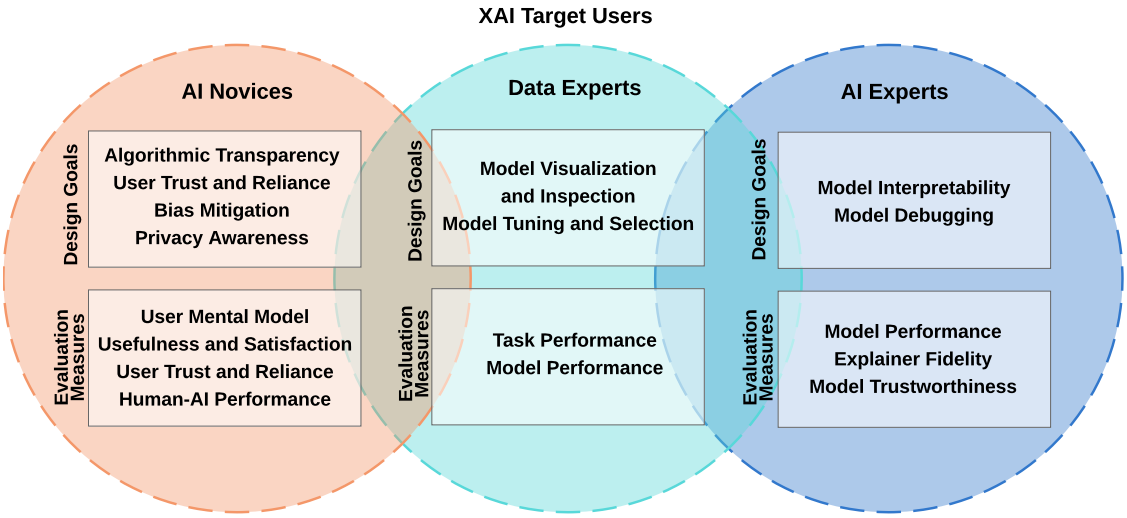


Figure 3.2: A summary of my categorization of XAI design and evaluation measures between user groups. Note that although there is overlap of XAI goals from different user groups, they require different design methods and elements for their target users. Reprinted from Mohseni et al. [3].

verify the effectiveness of the explainable systems for their initial goals.

After careful review and analysis of XAI goals and their evaluation methods in the literature, I recognized the following two attributes to be most significant for my purposes of interdisciplinary organization of XAI design and evaluation methods:

- **Design Goals**. The first attribute in my categorization is the design goal for interpretable algorithms and explainable interfaces in XAI research. I obtained XAI design goals from multiple research disciplines: machine learning, data visualization, and HCI. To better understand the differences between various goals for XAI, I categorized types of users of XAI systems into three groups: AI novices (i.e., general AI product end-user), data experts (experts in data analytics and domain experts), and AI experts (machine learning model designers).

- **Evaluation Measures**. I review evaluation methods and discuss measures used to evaluate machine learning explanations. The evaluation measures include user mental model, user trust and reliance, explanation usefulness and satisfaction, human-machine task perfor-

mance, and computational measures. The detailed review will emphasize more on evaluation measures of XAI as I found that this category is relatively less explored.

Figure 3.2 presents the pairing between XAI design goals and their evaluation measures for each user group. The overlap between XAI user groups shows similarities in the design and evaluation methods between different targeted user groups. To help summarize my characterization along with example literature, Table 3.1 presents a cross-reference table of XAI evaluation literature to emphasize the importance of design goals, evaluation measures, and user types. I review details design goals (eight XAI goals divided into their three user groups) and evaluation measures and methods (six main measures and their methods) at each framework layer in their appropriate design or evaluation step in Sections 3.2, 3.3, and 3.4.

3.1.3 A Nested Model for Design and Evaluation of XAI Systems

The variety of different XAI design goals and evaluation methods from our review (Section 3.1.2) suggests the need for diverse sets of techniques to build end-to-end XAI systems. However, it is generally insufficient to take design practices and evaluation methods separately. A holistic and more actionable vantage will require consideration of dependencies between design goals and evaluation methods and will inform when to choose between them during the design cycles. Previously, various models and guidelines for the design and evaluation of AI-infused interactive user interfaces [156, 157] and visual analytics systems [158] have been proposed to help designers through the design process. However, challenges in generating useful machine learning explanations and presenting them through an appropriate interface that aligns with target outcomes call for a multidisciplinary workflow framework.

Thus, based on our analysis of prior work, I propose a design and evaluation framework for XAI systems. The impetus for this framework is the desire to organize and relate the diverse set of existing design guidelines and evaluation methods in a unified model. The framework is intended to give guidance on what evaluation measures are appropriate to use at which design stage of the XAI system design. Figure 3.3 summarizes the framework as a nested model for end-to-end XAI

Table 3.1: Tabular summary of our XAI design goals and evaluation measures dimensions. The table includes 40 papers that represent a subset of the surveyed literature organized by the two dimensions. Reprinted from Mohseni et al. [3].

	Design Goals								Evaluation Measures				
	Novice Users				Data Experts		AI Experts						
Work	**G1: Algorithmic Transparency**	**G2: User Trust and Reliance**	**G3: Bias Mitigation**	**G4: Privacy Awareness**	**G5: Model Visualization and Inspection**	**G6: Model Tuning and Selection**	**G7: Model Interpretability**	**G8: Model Debugging**	**M1: Mental Model**	**M2: Usefulness and Satisfaction**	**M3: User Trust and Reliance**	**M4: Human-AI Task Performance**	**M5: Computational Measures**
Herlocker et al. 2000 [131]		◆								◆	◆	◆	
Kulesza et al. 2012 [132]	◆								◆	◆		◆	
Lim et al. 2009 [20]	◆								◆	◆			
Stumpf et al. 2018 [133]	◆	◆							◆		◆		
Bilgic et al. 2005[134]		◆								◆	◆		
Bunt et al. 2012 [135]	◆									◆			
Gedikli et al. 2014 [136]		◆								◆			
Kulesza et al. 2013 [137]	◆	◆							◆		◆		
Lim et al. 2009 [138]	◆	◆							◆	◆	◆	◆	
Lage et al. 2019 [96]	◆									◆		◆	
Schmid et al. 2016 [139]	◆											◆	
Berkovsky et al. 2017 [140]		◆								◆	◆		
Glass et al. 2008 [141]		◆								◆	◆		
Haynes et al. 2009 [142]		◆								◆	◆		
Holliday et al. 2016 [143]	◆	◆							◆		◆		
Nothdurft et al. 2014 [144]	◆	◆							◆		◆		
Pu and Chen et al. 2006 [145]		◆									◆	◆	
Bussone et al. 2015 [146]	◆	◆									◆		
Groce et al. 2014 [147]	◆								◆			◆	
Myers et al. 2006 [148]	◆								◆			◆	
Binns et al. 2018 [32]	◆		◆						◆				
Lee et al. 2019 [149]	◆		◆						◆	◆			
Rader et al. 2018 [33]	◆			◆					◆	◆			
Datta et al. 2015 [12]				◆									◆
Kulesza et al. 2015 [150]	◆				◆	◆			◆			◆	
Kulesza et al. 2010 [151]	◆				◆	◆			◆			◆	
Krause et al. 2016 [67]					◆	◆						◆	
Krause et al. 2017 [152]					◆	◆						◆	
Liu et al. 2014 [68]					◆							◆	
Ribeiro et al. 2016 [86]							◆		◆		◆	◆	◆
Ribeiro et al. 2018 [87]							◆		◆		◆	◆	◆
Ross et al. 2017 [153]							◆						◆
Adebayo et al. 2018 [110]							◆						◆
Samek et al. 2017 [104]							◆						◆
Zeiler et al. 2014 [107]							◆					◆	◆
Lakkaraju et al. 2016 [154]							◆					◆	
Kahng et al. 2018 [76]								◆		◆		◆	
Liu et al. 2018 [74]								◆		◆		◆	
Liu 2017 et al. 2009 [155]								◆		◆		◆	
Ming et al. 2017 [80]								◆		◆		◆	

system design and evaluation. The formulation of the model as layers relates to the core design goals and evaluation interests from the different research communities (as identified from the literature review) to help promote interdisciplinary progress in XAI research. The model is structured to support system design steps by starting from the outer layer (*XAI System Goals*), then addressing end-user needs in the middle layer (*Explainable Interface*), and finally focusing on underlying interpretable algorithms in the innermost layer (*Interpretable Algorithms*). The nested model is organized with a *Design Pole* focusing on design goals and choices, and an *Evaluation Pole* presenting appropriate evaluation methods and measures for each layer. Our framework suggests iterative cycles of design and evaluation to cover both algorithmic and human-related aspects of XAI systems. In this section, I elaborate on details of the nested framework and provide guidelines on using it for multidisciplinary XAI system design.

System design frameworks and models are intended to guide designers and developers to create interactive systems. However, frameworks can be more operational than a fixed road map for system design. I adapt Beaudouin-Lafon's [159] three dimensions of interaction models to XAI system design and evaluation process and present three goals for the XAI framework. First, *Generative Function* to help designers shape design thinking through guidelines. A multi-step framework would have between-steps guidelines to enhance multi-disciplinary team work and within-steps guidelines to providing design actions and evaluation measures at each step. Next is *Descriptive Function* to analyze and demonstrate an existing XAI system for post-hoc analysis. Such analysis of XAI design process helps in finding new insights and enhances communication. Lastly, the *Evaluative Function* helps to assess other design alternatives in the design process. The evaluation function provides recommendations for diagnosing XAI systems and identifying the next design iterations.

To showcase a practical example of using the framework, I also include a case study of a collaborative design and development effort for an XAI system. In the scenario of the case study, a multidisciplinary team of researchers designed a XAI system for fake news detection for non-expert (not AI experts or news analysts) daily newsreaders. The design team planned to add a

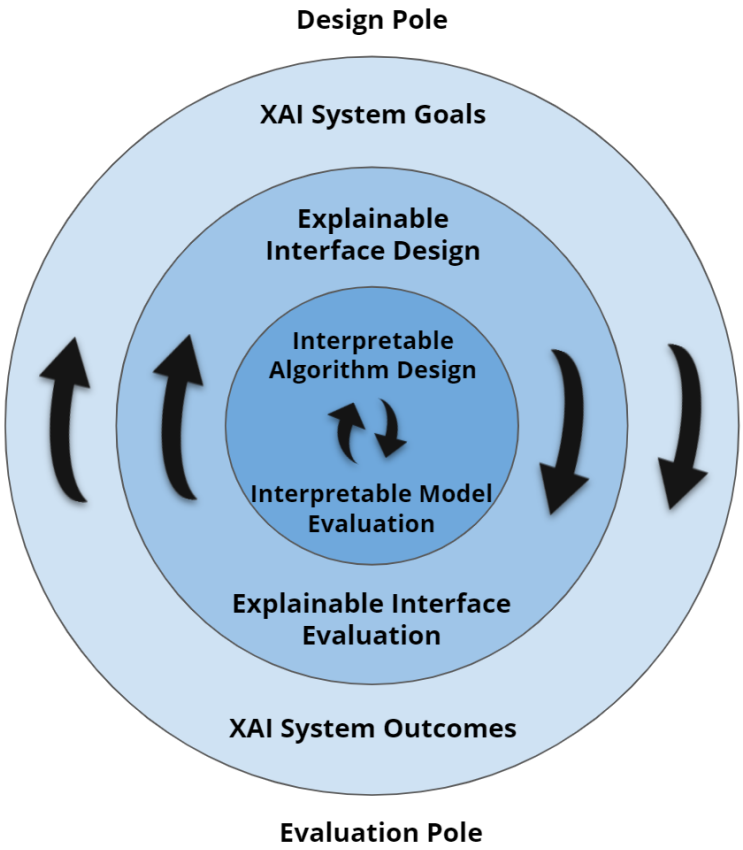


Figure 3.3: XAI design and evaluation framework. My proposed nested model for design and evaluation of explainable machine learning systems. The *outer layer* demonstrates system-level design goals which are paired with evaluation of high-level XAI outcomes. The *middle layer* shows explainable user interface and visualization design step paired with appropriate user understandability and satisfaction evaluation measures. The *innermost layer* presents design and evaluation of trustworthy interpretable machine learning algorithms. Reprinted from Mohseni et al. [3].

XAI Assistant feature to a news reading and sharing website to perform fake news detection. The system design consisted of a news reading interface equipped with the XAI news assistant (news assistant) to help the user identify fake news while reviewing news stories and articles. The presented case is summary of an ongoing research done over a one-year period by a team of eight university researchers with HCI, Visualization, and AI backgrounds. The details of design steps in this case study with comprehensive results and analysis will be reviewed in Section 4.2. During the following subsections, each framework guideline is followed by an example of its application in our case study.

3.2 Layer 1: System Design

As team members in a multidisciplinary team have different roles and priorities in building a XAI system, I suggest beginning the system design cycle from the *XAI System Goal* layer (the outer layer of Figure 3.3) to characterize design goal and system expectations. Specifically, this step involves identifying the *purpose for explanation* and choosing *what to explain* for the targeted end-user and dedicated application. The iterative refinements between XAI goal (top pole) and system outcome evaluation (bottom pole) present how the paired evaluation measures help to improve system design. Before reviewing the guidelines in the first layer of XAI framework, I categorize XAI goals for different users and review what can be explained from machine learning model in the next two subsections.

3.2.1 XAI Design Goals

Research efforts have explored many goals for XAI systems. Doshi-Velez and Kim [112] reviewed multiple high-level priorities for XAI systems with examples including safety, ethics, user reliance, and scientific understanding. Later, Arrieta et al. [40] presented a thorough review of XAI opportunities in different application domains. Accordingly, different design choices such as explanation type, scope, and level of detail will be affected by the application domain, design goal, and user type. For example, while machine learning experts might prefer highly-detailed visualizations of deep models to help them optimize and diagnose algorithms, end-users of daily-used AI products do not expect fully detailed explanations for every query from a personalized agent. Therefore, XAI systems are expected to provide the right type of explanations for the right group of users, meaning it will be most efficient to design an XAI system according to the user's needs and levels of expertise.

To this end, I distinguish XAI design goals based on the designated end-user and evaluation subjects, which I categorize into three general groups of AI experts, data experts, and AI novices. I emphasize that this separation of groups is presented primarily for organizational convenience, as goals are not mutually exclusive across groups, and specific priorities are case dependent for

any particular project. The XAI design goals also extend to the broader goal of *Responsible AI* by improving transparency and explainability of intelligent systems. Note that although there are overlaps in the methods used to achieve these goals, the research objectives and design approaches are substantially different among distinct research fields and their user groups. For instance, even though leveraging interpretable models to reduce machine learning model bias is a research goal for AI experts, bias mitigation is also a design goal for AI novices to avoid adverse effects of algorithmic decision-making in their respective domain settings. However, interpretability techniques for AI experts and bias mitigation tools for AI novice require different design methods and elements. In the following subsections, I review eight design goals for XAI systems organized by their user groups.

3.2.1.1 XAI Goals for AI Novices

AI novices refer to end-users who use AI products in daily life but have no (or very little) expertise on machine learning systems. These include end-users of intelligent applications like personalized agents (e.g., home assistant devices), social media, and e-commerce websites. In most smart systems, machine learning algorithms serve as internal functions and APIs to enable specific features embedded in intelligent and context-aware interfaces. Previous research shows intuitive interface and interaction design can enhance users' experience with the system through improving end-users' comprehension and reliance on the intelligent systems [160]. In this regard, creating human-understandable and yet accurate representations of complicated machine learning explanations for novice end-users is a challenging design trade-off in XAI systems. Note that although there are overlaps among goals for *AI Novices* and AI experts who build interpretable algorithms, each user group requires a different set of design methods and objectives that are being studied in different research communities.

The main design goals for AI novice end-users of XAI system can be itemized as the following:

G1: Algorithmic Transparency: An immediate goal for a XAI system – in comparison to an inexplicable intelligent system – is to help end-users understand how the intelligent system works.

Machine learning explanations improve users' mental model of the underlying intelligent algorithms by providing comprehensible transparency for the complex intelligent algorithms [118]. Further, transparency of a XAI system can improve user experience through better user understanding of model output [46], thus improving user interactions with the system [150].

G2: User Trust and Reliance: XAI system can improve end-users trust in the intelligent algorithm by providing explanations. A XAI system lets users assess system reliability and calibrate their perception of system accuracy. As a result, users' trust in the algorithm leads to their reliance on the system. Example applications where XAI aims to improve user reliance through its transparent design include recommendation systems [140], autonomous systems [161], and critical decision making systems [146] .

G3: Bias Mitigation: Unfair and biased algorithmic decision-making is a critical side effect of intelligent systems. Bias in machine learning has many sources, including biased training data and feature learning that could result in discrimination in algorithmic decision-making [162]. Machine learning explanations can help end-users to inspect if the intelligent systems are biased in their decision-making. Examples of cases in which XAI is used for bias mitigation and fairness assessment are criminal risk assessment [149, 32] and loan and insurance rate prediction [163]. It is worth mentioning that there is overlap between the biased decision-making mitigation goal for AI novices and the goal of dataset bias for AI experts (Section 3.2.1.2), which results in shared implementation techniques. However, the two distinct user groups require their own sets of XAI design goals and processes.

G4: Privacy Awareness: Another goal in designing XAI systems is to provide a means for end-users to assess their data privacy. Machine learning explanations can disclose to end-users what user data is being used in algorithmic decision-making. Examples of AI application examples in which privacy awareness is primarily important include personalized advertisements using users'

online advertisement [12] and personalized news feed in social media [33, 26].

Aside from major XAI goals, interactive visualization tools have also been developed to help AI novices to learn machine learning concepts and models by interacting with simplified data and model representations. Examples of these educative tools include TensorFlow PlayGround [164] for teaching elementary neural networks concepts and Adversarial Playground [165] for learning concept of adversarial examples in DNNs. These minor goals cover XAI system objectives that have limited extent compared to main goals.

3.2.1.2 XAI Goals for Data Experts

Data experts include data scientists and domain experts who use machine learning for analysis, decision-making, or research. Visual analysis tools can support interpretable machine learning in many ways, such as visualizing the network architecture of a trained model and training process of machine learning models. Researchers have implemented various visualization designs and interaction techniques to understand better and improve machine learning models.

Data experts analyze data in specialized forms and domains, such as cybersecurity [64, 65], medicine [66, 67], text [68, 69], and satellite image analysis [70]. These users might be experts of certain domain areas or experts in general areas of data science, but in my framework, I consider users in the *data experts* category to generally lack expertise in the technical specifics of the machine learning algorithms. Instead, this group of users often utilize intelligent data analysis tools or visual analytics systems to obtain insights from the data. Notice that there are overlaps between XAI goals in different disciplines and visual analytics tools designed by *Data Experts* could be used by both model designers and data analysts. However, design needs and approaches for these XAI systems may be different across research communities. The main design goals for data experts users of a XAI system are as follows:

G5: Model Visualization and Inspection: Similar to AI novices, data experts also benefit from machine learning interpretability to inspect model uncertainty and trustworthiness [71]. For in-

stance, machine-learning explanations help data experts to visualize models [78] and inspect for problems like bias [72]. Another important aspect of model visualization and inspection for domain experts is to identify and analyze failure cases of machine learning models and systems [166]. Therefore, the main challenge for data-analysis and decision-support systems is to improve model transparency via visualization and interaction techniques for domain experts [167].

G6: Model Tuning and Selection: Visual analytics approaches can help data experts to tune machine learning parameters for their specific data in an interactive visual fashion [68]. The interpretability element in XAI visual analytic systems increase data experts' ability to compare multiple models [168] and select the right model for the targeted data. As an example, Du et al. [169] present EventAction, an event sequence recommendation approach that allows the users to interactively select records that share their desired attribute values. In the case of tuning DNN networks, visual analytics tools enhance designers' ability to modify networks [75], improve training [74], and to compare different networks [170].

3.2.1.3 XAI Goals for AI Experts

In my categorization, *AI experts* are machine learning scientists and engineers who design machine learning algorithms and interpretability techniques for XAI systems. Machine learning interpretability techniques either provide model interpretation or instance explanations. Examples of model interpretation techniques include inherently interpretable models [171], deep model simplification [172], and visualization of model internals [173]. Instance explanations techniques, however, present feature importance for individual instances such as saliency map in image data and attention in textual data [174]. AI engineers also benefit from visualization and visual analytics tools to interactively inspect model internal variables [74] to detect architecture and training flaws or monitor and control the training process [76], which indicates possible overlaps among design goals. I list main design goals for AI Experts into two following items:

G7: Model Interpretability: Model interpretability is often a primary XAI Goal for AI ex-

perts. Model interpretability allows getting new insights into how deep models learn patterns from data [175]. In this regard, various interpretability techniques for different domains have been proposed to satisfy the need for explanation. For example, Yosinski et al. [173] created an interactive toolbox to explore CNN's activation layers in real-time that gives an intuition about "how the CNN works" to the user.

G8: Model and Training Debugging: AI researchers use interpretability techniques in different ways to improve model architecture and training process. For example, Zeiler and Fergus [107] present a use case of which visualization of filters and feature maps in CNN leads to revising training hyper-parameters and, therefore, model performance improvement. In another work, Ribeiro et al. [86] used model instance explanations and human review of explanations to improve model performance through feature engineering.

Other than main XAI goals for AI experts, machine learning explanations are used for other purposes including detecting dataset bias [176], adversarial example detection [177], and model failure prediction [178]. Also, visual saliency maps and attention mechanisms have been used as weakly supervised object localization [106], multiple object recognition [179], and knowledge transfer [180] techniques.

3.2.2 What to Explain

When users face a complex intelligent system, they may demand different types of explanatory information and each explanation type may require its own design. Here I review the two main categories of machine learning explanations (Global and Local Explanations) followed by six common types of explanations used in XAI system designs.

- **Global and Local Explanations** One way to classify explanations is by their interpretation scale. For instance, an explanation could be as thorough as describing the entire machine learning model. Alternatively, it could only partially explain the model, or it could be limited

to explaining an individual input instance. *Global explanation* (or *model explanation*) is an explanation type that describes how the overall machine learning model works. Model visualization [68, 155] and decision rules [154] are examples of explanations falling in this category. In other cases, interpretable approximations of complex models serve as the model explanation. Tree regularization [172] is a recent example of regularized complex model to learn tree-like decision boundaries. Model complexity and explanation design are the main factors used to choose between different types of global explanations.

In contrast, *local explanations* (or *instance explanations*) aim to explain the relationship between specific input-output pairs or the reasoning behind the results for an individual user query. This type of explanation is thought to be less overwhelming for novices, and it can be suited for investigating edge cases for the model or debugging data. Local explanations often make use of saliency methods [181, 107] or local approximation of the main model [86, 87]. Saliency methods (also as known as attribution maps or sensitivity maps) use different approaches (e.g., perturbation-based methods, gradient-based methods) to show what features in the input strongly influence the model's prediction. Local approximation of the model, on the other hand, trains an interpretable model (learned from the main model) to locally represent the complex model's behavior.

- **How Explanations** demonstrate a holistic representation of the machine learning algorithm to explain *how* the model works. For visual representations, model graphs [154] and decision boundaries [182] are common design examples for *How* explanations. However, research shows users may also be able to develop a mental model of the algorithm based on a collection of explanations from multiple individual instances [183].

- **Why Explanations** describe *why* a prediction is made for a particular input. Such explanations aim to communicate what features in the input data [86] or what logic in the model [87, 154] has led to a given prediction by the algorithm. This type of explanation can have either model agnostic [86, 102] or model dependent [184] solutions.

- **Why-not Explanations** help users to understand the reasons why a specific output was not in the output of the system [185]. *Why-not* explanations (also called *contrastive explanations*) characterize the reasons for differences between a model prediction and the user's expected outcome. Feature importance (or feature attribution) is commonly used as an interpretability technique for *Why* and *Why-not* explanations.

- **What-If Explanations** involve demonstration of how different algorithmic and data changes affect model output given new inputs [186], manipulation of inputs [138], or changing model parameters [54]. Different what-if scenarios may be automatically recommended by the system or can be chosen for exploration through interactive user control. Domains with high-dimensional data (e.g., image and text) and complex machine learning models (e.g., DNNs) have fewer parameters for users to directly tune and examine trained model compared to simpler data (e.g., low-dimensional tabular data) and models.

- **How-to Explanations** spell out hypothetical adjustments to the input or model that would result in a different output [138, 187], such as a user-specified output of interest. Techniques to generate *How-to* (or *counterfactual*) explanations are ad-hoc and model-agnostic considering that model structure and internal values are not a part of the explanation [188]. Such methods can work interactively with the user's curiosity and partial conception of the system to allow an evolving mental model of the system through iterative testing.

- **What-else Explanations** present users with similar instances of input that generate the same or similar outputs from the model. Also called *explanation by example*, *what-else* explanations pick samples from the model's training dataset that are similar to the original input in the model representation space [189]. Although very popular and easy to achieve, research shows example-based explanations could be misleading when training datasets lack uniform distribution of the data [190].

3.2.3 XAI Design Guidelines

I organize the following guidelines for the XAI goal layer. At the beginning of the system design process, the team will need to specify explainability requirements for each framework layer based on the evaluation metrics. The explainability requirements are intended to satisfy the main system goals defined by user (or customer) needs, and sometimes regulations, laws, and safety standards. Later, the evaluation step in each design cycle will have the team revisit the initial XAI system requirements. The sufficiency of the evaluation results in comparison to the initial design requirements serves as a key indicator of whether to stop or continue design iteration. However, since many subjective measures are used in the process, it is important to choose an appropriate evaluation baseline to track progress during design cycles.

3.2.3.1 Guideline 1: Determine XAI System Goals

Identifying and establishing clear goals and expectations from XAI system is the first step in the design process. XAI Design goals could be driven by many motivations like improving user experience on an existing system, advancing scientific findings [67, 191], or adhering to new regulations [192]. In Section 3.2.1 I reviewed eight main goals (G1-G8) for XAI systems. Also, ordering the priority of goals in cases with multiple design goals can be beneficial in the next steps of the process (see Guideline 2). Given the fact that different XAI user types and applications are interested in various design goals, it is important to establish these goals early in the design process to identify and align with appropriate design principles. A pitfall in this stage is to pick XAI goals without considering the end-user group, algorithmic limitations, and user preferences in the context of the application. Overshooting XAI goals could hurt evaluation results moving forward in the design process.

Application in Case Study: In the first step of our case study with a news curation application, the team started with identifying the main goals and expectations for the XAI news assistant. The design focused on novice end-users without any particular expertise. The XAI design goal was to improve user reliance and mental model of news predictions through explainable design. The team hypothesized that end-users would trust and rely on the fake news detection assistant, given that the new XAI is capable of providing explanations for each news story. Also, the team hoped that users would be able to use the explanations to learn model weaknesses and strengths to provide feedback to the developer team.

3.2.3.2 Guideline 2: Decide What to Explain

The second step in the XAI system design is to identify "what to explain" to the user in order to achieve the initial XAI goals (see Guideline 1) of the system. I reviewed multiple machine learning interpretability techniques and explanation types in Section 3.2.2 which can provide different types of information to the user. Although theory-driven design frameworks discuss explanation mechanisms driven by human reasoning semantics [187], user-centered methods to identify useful explanations such as online surveys, interviews, and user observations (e.g., [193, 146]) to understand *when* and *what* needs to be explained for the users to understand better and trust intelligent systems. Preliminary experiments are valuable in the early steps of the design cycle to identify and narrow down explanation options for the user in order to satisfy design goals. A typical approach for evaluating the effectiveness and usefulness of explanation choice in user-centric experiments is to compare the user's mental model of the system with and without explanation components. On this subject, Lim and Dey [20] conducted experiments to discover what type of information users are interested in different real-world context-aware application scenarios. Stumpf et al. [133] also performed end-user interviews to identify user perceptions and expectations from an interpretable interface as well as finding main decision points where users may need explanations. In another work, Haynes et al. [142] provide a review and studies incorporating different explanations (oper-

ational, ontological, mechanistic, and design rationale explanations) in intelligent systems. Similarly, visualization design involves expert interviews and focus groups in the design path to identify design goals [158].

The design process in this step involves algorithmic implementation constraints like "what can be explained" to the user. For example, model explanation of a DNN is not feasible and comprehensible due to the large number of variables in the graph. Additionally, research shows instance explanations from a DNN lack completeness and may fail to present salient features in cases [110]. Such constraints and decision points shall be solved through focused groups, brainstorming, and interviews between model designers and interface designers in the team. Therefore, a design pitfall for explanation choices is not to take limitations of interpretability techniques into account.

Application in Case Study: In our scenario, efficient news curation required fake news detection with the help of our XAI assistant. In the analysis of what the system should explain, the design team decided to identify candidate useful and impactful explanation options. I started with reviewing machine learning research on false information (e.g., rumor, hoax, fake news, clickbait) detection as well as HCI research on news feeds and news search systems to identify key attributes for news veracity checking [194]. Given the non-expert target end-users, explanatory information needed to limit technical details. Next, the user interface designers and machine learning designers in the team discussed candidate explanation choices and algorithmic constraints in interpretability techniques. That is, some options for what to explain may not be entirely possible given the interpretability of existing models, and the team needed to consider whether alternative learning techniques could provide better explanations or if the design team would need to figure out meaningful ways to explain the information that was available from the model.

3.2.4 XAI Outcomes Evaluation

Model explanations are designed to answer different interpretability goals, and hence different measures are needed to verify explanation validity for the intended purpose. For example, experimental design with human-subject studies is a common approach to perform evaluations with AI novice end-users. Various controlled in-lab and online crowdsourced studies have been used for XAI evaluation. Also, case studies aim to collect domain expert users' feedback while performing high-level cognitive tasks with analytics tools.

In this section, I review the main measures to evaluation XAI systems' outcome as presented in Table 3.1. I also provide summarized and ready-to-use XAI evaluation measures and methods extracted from the literature in Tables 3.2 and 3.3.

3.2.4.1 User Trust and Reliance

User trust in an intelligent system is an affective and cognitive factor that influences positive or negative perceptions of a system [195, 47]. Initial user trust and the development of trust over time have been studied and presented with different terms such as *swift* trust [196], *default* trust [197] and *suspicious* trust [198]. Prior knowledge and beliefs are important in shaping the initial state of trust; however, trust and confidence can update in response to exploring and challenging the system with edge cases [199]. Therefore, the user may have different feelings of trust and mistrust during different stages of experience with any given system.

Researchers define and measure trust in different ways. User knowledge, technical competence, familiarity, confidence, beliefs, faith, emotions, and personal attachments are common terms used to analyze and investigate trust [195, 201]. For these outcomes, user trust and reliance can be measured by explicitly asking about user opinions during and after working with a system, which can be done through interviews and questionnaires. Related to this, Ming et al. [52] studied the importance of model accuracy on user trust. Their findings show that user trust in a system was affected by both the system's stated accuracy and users' perceived accuracy over time. Ad-

Table 3.2: Evaluation measures and methods used in measuring user trust in XAI studies.

Trust Measures	**Evaluation Methods**
Subjective Measures	Self-explanation and Interview ([200, 146])
	Likert-scale Questionnaire ([200, 140, 53, 146])
Objective Measures	User Perceived System Competence ([52, 145, 53])
	User Compliance with System ([56])
	User Perceived Understandability ([52, 144])

ditionally, trust assessment scales could be specific to the systems application context and XAI design purposes. Similarly, Nourani et al. [53] explored how explanation inclusion and level of meaningfulness would affect the user's perception of accuracy. Their controlled experiment results show that whether explanations are human-meaningful can significantly affect perception of system accuracy independent of the actual accuracy observed from system usage. For example, multiple scales would assess user opinion on systems reliability, predictability, and safety separately. An example of a detailed trust measurement setup is presentation in the paper by Cahour and Forzy [200], which measures user trust with multiple trust scales (constructs of trust), video recording, and self-confrontation interviews to evaluate three modes of system presentation. Also, to better understand factors that influence trust in adaptive agents, Glass et al. [141] studied which types of questions users would like to be able to ask an adaptive assistant. Others have looked at changes to user awareness over time by displaying system confidence and uncertainty of the machine learning outputs in applications with different degrees of criticality [202, 203].

Multiple efforts have studied the impact of XAI on developing justified trust in users in different domains. For instance, Pu and Chen [145] proposed an organizational framework for generating explanations. They measured perceived competence and intention to return as measures for user trust. Another example compared user trust with explanations for different goals like transparency and justification explanation [144]. They considered *perceived understandability* to measure user trust and show that transparent explanations can help reduce the negative effects of trust loss in

unexpected situations.

Evaluating user trust in real-world applications, Berkovsky et al. [140] evaluated trust with various recommendation interfaces and content selection strategies. They evaluated user reliance on a movie recommender system with six distinct constructs of trust. Also on recommender algorithms, Eiband et al. [56] repeats the Langer et al.'s experiment [204] on the role of "placebic" explanations (i.e., explanations that convey no information) in mindlessness of user behavior. They studied if providing placebic explanations would increase user reliance on the recommender system. Their results suggest that future work on explanations for intelligent systems may consider using placebic explanations as a baseline for comparison with machine learning generated explanations. Also concerned with expert trust, Bussone et al. [146] measured trust by Likert-scale and think-alouds. They found explanations of facts that lead to higher user trust and reliance in a clinical decision-support system. Table 3.2 summarizes a list of subjective and objective evaluation methods to measure user trust in the machine learning system and explanations.

Many studies evaluate user trust as a static property. However, it is essential to take user's experience and learning over time into account. Collecting repeated measures over time can help in understanding and analyzing the trend of users' developing trust with the progression of experience. For instance, in their study, Holliday et al. [143] evaluated trust and reliance in multiple stages of working with an explainable text-mining system. They showed the level of user trust in the system varied over time as the user gained more experience and familiarity with the system.

I note that although my literature review did not find a direct measurement of trust to be commonly prioritized in analysis tools for data and machine learning experts, users' reliance on tools and the tendency to continue using tools are often considered as a part of the evaluation pipeline during interviews and case studies. In other words, my summarization is not meant to claim that data experts do not consider trust, but rather I did not find it to be a core outcome explicitly measured in the literature for this user group.

3.2.4.2 Human-AI Task Performance

A key goal of XAI is to help end-users to be more successful in their tasks involving machine learning systems [120]. Thus, human-AI task performance is a measure relevant to all three groups of user types. For example, Lim et al. [138] measured user performance in terms of test accuracy and task completion time to evaluate the impact of different types of explanations. They showed machine explanations to have a significant impact on users' accuracy in determining the way the machine learning system works. They use a generic test interface that can be applied to various types of sensor-based context-aware systems, such as weather prediction.

Also, explanations can assist users in adjusting the intelligent system to their needs. Kulesza et al. [132] study of explanations for a music recommender agent found a positive effect of explanations on users' satisfaction with the agent's output, as well as on users' confidence in the system and their overall experience.

Another use case for machine learning explanations is to help users judge the correctness of system output [147, 152, 34]. Explanations also assist users in debugging interactive machine learning programs for their needs [150, 151]. In a study of end-users interacting with an email classifier system, Kulesza et al. [150] measured users' mental model accuracy and classifier performance to show that explanatory debugging benefits both user and machine performance. Similarly, Ribeiro et al. [86] found users could detect and remove wrong explanations in text classification, resulting in training better classifiers by rewiring the algorithms and changing its logic. To support these goals, Myers et al. [148] designed a framework that users can ask *why* and *why not* questions and expect explanations from the intelligent interfaces. Table 3.3 summarizes a list of evaluation methods to measure task performance in human-AI collaboration and model tuning scenarios.

Visual analytics tools also help domain experts to better perform their tasks by providing model interpretations. Visualizing model structure, details, and uncertainty in machine outputs can allow domain experts to diagnose models and adjust parameters to their specific data for better analysis. Visual analytics research has explored the need for model interpretation in text [205, 206, 69] and multimedia [207, 208] analysis tasks. This body of work demonstrates the importance of integrat-

Table 3.3: Evaluation measures and methods used in measuring human-machine task performance in XAI studies.

Performance Measures	Evaluation Methods
User Performance	Task Performance ([151, 138, 76, 147])
	Task Throughput([151, 138, 154])
	Model Failure Prediction ([147, 152, 34])
Model Performance	Model Accuracy ([86, 150, 34, 155, 75])
	Model Tuning and Selection ([68])

ing user feedback to improve model results. An example of a visual analytics tool for text analysis is TopicPanaroma [68], which models a textual corpus as a topic graph and incorporates metric learning and feature selection to allow users to modify the graph interactively. In their evaluation procedure, they ran case studies with two domain experts: a public relations manager used the tool to find a set of tech-related patterns in news media, and a professor analyzed the impact of news media on the public during a health crisis. In analysis of streaming data, automated approaches are error-prone and require expert users to review model details and uncertainty for better decision making [209, 65]. For example, Goodall et al. [64] presented Situ, a visual analytics system for discovering suspicious behavior in cyber network data. The goal was to make anomaly detection results understandable for analysts, so they performed multiple case studies with cybersecurity experts to evaluate how the system could help users to improve their task performance. Ahn and Lin [72] present a framework and visual analytic design to aid fair data-driven decision making. They proposed FairSight, a visual analytic system to achieve different notions of fairness in ranking decisions through visualizing, measuring, diagnosing, and mitigating biases.

Other than domain experts using visual analytics tools, machine learning experts also use visual analytics to find shortcomings in the model architecture or training flaws in deep neural networks to improve the classification and prediction performance [155, 75]. For instance, Kahng et al. [76] designed a system to visualize instance-level and subset-level of neuron activation in a long-term investigation and development with machine learning engineers. In their case studies, they interviewed three Facebook engineers and data scientists who used the tool and reported the key

observations. Similarly, Hohman et al. [78] present an interactive system that scalably summarizes and visualizes what features a DNN model has learned and how those features interact in instance predictions. Their visual analytic system presents activation aggregation to discover important neurons and neuron-influence aggregation to identify interactions between important neurons. In the case of recurrent neural networks (RNN), LSTMVis [79] and RNNVis [80] are tools to interpret RNN models for natural language processing tasks. In a recent example, Wang et al. [77] presented DNN Genealogy, an interactive visualization tool that offers a visual summary of DNN representations.

Another critical role of visual analytics for machine learning experts is to visualize model training processes [81]. An example of a visual analytics tool for diagnosing the training process of a deep generative model is DGMTracker [74], which helps experts understand the training process by visually representing training dynamics. An evaluation of DGMTracker was conducted in two case studies with experts to validate efficiency of the tool in supporting understanding of the training process and diagnosing a failed training process.

3.2.5 XAI Evaluation Guidelines

Evaluation of XAI system outcomes is the final step in the evaluation process. Figure 3.3 shows how the final system outcome evaluation is paired with the initial design goals in the outer layer of my framework.

3.2.5.1 Guideline 3: Evaluate System Outcomes

The main goal of this stage is to quantitatively and qualitatively assess the effectiveness of the XAI system for the initially established system-level XAI goals. Clearly, evaluation of final system outcomes could be influenced by the design of the explainable user interface (intermediate layer) and the design of interpretable algorithms (innermost layer). For example, evaluating a newborn interpretable machine learning algorithm's output using human subjects through a weak in-lab or crowdsourced user study may not be meaningful or productive for XAI system outcomes if core computational changes are still in progress and could ultimately change the entire model

interpretability and explanation format later. Also, changes in the targeted user could affect evaluation results at this stage. For example, a system designed for novices may not satisfy the needs of an expert user and hence would not improve performance as expected. Evaluation measures in this layer depend on the design goals, application domain, and targeted users. Example evaluation measures for final system outcomes include user trust [145] and reliance on the system [140], human-machine task performance [62], user awareness [203], and user understanding of their personal data [33]. An effective process for evaluation of high-level XAI outcomes is to break down the evaluation goal into multiple well-defined measures and metrics. This way, the team can perform evaluation studies on different steps using valid methods in controlled setup. For example, in the evaluation of XAI systems for trustworthiness, several factors of human trust could be measured during and after a period of user experience with the XAI system. In addition, computational measures (Section 3.4.3.1) are used to examine the fidelity of interpretability methods and trustworthiness of the model with objective metrics. A possible pitfall in evaluation of the XAI system outcomes is performing the evaluation without considering the model trustworthiness and explanations' correctness from the interpretable model layer (see Guideline 7) and explanation understandability and usefulness from the user interface layer (see Guideline 5).

Application in Case Study: In our case study with news review and curation, we needed to evaluate our XAI news assistant with non-expert users who would gather news stories while flagging fake news articles. In the evaluation step, the team ran multiple large-scale human-subject studies with novice participants recruited through Amazon Mechanical Turk to work with our news reading system. Note that both the explainable interface and interpretable algorithm passed multiple design and testing iterations before this evaluation step. Major decisions for this evaluation was how to structure the duration and complexity of the user task while appropriately testing the system's full range of functionality. The task was designed with questions built in to help collect subjective data in addition to the objective user performance data. Multiple evaluation measures were chosen for system outcomes, including: subjective user trust in the news assistant, user agreement rate with the news assistant, veracity of user-shared news stories, and user accuracy in guessing the news assistant output. Both qualitative and quantitative analysis of user feedback and interaction data were valuable to the evaluation of system outcomes. The results and analysis from these evaluations helped the team to understand the effectiveness of the XAI elements (in both the algorithm and the interface) for the initial system goals.

3.3 Layer 2: Interface Design

The middle layer of my framework is concerned with designing and evaluating an explainable interface or visualization for the user to interact with XAI system. Interface design for explanations consists of presenting model explanations from interpretable algorithms to end-users in terms of their *explanation format* and *interaction design*. The importance of this layer is to satisfy design requirements and needs to be determined in the XAI system design layer (see Guideline 2). Hence, the iterative movement between *Design pole* and *Evaluation pole* in this layer presents design refinement in pursuit a desired goal state. An elegant translation of machine generated explanations (e.g., verbal, numeric, or visual explanation) needs carefully designed human-understandable and

satisfying explanations in the user interface. In the following I review multiple types of explanation formats for integrating XAI elements into the user interface.

3.3.1 How to Explain

In all types of machine learning explanations, the goal is to reveal new information about the underlying system. In this survey, I mainly focus on human-understandable explanations, though I note that research on interpretable machine learning has also studied other purposes such as knowledge transfer, object localization, and error detection [175, 177].

Explanations can be designed using a variety of formats for different user groups [167]. *Visual explanations* use visual elements to describe the reasoning behind the machine learning models. Attention maps and visual saliency in the form of saliency heatmaps [107, 106] are examples of visual explanations that are widely used in machine learning literature. *Verbal explanations* describe the machine's model or reasoning with words, phrases, or natural language. Verbal explanations are popular in applications like question-answering explanations and decision lists [154]. This form of explanation has also been implemented in recommendation systems [140, 131] and robotics [210]. Explainable interfaces commonly make use of multiple modalities (e.g., visual, verbal, and numerical elements) for explanations to support user understanding [148]. *Analytic explanation* is another approach to view and explore the data and the machine learning models representations [125]. Analytic explanations commonly rely on numerical metrics and data visualizations. Visual analytics tools also allow researchers to review model structures, relations, and their parameters in complex deep models. Heatmap visualizations [79], graphs and networks [64], and hierarchical (decision trees) visualizations are commonly used to visualize analytic explanations for interpretable algorithms. Recently, Hohman et al. [211] implemented a combination of visualization and verbalization to communicate or summarize key aspects of a model.

From a different perspective, Chromik et al. [212] extends the idea of "dark patterns" from interactive user interface design [213] into machine learning explanations. They review possible ways that phrasing of explanations and their implementation in the interface could deceive users for the benefit of *other parties*. They review negative effects such as lack of user attention to

explanations, formation of an incorrect mental model, and even algorithmic anxiety [214] could be among the consequences of such deceptive presentations and interactions of machine learning explanations.

3.3.2 User Interactions with XAI

Another important consideration in designing the XAI interface is if and how to leverage user interactions to better support system understandability. The benefits of interactive system design have been previously explored in the topic of interactive machine learning [119, 157] for novice end-users. AI and data experts also benefit from interactive visual tools to improve model and task performance [123]. Here, I discuss multiple examples of interaction design that support user understanding of the underlying black-box model.

Focusing on interactive design for AI-based systems for AI novices, Amershi et al. [119] reviewed multiple case studies that demonstrate the effectiveness of interactivity with a tight coupling between the algorithm and the user. They emphasize how interactive machine learning processes allow the users to instantly examine the impact of their actions and adapt their next queries to improve outcomes. Such interactions allow users to test various inputs and learn about the model by creating *What-If* explanations [49]. Particularly, user-led cyles of trial and error help novices to understand how the machine learning model works and how to steer the model to improve results. In the context of XAI, Jongejan and Holbrook [186] present a study in which users draw images and see whether an image recognition algorithm can correctly recognize the intended sketch. Their system and study allows for interactive trial-and-error to explore how the algorithm works. In addition, their system provides example-based explanations in cases where the algorithm fails to correctly classify drawings. Another approach is to allow users to control or tune algorithmic parameters to achieve better results. For example, Kocielnik et al. [54] present a study in which users were able to freely control detection sensitivity in an AI assistant. Their results showed a significant effect on user perception of control and acceptance.

Visual analytics tools also support model understanding for expert users through interaction with algorithms. Examples including allowing data scientists and model experts to interactively

explore model representations [78], analyze model training processes [74], and detect learning biases [215]. Also, embedded interaction techniques can support the exploration of very large deep learning networks. For instance, Hohman et al. [78] present multiple interactive features to select and filter of neurons and zoom and pan in feature representations to support AI experts in interpreting and reviewing trained models.

3.3.3 Interface Design Guidelines

The guidelines in this layer are helping to execute design requirements that are determined in the XAI system design layer (see Guideline 2). After reviewing multiple types of explanation formats and interaction designs for integrating XAI elements into the user interface in Section 3.3.1, I review the internal steps of this layer in the following guideline. The iterative movement between *Design pole* and *Evaluation pole* in this layer presents design refinement to achieve explainable interface.

3.3.3.1 Guideline 4: Decide How to Explain

Identifying candidate explanation formats for the targeted system and user group is the first step to deliver machine learning explanations to end-users. The design process can account for different levels of complexity, length, presentation state (e.g., permanent or on-demand), and interactivity options depending on the application and user type. The explanations format in the interface is particularly important to improve user understanding of underlying algorithms. Studies show that while detailed and complex interactive representations may aim to communicate the explanations to the expert users, AI-novice users of XAI system prefer more simplified explanation and representation interfaces [96]. User satisfaction of interface design is also another critical factor in user engagement of the interface components [160]. Additionally, interaction design for explainable interfaces can allow a user to communicate with the system to adjust explanations and could better support user inspection of the system [151].

Research of intelligent interface design presents multiple design methods such as wireframing and low-fidelity prototyping (e.g., [193, 146]) that could also be adapted to the explainable

interface design. Also, existing design guidelines and best-practice knowledge for AI-infused interfaces (e.g., [157]) and visualizations (e.g., [216]) could be used in this stage to leverage similar systems for explainable interface design. Aside from model explanations, providing prediction uncertainty also has been identified as an important factor for both general end-users and data expert users [71]. For example, Kay et al. [203] presented the full design cycle for an uncertainty visualization interface in a bus arrival time application. Their design process included surveying to identify usage requirements, developing alternative layouts, running user testing, and final evaluation of user understanding of machine learning output.

Application in Case Study: To determine *how* to explain news classification results to non-expert end users, the user interface design team started the process by reviewing the initial system goals and explanation types. The team then continued with multiple interface sketches that matched the intended application and user tasks. During the initial design steps, the team tried to keep a balance between interface complexity and explanation usefulness by choosing among available explanation types from our interpretable machine learning algorithms. Next, mock-ups from the top three designs were implemented for testing with a small number of participants. Each mock-up had a different arrangement of data, user task flow, and explanation format for the news assistant interface. Our human-subject experiments in this stage were based on user observations and post-usage interviews to collect qualitative feedback regarding participant understanding and subjective satisfaction of explanation components and interface arrangements. Interviews resulted in the selection of the most comprehensible and conclusive design among the available options to continue with (see Guidelines 5).

3.3.4 Explainability Evaluation

Following the evaluation measures for XAI system's outcomes in Section 3.2.4 In this section, I review the main measures to evaluation XAI systems' outcome as presented in Table 3.1. I also provide summarized and ready-to-use XAI evaluation measures and methods extracted from the

Table 3.4: User satisfaction measures and study methods used in measuring user satisfaction and usefulness of explanations in XAI studies.

Satisfaction Measures	Evaluation Methods
User Satisfaction	Interview and Self-report ([20, 136, 138, 135])
	Likert-scale Questionnaire ([218, 96, 20, 136, 138])
	Expert Case Study ([76, 79, 69, 219, 155])
Explanation Usefulness	Engagement with Explanations ([218])
	Task Duration and Cognitive Load ([138, 96, 136])

literature for explanations usefulness and satisfaction (Tables 3.4) and user mental model (3.3).

3.3.4.1 Explanation Usefulness and Satisfaction

End-user satisfaction and usefulness of machine explanation are also of importance when evaluating explanations in intelligent systems [134]. Researchers use different subjective measures for understandability, usefulness, and sufficiency of details to assess explanatory value for users [117]. Although there are implicit methods to measure user satisfaction [217], a considerable part of the literature follows qualitative evaluation of satisfaction in explanations, such as questionnaires and interviews. For example, Gedikli et al. [136] evaluated ten different explanation types with user ratings of explanation satisfaction and transparency. Their results showed a strong relationship between user satisfaction and perceived transparency. Similarly, Lim et al. [138] explore explanation usefulness and efficiency in their interpretable context-aware system by presenting different types of explanations such as "why", "why not" and "what if" explanations and measuring users response time.

Another line of research studies whether intelligible systems are always appreciated by the users or it is a conditional fact. An early work from Lim and Dey [20] studied user understanding and satisfaction of different explanation types in four real-world context-aware applications. Their findings show that, when considering scenarios involved with criticality, users want more

information explaining the decision making process and experience higher levels of satisfaction after receiving these explanations. Similarly, Bunt et al. [135] considered whether explanations are always necessary for users in every intelligent system. Their results show that, in some cases, the cost of viewing explanations in diary entries like Amazon and YouTube recommendations could outweigh their benefits. To study the impact of explanation complexity on users' comprehension, Lage et al. [96] studied how explanation length and complexity affect users' response time, accuracy, and subjective satisfaction. They also observed that increasing explanation complexity resulted in lowered subjective user satisfaction. In a recent study, Coppers et al. [218] also show that adding intelligibility does not necessarily improve user experience in a study with expert translators. Their experiment suggests that an intelligible system is preferred by experts when the additional explanations are not part of the translators readily available knowledge. In another work, Curran et al. [220] measured users' understanding and preference of explanations in an image recognition task by ranking and coding user transcripts. They provide three types of instance explanations for participants and show that although all explanations were coming from the same model, participants had different levels of trust in explanations' correctness, according to explanations clarity and understandability.

Table 3.4 summarizes the study methods used to measure user satisfaction and usefulness of machine learning explanations. Note that the primary goal of XAI system evaluations for domain and AI experts is through direct evaluation of user satisfaction of explanation design during the design cycle. For example, case studies and participatory design are common approaches for directly including expert users as part of the system design and evaluation processes.

3.3.4.2 Mental Model

Following cognitive psychology theories, a mental model is a representation of how a user understands a system. Researchers in HCI study users' mental models to determine their understanding of intelligent systems in various applications. For example, Costanza et al. [221] studied how users understand a smart grid system, and Kay et al. [203] studied how users understand and adapt to uncertainty in machine learning prediction of bus arrival times.

Table 3.5: Evaluation measures and methods used in studying user mental models in XAI systems

Mental Model Measures	**Evaluation Methods**
User Understanding of Model	Interview ([221]) and Self-explanation ([222, 223, 32])
	Likert-scale Questionnaire ([183, 224, 20, 137, 225, 154])
Model Output Prediction	User Prediction of Model Output ([203, 86, 87])
Model Failure Prediction	User Prediction of Model Failure ([62, 226])

In the context of XAI, explanations help users to create a mental model of *how the AI works*. Machine learning explanation is a way to help the user in building a more accurate mental model. Studying users' mental models of XAI systems can help verify explanation effectiveness in describing an algorithm's decision-making process. Table 3.5 summarizes different evaluation methods used to measure user mental model of machine learning models.

Psychology research in human-AI interactions has also explored structure, types, and functions of explanations to find essential ingredients of ideal explanation for better user understanding and more accurate mental models [227, 228]. For instance, Lombrozo [183] studied how different types of explanations can help structure conceptual representation. In order to find out how an intelligent system should explain its behavior for non-experts, research on machine learning explanations has studied how users interpret intelligent agents [222, 223] and algorithms [224] to find out what users expect from machine explanations. Related to this, Lim and Dey [20] elicit types of explanations that users might expect in four real-world applications. They specifically study what types of explanations users demand in different scenarios such as system recommendation, critical events, and unexpected system behavior. In measuring user mental model through model failure prediction, Bansal et al. [62] designed a game in which participants receive monetary incentives based on their final performance score. Although experiments were done on a simple three-dimensional task, their results indicate a decrease in users' ability to predict model failure as data and model get more complicated.

A useful way of studying user comprehension of intelligent systems is to directly ask the user about the intelligent system's decision-making process. Analyzing user interviews, think-alouds,

and self-explanations provides valuable information about the users' thought processes and mental models [151]. On studying user comprehension, Kulesza et al. [137] studied the impact of explanation soundness and completeness on fidelity of end-users mental model in a music recommendation interface. Their results found that explanation completeness (broadness) had a more significant effect on user understanding of the agent compared to explanation soundness. In another example, Binns et al. [32] studied the relation between machine explanations and users' perception of justice in algorithmic decision-making with different sets of explanation styles. User attention and expectations may also be considered during the interpretable interface design cycles for intelligent systems [133].

Interest in developing and evaluating human-understandable explanations has also led to interpretable models and ad-hoc explainers to measure mental models. For example, Ribeiro et al. [86] evaluated users' understanding of the machine learning algorithm with visual explanations. They showed how explanations mitigate human overestimation of the accuracy of an image classifier and help users choose a better classifier based on the explanations. In a follow-up work, they compared the global explanations of a classifier model with the instance explanations of the same model and found global explanations were more effective solutions for finding the model weaknesses [87]. In another paper, Kim et al. [225] conducted a crowdsourced study to evaluate feature-based explanation understandability for end-users. Addressing understanding of model representations, Lakkaraju et al. [154] presented interpretable decision sets, an interpretable classification model, and measured users' mental models with different metrics such as user accuracy on predicting machine output and length of users' self-explanations.

3.3.5 Interface Evaluation Guidelines

3.3.5.1 Guideline 5: Evaluate Explanation Usefulness

This mid-layer evaluation step can be used along with various measures to help assess user understanding of the XAI underlying intelligent algorithms. A series of user-centered evaluations of explainable interface with multiple goals and granularity levels could be performed to measure:

1. User understanding of explanation.

2. User satisfaction of explanation.

3. User mental model of the intelligent system.

Evaluations in the middle layer are particularly important due to the impact on XAI system outcomes (outer layer) and being affected by interpretable model outputs (inner-most layer). Specifically, evaluation measures in this stage can inform how well users understand the interpretable system, however, the design validity at this step also may be reflected by higher-level XAI outcomes (i.e., outer-layer evaluation) such as user trust and task performance. Note that user understanding of an XAI system could be limited to parts of the system rather than the entire system; similarly, understanding may be limited to a subspace of scenarios rather than all possible scenarios.

The three evaluation measures introduced for this step could be used on multiple iterative cycles to improve overall explainable interface design. For example, Saket et al. [229] studies users understanding of visualization encoding and effectiveness of interactive graphical encoding for end-user. On the other hand, user satisfaction of explanation type and format depends on factors such as targeted application criticality and user-preferred cognitive load [112]. Evaluating user mental model is also an effective way to measure usefulness of explainable interfaces. Tables 3.5 and 3.4 present a list of measures for evaluating explainable interfaces in this step. The choice of baseline is another important factor in evaluating explainable interfaces. Typically, a combination of qualitative and quantitative analysis are used to measure effects of explanation components (in comparison to non-explainable system) or to compare multiple explanations types. However, the choice of placebic explanations has been proposed as the evaluation baseline for more accurate measurement of explanation content [56]. In the case of expert review, evaluation of a domain expert's mental model commonly involves comparison with the AI expert's mental model and description of "how the model works". Section 3.5.3 reviews common choices of ground-truth baselines in XAI evaluation studies. With all approaches, updates in explanation components of the interface require assessment of their impact on user experience and understandability. How-

ever, the metrics and depth of evaluation vary during the evaluation cycles as the team narrows down specific needs. Finally, a possible evaluation pitfall for explainable interfaces is going after broad measures of XAI outcomes (See Guideline 3) rather than focusing on a narrower scope of explanation components and interactions.

Application in Case Study: In our case study, interface designers started evaluation of candidate explanation components by a series of small studies with a repeated-measures design so that the same study participant could experience different explanation designs in one session. Next, we analyzed quantitative and qualitative data collected from the end-users to choose candidate designs and routes to further improve the interface for explainable components. Discussions with the machine learning team also helped to find sources of limitations in the interpretability technique that could possibly affect user satisfaction. After the initial cycles of revision, we collected a round of external and internal expert reviews to update the study methodology and data collection details according to project progress.

3.4 Layer 3: Algorithm Design

The innermost layer of my framework involves designing interpretable algorithms that are able to generate explanations for the users. The last design step in my XAI system framework is the choice of interpretability technique (design pole) to generate the outlined explanation types. However, evaluating the generated explanation (evaluation pole) is the first evaluation step before human-subject evaluations in the explainable interface.

Ideally, the interpretability techniques should generate explanations in accordance with the requirements in the explainable interface design step (see Guideline 4); however, the choice of interpretability technique depends on domain and carries implementation limitations. For example, while shallow models are desired for their high interpretability, these models typically do not perform well in cases of complex and high dimensional data like image and text. On the other hand, highly accurate predictions in black-box models (e.g., deep neural networks and random

forest models) require post-processing and ad-hoc algorithms to generate explanations. The ad-hoc approach also has limitations on both choice of explanation type and need for completeness [110] and fidelity [86] validation compared to the original model. This shows not only machine learning designers should consider the trade-off between model interpretability and performance but also should consider the fidelity of the ad-hoc explainer to black-box model.

3.4.1 Interpretability Techniques

The human interpretability of a machine learning model is inversely proportional to the model's size and complexity. Complex models (e.g., deep neural networks) with high performance and robustness in real-world applications are not interpretable by human users due to their large variable space. Linear regression models or decision trees offer better interpretability but have limited performance on high-dimensional data, whereas a random forest model (ensemble of hundreds of decision trees) can have much higher performance but is less understandable. This trade-off between model interpretability and performance led researchers to design ad-hoc methods to explain any black-box machine learning algorithm such as deep neural networks. Ad-hoc explainers (e.g., [86, 102]) are independent algorithms that can describe model predictions by explaining "why" a certain decision has been made instead of describing the whole model. However, there are limitations in explaining black-box models with ad-hoc explainers, such as the uncertainty of the fidelity of the explainer itself. I briefly reviewed different techniques and their limitations in generating explanations from black box models in Section 2.4. For example, similar to the black-box model itself, the explanations could too complex or nonsensical to understand for end-users. In the next section, I'll review considerations to choose the right interpretability technique for the XAI system.

3.4.2 Model Design Guidelines

The interpretable model layer includes a design pole (top) and an evaluation pole (bottom) to improve the interpretability technique during the iterative design steps. We suggest the following design guideline for this layer:

3.4.2.1 Guideline 6: Design Interpretability Technique

Designing interpretable decision-making algorithms starts with the choice of machine learning model. Shallow machine learning models (e.g., linear models and decision trees) have intrinsic interpretability due to low number of variables and model simplicity. For more complex models (e.g., random forest and DNN), ad-hoc explainer technique (see Section 2.4) are needed to generate explanations. However, the choice of machine learning model (i.e., shallow vs. deep) is bounded by model's performance on data domain. Secondly, ad-hoc explainer techniques have certain limitations in their explanation type. The importance of choosing the right combination of model and explainer is in their impact on providing useful (See Guideline 4) and trustworthy explanations for end-users.

Machine learning research has proposed various ad-hoc explainers to generate "Why" explanations (e.g., feature attribution [225, 102]), "How" explanations (e.g., rules list [171, 230]), "What else" explanation (e.g., similar training instance [182, 190]), and "What if" (e.g., sensitivity analysis [107]) explanation types. However, despite substantial research in interpretable machine learning techniques, a core issue in model explanations is the difference between machine learning model's decision-making logic and human sense-making as the receiver [231, 232].

Application in Case Study: In our fake news detection case study, the explainable interface design team had previously discussed candidate explanation choices with the machine learning design team (see Guidelines 2 and 4). Therefore, a final review of model-generated explanations and an assessment of implementation limitations were performed in this step. For example, removing noise-like features from saliency maps, normalizing attributions scores, and resolving contradicting explanations between an ensemble of models were primary implementation bottlenecks that were resolved in this step. Specifically, as a decision point for trade-offs between clarity and faithfulness of explanations, the team decided on using heuristic filters to eliminate features with a very low attribution score for the sake of presentation simplicity.

3.4.3 Interpretable Algorithm Evaluation

Following the review of background in evaluation measures for fidelity of explainer in Section 2.4.2 and truthfulness of explanations in Section 2.4.1, in this section, I provide summarized and ready-to-use computational methods (as opposed to used study based methods) for evaluating interpretability techniques (Tables 3.6).

3.4.3.1 Computational Methods

Computational measures are common in the field of machine learning to evaluate interpretability techniques' correctness and completeness in terms of explaining what the model has learned. Herman [111] notes that reliance on human evaluation of explanations may lead to persuasive explanations rather than transparent systems due to user preference for simplified explanations. Therefore, this provides an argument that explanations' fidelity to the black-box model should be evaluated by computational methods instead of by human-subject studies. Fidelity of an ad-hoc explainer refers to the correctness of the ad-hoc technique in generating the true explanations (e.g., correctness of a saliency map) for model predictions. This leads to a series of computational methods to evaluate correctness of generated explanations, consistency of explanation results, and fidelity of ad-hoc interpretability techniques to the original black-box model [233].

However, in many cases, machine learning researchers often consider model consistency, computational interpretability, and self-interpretation of results as evidence for explanation correctness [175, 234, 235]. For example, Zeiler and Fergus [107] discuss fidelity of the visualization for CNN network by its validity in finding model weaknesses resulted in improved prediction results. In another case, Yosinski et al. [173] created an interactive tool to explore the CNN's activation layers in real-time to provide an intuition about "how the CNN works" to the user. On the other hand, intrinsic interpretable machine learning models (e.g., linear regression and decision trees) are considered as white-box models and do not need additional interpretability techniques.

In some cases, comparing a new explanation technique with existing state-of-the-art expla-

Table 3.6: Evaluation measures and methods used for evaluating fidelity of interpretability techniques and reliability of trained models. This set of evaluation methods is used by machine learning and data experts to eighter evaluate the correctness of interpretability methods or evaluate the training quality trained models beyond standard performance metrics.

Computational Measures	**Evaluation Methods**
Explainer Fidelity	Simulated Experiments ([87, 86])
	Sanity Check ([175, 234, 235, 173, 109, 153]
	Comparative Evaluation ([104, 103])
Model Trustworthiness	Debugging Model and Training ([107])
	Human-Grounded Evaluation ([236, 102, 97, 174])

nation techniques is a way to verify explanation quality [100, 101, 102]. For instance, Ross et al. [103] designed a comprehensive set of empirical evaluations and compared their explanations' consistency, features, and computational cost with the LIME technique [86]. In a comprehensive setup, Samek et al. [104] proposed a framework for evaluating saliency explanations for image data that quantify the importance of pixels with respect to the classifier prediction. They compared three different saliency explanation technique for image data (sensitivity-based [106], deconvolution [107], and layer-wise relevance propagation [108]) and investigated the correlation between saliency map quality and network performance on different image datasets under input perturbation. On the contrary, Kindermans et al. [109] show interpretability techniques have inconsistencies on simple image transformations, hence their saliency maps can be misleading. They define an input invariance property for reliability of explanations from saliency methods. To extend a similar idea, Adebayo et al. [110] propose three tests to measure adequacy of interpretability techniques for tasks that are sensitive to either data or model.

Other evaluation methods include assessing explanation's fidelity in comparison to inherently interpretable models. For example, Ribeiro et al. [86] compared explanations generated by their ad-hoc explainer to explanations from an interpretable model. They created gold-standard explanations directly from the interpretable models (sparse logistic regression and decision trees) and

used these for comparisons in their study. A downside of this approach is that the evaluation is limited to generating a gold standard by an interpretable model. User-simulated evaluation is another method to perform computational evaluations of machine-generated explanations. Ribeiro et al. [86] simulated user trust in explanations and models by defining "untrustworthy" explanations and models. They tested a hypothesis on how real users would prefer more reliable explanations and choose better models. The authors later repeated similar user-simulation evaluations in the *Anchors* explanation approach [87] to report simulated users' precision and coverage in finding the better classifier by only looking at explanations.

A different approach in quantifying explanations quality with human intuition has been taken by Schmidt and Biessmann [97] by defining an explanation quality metric based on user task completion time and agreement of predictions. Another example is the work of Lundberg and Lee [102], who compared the SHAP ad-hoc explainer model with LIME and DeepLIFT [101] based on the assumption that good model explanations should be consistent with the explanations from humans who understand the model. Lertvittayakumjorn and Toni [237] also present three user tasks to evaluate local explanation techniques for text classification through revealing model behavior to human users, justifying the predictions, and helping humans investigate uncertain predictions. A similar idea has been implemented in [236] by feature-wise comparison of a ground-truth and model explanation. They provide a user-annotated benchmark to evaluate machine learning instance explanations. Later, Poerner et al. [84] use this benchmark as human-annotated ground truth in comparison to small-context (word level) and large-context (sentence level) explanation evaluation. Human benchmarks can be valuable when considering human meaningfulness of explanations, though the discussion by Das et al. [174] implies that machine learning models (visual question answering attention models in their case) do not seem to look at the same regions as humans. They introduce a human-attention dataset [238] (collection of mouse-tracking data) and evaluate attention maps generated by state-of-the-art models against human.

Interpretability techniques also enable quantitative measures for evaluating model trustworthiness (e.g., model fairness, reliability, and safety) through its explanations. Trustworthiness of a

model represents a set of domain specific goals such as fairness (by fair feature learning), reliability and safety (by robust feature learning). For example, Zhang et al. [176] present a case of using machine learning explanations to find representation learning flaws caused by potential biases in the training dataset. Their technique mines the relationships between pairs of attributes according to their inference patterns. Further, Kim et al. [225] presented quantitative testing of machine learning models by their explanations. In their concept activation vectors technique, the model can be tested for specific concepts (e.g., image patterns) and a vector score shows if the model is biased toward that concept. They later extended their concept-based global explanation of model representation learning for systematic discovery of concepts that are human-meaningful and important for the model prediction [239]. They use human-subject experiments to evaluate learned concepts. Table 3.6 summarizes a list of evaluation methods to measure fidelity of interpretability technique and model trustworthiness with computational techniques.

3.4.4 Model Evaluation Guidelines

I suggest the following evaluation guideline for interpretable algorithm design step.

3.4.4.1 Guideline 7: Evaluate Model Trustworthiness

Evaluating the interpretable machine learning is the first evaluation step in my framework due to its impact on outer layer evaluation measure. The high significance of this evaluation step stems from the possibility that any unreliability of interpretability at this inner layer will propagate to all other outer layers. Such unintended error propagation may lead to problematic outer-layer design decisions as well as misleading evaluation results. I discuss two main evaluation goals for the innermost layer:

1. Evaluating model trustworthiness.

2. Evaluating ad-hoc explainer fidelity.

The first evaluation goal aims to utilize interpretability techniques as a debugging tool to analyze the model's trustworthiness on learning concepts beyond general performance measures [225].

Examples of model trustworthiness validation include evaluating model reliability in financial risk assessment [240], model fairness in social influencing applications [176], and model safety for its intended functionality [241]. Researchers have also proposed various regularization techniques for enhancing trustworthy feature learning in machine learning models [103, 242]. Next, the second evaluation goal targets fidelity of ad-hoc explainer techniques to the black-box model. Research shows that different ad-hoc interpretability techniques have inconsistencies and can be misleading [110]. Evaluating explanation trustworthiness can verify explainer fidelity in terms of how well it represents the black-box model (see Section 3.4.3.1).

Application in Case Study: In our case study, we paid careful attention to qualitative reviewing of the model explanations after each design iteration. Our initial qualitative review of model explanations led to dataset cleaning through a heuristic search aimed at the removal of mislabeled examples and unrelated news articles. An improvement to model performance was achieved after dataset cleaning. Then, after the first round of human-subject evaluation of the explainable interface (see Guideline 5), the team identified negative effects of keyword explanations with low attention scores from end-users. The team decided on using a lower threshold for visualizing attention maps to reduce clutter and "noisy explanations" for end users. Finally, after one round of XAI outcome evaluation (see Guideline 3), analysis of users' mental models revealed that a dataset imbalance between the "fake news" and "true news" was causing a bias for the model in that the model was usually more confident in predicting fake news over true news.

3.5 Discussion

In my review, I discussed multiple XAI design goals and evaluation measures appropriate for various targeted user types. Table 3.1 presents my categorization of selected existing design and evaluation methods that organizes literature along three perspectives: *design goals*, *evaluation methods*, and the *targeted users* of the XAI system. The categorization revealed the necessity of an interdisciplinary effort for designing and evaluating XAI systems. To address these issues, I

proposed a design and evaluation framework that connects design goals and evaluation methods for end-to-end XAI systems design, as presented through a model (Figure 3.3) and guidelines. In this section, I discuss further considerations for XAI designers to benefit from the body of knowledge of XAI system design and evaluation. The following recommendations support and promote different layers of the proposed evaluation model as well.

3.5.1 Pairing Design Goals with Evaluation Methods

It is essential to use appropriate measures for evaluating the effectiveness of design elements. A common pitfall in choosing evaluation measures in XAI systems is that the same evaluation measure is sometimes used for multiple design goals. A simple solution to address this issue is to distinguish between measurements by using multiple scales to capture different attributes in each evaluation target. For example, the concept of *user trust* consists of multiple constructs [200] that could be measured with separate scales in questionnaires and interviews (see Section 3.2.4.1). User satisfaction measurements could also be designed for various attributes such as understandability of explanations, usefulness of explanations, and sufficiency of details [48] to target specific explanation qualities (see Section 3.3.4.1).

An efficient way to pair design goals with appropriate evaluation measures is to balance different design methods and evaluation types in iterative cycles of design. Managing the trade-offs between qualitative and quantitative methods in the design process can allow designers to take advantage of different approaches, as needed. For example, while focus groups and interviews provide more detailed and in-depth feedback on the users' mental model [132], remote measurements are highly valuable due to the scalability of the collected data even though they provide less detail for drawing conclusions [96]. Thus, one successful approach could be to start with multiple small-scale prototyping and formative studies collecting qualitative measures at the earlier stages of the design (e.g., for XAI system goals layer in the framework) and continue with larger-scale studies and quantitative measures in the later stages (e.g., for interpretable model and interface evaluations in the framework).

3.5.2 Role of User Interactions in XAI

Another important consideration in designing XAI systems is how to leverage user interactions to better support system understandability. The benefits of interactive system design have been previously explored in the topic of interactive machine learning [119, 157] for novice end-users. AI and data experts also benefit from interactive visual tools to improve model and task performance [123]. In this section, I discuss multiple examples of interaction design that support user understanding of the underlying black-box model.

Focusing on interactive design for AI-based systems for AI novices, Amershi et al. [119] reviewed multiple case studies that demonstrate the effectiveness of interactivity with a tight coupling between the algorithm and the user. They emphasize how interactive machine learning processes allow the users to instantly examine the impact of their actions and adapt their next queries to improve outcomes. Such interactions allow users to test various inputs and learn about the model by creating *What-If* explanations [49]. Particularly, user-led cycles of trial and error help novices to understand how the machine learning model works and how to steer the model to improve results. In the context of XAI, Jongejan and Holbrook [186] present a study in which users draw images to see whether an image recognition algorithm can correctly recognize the intended sketch. Their system and study allows for interactive trial-and-error to explore how the algorithm works. In addition, their system provides example-based explanations in cases where the algorithm fails to correctly classify drawings. Another approach is to allow users to control or tune algorithmic parameters to achieve better results. For example, Kocielnik et al. [54] present a study in which users were able to freely control detection sensitivity in an AI assistant. Their results showed a significant effect on user perception of control and acceptance.

Visual analytics tools also support model understanding for expert users through interaction with algorithms. Examples including allowing data scientists and model experts to interactively explore model representations [78], analyze model training processes [74], and detect learning biases [215]. Also, embedded interaction techniques can support the exploration of very large deep learning networks. For instance, Hohman et al. [78] present multiple interactive features to

select and filter of neurons and zoom and pan in feature representations to support AI experts in interpreting and reviewing trained models.

3.5.3 Evaluation Ground Truth

Research on XAI systems study various goals with different measures across multiple domains. The breadth of XAI research makes it challenging to interpret and transfer findings from one task and domain to another. Knowing key factors for interpreting implications of evaluation results is essential to aggregate findings across domains and disciplines. An important factor in understanding XAI evaluation results and comparing results among multiple studies is the choice of ground truth. In the following, I review common choices of ground truth for both human-subject and computational evaluation methods.

Human-subject experiments often take the form of controlled studies to examine the effects of machine learning explanations on a control group in comparison to a baseline group. In these setups, the choice of the baseline could affect results implications and significance. My review of papers in the space of XAI evaluation shows the majority of study designs use a *no explanation* condition as the baseline condition to measure the effectiveness of model explanations in an explanation group. Examples for the baseline include approaches that remove model explanations related components and features form the interface in the baseline condition [54, 53]. In other work, Poursabzi et al. [58] also included a *no AI* baseline to measure participants' performance without the help of model predictions. Another way is to compare the effects of explanation type or complexity between study conditions without the *no explanation* baseline. For instance, Lage et al. [96] present a study to evaluate the effects of explanation complexity on participants' comprehension and performance. They used linear and logistic regression to estimate the effects of explanation complexity on participants' normalized response time, response accuracy, and subjective task difficulty rating.

Though the above mentioned studies are controlled experiments, there may still be unaccounted human behavioral implications due to differences in the complex process of explaining worthy of consideration. Langer et al. [204] present an experiment on "placebic" explanations that shows

people's mindless behavior when facing explanations for actions. In a simple setup, their study showed that when asking a request, inclusion of explanations and justifications increased user's willingness to comply even if the explanations convey no meaningful information. Recently, Eiband et al. [56] proposed using *placebic explanations* instead of a *no explanation* condition as the baseline for XAI human subject studies. Therefore, using non-informative or even randomly generated explanations as the baseline condition could potentially counteract a participant's positive tendency toward explanations and improve study results.

Considering other approaches, a commonly accepted computational technique for quantitatively evaluating instance explanations is to create a ground truth based on the input features that semantically contribute to the target class. For example, image segmentation maps (annotations of objects in images) are used to evaluate model generated saliency maps in weakly supervised object localization tasks [92]. Mohseni et al. [236] proposed a multi-layer *Human-Attention* baseline for feature-level evaluation of machine learning explanations. Their *Human-Attention* baseline provides a human-grounded feature attribution map with a higher level of granularity compared to object segmentation maps. Similarly, feature-level annotations have been used as the explanation ground truth in the text classification domain [243]. Other less accurate means of feature attribution like bounding box in images datasets have been used for quantitative evaluation of saliency maps. For instance, Du et al. [93] evaluated saliency maps generated from a CNN model by calculating pixel-wise IOU (intersection over union) of model-explanation bounding boxes and ground truth bounding boxes.

3.5.4 System Evaluation Over Time

An important aspect in evaluating complex AI and XAI systems is to take the user learning into account. Learnability is even more critical when measuring mental models and user trust in the system. A user learns and gets more familiar with the system over time with continued interaction with the system. This brings the importance of repeated temporal data capture (in contrast to static measurements) for XAI evaluations. Holliday et al. [143] present an example of multiple trust assessments during the user study. They measured user trust at regular intervals

during the study to capture changes in user trust as the user interacts more with the system. Their results indicates an XAI system outperformed a non-XAI counterpart in maintaining user trust over time. Time-based measurements, also referred to as *dynamic measurements*, allows designers to monitor explanation usability and effectiveness in various contexts and situations [244, 245]. For instance, Zhang et al. [246] explore the effect of intelligent system explanations in user trust calibration. In their experiments, they observe significant effect on calibration of trust when model prediction confidence score was shown to participants. In another example, a study by Nourani et al.[247] controlled whether users' early experiences with an explainable activity recognition system had better or worse model outputs, and the first impressions significantly affected both task performance and user confidence in understanding how the system works. In a study with a news review task, Mohseni et al. [248] identified different user profiles for changes in trust over time (trust dynamics) while working with the assistance of an explainable fake news detector. Their analysis of results revealed a significant effect of machine learning explanations on user trust dynamics.

Long-term evaluation of XAI systems can also allow designers to estimate valuable user experience factors such as over-trust and under-trust on the system. User-perceived system accuracy [151] and transparency [224] are examples of long-term measures for explanation usability that depend on building user trust in the system's interpretability. As more information is provided by explanations over time, reasoning and mental strategies may change as users create new hypotheses about system functionality. Therefore, it is essential to also consider users' mental models and trust in extended studies to evaluate all aspects of the XAI system.

Another use case of long-term measurements is to evaluate the effects of intelligent system's non-uniform behaviors in real-world scenarios. This means, although in a controlled study setup, a balanced set of input examples will present the system to the user, in real-world scenarios, users may face alterations in system performance in long-term interaction with the system. Long-term measurements will identify user's unjust trust in the system due to a limited or biased set of interactions with the system. For example, in the context of autonomous vehicles, Kraus et al. [249]

presented a model of trust calibration and presented studies on trust dynamics in the early phases of user interaction with the system. Their results indicate the effects of error-free automation in steady increase of user trust as well as the effects of user a priori information in eliminating the decrease of trust in case of system malfunction.

3.5.5 Generalization and Extension of the Framework

Our framework is extendable and compatible with existing AI-infused interface design and interaction design guidelines. For example, Amershi et al. [157] propose 18 design guidelines for human-AI interaction design. Their guidelines are based on a review of a large number of AI-related design recommendation sources. They systematically validated guidelines through multiple rounds of evaluations with 49 design practitioners in 20 AI-infused products. Their design guidelines provide further details within the user interface design layer of our framework (Section 3.3.3) to guide the development of appropriate user interactions with model output and interactions. In other work, Dudley and Kristensson [156] present a review and characterization of user interface design principles for interactive machine learning systems. They propose a structural breakdown of interactive machine learning systems and present six principles to support system design. This work also benefits our framework by contributing practices of interactive machine learning design to the XAI system goals layer (Section 3.2) and the user interface design layer (Section 3.3.3) From the standpoint of evaluation methods, Mueller and Klein [250] discuss how common usability tests cannot address intelligent tools where software replicates human intelligence. They suggest new solutions are needed to allow the users to experience an AI-based tool's strengths and weaknesses. Likewise, our nested framework highlights the potential for error propagation from the inner layers (e.g., interpretable algorithms design) to the outer layers (e.g., system outcomes) in the XAI system evaluation pole. The iterative back-and-forth between layers in the nested model encourages expert review of system outcomes, user-centered evaluation of the explainable interface, and computational evaluation of machine learning algorithms.

3.5.6 Overlap Among Design Goals

In our categorization of XAI systems, we chose two main dimensions to organize XAI systems by their *Design Goals* and *Evaluation Measures* in Section 3.1.2. The XAI design goals (G1–G8) were based on the goals extracted from the surveyed papers, and since the XAI design goals are primarily derived from their targeted user groups, we note that overlaps among goals do exist across disciplines. For instance, there is overlap of the goals of *G1: Algorithmic Transparency* for novice users in HCI research, *G5: Model Visualization* for data experts in visual analytics, and *G7: Interpretability Techniques* for AI experts in machine learning research. While overlapping, these similar goals are studied with different objectives across the three research disciplines leading to diverse sets of design requirements and implementation paths. For example, designing XAI systems for AI novices requires processes and steps to build human-centered explainable interfaces to communicate model explanations to the end-users, whereas designing new interpretability techniques for AI experts has a different set of computational requirements. Another example of overlap in XAI goals is between the goal for *G6: Model Visualization and Inspection* for data experts and *G8: Model Debugging* for AI experts, in which different sets of tools and requirements are used to address different research objectives.

To address the overlap between XAI goal among research disciplines, we used the XAI *User Groups* as an auxiliary dimension to organize XAI goals in this cross-disciplinary topic (Section 3.2.1) and emphasize the diversity of diverse research objectives. The three user groups were chosen to organize research objectives and efforts into HCI (for AI novices), visual analytics (for data experts), and machine learning (for AI experts) research fields. Additionally, as described in the framework, the three user groups prioritize design objectives in the design process for the XAI system rather than absolute separation of design goals. For example, the objectives and priorities in XAI system design for algorithmic bias mitigation for domain experts in a law firm are certainly different from those of model training and tuning tools for AI experts. However, by following the multidisciplinary design framework, a design team can translate XAI system goals into design objectives for explainable interface and machine learning techniques to improve the design process

in different layers. Therefore, in the above example, the design team can focus on diverse interface design and interpretability technique objectives to achieve the primary XAI goal of bias mitigation for the domain experts. Note that the specifics of any particular system will determine the priorities of different objectives.

3.5.7 Limitations of the Framework

Our framework provides a basis for XAI system design in interdisciplinary teamwork and we have described our case study example to validate and improve the framework. The presented case study serves as a practical example of using our framework in a multidisciplinary collaborative XAI design and development effort. Our use case is the result of a year-long (and ongoing) research done by a team of eight university researchers with diverse backgrounds. The lessons learned and pitfalls in our end-to-end implementation case study are incorporated in the presented design guidelines. However, no framework is perfect or entirely comprehensive. We acknowledge that the validity and usefulness of a framework are to be proven in practice with further case studies. In our future work, we plan to run multiple validation case studies to examine practicality and usefulness of this framework.

Moreover, this framework has a common limitation of many multidisciplinary design frameworks of being light on specific details at each step. Rather than contributing detailed guidelines for each framework layer, the framework is intended to pave the path for efficient collaboration among and within different teams, which is essential for XAI system design given the inherently interdisciplinary nature of the area. The diversity of design goals and evaluation methods at each layer can help maintain the balance of attention from the design team to different aspects of XAI system. This higher level of freedom allows for extendability with other design guidelines (see the discussion in Section 3.5.5) to integrate with more tailored approaches for specific domains.

4. CASE STUDY AND EXAMPLES

4.1 Introduction

I present a case study and two XAI system design examples in this section to demonstrate benefits of the proposed XAI framework and present how to use it. The first case study demonstrates the *Generative Function* of the framework and provides a step-by-step review of an example XAI design process. The case study is a one-year long project with a multidisciplinary team of researchers working on a XAI system for fake news detection for non-expert (not AI experts or news analysts) daily newsreaders. The presentation of my case study reviews detailed design goals at each step and evaluation methods for system components. In the end, I present comprehensive results and analysis for the assessment of XAI system outcomes.

The following two XAI system design examples are analysis of two existing XAI system designs for interpretable video analysis [1] and model interpretability analysis [2] tasks. The goal for these two design examples is to present the XAI framework's *Descriptive Function* to describe (for communication purpose) and analyze (to assess design alternative) in existing XAI systems from different domains and applications. My review would focus on both design and evaluation steps and emphasize on the process and pitfalls in the two examples in comparison to the guidelines from my framework.

4.2 Case Study: Fake News Detection

4.2.1 Introduction

Intelligent algorithms are used in a variety of online applications, from product recommendation and targeted advertisement to loan and insurance rate prediction. However, as AI-based decision-making is directly affecting people's lives, the accountability and fairness of advanced AI algorithms are under question [162]. In recent years, the need for algorithmic transparency is gaining more attention to enable accountable AI-based decision-making systems, and XAI techniques have been introduced to annex transparency into black-box machine-learning algorithms. Inter-

pretability can help users to build a mental model of how algorithms work and build appropriate trust in intelligent systems Rader et al. [33].

In the social media domain, news feed and search algorithms function similar to decision-making algorithms, as users are exposed to algorithmically selected content. Blindly trusting algorithmically-curated news could potentially lead to unintentional large-scale propagation of false and fabricated information with users being exposed to malicious content and its re-sharing through social media. Human review of news and data mining techniques for fake-news detection and debunking are commonly being practiced as primary approaches in reducing fake news in social media. However, reviewing the life cycle of news in social media reveals opportunities to combat the propagation of fake news within news-feed platforms [194]. For example, AI-based news review assistant tools can be embedded in news feed platforms and have the potential to benefit users by providing direct suggestions related to news credibility rather than automatic organizational news debunking.

In a case study, I demonstrates the *Generative Function* of the framework that provides step-by-step guidelines for design and evaluate a XAI system. A team of researchers with machine learning, data visualization, and HCI backgrounds, design an explainable fake news detection algorithm to study the effects of algorithmic transparency for news review applications and social media. In a close collaboration, I investigate whether the interpretability of the fake news detector algorithm could enhance users overall experience and result in increased credibility of user-shared news. I also aim to examine whether model explanations can help users to avoid overtrusting the fake news detector when explanations are nonsensical to users. I formulate our research goals into the following questions for XAI system outcome:

- RQ1: Do AI and XAI assistants help end-users share more credible news?
- RQ2: How do AI explanations affect users' mental models of intelligent assistants?
- RQ3: How do AI explanations affect end-user trust and reliance in intelligent assistants?

The following sections, I review design steps for a news reviewing and sharing interface with a

built-in interpretable fake news detector for end-users and run a series of evaluation experiments. With this system, I conducted a series of crowdsourced experiments to evaluate potential benefits and limitations of machine learning explanations through our intelligent assistant. The study results indicate the complexity of the fake news detection problem and the limitations of current model interpretability techniques for this task. Though the addition of explanations to our system did not improve user task performance, I observed that explanations helped participants' to build appropriate mental models of the intelligent assistants in different conditions and adjust their trust accordingly for the model logic.

4.2.2 Background

Machine learning algorithms are heavily used in online platforms and social media to analyze user data for improving user experience and increasing corporate profit. However, the lack of transparency can raise data privacy and model trustworthiness concerns in critical domains, and hence potentially decreases user trust and confidence in the long run [3]. In this regard, researchers study the communication of algorithmic processes in various domains such as online advertising [251], social media feeds [26], and personalized news search engines [252]. In this section, I briefly review machine learning and human-computer interaction papers related to the explainable news feed and fake news detection systems.

4.2.2.1 Fake News in Social Media

In this section, I briefly review various techniques to combat fake news in social media from the perspective of *News Life Cycle in Social Media*, shown Figure 4.1. Specifically, I want to emphasize on a research gap in studying the social impact of news feed algorithms on the spread of fake content and crediting unreliable sources. Different surveys provide comprehensive reviews of fake news problem characterization [253] and data mining methods [254, 255, 256] for fake news detection. However, these works mainly focus on machine learning techniques in feature learning and news classification, and therefore lack to address the importance of user involved news distribution stage in the news life cycle.

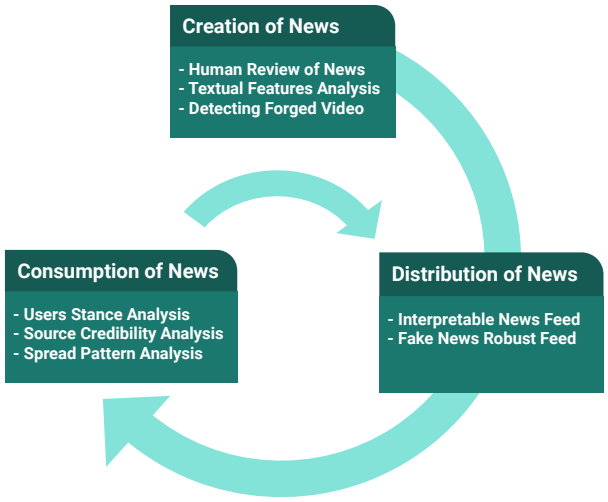


Figure 4.1: A summary of misinformation detection methods at different stages of the news life in social media. While natural language processing and social media data mining methods are popular in fake news at creation and consumption stages, there is a limited amount of research on fake news robust news feed algorithms to mitigate the propagation of fake content. The inner arrow shows how users' social data is used to curate personalized news feed.

The first stage is to detect fake content at the news creation step, which traditionally was done with human review through the expert review or crowdsourcing techniques at the early stages. Experts review the truthfulness of the news by evidence and determine whether claims are accurate or false (partially or entirely). Fact-checking is a knowledge-based approach usually done by fact-checking organizations (e.g., Politifact and Snopes) to judge the veracity of news pieces with external references. However, human review fact-checking methods are time-consuming, expensive, and not scalable for stopping the spread of fake content in social media. Machine learning solutions to detect false information in the news and social media use various knowledge-based methods, natural language processing, and social media data mining techniques. One approach, for instance, is to use linguistic features to analyze writing styles to detect possible false content [257]. Other style-based approaches, such as recognizing deception-oriented [258] and hyper-partisan content [259] can be used as a basis for detecting intentionally falsified information. In the case of clickbait detection algorithms, the inconsistency between headlines and content of the news has been used for possible fake news detection [260]. Additionally, fact-checking is not limited

to the correctness of textual content; forged images and videos researchers use computer vision methods (e.g., [261]) to detect falsified contents. New methods like provenance analysis have also been utilized for content validation via generating provenance graph of images as the same content is shared and modified over time [262].

The next stage of news life in social media is the distribution of content via news feed algorithms and search engines. Although the distribution of news can be another practical stage to combat fake news distribution, due to the interdisciplinary nature of social media, the current state of machine learning research is short on studying the social impact of news feed algorithms in the distribution of false content and their vulnerability for being misused. Others also examined the use of new designs news feed in virtual reality environment as opposed to linear feeds to reduce the creation of filter bubbles [263]. For example, echo chambers are from social media vulnerability points that create and propagate false information. Multiple sources of evidence show personalized algorithms (e.g., news feed, search engines, and personalized advertisement) can have drastic effects on information diversity and cause the creation of echo chambers in social media, filter bubbles in search engines, and discrimination in information access. For instance, in a recent study Geschke et al. [264] presented an agent-based simulation of different information filtering scenarios can boost social polarization and lessen the interconnections of social media echo chambers. Researchers also propose using diversity metrics and bias quantification methods [265, 266] as other ways to study bias and discrimination in recommendation algorithms. For example, Kulshrestha et al. [267] proposed a framework to quantify bias in ranked search results in political-related queries on Twitter. Their framework can distinguish bias from news content and ranking algorithm, and they found evidence of significant effects of both input content and search algorithms in producing bias.

The final stage of analyzing fake news is to use social media data including user stance [268], news propagation patterns [269], and news source credibility estimation [270] to detection fake news. Research on social media data mining show dominant results in detecting fake content and malicious account, however, utilizing social media data implies standing by until the fake news is

already exposed to the users. This shows a trade-off between leveraging rich social data for fake news detection and waiting until a group of users is exposed to fake content. Such social features can be applied to user groups to evaluate the credibility of specific news pieces by considering the stance of a group of users for the news topics [271]. Similarly, rumor detection methods aim to detect a track of posts discussing a specific topic [272]. To increase fake news detection accuracy and model generalizability, training on multi-source and multi-modal datasets are also studied. For example, Shu et al. [273] explored the correlation between news publisher bias, user stance, and user engagement together in their Tri-Relationship fake news detection framework. In their following work, Shu et al. [274] proposed a training dataset to include news content and social context along with dynamic information of news. Although most aforementioned data mining methods do not perform direct fake news detection, these methods can leverage both social and textual feature to identify suspicious news pieces for human review.

4.2.2.2 Interpretable Fake News Detection

Machine learning solutions to detect false information in the news and social media take diverse directions like knowledge-based methods, Natural Language Processing (NLP), and social media data mining techniques. Research shows that NLP methods can learn various features related to misinformation including linguistic features to analyze writing styles to detect possible misinformation content [257]. Style-based approaches are also used to recognize deception-oriented [258] and hyper-partisan content [259] as a basis for detecting intentionally falsified information. Other techniques, such as training on multi-source, multi-modal, and noisy-labeled datasets are also studied to increase fake news detection accuracy and model generalizability. For instance, Shu et al. [273] incorporated multi-source data from news publisher bias, user stance, and user engagement in their fake news detection framework. In another work, Popat et al. [275] used the Google search engine to directly collect similar instances from the web to leverages external news articles as a training source.

Interpretation methods to explain predictions of natural language processing models could generally be grouped into four categories [82]. The first category is the back-propagation based meth-

ods, which calculate the gradient or variants of gradients of a model prediction in terms of the model input [83]. Those words in the input with large gradient values would have more significant contribution to the model prediction. The second category is perturbation based methods in which the key idea is to perturb the input text and those words having more contributions once perturbed would cause more dramatic changes in model prediction [84]. Thirdly, local approximation based methods could be employed to explain model predictions. Although the whole model behavior is highly intricate, the local behavior around an input instance could be approximated and well explained. Local model behavior for an input instance either could be approximated using a linear model (such as sparse linear model [86]), or an interpretable non-linear model (such as if-then rules [87]), depending on the property and complexity of the complex NLP model at hand. The last category is decomposition-based methods [88]. For instance, Du et al. [89] present a technique for recurrent neural networks to decompose predictions into the additive contribution of each input word by modeling the information flow process from the input text to the model output. Note that the former three categories are mainly based on heuristics or approximations and thus result in explanations that might not be faithful to the original model. In contrast, this kind of decomposition could more faithfully reflect the decision-making process of the original DNN model.

Enabling interpretability in fake news detection algorithms could enhance users' ability to find model weaknesses resulting in the appropriate user trust level in AI predictions. For example, Shu et al. [276] present dEFEND framework to discover news sentences and user comments that can explain model prediction. For instance, XFake detector in [277] uses various NLP news attribution explanations and a tree-based visualization of their ensemble model to explain the decision paths for each input news sample. However, since these models only achieve moderate detect performance (i.e., in the range of 80% accuracy) in the binary fake news detection task, it remains uncertain that how would model explanations effect on end-users trust in these models. To gain insight into whether news recommendation algorithms should be transparent about their decisions, Hoeve et al. [252] run a survey and learn that a vast majority of respondents prefer explanations. However, in a follow up A/B testing, they find participants are not opening (via click count) model

explanations. This could be due to the low urgency of explanations in news recommendation and/or their study news test set. In human studies for AI-based news fact-checking, Horne et al. [51] run an experimental human subject study and find that feature-based explanations in AI assistant significantly improve users perception of news bias. However, their measured effect size was much larger for participants who were frequent newsreaders and those familiar with politics. In another paper, Nguyen et al. [278] present design and evaluation of a mixed-initiative fact-checking system to blend human knowledge with machine learning algorithms. They also conclude that transparency and interactivity significantly affect users' ability to predict the veracity of given claims. To continue this line of research, I investigate how different types of model explanations affect the credibility of news shared by users in social media like scenario. I also measure a wider range of explicit and implicit user feedback to study interactions among key XAI design goals in the explaining process.

4.2.3 XAI System Goals

As the first step in the XAI framework, the team followed guidelines from the framework to (1) decide on the main system goals and (2) identify impactful explanation types, and (3) decide on appropriate measures for evaluation of system outcomes. We started with identifying candidates for useful and impactful explanations for fake news detection such as keyword attention, supporting evidence, and source credibility based on machine learning research on misinformation detection and human-computer interaction research on news feed systems [194]. Also, the design process in this step involved reviewing algorithmic implementation constraints such as "what can be explained" to the user.

4.2.3.1 Guideline 1: XAI System Goals and Users

As the first step in design process of the XAI system, the team started with identifying the main goals and expectations for the XAI news assistant. Our system's targeted users are the general public who read daily news and are not AI experts nor news analysts. The XAI design goal was to improve user reliance and mental model of news predictions through explainable design. The team

hypothesized that end-users would trust and rely on the fake news detection assistant, given that the new XAI is capable of providing explanations for each news story. Also, the team hoped that users would be able to use the explanations to learn model weaknesses and strengths to provide feedback to the developer team. The system allows us to study the role of interpretable models in fake news detection as the main project research goal.

4.2.3.2 Guideline 2: What to Explain

In the second step of our XAI system design, the team identified "what to explain" to the user in order to achieve the initial XAI goals (see Guideline 1) of the system. In our case study, efficient news curation required fake news detection with the help of our XAI assistant. In the analysis of what the system should explain, the design team decided to identify candidate useful and impactful explanation options. I started with reviewing machine learning research on false information (e.g., rumor, hoax, fake news, clickbait) detection as well as HCI research on news feeds and news search systems to identify key attributes for news veracity checking [194]. Given the non-expert target end-users, explanatory information needed to limit technical details. Next, the user interface designers and machine learning designers in the team discussed candidate explanation choices and algorithmic constraints in interpretability techniques. That is, some options for what to explain may not be entirely possible given the interpretability of existing models, and the team needed to consider whether alternative learning techniques could provide better explanations or if the design team would need to figure out meaningful ways to explain the information that was available from the model.

Give the training set and types of models the team was planning to use in the ensemble approach, the system was expected to provide *why-type* explanations for each news veracity prediction. Specifically, the explanations could describe the attribution of different news features for each news veracity prediction. Therefore, attribution scores for the news headline, the article text, and article sources is used to explain why the model is arrived to its prediction. More details about these explanations are presented in the interface and model design sections.

4.2.3.3 Guideline 3: System Evaluation

Although evaluation of system outcomes is the last step in the XAI design and evaluation cycle, identifying main evaluation measures early on helps to clarify the evaluation path. I formulate system evaluation goals into the following questions for XAI system outcome:

- RQ1: Do AI and XAI assistants help end-users share more credible news?
- RQ2: How do AI explanations affect users' mental models of intelligent assistants?
- RQ3: How do AI explanations affect end-user trust and reliance in intelligent assistants?

Note that both the explainable interface (Section) and interpretable algorithm (Section) passed multiple design and testing iterations before the system outcome evaluation step. Our system evaluation step consist of a human-subject study with non-expert participants our fake news detection system. Major decisions for this evaluation was how to structure the duration and complexity of the user task while appropriately testing the system's full range of functionality. Multiple evaluation measures are chosen for system outcomes, including: (1) subjective user trust in the news assistant, (2) user agreement rate with the news assistant, (3) veracity of user-shared news stories, and (4) user accuracy in guessing the news assistant output. Both qualitative and quantitative analysis of user feedback and interaction data were valuable to the evaluation of system outcomes. The results and analysis from these evaluations helped the team to understand the effectiveness of the XAI elements (in both the algorithm and the interface) for the initial system goals, see Section 4.2.6

4.2.4 Explainable Interface Design

The explainable interface design step starts with an interactive news reading interface and continues with model explanations components (Guideline 4). We also performed preliminary rounds of user testing for interface complexity and explanations' understanding (Guideline 5).

4.2.4.1 News Review Interface

We designed an interface for users to review a queue of news stories, share true news for other users, and report fake news stories. The interface design process started with multiple interface

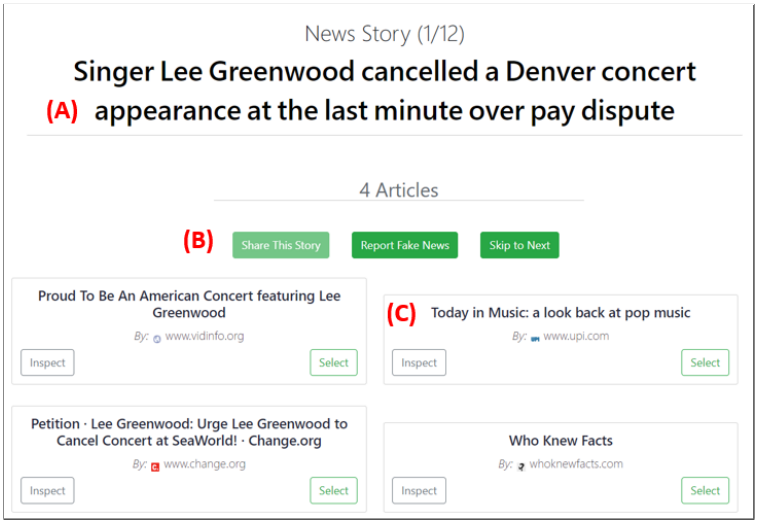
sketches that suit the news reviewing task. We aimed to design a simple interface with useful explanations for fake news detection. The team tested mock-up implementations from the top design choices with a small number of participants. After reviewing feedback from user observations and interviews, we selected the most comprehensible and conclusive design for the human-subject experiments.

Figure 4.2-Top shows the baseline interface that enables the participants' news review task. The interface shows a news headline for a news story on the top (Figure 4.2-A) followed by a list of related articles below (Figure 4.2-C). The related articles provide context and article sources for the news headline, and they can help the user to understand contributing information and factors for model prediction. The system allows users to open and read the related articles, but for the sake of user study, it was not required for sharing the news headline. The system was designed to allow users to review news stories one-by-one and decide if 1) the story is true to be shared with other users, or 2) it is fake news to be reported, or 3) they want to skip to the next story due to their unfamiliarity with the topic or lack of confidence (see Figure 4.2-C).

4.2.4.2 Guideline 4: How to Explain

The fake news detection assistant is embedded in the interface which provides the model prediction (with or without explanation) about the news stories' credibility. Our core design rationale for the four explanation types was to embed model attribution explanations using visual elements for each news feature. Figure 4.2-Bottom shows the interface with the XAI assistant. Both the AI assistant prediction and its explanations are in the form of on-demand recommendations for the user, which are collapsible on user click. Each visual explanation element was tested during pilot studies and refined through iterative design.

These explanations describe the attribution of different news features (i.e., news headline, article text, and article source) for each news veracity prediction. These different explanations are presented to the user with the following visual elements: (1) A heatmap of keyword attribution score that explains how the XAI assistant learned word-level features in the news headline (Figure 4.2-D) and its related articles (in the news article page). (2) A single bar chart for each related



Interface in Baseline condition.

Interface in XAI-all condition.

Figure 4.2: Our news review interface with AI and XAI assistants. **Top:** Baseline interface without AI assistant. (A) news headline. (B) user selecting to share, report, or skip the news story. (C) a list of related news articles for the headline. **Bottom:** Interface with clickable model prediction and different feature attribution explanations. (D) a heatmap of word level feature attribution explanation for news headline. User can see the attribution score values in tooltips when hovering mouse over the keywords. (E) fake news prediction and confidence. (F) confidence for the headline and article separately. (G) a bar chart for each article attribution score in comparison to other related articles. Bar charts show lower values when the articles are less related to the headline or less significant for model prediction. (H) A donut chart for each news article for source attribution scores compared to headline (Claim) and article (Text) content.

article (Figure 4.2-G) explaining each related article's attribution score in comparison to other articles for the model prediction. (3) A pie chart to present attribution score for the articles' source in comparison to the articles' content attribution and news headline attribution. (4) A list of top-3 important sentences for the article is shown when reviewing news articles to explain sentence-level feature learning of the models.

4.2.4.3 Guideline 5: Interface Evaluation

We used preliminary user testing to evaluate interface components and presentation of explanations. We tried to keep a balance between interface complexity and explanation usefulness by choosing among available explanation types from our interpretable machine learning algorithms. Next, mock-ups from the top designs were implemented for testing with a small number of participants. Each mock-up had a different arrangement of data, user task flow, and explanation format for the news assistant interface. Our human-subject experiments in this stage were based on user observations and post-usage interviews to collect qualitative feedback regarding participant understanding and subjective satisfaction of explanation components and interface arrangements. Interviews resulted in the selection of the most comprehensible and conclusive design among the available options to continue with.

4.2.5 Interpretable Algorithm Design

In this section, I briefly review the training data, fake news detection models, and interpretability techniques used in the XAI system as a part of Guideline 6 and to provide context about the underlying algorithms.

- *Fake News Data:* Training data for our models come from two sources: a) news story headlines and labels from Snopes (www.snopes.com) and b) related articles crawled from Google search results (top 16). The related articles were collected for each Snope news headline separately and labeled the same as their respective Snopes news story statements with noisy label assumption for the purpose of model training. The training data includes 4638 news story headlines with an average length of 15 words and 30599 related articles

with an average length of 1012 words. We used 80% of data for model training, 10% data for validation, and 10% data for testing. We took news samples and model predictions from our test set to feed the interface for human-subject studies. Our dataset consists of news, rumors, and hoax which covers a range of different topics, including politics (725 stories), business (224 stories), health (192 stories), and crime (141 stories).

4.2.5.1 Guideline 6: Interpretable Models

Following the previous NLP algorithms for fake news detection, the team planned to implement an ensemble of four classifiers for fake news detection to generate different types of explanations. Our purpose in choosing the ensemble model approach was to study the effects of different explanation types later in the evaluation experiments. The final prediction score is obtained through averaged ensemble results with 73.65% detection accuracy.

Our first model is a Bi-LSTM network [279] with an additional self-attention layer to extract attention scores for instance explanations. This model is trained on news headlines only and generates attention explanations for its predictions. In our empirical tests, Bi-LSTM network outperformed similar networks (e.g., RNN, LSTM. RCNN) for our dataset by capturing both forward and backward states. We also use Word2Vec [280] to embed each word into an embedding vector before feeding into the network. This trained model achieves 72.00% fake news detection accuracy on our test set.

The second model performs fake news detection based on both the news story headlines and the set of related articles for each. The article set representation is constructed using hierarchical attention at sentence level and article level. We use the hierarchical attention network (HAN) [281] to help our model focus on the salient sentences and articles at two levels. HAN scores each article and selects the most important sentences in each article. Each sentence representation of an input article is generated by taking an average of the word embedding of all the words therein. Our design allows us to get the attribution score for each article and select the three most important sentences in each article using attention weighs. For the news story representation, similar to our first model, we used a Bi-LSTM network. Finally, a weighted sum is performed over all articles to

build the article representation, which is combined with the news story representation to form the final vector representation for news story classification. This model achieved 76.04% classification accuracy on the test set.

For the third model, we use a knowledge distillation approach [282] to approximate a deep architecture (teacher) with a random forest (student) model. This model takes news stories, related articles, and article source as the input, and with the mimic learning framework, we can leverage the performance of a deep model and analyze the attribute importance of news stories, articles, and their sources for each prediction. We first train a Bi-LSTM teacher model using Glove word embedding [283] and then train a 60 trees XGBoost [284] student model. The XGBoost student model provides attribute importance (news story headline, article content, and article source) as for instance explanations. Our third model achieved *72.08%* prediction accuracy in the test set.

For the last model, we use both news headlines and related articles to train a BiLSTM network with Word2Vec word embedding. We use an attention mechanism to focus on parts of the articles that are more relevant to the news story. In order to do so, we calculate a weighted average of the hidden state representation based on the attention score corresponding to all the article tokens [285]. Our method then aggregates all the information about the news story, article context, and attention weights to predict the story's credibility. Finally, to generate an overall credibility label for the classification task, the final representation is processed using the final fully connected layer. The attention mechanism also generates keyword attribution explanations for each article.

4.2.6 System Outcome Evaluation

I designed a controlled human subject studies in order to test the hypothesis regarding the effectiveness of AI assistance and its explanation in news review task. The following presents my study design details in terms of study conditions, evaluation measures, and participants' task.

4.2.6.1 Study Design

I conducted human-subjects studies for controlled comparison of elements of the AI assistant and its explanations. The study followed a between-subjects design with five different conditions,

Table 4.1: Study conditions and intelligent assistant components to detect fake news and explain its prediction.

Study Condition	Model Output	Model Explanations
Baseline	–	–
AI Assistant	Prediction and Confidence	–
XAI Assistants (3 conditions)	Prediction and Confidence	XAI-attention: Keywork importance heatmap for news headline and articles.
		XAI-attribution: News attribute and article importance for related articles.
		XAI-all: Explanations from both XAI-attribute and XAI-attention conditions.

where each participant used one variation of the news reviewing system as described in the following and summarized in Table 4.1.

- **Baseline Condition:** For the *Baseline* condition, I remove AI prediction and its explanations in the interface. The baseline interface (Figure 4.2, top) allows the user to review and share news headlines without any machine learning support. This condition serves as the baseline for human-alone performance in comparison to human-AI collaboration. Also, since the *Baseline* condition did not include AI or XAI elements, the condition did not measure user mental model and trust in AI or XAI.

- **AI Assistant:** My interface in *AI Assistant* condition includes AI prediction and confidence for news headline credibility. The prediction and confidence from the ensemble model (without explanations) are used in this condition. The AI predictions are in form of on-demand using a collapsible menu on user click. This condition serves as the baseline for user mental model and trust measurements in the AI without explanation. Figure 4.2 shows model prediction and confidence at (E) and models confidence for the headline and articles separately at (F).

- **XAI Assistants:** The user interface in *XAI assistant* conditions provides instance explanations in addition to news credibility prediction. I design three *XAI Assistant* conditions to study how different types of explanations affect Human-AI collaboration. I use two interpretable models in each *XAI Assistant* condition. The *XAI-attention* condition presents a

heatmap of keywords using attention weights for news headline (Figure 4.2-D) and each related news articles. The *XAI-attribution* condition shows news attribution explanations for related articles and news sources. The hierarchical attention network generates articles importance score (Figure 4.2-G) and top-3 important sentences from each article. The mimic model generates source, article, and news story attribution score (Figure 4.2-H) to present instance explanations. The *XAI-all* condition is the combination of explanations in the *XAI-attribution* and *XAI-attention* conditions. The purpose for designing *XAI-all* condition was to study the effect of variety of explanation types on users.

4.2.6.2 Study Procedure

Figure 4.3 presents the overall study procedure. Participants started the task by accepting the information sheet including the approved IRB number and study contact points information. Next, participants saw step-by-step task instructions with visual guides for all interface components. Visual instructions include descriptions for the headline and article attribution explanations from XAI assistant. Next, participants answered the pre-study questionnaire including text entry and multiple-choice questions. Participants then started the main task by reviewing news stories.

Participants were prompted to review a queue of news stories and share 12 true news for social media users. To engage participants to review news articles and their explanations, users had to select at least one article that represents the news headline for each news story they chose to share. They could always skip to the next news story (as many times as needed) if they were not familiar with the topic. The choice of the sharing task and ability to skip unfamiliar topics (unlike work that assumes participants are familiar with a short curated list of news stories e.g., [51, 278]) improves the fake news detection task by allowing participants to interact and examine the AI/XAI assistant rather than focusing on news analysis. Participants also had the chance to flag news stories as fake if they found headlines to be fake; however, these were not counted toward the required number of shared stories needed for task completion. Also, in contrast with previous work, my interface gives a list of related news articles to provide the context of news stories for users. Further, unlike [278], participants did not receive feedback of the ground truth after each instance (i.e., whether the model

made a correct or wrong prediction) to simulate a real-world scenario in which users do not have immediate access to the credibility of their daily news. During the last four news stories (the last third of the study), participants were asked pop-up questions about the AI assistant's prediction before revealing the model prediction; This was done to collect data to estimate user ability to predict the AI's output.

In the end, participants answered a final questionnaire of Likert-scale and slider questions about the AI assistant followed by four open-ended response forms.

4.2.6.3 Participant Pool

The XAI system and user task were designed for non-expert end users with little knowledge of AI. I recruited remote participants from Amazon Mechanical Turk "Master" users with above 90% acceptance rate. To encourage participants to spend enough time on the task, I measured task duration and paid flexible time-based compensations. The payment was set to $10 per hour and each participant could only participate once in the HIT. To further ensure data quality for analysis, I filtered data samples based on collected user engagement measures including task duration, number of clicks, and character counts in the final questionnaire form.

4.2.6.4 Study Measures

I take users' mental model, human-AI performance, and trust as the primary measures in the studies. I mainly use quantitative methods for the measurements to aim for investigating the initial research questions (RQ1 – RQ3).

- **Task Performance:** I calculate the veracity of participants' final shared and reported news as the main performance metric. I take the credibility score of user shared news as the number of shared true news divided by total shared news (equal to 12 in all experiments). I also review and analyze results for the incredibility score (calculated as 1.0 – *credibility score*) of all reported fake news as the secondary performance measures.

- **Mental Model:** I take participants' accuracy in guessing model output (similar to Poursabzi et al. [58]) as representative for model predictability and user mental model. For the mea-

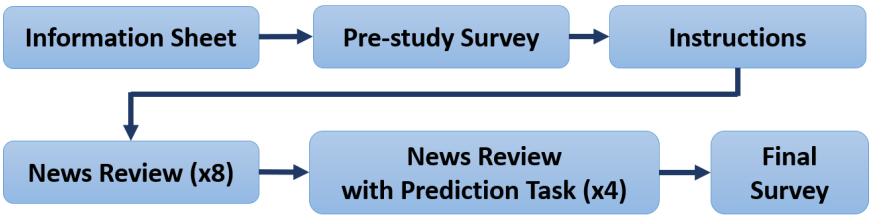


Figure 4.3: Overview of study Procedure. The core user tasks involve the main news review task (8 samples) to allow users to build a mental model of the AI assistant, and additional prediction task (4 samples) where the user guesses the model output for new instances.

surement of this prediction task, I use short pop-up questions during the study to ask "what would the AI fake news detector predict for this news story?" from participants. Participants could response with short "True" or "Fake" answers. Since I expect participants to interact and understand the intelligent assistant during the early stages of the study, the pop-up questions for mental model measurements were limited to the final third of the study (i.e., the last four news review instances). I also calculate response time for each participant as the average time (in minutes) to review news stories for sharing or reporting. Somewhat similar to Lage et al. [96], I aimed to see if explanations might cause longer response time or even information overload for users.

- **User Trust and Reliance:** I measure user trust using a subjective rating of participants' perceived accuracy of AI assistant. Specifically, participants answer "What was the accuracy of the AI fake news detection?" using a continuous slider bar (between 0–100%) to indicate their perception of AI or XAI assistant's accuracy in the post-study survey.

  I also measure user reliance using participants' agreement rate with AI assistant predictions. To quantitatively measure participants' reliance on model predictions, similar to [52], I calculate user agreement rate as the number of news stories which the participant inspected and agreed with the model prediction (either true or fake news), divided by total number of shared or reported news stories.

4.2.7 Experiments and Results

I ran five between-subject experiments in different interface conditions for hypothesis testing. The study had a total of 220 Amazon Mechanical Turk participants with equal participants in each condition of which 47% were female, with 37.8% between 30-39 years old, 30.3% between 40-59 years old, 15.9% between 20-29 years old, and 3% between 50-59 years old; 51% had a bachelor's degree, 23% had a college degree less than bachelor, 14% had graduate school degree, 14% had high school education, and 1% had less than high school education. I removed data from 19 participants who spent less than 10 minutes or had especially low interaction behavior during the task. A total of 122.8 hours of study time was recorded for the remaining 201 participants, who on average spent 32.1 minutes (range = [10.3, 90.6] with SD = 20.3) on news review and selection, and 6.5 minutes answering surveys and reading instructions.

For statistical analysis, inferential tests used one-way independent ANOVAs to compare the conditions for each measure. In the end, I briefly review participants' qualitative feedback to see if they support the quantitative findings.

4.2.7.1 Human-AI Performance

To answer my first research question, I review and analyze the user performance measure for participants' news reviewing and sharing. I run a between-subject experiment with 40 participants in three primary interface conditions: 1) *Baseline* without any intelligent assistant, 2) Interface with the *AI Assistant*, and 3) Interface with the *XAI-all* Assistant.

Hypothesis 1: Users can share more true news stories with the help of XAI Assistant.

I report the credibility score of participants' shared news as the primary performance measure. Results show the average credibility score is higher than the original news feed (50% credibility) in all three groups that indicates the overall ability of participants in news review and their engagement with the task. Participants shared news in *XAI assistant* condition had the highest average of 75.05% (range = [61%, 92%] with SD = 10.06%) credibility and *Baseline* had the least credibility

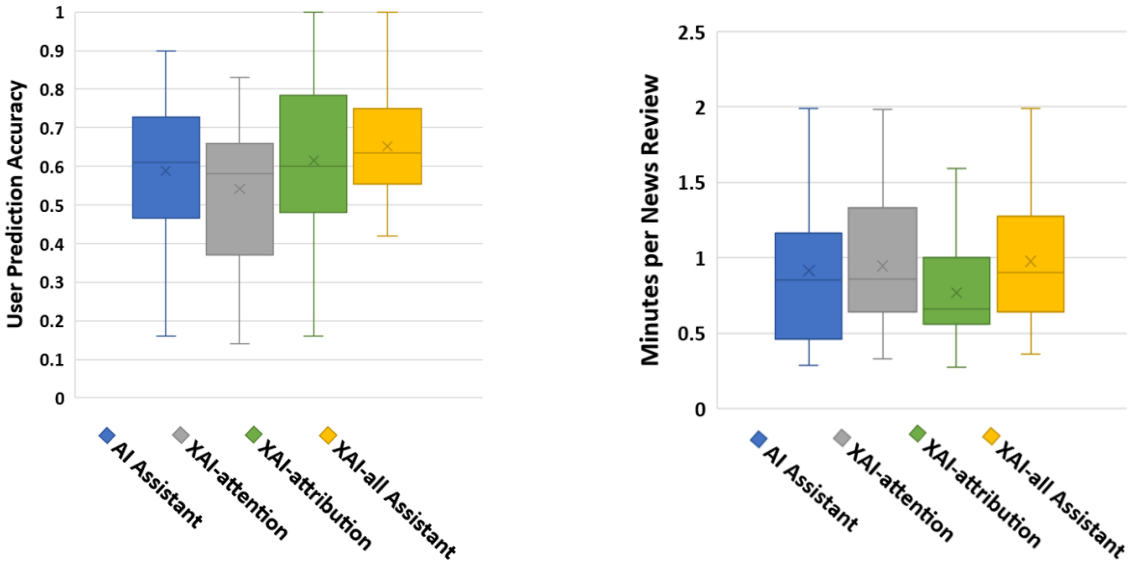
with 68.4% (range = [46%, 88%] with SD = 11.5%) credibility. The data met the assumptions for parametric testing for all groups with validation checks passing for data normality (Shapiro-Wilk) and homogeneity of variance (Levene's) tests. A significant effect was observed by an ANOVA test with $F(2, 107) = 3.32$ and $p = 0.04$ for the news credibility measure among all three conditions. A post-hoc Tukey test showed borderline significance ($p = 0.050$) between the *XAI assistant* and *Baseline* conditions, with higher news credibility scores for participants in *XAI-all* group compared to *Baseline* group.

I use incredibility score (calculated as *1.0 – credibility score*) of all reported fake news as the secondary performance measures. Similar to credibility scores for shared news, the *XAI* group has the highest average incredibility of reported fake news stories with 73.8% reporting fake news (range = [53%, 100%] with SD = 10.7%). An ANOVA test revealed a significant main effect with $F(2, 107) = 3.78$ and $p = 0.026$. A Tukey post-hoc test showed participants in the *XAI assistant* condition had ($p = 0.019$) reported fake news significantly more than the *Baseline* condition, even though reporting fake news was not the user's primary task during the study.

Implications of Results: The study results show that the *XAI assistant* improved user performance compared to the *Baseline* interface without any intelligent agent. However, model explanations did not significantly improve user performance over the *AI assistant* condition. Given the unique design challenges in misinformation detection models, this is a positive indicator that an intelligent agent together with model explanations can potentially improve collaborative human-AI news reviewing.

4.2.7.2 Mental Model

My second experiment is designed to answer RQ2 by studying the effects of model explanations on users' mental model and response time. I recruited new participants to ran studies for hypothesis testing through comparison of *AI assistant* condition (as the baseline) with three *XAI assistant* conditions (as treatments) in our interface.



(a) User Prediction Accuracy

(b) Task Cognitive Load

Figure 4.4: Evaluation of user (a) mental model through guessing model output and (b) cognitive load calculated as time per news review for AI Assistant, and three XAI Assistant conditions.

Hypothesis 2: Different types of explanations have different effects on user understanding of intelligent assistants.

My measure for quantitative evaluation of mental model is through user prediction task (user guessing of model output). Figure 4.4a shows user prediction task results from four study groups. User accuracy in their prediction task was highest ($M = 62.20\%$) in the *XAI-all* group and the worst ($M = 54.65\%$) in the *XAI-attention* group. The data passed parametric tests for normality (Shapiro-wilk test) and homogeneity of variances (Levene's test). An ANOVA test detected a significant main effect with $F(3, 149) = 3.16$ and $p = 0.026$ for participants between all four conditions with intelligent assistant. A Tukey post-hoc test yielded a significant difference ($p = 0.017$) between the *XAI-attention* and *XAI-all* groups. However, no significant pairwise difference was detected between the *AI* group and any of *XAI* groups. Note that average user prediction task accuracy in the *XAI-attention* group was lower than the *AI assistance* group, indicating the negative effect of explanations in participants' ability to predict model output.

Hypothesis 3: Model explanation increases users response time.

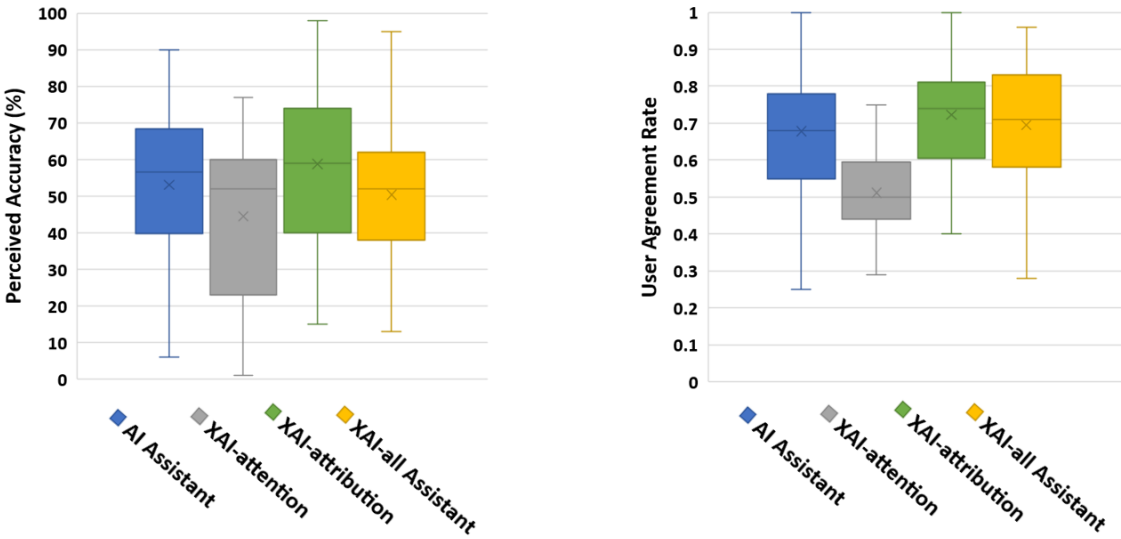
Figure 4.4b presents participants response time for *AI assistant* and three *XAI* groups. Averaged response time was the lowest in the *AI assistant* group and the highest in the *XAI-attention* group. An ANOVA test detected a significant main effect with $F(3, 149) = 3.34$ and $p = 0.021$ for participants response time between conditions in support of our initial hypothesis. A post-hoc Tukey test revealed a significant difference ($p = 0.046$) between *XAI-attention* and *AI assistant* group. This clearly indicates understanding and remembering the relation between attention map explanations (keywords importance in news headline and supporting articles) and model weaknesses is a demanding task for users. Additionally, Tukey's test recognized a larger but not significant ($p = 0.058$) between *XAI-attention* and *XAI-attribution* groups. Similar to [96], our results also show participants can better process low dimensional news attribute explanations (4 attributions: articles importance, news headline, news source, article content) compared to often lengthier attention explanations.

Implications of Results: The results show a significant effect of explanation types on user mental model based on the user-prediction task measure. However, none of the model explanation conditions improved users' accuracy in prediction. Notably, word level attention map explanations for news headline and articles (in the *XAI-attention* condition) had a negative effect on user mental model, potentially due to lower user satisfaction and engagement with the AI assistant. The discrepancy between user prediction task accuracy between the three XAI conditions indicates that not all explanations are informative or meaningful for end users to be able to predict model behavior. Additionally, even though I embedded on-demand visualization of model explanations in our interface, explanations impose more response time and require more time for user comprehension.

4.2.7.3 Trust and Reliance

To address RQ3, I review and analyze user trust and reliance measures in my experiments.

Hypothesis 3: Users have higher perceived accuracy in XAI assistant compared to AI Assistant.



(a) User Perceived Accuracy of Intelligent Assistant (b) User Agreement Rate with Intelligent Assistant

Figure 4.5: User trust measures for *AI Assistant* and three *XAI Assistants* conditions.

My primary measures for user trust in the AI and XAI assistant is the participants' perceived accuracy of the intelligent assistant. Figure 4.5a shows a box-plot of participants' perceived accuracy of the AI and three XAI assistants conditions. The results show participants had the highest rate of perceived accuracy in the *XAI-attribution* group (with the visualization of news feature attribution) and lowest in the *XAI-attention* group (with the heatmap of word feature attribution) on average. Using an ANOVA test, I found a significant difference ($F(3, 155) = 2.86$ and $p = 0.039$) between perceived accuracy in the four groups. For pair analysis, a post-hoc Tukey test revealed participants' perceived model accuracy of the *XAI-attribution* condition ($M = 58.70\%$) was significantly ($p = 0.024$) higher than *XAI-attention* ($M = 45.55\%$). Interestingly, participants in the *XAI-all* group responded with lower perceived accuracy ($M = 50.38\%$) compared to AI Assistant ($M = 53.05\%$) with no explanation.

Hypothesis 4: Users will agree more with XAI assistant predictions compared to AI assistant.

I measure user reliance on algorithms via the user agreement rate with AI and XAI assistants predictions. Figure 4.5b presents results for participants agreement rate with the *AI assistant* and

three *XAI assistant* groups. Overall, participants had near 0.70 agreement rate with model prediction in all groups except for the *XAI-attention* group with 0.51 agreement rate. I observed a significant main effect using an ANOVA test with $F(3, 149) = 16.44$ and $p < 0.001$ for participants agreement rate with intelligent assistants prediction. From the pairwise Tukey post-hoc analysis, participants had a significantly lower agreement rate in the *XAI-attention* group compared to all three other groups ($p < 0.001$ for all pairwise comparisons). Similar to participants' perceived accuracy, tests did not detect a significant increase in user agreement from model explanations.

Implications of Results: The study results indicate that model explanations helped users to adjust their trust and reliance on the intelligent assistant. I did not observe improvements in user trust or reliance for the XAI assistants over the AI assistant. In fact, participants actually lost trust in the *XAI-attention* assistant when—despite their initial expectations—they found the system was detecting fake news only based on news keywords. This could be considered an appropriate result given the limitations of the model logic. The lower user trust in the *XAI-attention* condition coincides with participants' mental model results and might suggest the effectiveness of explanations in helping users avoid overtrusting the intelligent assistant in cases when model logic may not be optimal or meaningful based on human logic.

4.2.7.4 Qualitative Feedback

Reviewing participants' written feedback in the post-study survey reflects their reasoning about AI assistant that provides further insight into participants' mental models of the AI/XAI assistants. Participants answered two descriptive questions regarding their mental model of the AI assistant's reasoning ("*How do you describe this AI's reasoning to find fake news?*") and AI assistant's limitations ("*In your opinion, what are the biggest limitations of this AI fake news detector?*"). The mean participant response length was 77.8 words (range = [338, 28], SD = 46.4) for all descriptive response forms. Two team member separately reviewed participants' qualitative feedback and performed open coding to extract themes in participants' notes and comments. Then, the two students coded participants' free response questions to identify salient themes. Over three sessions of coding and discussion, we identified 19 codes with an inter-rater reliability of 0.82. I use codes to

from three main categorizes of responses: AI reasoning, AI limitations, and participant-strategy.

Regarding participant mental models of AI assistants, we observed that explanations clearly improved their understanding of AI reasoning. On average, 63.5% of participants in the *XAI-attention* and 52.8% of in the *XAI-all* group pointed out the importance of keywords in the news; example comments include "*I think it looked for certain key words*" and "*The AI compares relevant phrases in the headline to relevant keywords in the supporting stories.*" In contrast, only 17.9% of participants in the *AI assistant* condition had expressed such understanding. I also found 62.0% participants in the *XAI-attribution* group mentioned related articles and their sources as key features for AI reasoning compared to 31.7% in the *XAI-attention* group. For example, one participant in this group commented "*It tries to pull related articles from the web to prove or disprove the headline*", and another participant said "*I think it went by how many article below matched the news.*

I found interesting feedback on participants' subjective opinions on the limitations of the AI assistant. I saw a clear theme in responses of the need for common sense to distinguish fake and true news. On average, from 20% of participants in all conditions (except *XAI-all* with 11.1%), I received comments such as "*it doesn't have human judgment*", "*I guess they will not see common sense*", and "*The AI doesn't have the experience that a real person has in dealing with the fake news out there.*". Also, participants in *XAI-attention* group paid more attention to the quality and combination of articles in each news story with 43.1% of them expressing comments like "*AI doesn't have enough information*" and "*It doesn't see multiple sides of the story*" compared to other conditions with the average of 19.3%. Additionally, 27.2% of all participants expressed concern about AI ability in understanding the context of the news or recognizing sarcasm. For instance, one said "*I think it can't detect sarcasm satire or parody so it has some limitations*" and another mentioned that "*The AI isn't able to understand the context of the text. It's not able to actually understand the story or [its] plausibility.*". In another example, the participant said:

> "*It can't seem to discern between junk news and the real deal. It can't discern that a site is biased. It could pull the keywords that's for certain but any site could have the*

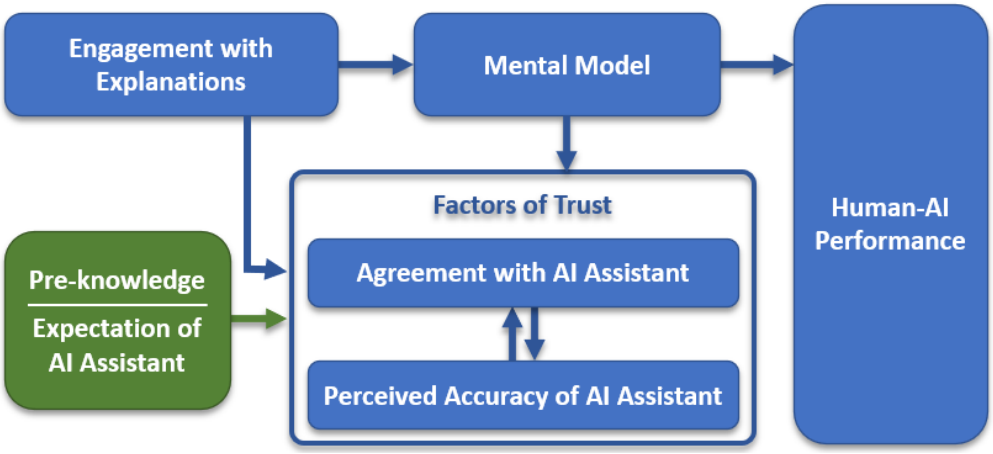


Figure 4.6: Conceptual model of relationships among user engagement, mental model, trust, and human-AI performance in XAI systems. Figure created based on a model of the "process of explaining" in XAI context from [48].

keywords. There's more to it than that and it can't do it."

Challenges participants encountered in learning the model behavior was also reflected in 13.5% of participants' comments in all groups, for example one said:

> *I said [to myself] twice that I thought I understood how it worked but when asked to predict the AI's inference about a given headline in the last portion of the study I believe I only matched one out of four so maybe I didn't understand anything that well.*

Overall, the qualitative user feedback complement the quantitative findings in showing which model explanations helped participants to observe model limitations and adjust their trust and reliance accordingly.

4.2.8 Implications of Results

In this section, I summarize different implications of my study results from our XAI system and how machine learning explanations and fake news detection. Following Hoffman et al.'s [48] conceptual model (Figure 4.6), I look for correlations between my measurements of user engagement, mental model, performance, and trust to investigate the interplay between these factors.

4.2.8.1 User Expectations of AI Assistant

I first analyze the relation between user expectations of AI before the study and their perceived algorithm accuracy after the study. Research shows that various external and internal factors can interact with user trust, with examples including user pre-knowledge [51], model stated performance [52], and model observed performance [53]. In the pre-study questionnaire, I measured 1) participant expectation of AI assistant accuracy and (with "If you had an Artificial Intelligence (AI) algorithm to review your daily news for fake news detection, what would be your expectation of AI accuracy to do a good job?" question) and 2) participant estimation of fake news rate in media (with "In your experience, what percentage of news that you read daily is false news? e.g., fake news, hoax, rumors, made up stories, misinformation" question).

As expected, a Pearson test shows a positive correlation ($r = 0.223$, $p = 0.005$) between participants' perceived accuracy at the end of the study and their initial expectation of AI accuracy. Regarding participants' expectations of fake news occurrence in daily news, I expected to see more user engagement for participants with higher anticipation of fake news. However, I surprisingly found that participants expectation on fake news occurrence has a negative correlation with their engagement with AI assistant ($r = -0.189$, $p = 0.018$). This could be due to participants underestimating the AI assistant or choosing their intuition rather than model suggestions.

4.2.8.2 Engagement with Intelligent Assistants

As an objective measure of user engagement with intelligent assistants, I consider total continued usage based on the frequency of user interactions (clicks count) with the AI and XAI assistant predictions. Overall average results show that participants in the *XAI-all* group had the highest engagement rate with the XAI assistant (0.95 prediction check rate) for shared or reported news stories. An ANOVA test of user engagement with the AI assistant found a significant difference with $F(3, 156) = 2.773$ and $p = 0.046$ between conditions, and a Tukey post-hoc test shows participants were significantly ($p = 0.034$) more engaged in the *XAI-all* condition compared to *XAI-attention* condition.

The conceptual model of the process of explaining [48] suggests that explanations in XAI system revise mental model and can engender appropriate trust, see Figure 4.6. To test the interplay between user engagement and their mental model of XAI assistants, I performed a bivariate Pearson correlation test between user engagement rate and prediction task accuracy as the mental model measure. Despite the initial hypotheses, a Pearson correlation did not show a positive relation between engagement and mental model ($r = 0.099, p = 0.215$). This could be due to the narrow scope of mental-model measurement in my study being limited to the user prediction task (model predictability for users). However, user engagement had a significant positive correlation ($r = 0.228$, $p < 0.001$) with user agreement with the intelligent assistant. This shows as more participants got involved with the AI or XAI predictions, the more they agreed with its predictions.

4.2.8.3 Mental Model Affecting Performance and Trust

Next, I analyze how users' mental model interacts with trust and human-AI performance. A Pearson test between users' prediction task accuracy (mental model measure) and perceived accuracy of AI assistant (my first user trust measure) showed a positive significant correlation ($r = 0.212$, $p = 0.008$). A correlation test between user prediction accuracy and user agreement with the AI assistant (my second user-trust measure) also showed a positive significant correlation ($r = 0.280$, $p < 0.001$) between participants' mental model and trust. Positive correlations of mental model with both trust measures demonstrate the relation between predictability of the intelligent agent and trust.

As hypothesized, user prediction task accuracy was positively correlated with credibility of shared news ($r = 0.305$, $p < 0.001$) as well as incredibility of reported fake news ($r = 0.283$, $p < 0.001$). This finding suggests users with a more accurate mental model could better guess model failure cases, and by avoiding those cases, they could improve their performance.

4.2.8.4 Interactions Between Trust Measures

Another interesting finding from my study is that I observed interactions between multiple measures of user trust. Previous research studies have utilized various independent trust measures

such as perceived algorithm performance [53], perception of control over the system [54], and the rate of user agreement with an algorithm's recommendations [52]. In my studies, I measured two different trust factors to examine how they may interact. A Pearson correlation test between the two trust measures shows a positive significant correlation between the perceived accuracy and user agreement rate ($r = 0.482$, $p < 0.001$). This positive correlation suggests that as users feel more confident about AI competence, they tend to agree more with its predictions.

4.2.9 Lessons Learned

In this case study, I evaluated model explanations from multiple models for intelligent assistance in the fake news detection task. The case study allowed to validate the usefulness of my framework and its guidelines for designing an end-to-end XAI system. Plus, this case study was an opportunity to study how different types of explanations affect users in fake news detection. To analyze the study results, I first used analysis of means for hypothesis testing based on the initial research questions, then performed correlation analysis for meta-analysis of the results based on the conceptual model for XAI process.

In conclusion, my research revealed multiple challenges in designing effective XAI systems in the fake news detection domain. In particular, I observe challenges rising from the inherent difference between models' feature learning (word-level features in our case) and human understanding of news and information. Overall, users' interaction with the AI and XAI assistants affected their performance, mental model, and trust. However, model explanations in my studies did not improve task performance or increase user trust and mental model. Instead, the quantitative results and qualitative feedback indicate that explanations helped users' to build an appropriate mental model of intelligent assistants and adjust their trust accordingly, given the limitations of the models. For example, participants in the *XAI-attention* group that was significantly less successful in guessing model outputs also showed significantly lower trust (in both trust measures) compared to the *XAI-all* condition. Likewise, reviewing user engagement results showed that *XAI-attention* explanations were not appreciated by the users. Similarly, reviewing qualitative comments showed that the majority of users did not appreciate the keyword-based explanations as reliable. There-

fore, I conclude that improving transparency of the model helped users to appropriately avoid overtrusting the fake news detector when they found the AI reasoning was not trustworthy or simply explanations were nonsensical. Future research is needed to assess the effectiveness of other types of explanations, such as knowledge graphs and multi-modal evidence retrieval on users in fake news detection assistants.

4.3 Example 1: Video Activity Recognition

4.3.1 Introduction

Following the case study presented in Section 4.2, I use my framework to analyze an example XAI system from perspectives of design process workflow (between-layers) and design and evaluation choices (within each layer). This analysis is aiming to find insights from their work and intended to suggest future design iterations. In this example, I analyze Nourani et al.'s [1] paper in which authors present an XAI system to support AI novice users tasked with activity recognition in a series of videos. This XAI research focuses on comparing variations of explanation veracity for users in their video review and querying task. The authors explore the importance of explanation veracity for user performance and agreement with the intelligent system through a controlled user study. To provide an in depth analysis, I conducted an interview with the first author to review their design step and main considerations during the process including interactions between machine learning designers and interface designers in the team. The following descriptive analysis will emphasize on the design process in Nourani et al.'s [1] XAI system as compared to my framework.

4.3.2 Analysis of Workflow

I present a descriptive analysis in this section to review the design process (between-layer) and decision choices (within-layer) in the Nourani et al.'s [1] XAI system. Figure 4.8 shows the result of my breakdown of the design and evaluation steps in Nourani et al.'s [1] XAI system in terms of my nested framework. This visual presentation presents transitions between design and evaluation steps during the multidisciplinary team work. The following subsections reviews the design and

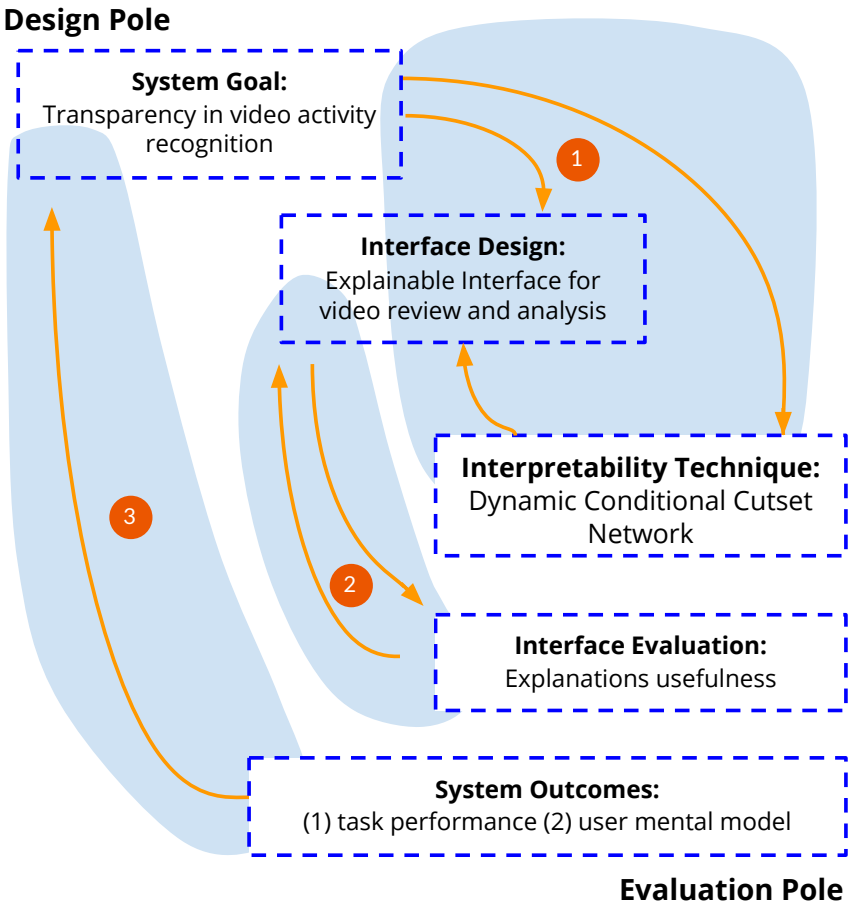


Figure 4.7: My analysis of Nourani et al.'s paper [1] in terms of my proposed nested framework. Boxes represent design and evaluation steps and arrows show transitions between the steps. The main system design and development includes: (1) Translating system goals into interface components and interpretability requirements while taking machine learning limitations into consideration (2) Interface usability and user task testing (3) Evaluating system outcomes and revisiting design goals.

evaluation details for their between step and within step activities.

4.3.2.1 System Goals

The main system goal for this system is to improve algorithmic transparency with the help of model explanations. Model explanations are aiming to help users understand how the activity recognition model works and therefor perform better in their task. The system is designed for novices (a non-specialist population) without any particular domain expertise or AI knowledge.

To adjust the task difficulty to their targeted user type, authors chose to use cooking videos in a kitchen setting as the main system input and user were to identify cooking activities in the videos. Following a previous work [53], authors decided the video activity recognition task and interactive interface design would be suitable real-world scenario for studying XAI systems.

The choice of what to explain to users as the explanations from the machine learning model had multiple bottlenecks. The bottlenecks included different explanations type design requests from HCI designers in the team that were design bottlenecks for machine learning algorithms. For example, the idea of presenting a global explanation using a tree summarization visualization for the activities recognized in the whole cooking videos was not possible due to model constraints. Additionally, the HCI designer have requested for instance explanations for each component (i.e., objects, actions and location) in each frame in form of saliency map or bounding box. However, machine learning model design limitations did not allow for high-quality frame-level (i.e. in each image) explanatory information for model predictions. Therefor, the main explanations to was planned as video segments that attributed to model prediction.

4.3.2.2 Interface Design

An interface is designed to present the videos, key video segments as model explanations, and top-3 model predictions for each video segment. The implementation is an interactive web-based interface in the front end. Figure 4.8 shows the components of the final interface.

Design iterations involved an initial mock up sketch step to review with all team members and discuss the necessary components and details. In the next step, authors ran a round of user testing and interview using sample videos and wizard-of-oz intelligent assistant to collect quantitative (system logs) and qualitative feedback regarding interface usability. The user testing was done with 10 participants including design team members (familiar with the system) and fresh participants. Then, authors performed a round of pilot testing after refining the interface based on feedback from first user testing. The pilot testing included the model's predictions and explanations for the main video dataset. The main goal for this interface testing round was to refine user task design, clarity of task instructions and steps for the main study. Authors used think-aloud method to collect user

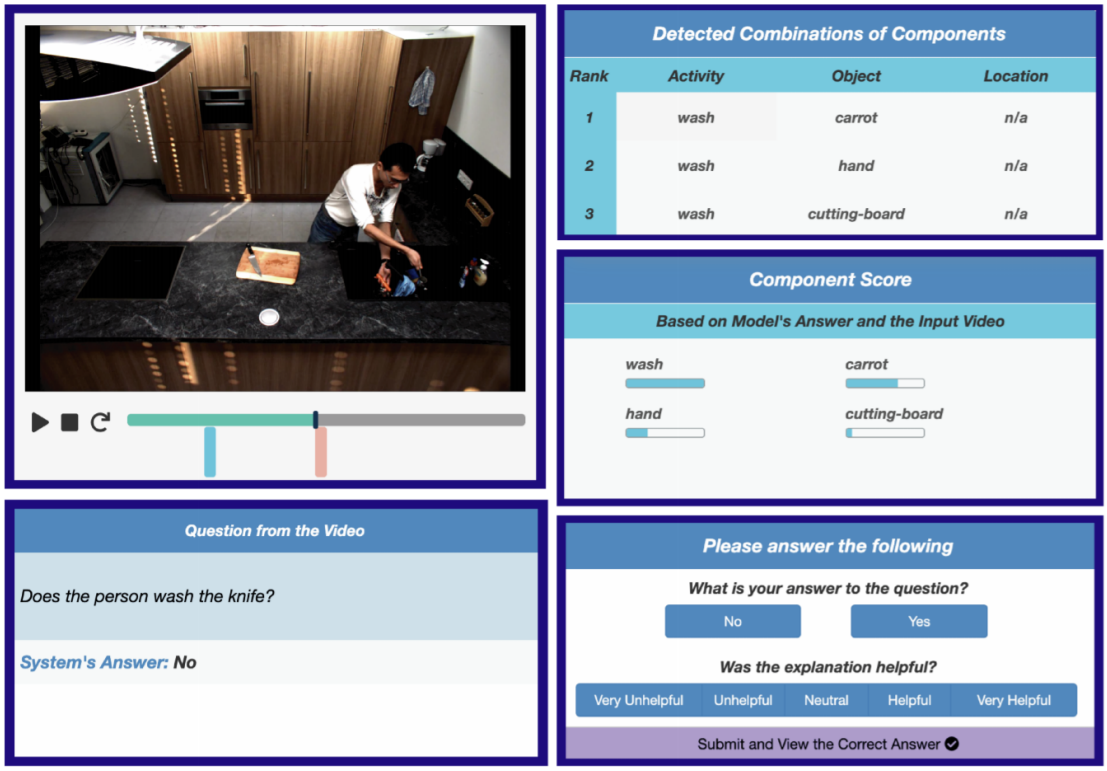


Figure 4.8: The XAI user interface design by Nourani et al. [1]. Different panels and components including a video player panel with visualization of attributed video segments (top-left) and the query panel that presents the XAI's prediction (bottom-left).

feedback during the pilot study. Pilot tests focused on understanding the usefulness and helpfulness of interface components to perform the task.

The pilot studies also allowed authors to review model outputs prior to the main system outcome evaluations. For example, considering the system outcome evaluation goals and study hypothesis, the model performance (96% prediction accuracy) on the test samples was higher than expected to conduct the user studies. Authors selected 20 videos from scenarios which the model has lower accuracy with a combination that results in 16 correct and 4 wrong model predictions. Another interesting observation during iterative rounds of pilot testing with new models was that the model outputs and explanations were changing at each training round. While this is part of the machine learning design cycles, it was not expected in the interface design cycles.

Another example of limitations from machine leaning explanations that were identified during

pilot testing was the inappropriate length of attribute video segments. The interpretability technique was generating video segments with less than 1 second length which was too short for user review in the interface. After merging short and adjacent video segments, the video progress bar was showing longer video segments that were appropriate for user review. However, the team had to eliminate merging of any short attribution video segment without adjacent segments. This was an example of trade-off between presenting trustworthy explanations and user understanding of explanations during the design cycles.

4.3.2.3 Algorithm Design

The training dataset for the model is a publicly available cooking videos (TACoS dataset [286]) that has labels for supervised learning. The cooking videos in the TACoS dataset are recorded in a simple kitchen setting which a person is performing short cooking-related activities in each video clip with no visual occlusion or any co-occurrence activities. The choice of dataset enables non-expert users to easily review video data and identify requested activities. Further to limit the training domain and simplify the user task, authors chose 28 components (including actions, objects, and locations) from the dataset for the model train.

As for the machine learning solution, authors chose a two layer approach in which a temporal probabilistic model mimics the relationship between DNN predictions and the ground truth labels. The main DNN model is based on a backbone model pre-trained on large image datasets and performs well after training on the target cooking video dataset. The interpretable model enables complex temporal inference queries to explain its predictions by providing video segments for the predicted activities. The queries include the three elements of action, object, and location. For example, users can ask "Does the person cut the orange on the plate?" or "Does the person wash the knife?". However, authors use a set of pre-defined queries are selected and implemented in the interface for the sake of demo and user studies. Additionally, the interpretability algorithm generates the top-3 activities for each key video segment.

The iterative interface design and evaluation steps revealed bugs in machine learning training. For instance, after implementing a query tool for users to search objects and actions in the video,

the preliminary user testing results showed that the model does not identify any edible object in the videos other than the "carrot". Then, the machine learning team updated the training set to to fix this bug.

4.3.2.4 System Outcome Evaluation

In the last evaluation iteration, the authors chose a range of measures to evaluate the system outcome in accordance to the original design goals. In a controlled study with Amazon Mechanical Turk participants, authors studied the effects of model explanations and veracity of explanations on end users. The main objective measures included user task performance with the help of model explanations and user understanding of model via predicting model outputs. The main subjective measures included user's perceived accuracy of the AI and subjective rating of their trust on the system. Author also measured explanations usefulness and helpfulness using likert-scale questions.

Furthermore, authors paid attention to improving their measurements by calculating class-based user task performance and user prediction performance in contrast to overall task performance. Analysis of results showed it was more difficult for users to understand model weaknesses compared to learning system strength. The series of user studies also confirmed the strong effect of study condition's ordering when studying AI-based systems in which users learning significantly affects the results in within-subject study design.

4.3.3 Lessons Learned

I presented a descriptive analysis for design and evaluation workflow of the Nourani et al.'s [1] XAI system. When compared to my proposed nested framework, my analysis (figure 4.7) shows missing evaluation and design iterations and design steps from the guidelines. I find this as an opportunity to continue the design cycles for addressing the identified limitations in the final system evaluation. Looking into the system outcome evaluation results, the authors also suggest the need for more refinement of the machine learning algorithm. The descriptive analysis example indicated the importance of iterations between framework layers to identify design bottlenecks and improve system outcomes. The authors mention communication barriers between the HCI and machine

learning teams during the team work, which is another common issue in multidisciplinary team work.

4.4 Example 2: Interactive Naming for DNN Visual Concepts

4.4.1 Introduction

In the analysis of second XAI system example, I use my framework to analyze Hamidi-Haines et al.'s [2] paper in which authors present a XAI system for interactive clustering of visual concepts in model explanations. In this analysis, I am aiming to find insights from their XAI system design by structuring their design process based on my framework. Their system design is followed by a systematic study of the visual concepts created by participants when using the interactive naming interface. Authors studied the problem of users' mental model and understanding of DNNs decision-making in terms of human-recognizable visual concepts.

4.4.2 Analysis of Workflow

I present a descriptive analysis in this section to review the design process (between-layer) and decision choices (within-layer) in the Hamidi-Haines et al.'s [2] XAI system. The analysis starts with identifying system-level goals, interface design steps, and algorithm implementations. Then, I present the human-subject study and results to evaluate the system outcomes.

4.4.2.1 System Goals

The main system goal in this paper is to develop a tool to help users understand the decisions of a DNN trained for multi-class image recognition with supervised learning. Such a system can provide insight into the strengths and weaknesses of the network's decision making that may not be observed by common performance metrics like test set precision and recall. For example, one might discover cases in which the DNN is making the right prediction by looking at the wrong or nonsensical reasons, which would identify potential future mispredictions. Authors formulate their main research questions (RQs) as following:

- What fraction of DNN activation maps are explainable using human recognizable visual

concepts?

- Is there a strong relation between DNN activation maps and human recognizable visual concepts?
- Is there strong consistency between users' defined visual concepts?

For the choice of what to explain, authors chose to visualize neural network activation maps that represents the most important features that activate each node for the input instance. The heatmap visualization of activation maps on each image would represent which features (i.e., pixels) are attributed to the node's activation.

4.4.2.2 Interface Design

Authors designed an interface to present images examples, visualize activation maps, and enable users interactions with samples. The interface is designed to support interactive clustering of visual representations to establish semantics to the DNN's activation maps limited to the test set. The web-based implementation of the interface as shown in Figure 4.9. The set of activation maps is presented to the users in the unlabeled examples panel (top panel) section of the interface. A list of clusters for visual concept underneath is where users can drag and drop examples (from unlabeled examples panel) with similar prediction reasoning. Users can give a textual label or name to each cluster of visual concepts.

Users' task is to review and cluster visual representations (X-feature activation maps) from the top unlabeled panel into visual concepts cluster (bottom panel) and label each cluster. The interface allows the users to review and compare all images in the test set. After review and clustering of instances with meaningful heatmap explanations, the remaining instances in the unlabeled panel will remain will be eliminated from the system as less representative features.

4.4.2.3 Algorithm Design

The model explanation in this work is based on DNN activation maps with a multi-step approach to improve human interpretability of features. First, authors reduce the number of maps

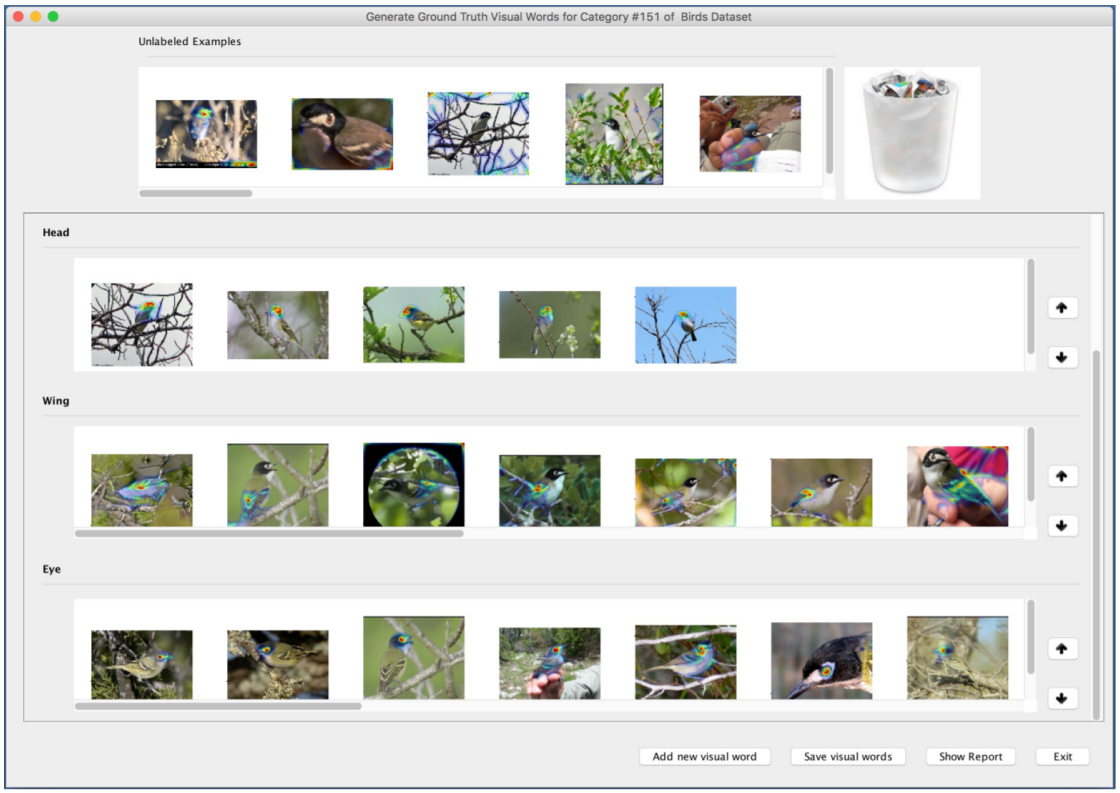


Figure 4.9: The user interface used for visualization of feature activations and interactive naming reprinted from Hamidi-Haines et al. [2]. System helps users to create clusters (bottom panel) of similar activation maps for understanding DNN features and identifying human understandable visual concepts.

using a mimic learning technique and then visualize the reduced activation maps with a gradient-based visualization technique.

For the first step, they utilize explanation modules [287] to the fully-connected layer of the original DNN to reduce the number of activation maps to a handful of X-features. The new explanation module forms a low-dimensional explainable concept space for the original deep networks which is desired to maintain the following properties (1) faithfulness of X-features to the original model, (2) sparsity of X-features, and (3) orthogonality of X-features. Additionally, the authors define the significant X-features for each input instance to be a subset of X-features that account for at least 90% of the prediction score. We call these maps the significant activation maps or simply the significant activations. Lastly, activation maps of each X-feature on input images are

visualized using Excitation Back Propagation [288] (ExcitationBP), a gradient-based visualization approach.

4.4.2.4 System Outcome Evaluation

The evaluation of this system was done for an image classification task on the Caltech-UCSD Birds [289] dataset which includes 12 labeled categories of birds. After training the main DNN with explanation modules to obtain X-features, authors applying ExcitationBP on each X-features and present feature maps with higher than %90 importance to the model prediction. This approach reduces the feature dimensionality without significant loss of accuracy from 4096 features (in the main DNN) to only 5 X-features (from the explanation module).

The authors run a user study with five participants to work with the XAI interface for reviewing images with the activation heatmap overlay. The participants were instructed cluster images with similar explanations and label each cluster with an appropriate visual concept. Although not all images were supposed to be reviewed and clustered, participants' were instructed to add as many images to each cluster during the study. The evaluation results were analyzed with three measures (M1-M3) for quantitative and qualitative assessment of initial goals.

Completeness of Visual Concepts (M1): For the first evaluation measure, authors measure how well the participants cover the activation maps. Authors define the partial coverage and complete coverage metrics for activations that have been labeled by at least one or all participants. Results show that approximately between 20% to 40% of activation maps were not labeled with visual concept clusters. However, we see that partially covered activations are quite high for most annotators. The comparison between each individual participant and union of all participants show consistency between them.

Looking into the study post-interview results show participants were unsure about activation maps with unclear semantic information such as case in which activation heatmap showed the edges of the image or mainly looking at the background.

Visual Concepts Correctness (M2): Next, authors measure if activation maps generated by the system can fully represent human-understandable visual concepts. Specifically, the goal is to observe

whether each activation map can represent a particular semantic visual concept. Authors adopt the metric of cluster purity [290] to measure the quality of each cluster of visual concept created by participants.

The purity of a visual concept cluster is calculated as the number of activation maps that belong to the major map in each cluster divided by the total number of examples in that cluster. Results show that the purity rate for different clusters varies between 0.52 to 0.90, suggesting that the mapping from activation maps to visual concepts is not a one-to-one relation. However, all participants show a relatively consistent purity rate in their visual concept clustering task, indicating the mutual perception of users on DNN features.

Participants Agreement (M3): For the last evaluation measure, authors review similarity in the labels that participants used for each visual concept cluster. The similarity metric between cluster labels helps to how well participants agree on semantic visual concepts. Note that participants were free in creating any number of clusters and labeling these clusters during the task.

Analysis of user labeling results shows the largest fraction of activation maps are annotated by all participants, indicating an overall agreement in cluster labels. Next, authors look into activation maps in similar clusters to find potential translations between the labels among participants. Overall, many of the mismatched cluster labels have the same are sensible based on descriptions provided by the annotators. Also, in many cases, different choices of labels show a difference in the resolution of choosing labels of objects (i.e., birds) parts in the image. This problem could be potentially resolved by using dictionaries to reduce the total number of labels used for visual concepts.

4.4.3 Lessons Learned

I present a descriptive analysis of Hamidi-Haines et al.'s [2] XAI system for interactive naming of DNN activation map to create human recognizable visual concepts. This XAI example also showed the importance of following design guidelines in for tools dedicated to the analysis and annotation of DNNs involved with end-users. When compared to my proposed nested framework, I found the authors' attention and focus on supporting user understanding of DNN features for

creating visual concepts. This is distinct from the previous case study and example in which the primary goal was to design an XAI assistant for end-users. Authors mostly relied on computational methods to analyze participants' data and draw conclusions about visual concepts. I find this XAI system design examples as an opportunity to identify new design considerations focused on user understanding and perception of raw model explanations. In comparison to my proposed XAI framework, the promising results reported from system evaluation are mainly derived from the interpretability technique (Layer 3) and user understanding of explanations (Layer 2) that address the initial research questions. However, looking into the evaluation process, I find the evaluation of system outcomes (Layer 1) to be a missing piece that could have improved system evaluation for obtaining conclusive results, necessary to proceed with the system design steps.

4.5 Findings and Conclusion

I presented a case study and two XAI system design examples to demonstrate and validate different functionalities of the proposed XAI framework. Specifically, the first case study demonstrated the *Generative Function* of the framework with step-by-step review of a fake news detection XAI system. Then, for the *Descriptive Function* of the framework, the design process of two existing XAI systems were analyzed (e.g., for communication purpose and to assess design alternative) and recommendations for next design cycles were made.

My case study and reviews focused on both design steps and evaluation steps of these XAI systems to identify their process and pitfalls in comparison to the guidelines from my framework. The lessons learned from the case study and examples showed that there are opportunities to improve the framework by introducing new design and evaluation methods (within-step contribution) specific for XAI systems as well as by adapting guidelines from other frameworks (between-step contribution). For example, the interactive naming in Hamidi-Haines et al.'s [2] XAI system showed the extend in which the annotation data collected users could be used for insights on model explanations meaningfulness. Also, the Nourani et al.'s [1] XAI system showed the difficulties in users' learning of model behavior on complex domains and tasks. Accordingly, in the next two chapter I will introduce two evaluation methods for specific to XAI systems as contributions to

the my framework. Additionally, future work and case studies are needed to further validate the usefulness of this framework in XAI system design.

5. USER TRUST DYNAMICS IN EXPLAINABLE AI

5.1 Introduction

Research shows user trust and reliance on AI predictions could enhance human-AI performance in a collaborative setup when learning a mental model of AI error boundaries [62]. However, building a correct mental model to achieve justified trust can be difficult in situations involving complex and demanding tasks, which often results in users over-trusting or under-trusting the intelligent system in the process of learning and revising their understanding [291]. Consequently, the evaluation of intelligent interfaces can be difficult due to the longitudinal nature of user experience and learning in cognitive tasks. As a common case within complex AI-based systems, user responses to insights may trigger long after exposure and affect system evaluation results [10]. Such examples indicate the necessity of repeated measurements during human-subject studies and look into relations between multiple measures in complex systems. Studying dynamics of user behavior is particularly important to understand users temporal patterns of trust and reliance on the intelligent agent and improve system design and evaluations accordingly. Following the discussion in Section 3.5.4, the proposed XAI design and evaluation framework also lacks the explicit emphasize on evaluating dynamics of user behavior in evaluation cycles.

To improve evaluation techniques for XAI system outcomes (Layer 1, Section 3.2.4), I analyze the human-subject evaluation data from our case study in Section 4 and investigate types of user trust dynamics in an explainable intelligent assistant. Specifically, I investigate the effects of interpretability on trust evolution over time in a human-AI collaborative setup for fake news detection. My study results show not only model explanations effect on user trust level (see Section 4), but also trust morphs over time. I cluster user trust changes over time into five types of trust dynamics and look into each cluster for insights on user behavior trends. Analysis of results revealed a positive interaction between two constructs of trust and positive effect of initial user expectation of intelligent assistant on their trust journey.

5.2 Method

My analysis goal is to study the effects of model explanations on temporal dynamics of user behavior which is an addition to the existing static measurements (Table 3.2) I aim to investigate how user trust and reliance in the explainable intelligent agent changes over time as the user is interacting with model predictions and its explanations. I formulate the main research goal with the following research question: How does user trust evolve over time, and how is the evolution affected by the presence of AI explanations?

To answer this research questions, I analyze user study data from the fake news detection case study presented in Section 4 in which novice users worked with a news review interface with a build-in intelligent assistant. I periodically measured participants' subjective trust and calculated their reliance on the AI predictions based on study logs. I exposed participants to different types of intelligent assistants (with and without explanations) in each study condition. In the following I briefly revisit the user task, study design, and its periodical measures.

5.2.1 Experimental Design

Considering I am using the same collected data as the fake news detection case study in Section 4, I only review the general study conditions, dynamic trust and reliance measurements, and leave the details to the Section 4.2.6.1.

I run a controlled human-subject experiments with a baselines and three control conditions to study user behavior with the AI assistant and its explanations. The study followed a between-subjects design with four different conditions, in which each participant used one variation of the system as described in the following:

The *AI Assistant* baseline condition includes the AI prediction and its confidence for the credibility of the news story. The AI predictions are in the form of on-demand using a collapsible menu on user click. I used three *XAI Assistant* conditions to study how different explanations types might affect user trust. The user interface in the three XAI conditions provides instance explanations in addition to the news credibility prediction. The *XAI-attention* condition presents a

heatmap of keywords using attention weights for the news headline and in each related news article (in the news article pages). The *XAI-attribution* condition shows news attribution explanations for related articles and their news sources. The hierarchical attention network generates an article's importance score and the top-3 important sentences from each article. The mimic learning model generates source, article, and news story attribution score for each article. The *XAI-all* condition is the combination of explanations in the *XAI-attribution* and *XAI-attention* conditions. The purpose of this *XAI-all* condition was to study the effect of variety of multiple explanation types together.

The study procedure is the same as the presentation in Section 4.2.6.2 with additional details on periodical measurements for user trust. Specifically, participants' were answering mini-questionnaires during the task as described in the next section.

5.2.2 Dynamic Measurement

I use subjective and objective measures of user expectation, subjective trust, and reliance to aim for investigating the initial research question.

- **Expectation of AI:** In the pre-study questionnaire, I measure participants' expectation of AI assistant accuracy by asking "If you had an Artificial Intelligence (AI) algorithm to review your daily news for fake news detection, what would be your expectation of AI accuracy to do a good job?" question. I intend to test the possible interactions between users' pre-knowledge and expectations and trust during their experience with the system.

- **User Trust:** I measure user trust using the subjective rating of participants' perceived accuracy of the AI assistant twice during the study (at 1/3 and 2/3 task progress) and once at the post-study questionnaire. Specifically, participants answer "What was the accuracy of the AI fake news detection?" using a continuous slider bar (between 0–100%) with the step size of 1.

- **User Reliance Rate:** My metric for user reliance on AI assistant is based on (1) user agreement and (2) engagement with AI predictions. Similar to [52], I count user agreement (or

disagreement) instances by news stories which the participant inspected and agreed (or disagreed) with the model prediction; either for the true or fake news samples. (1) User agreement rate is calculated as the difference between total agreement and disagreement instances divided by total instances which the user inspected the AI prediction. (2) User engagement (in 0 to 1 range) is calculated as total divided by the total number of user's model prediction inspection divided by shared and reported news. User reliance is calculated as the multiplication of user agreement and user engagement rates.

5.3 Results

In this section, I evaluate how different types of model explanations affect on dynamics of user trust and relianace. We hypothesize that explanations would help users to build appropriate trust with respect to observed system performance (set of %75 accurate predictions).

5.3.1 User Trust Dynamics

In this section, I analyze the repeated trust measurements during the study to investigate how user trust on intelligent assistant evolved over time.

Using the three perceived accuracy measurements during the study, I identified four profiles for trends of participants' trust in model prediction. I use rule-based clustering to classify participants into clusters of 1) progressive trust in which participants continually have higher perceived accuracy, 2) tentative trust with overshoot, 3) tentative trust with undershoot, and 4) digressive trust where participants continually lose trust in the AI. I analyzed the four clusters of participants with similar behavior for further insights.

Figure 5.1 shows the clustering results for trust measurements during the study (T1-T12). Overall, the most common trend (36.3%) was tentative trust with an overshoot, 23.6% gained trust continually, 21.0% lost trust continually, 10.8% did not change their subjective trust feedback, and 8.3% had trust undershoot during the task. A Pearson Chi-square test shows a significant relationship with $\chi^2(18.89, N = 139) = 18.895$ and $p = 0.026$ between study conditions and user trust types. Results show that the majority of participants 61.7% from the *XAI-attention* condition had

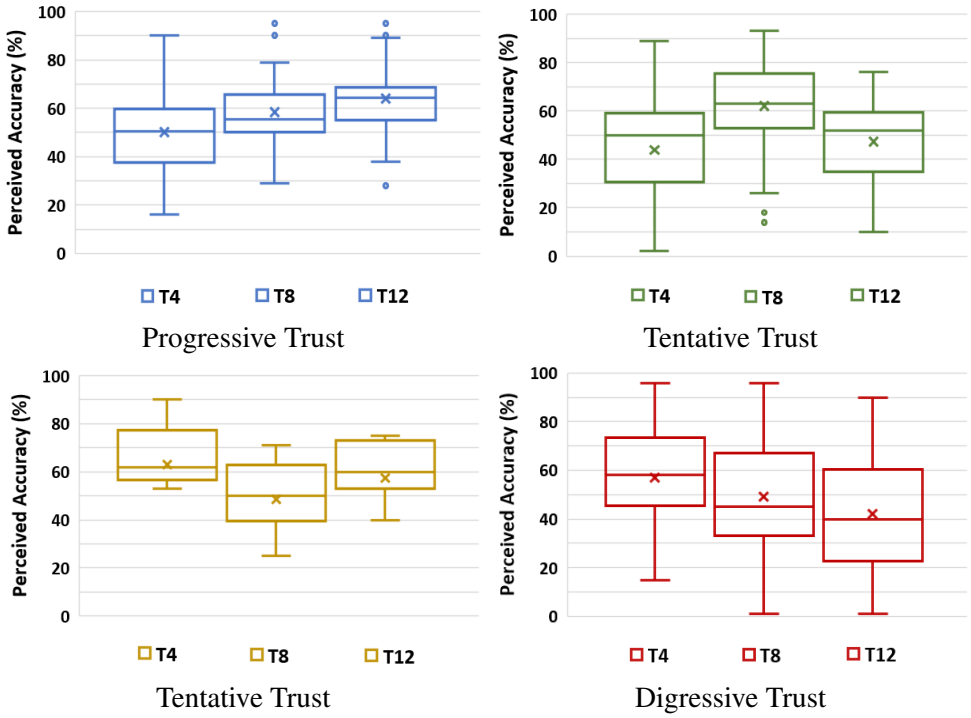


Figure 5.1: User trust dynamics: four profiles of participants' trust changes over time. Subjective perceived accuracy of news assistant is measured three times during the study (at T4, T8, and T12) in the range of 0-100.

overshoot in their second perceived accuracy measurement. In comparison, 34.2% of participants from the *AI* group and 39.4% of the *XAI-attribute* group were continuously gaining trust in the system.

Clustering participants with their trust evolution patterns shows model explanations effect on how user trust evolves over time. However, this effect was not observed for user reliance changes over time. Also, I did not detect dependency between the same clusters in reliance and trust measures. This could be because of the possible lag between user exposure and insights in complex interactive systems. This latency between users' interactions with the system and coming to their conclusions have also been reported in previous research [10].

5.3.2 User Reliance Dynamics

In this section, I analyze the repeated measurements during the study to investigate how user reliance on intelligent assistant evolved over time.

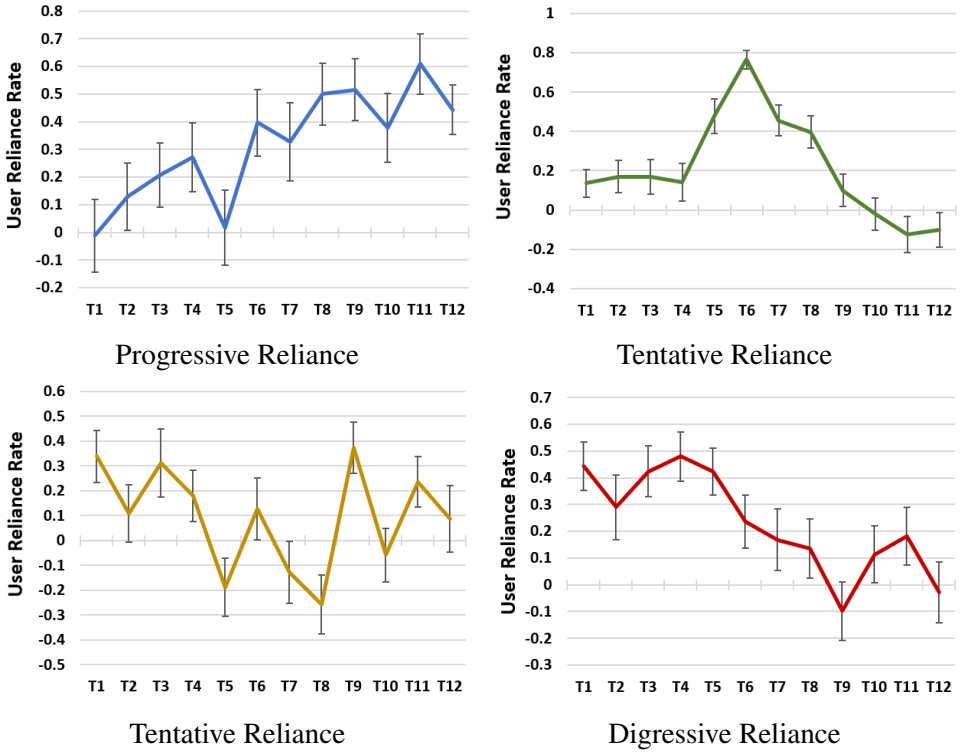


Figure 5.2: User reliance dynamics: four profiles of user reliance evolution in time in the range of -1.0 (complete independence) to 1.0 (complete dependence). Measurements are for all 12 news sharing instances and error bars represent standard error of the mean.

Similar to the previous section, I identified four profiles for trends of participant reliance on model predictions. I first summarized participants' news review progress into three average points of early (T1-T4), mid (T5-T8), end (T9-T12) segments. Then, using rule-based clustering I classified participants into four clusters of 1) progressive reliance with participants continually relying more on the AI, 2) tentative reliance with overshoots in reliance, 3) tentative reliance with an undershoot in reliance, and 4) digressive reliance where participant continually lost trust in the AI.

Figure 5.2 shows the clustering results for reliance measurements during the study (T1-T12). Overall, the cluster of participants with an overshoot in their reliance on AI assistant is the largest group with 42.3% of the total, only 16.5% gained trust continually, 21.1% lost trust continually, and 18.6% had trust undershoot during the task. A Pearson Chi-square test did not show any significant relationships between the explanation conditions and user reliance profile types.

Clustering participants with their trust evolution patterns shows model explanations effect on how user trust evolves over time. However, this effect was not observed for user reliance changes over time. Also, I did not detect dependency between the same clusters in reliance and trust measures. This could be because of the possible lag between user exposure and insights in complex interactive systems. This latency between users' interactions with the system and coming to their conclusions have also been reported in previous research [10].

5.4 Discussion

The following discussion presents findings drawn from the two trust measures in the experiments and recaps the main highlights. In the end, I review the limitations of this work and open questions to investigate.

5.4.1 Model Explanations Significantly Affect User Trust and Its Dynamics

The study results indicate that users working with the same intelligent system can perceive the system accuracy differently depending on how the model and its decision making is explained. In my experiments, the news keyword heatmap explanations (XAI-attention condition) significantly reduced user-perceived accuracy, as participants considered it an unreliable way of detecting fake

news. This finding is similar to Nourani et al.'s [53] finding that explanations that do not align with human rationale ("meaningless" explanations) reduced user trust. Following the related research on user trust in intelligent systems (e.g., [52, 292]), I conclude that AI transparency and machine learning explanations do not necessarily improve user trust, instead transparency empowers the user to build appropriate trust in the system.

I observed that explanations can shape how user trust is evolved by analyzing study measurements over time. Recurring measurements of user reliance revealed whether model explanations are persuasive (resulting in an increase of user overtrust) or implausible (resulting in a decrease of user trust) to the user. However, the findings suggest the dynamics of self-reported subjective performance measures were not aligned with the objective behavioral measures. This could be an indicator of possible lead or lag in reflections of trust between the two measurements of trust. This latency between users' exposure to the system, adapting their behavior, and coming to their conclusions have also been reported in previous research, see [10]. I conclude that the recurring measurements of user trust in complex systems (e.g., AI-based systems) is invaluable to understand the dynamics of user behavior and complement the limitations of self-report measurements.

5.4.2 User Expectations of AI Assistants

I looked into the relationship between participants' expectation of the AI (before the study) and their perceived detection accuracy of the AI at the end of the study to test for possible interactions. A Pearson test showed a small positive correlation ($r = 0.223$, $p = 0.005$) between participants' expectation and perceived accuracy rating. Comparison of this correlation among the four conditions shows the correlation is moderate ($r = 0.436$, $p = 0.006$) in the *XAI-attention* group.

5.4.3 User Trust and Reliance Changes Significantly Over Time

Analyzing profiles of user trust and reliance showed that users changed their thoughts and behavior about the intelligent agent over the study duration. To test for this pattern over the different study conditions, I divided the study duration into three early (T1-T4), mid (T5-T8), end (T9-T12)

segments to perform analysis on these changes. A one-way independent ANOVA test on reliance rate showed significant differences over time segments in the *AI* group ($F(2, 113) = 3.84$ and $p = 0.024$), the *XAI-attention* group ($F(2, 122) = 4.44$ and $p = 0.014$), and the *XAI-all* group ($F(2, 116) = 5.024$ and $p = 0.008$). The test did not detect statistically significant changes in participants' perceived accuracy in each condition; however, participants showed significant changes between their first and last trust measurements for the digressive trust cluster ($F(2, 98) = 3.634$ and $p = 0.030$), progressive trust cluster ($F(2, 110) = 5.154$ and $p = 0.007$). This is an indicator of participants' learning of AI limitations and strengths during their experience with the intelligent system. Therefore, it could be beneficial to take user learning phases into account for user study experiments of AI-based systems by allocating longer study duration proportional to the agent's degree of complexity.

5.4.4 User Reliance Variations Dampen Over Time

To understand the rate of change of reliance during user experience/interactions with the AI/XAI assistants, I looked into participants' reliance variance during the study. I observed a high magnitude zig-zag pattern of user reliance changes at the beginning of the study compared to lower variation towards the end of the study. I investigated variations in participants' reliance rate based on the standard deviation between the first and second half of the study. I observed less variance in user reliance in the second half of the study in all conditions, and a one-way independent ANOVA test found a marginally significant difference ($F(1, 77) = 4.00$ and $p = 0.049$) between the first and second half of the study for the *XAI-attribution*. Hence, based on the observations in our case study, I conclude that recurrent measurements can help to recognize users' learning phases and identify appropriate trust measurements time for reducing noise from user learning effects.

5.4.5 Trust Evolution Rate

Another finding in the study results was the difference between the rate of changes in participants' trust measurements. A correlation test between participants' final perceived accuracy and the absolute value of changes in participants' perceived accuracy showed a negative Pearson

correlation with $p = 0.003$. This negative correlation could indicate that participants with higher final perceived accuracy gained their trust in smaller steps (more skeptical) compared to participants with lower final perceived accuracy who lost trust in larger steps during the study. Also, participants' with tentative trust had larger steps sizes (mean = 15.5%) compared to participants with progressive or digressive trust (mean = 7.71%). The signal in the trust evolution rate could be further investigated as an opportunity to identify cautious users who gently build justified trust compared to users with more spontaneous swings in their feeling and perception of the system.

5.4.6 Effects of Early Impressions

Lastly, I examined the possible effects of participants' first impressions on the dynamics of their trust. I divided participants' into two groups of *positive first impressions* who were gaining trust in the system at the beginning of the study (i.e., first six news review instances), and *negative first impressions* who were losing trust in the system early on. I observed that participants' with *positive first impressions* were more likely (39.4% of total participants) to continue gaining trust until the end of their experience compared to participants' with a *negative first impression* were less likely (28.3% of total participants) to change their mind and gain trust. Similar to the effect of user expectations of the AI Assistant prior to study, this observation suggests the importance of users' first impression of the intelligent agent in their trust dynamics as participants' were more likely to keep their early perception of the system.

5.5 Findings and Conclusion

Overall, the study results showed the value of using recurring measurements in XAI system evaluation as suggested in Section 3.5.4. The dynamic measures of trust improves the XAI framework by introducing new techniques and considerations for XAI outcome evaluation (Layer 1) and interface (Layer 2). Also, dynamic measurements of user trust and other interactions with XAI systems motivate the potential design approaches such as the use of adaptive explanations to prevent users from overtrusting and undertrusting an intelligent agent.

Finally, I recognize a few limitations in our studies and analysis that could become more clear

in future work. Primarily, the fake news detection domain tackles a complicated problem. Though the presented study used the same curated list of news and articles for all conditions, it is not clear how participants' prior knowledge might have influenced the results. Second, I did not observe dependency or correlation between dynamics of user trust and user reliance over time, although the two measures showed positive correlation on their static measurements in Section 4.2.8.4. In addition to the latency between user learning and experience (reflected in the behavioral reliance measurements) and their coming to a conclusion about the system (reflected in the self-report trust measurements), other factors such as potential cognitive biases or users' lack of conscious awareness of behavior could have been affected the measurements which require further investigation.

6. HUMAN-ATTENTION BENCHMARK

6.1 Introduction

Recent and continuing advancements in model interpretability techniques unveiled new opportunities to enable human review of model reasoning and learning representations for their correctness in accordance to design goals, law and regulations, and safety requirements. Such evaluations could potentially prevent adverse outcomes of AI-based systems—such as unfair and discriminatory decision-making when performing real-world tasks. However, with the complexity of interpretability techniques and human cognitive biases, the question remains: how should we choose effective and efficient methods for the evaluation of machine learning explanations? Different approaches have been proposed for evaluating interpretable models and XAI systems at different stages of system design [112]. In machine learning research, various computational methods are used to measure the fidelity of interpretability techniques with respect to the underlying black-box model [110, 105]. On the other hand, in the field of human-computer interaction, human-grounded evaluation approaches measure human factors such as user satisfaction, mental model, and trust in XAI systems designed for different tasks. However, there are fundamental differences between these evaluation approaches. Computational methods set a precedent to objectively evaluate the model against a baseline ground truth, yet they lack the ability to quantify human interpretations. On the other hand, while more descriptive in nature, human subject studies tend to be more costly, imprecise, and subjective to the task. Another major difference between these evaluation methods is that once the human user is exposed to the evaluation study setup, she can not unlearn the experience for another round of evaluation. These differences raise the need to study the trade-off between objective ground-truth evaluation and subjective human-judgment of explanations.

Looking into the discussion on limitations of ground truths for model explanations in Section 3.5.3, I propose a human-attention baseline to quantitatively evaluate model saliency explanations. The proposed evaluation benchmark contributes to the computational methods for direct

evaluation of model explanations (Layer 3, Guideline 7, Section 3.4.4) in XAI evaluations steps. My publicly available human-grounded benchmark enables fast, replicable, and objective execution of evaluation experiments for saliency explanations. To foster the interest of the machine learning community, I demonstrate this benchmark's utility for quantitative evaluation of model explanations and compare it with the single-layer feature mask ground truth and human judgment rating evaluations. My study results reveal the efficiency of threshold-agnostic evaluation with a human-attention baseline as compared to previous methods with binary ground truth masks and labels. My experiments also reveal user biases in the subjective rating of saliency explanations.

6.2 Background

The evaluation of model explanations and interpretability techniques can be categorized in different ways [112, 3]. For instance, previous works have examined the fidelity of interpretability techniques to the black-box model [105, 110], evaluated correctness of model explanations with ground-truth [93], as well as the usefulness of explanations in different tasks and domains [54].

Following the review in Section 2.4.1, I review limitations in the two evaluation approaches, *human judgment evaluation* and *ground-truth evaluation*, for the trustworthiness of machine learning explanations and assess their advantages and limitations. Note that in this section, I focus on the trustworthiness of explanations with the assumption of having a high-fidelity ad-hoc explainer.

6.2.1 Objective Evaluation with Ground Truth

Ground truth baselines have been used as an objective way to quantify the correctness of model explanation is to examine it against. Ground truth is often annotated by users or synthesized in data to represent relevant features (i.e., a binary mask for features) and provide a baseline for quantitative evaluation of explanations quality. Quantitative similarity metrics like Intersection over Union (IoU) and mean Average Precision (mAP) are used to measure the model's saliency map explanations in comparison to the ground truth mask. However, the relationship between the evaluation of machine learning explanations and the auxiliary tasks, such as binary object

https://github.com/SinaMohseni/ML-Interpretability-Evaluation-Benchmark

localization and semantic segmentation, is not clear yet.

In a review of limitations in threshold-based evaluations for model saliency map, Choe et al. [293] present an evaluation protocol to include a hyperparameter search for the τ threshold for generating objects' "binary mask" from the saliency score map. However, unlike our proposed evaluation protocol, they do not consider the pixel-wise evaluation of saliency score maps in the first place. Apart from binary mask baselines that annotate entire features associated with the target class, perhaps closest work to our human attention benchmark is Das et al.'s [174] VQA-HAT baseline for evaluating saliency maps in visual question and answering models. They test multiple game-inspired, attention annotation methods to ask participants to sharpen regions of a blurred image to answer a question. The resulting baseline is a human attention map that enables object identification but does not indicate whether the necessary or sufficient features are annotated by individual participants.

6.2.2 Subjective Human Judgment

User review of model explanations for their subjective feedback is a common approach for evaluating machine learning explanations. Different papers have run user studies to evaluate the human understanding of saliency map explanations from DNNs as a proxy for explanations goodness and human interpretability of explanations. For example, Alqaraawi et al. [294] showed that instance explanations carry new information to users, but model behavior remained largely unpredictable for participants. In other work, Zhang et al. [232] compared saliency explanations from multiple networks with human explanations of objects in images. They performed a large crowdsourced study to directly compare machine learning and human explanations and human feedback on model explanations. Their results indicate that the features learned by some DNN models are more similar to human intuition. However, it is not clear from their study whether the model generalizability or the choice of interpretability technique was more effective on user satisfaction of explanations. To address the limitations in human judgment evaluation studies, Lertvittayakumjorn and Toni [237] defined a set of objective evaluation tasks for quantitative evaluation of model explanations with respect to different explanatory purposes. They used three human-grounded tasks

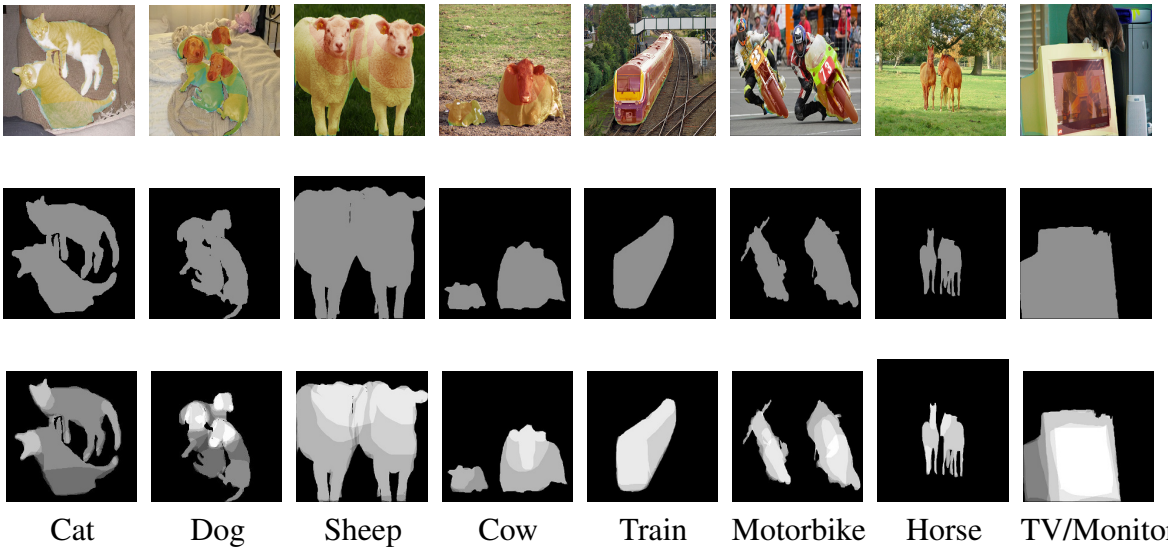


Figure 6.1: Examples of human annotations of salient features on images with the target class in the caption. **(Top)** Input images with human-attention mask heatmap overlay. **(Middle)** Single-layer object's segmentation mask for the target class. **(Bottom)** Resulting multi-layer human attention mask. Each image is annotated by 10 unique participants.

to evaluate local explanation methods for their ability to reveal model behavior, justify model predictions, and help users investigate uncertain predictions. The review of previous research indicates that the dissonance between machine learning models' goal to *learn discriminant features* and human expectation of *logical and common sense explanations* undermines the correctness and completeness of human judgment evaluation methods.

6.3 Human-Attention Benchmark

I captured the human annotation of salient features in order to create a human-grounded benchmark to evaluate model explanations. Participants were prompted to select relevant regions in images and phrases in text documents that they felt most representative of the target subject or topic, respectively. Figure 6.1 show examples from the resulting multi-layer ground truth from aggregating annotation from multiple unique annotators for each image. In comparison to the single-layer object's segmentation map, the human-attention benchmark allows for a higher level of granularity in the evaluation of saliency maps and reflects human attention to features. Also, compared to human judgment rating evaluations, the human attention benchmark enables reproducible and

cost-efficient evaluation. The following reviews the details of benchmark specification, annotation procedure, and data processing.

Table 6.1: Details of the evaluation benchmark for human-attention masks in different public datasets.

Domain	Image		Text	
Dataset	PASCAL VOC 2012	ILSVRC 2014	20 Newsgroup	IMDB 50K
Number of classes	20	20	2	2
Samples per class	50	5	100	100
Total annotation sample size	1000	100	200	200

6.3.1 Benchmark Specifications

The benchmark presents multi-layer masks representing what features humans expect to be the most important representations of a particular class. For each sample, I collect annotations from 10 unique annotators from Amazon Mechanical Turk platform that were instructed to select areas (in images) or words (in documents) that they deem most relevant to the target class. The multi-layer mask generated by aggregating annotations for each individual sample provides more granular representation of attributed features compared to the single-layer mask. Note that this method—collecting multiple user annotations for human-attention masks—balances the trade-off between objective annotation of precise feature-masks (i.e., segmentation mask) and subjective human judgment of the representative features. Also, it is important to mention that this human-attention baseline evaluates the explanations' correctness or trustworthiness of saliency explanations and does not intend to measure the fidelity of ad-hoc interpretability techniques to the black box models.

The development of this benchmark consists of a validation subset from *ImageNet* [295] and *PASCAL VOC2012* [90] image datasets and *20 Newsgroup* [296] and *IMDB* [297] text datasets.

Table 6.1 presents details for the number of classes and annotated samples from the four datasets in this explanation evaluation benchmark. For the PASCAL VOC dataset, 50 randomly selected samples from all 20 classes are annotated including Vehicles (airplane, bicycle, boat, boat, bus, car, motorbike, train), Households (bottle, chair, dining table, potted plant, sofa, TV/monitor), and Animals (bird, cat, cow, dog, horse, sheep) and other (person). To create a validation set from the ImageNet dataset, I randomly selected five images from 20 classes including living things (man, woman, cat, dog, bird, ant, elephant, shark, zebra, flower, tree), indoor objects (chair, computer, ball, book, phone), outdoor objects (car, ship, airplane, house). The set includes images covering broad considerations such as multi-object and complex scenes, co-occurrence of target object, target object in different scales, and lighting conditions.

For the text domain datasets, 100 randomly selected movie reviews from each positive and negative classes of *IMDB* dataset are selected. Similarly, 100 randomly selected text documents (with the headers removed from samples) from the *20 Newsgroup* dataset are selected from two categories of medical (*sci.med*) and electronic (*sci.elect*).

6.3.2 Annotation Interface and Procedure

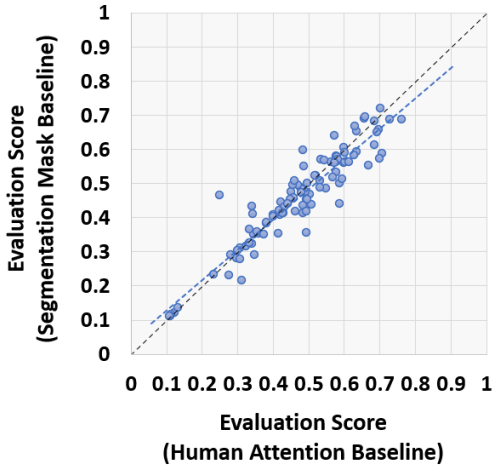
In order to generate multi-layer human-attention explanations, I ask annotators to provide their interpretations of the salient features that are most meaningful for the specific class from the data set. Each sample is annotated with 10 unique annotators recruited from Amazon Mechanical Turk (AMT). Recruitment advertisement for Human Intelligence Task (HIT) required participants to have at least 1000 previously approved HITs in AMT platform with the HIT approval rate of above 95%. Recruited participants were walked through a training slideshow of the task instructions and interface controls at the beginning of their HIT. As a control, each training slide was displayed on screen for two seconds before participants were able to continue to the next slide. Afterward, they were asked to agree to the IRB approved information sheet for data collection, and continued to a set of 12 images or documents for annotation. Participants were paid $0.40 for the image and text annotation HITs to reach an average hourly pay rate of $10 an hour.

I designed two fundamentally similar human annotation interfaces to capture human feedback

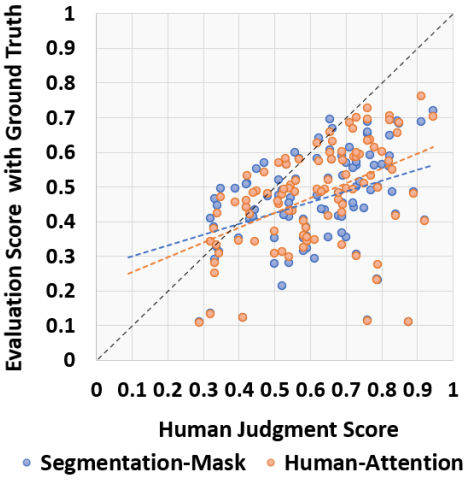
for all image and text datasets. Annotators were using an interface with basic annotation tools in which each document or image was presented individually. Each annotation HIT started with the same two samples to serve as attention check and help the annotator to get adjusted with the interface and task. These are then followed by 10 samples from the main validation set. Task instructions prompted participants to select relevant regions in images that they felt most representative to the target object that could be entire or parts of it but generally not the background scenery. For image annotations, the annotators were specifically asked to use their mouse to lasso "salient area(s) that explain target *"object"* in the image". Similarly, for text annotations, participants were prompted to select relevant words in text documents that they felt most representative of the target topic or class. For example, for the movie review IMDB dataset, the annotators were explicitly asked to "select words and phrases which explain the positive or negative sentiment of the movie review".

6.3.3 Data Processing and Storage

In order to generate multi-layer feature masks from multiple user annotations, I run a union operation on all individual annotation that displays what areas are most frequently selected by the annotators. Figure 6.1 presents examples of resulting human-attention masks from different images. Although specified in annotation task instructions, I also applied the exact segmentation mask of the target object's true pixels (only for image datasets) to remove the impact of participants' imprecision or hand jitter that might have included the background pixels. The exact segmentation masks for images are created by two authors and included in the benchmark. Human attention masks for image datasets are stored in the format of grayscale masks the same size as original images. The human attention masks for text datasets are JSON objects with lists of index-word pairs with human-attention scores in the range of 0 to 1.0. I did not perform any feature filtering for text annotation samples. The benchmark is stored in a public domain and free for research use.



(a) Relation between two ground truth measures



(b) Relation between subjective rating and baseline measures

Figure 6.2: Comparison of averaged evaluation scores (1.0− MAE) between two ground truth baselines and human judgment rating for each sample. Evaluation scores are not normalized and the black dashed lines shows the ideal regression line with the slope equal to 1.0 and intercept of zero. (a) Scatterplot of evaluation scores based on segmentation mask (vertical axis) and human-attention mask (horizontal axis). (b) Scatterplot of evaluation score based on two ground truth baselines and human judgment rating.

6.4 Evaluation of Saliency Explanations

In this section, I present multiple evaluation experiments to validate the proposed benchmark with empirical results. These experiments compare three baselines: 1) human-attention mask as the ground truth, 2) segmentation mask as the ground truth, and 3) human-judgment rating for evaluating model saliency explanations. My goal is to understand the relationship between the three evaluation methods and communicate the benefits of the proposed benchmark over other common evaluation methods in the literature. The series of experiments are based on saliency maps generated by the Grad-CAM [184] technique for a VGG-19 [298] image classifier on a subset of 100 validation samples from the two classes of cat and dog in PASCAL VOC dataset. The VGG network is pre-trained on ImageNet-1k and tuned on PASCAL VOC 2007 for the purpose of this evaluation. All evaluation scores are based on pixel-wise Mean Absolute Error (MAE) between

https://pytorch.org/docs/master/torchvision/models.html

model saliency score map and the ground truth baseline.

The saliency map error is calculated as the MAE between model saliency score map and the ground truth mask. I also looked into False Positive (FP) and False Negative (FN) saliency explanation errors individually. I calculate FP saliency error as pixel-wise MAE for the model saliency map scores outside the object's segmentation mask (i.e., error in background pixels) and FN error as the pixel-wise MAE for model saliency map scores inside the ground truth mask (i.e., error in target pixels). In the following subsections, I review details and share evaluation results from the three methods.

6.4.1 Comparison to Segmentation Mask

In the first evaluation experiment, I compare my proposed human-attention benchmark (multi-layer feature mask) with the segmentation mask (single-layer feature mask) as the evaluation ground truth for the set of saliency maps from Grad-CAM technique. Given the lack of granularity for distinguishing important features in the segmentation mask, I hypothesize that the two baselines would result in different evaluation scores for the same set of inputs.

Intuitively, the difference between the two baselines is that unlike the segmentation mask, which scores all target features equally, the human-attention mask gradiates the "salient" features more than others. To identify the difference between two evaluation baselines, I calculate evaluation scores using both baselines for direct comparison. Specifically, I first normalize both ground truth masks and model saliency maps and then calculate the pixel-wise MAE error between model saliency map and the ground truth baseline. For example, a saliency map identical to its human attention mask results in zero MAE error. In the opposite situation, with cases having no overlap between the ground truth mask and the model saliency map, the MAE error would be 1.0. Note that MAE is a threshold agnostic metric that—unlike Intersection over Union—does not require choosing the τ hyperparameter for generating objects' binary masks or bounding boxes, see [293] for more discussion. Also, evaluating the saliency score map (without converting to a binary mask) retains the granular information in the model explanation.

Results: Figure 6.2-(a) shows the scatter plot of evaluation scores ($1.0 - \text{MAE}$) between human-

attention and segmentation mask baselines. The two evaluation scores are statistically significantly ($r = 0.896$, $p < 0.001$) correlated, as expected. Using a linear regression test, I find a regression slope of $w = 0.896$ and intercept of $b = 0.48$. As seen in Figure 6.2-(a), this weight and bias result in different evaluation scores between the two ground truth, especially in the higher and lower range of scores. To examine the statistical significance of the difference between two ground truth evaluations, I use an ANCOVA test with a custom model to the test for homogeneity of regression slopes between the calculated regression model and the ideal of slope 1.0 with a zero intercept. The test for homogeneity of regression slopes fails with a significant difference ($p < 0.001$) between the two lines indicating that the two evaluation baselines are not equal. Next, I look into FP and FN saliency explanation errors individually. The results show that the difference between the two baselines is only due to FN errors being treated differently between the two baselines. This was expected since both baselines measure zero evaluation score for the saliency explanations outside the ground truth mask.

6.4.2 Comparison to Human Judgment

In the second evaluation experiment, I compare explanation evaluation scores using the two ground truth baselines with the human ratings of explanation goodness. Subjective human ratings of the model explanations are commonly used as a direct approach for evaluating machine learning explanations by providing a numerical rating of explanations goodness using a simple quantitative measure such Likert scales. However, subjective measures typically lack precision and may include user bias. I hypothesize that results from human-judgment scores will be significantly different for both (human-attention mask and object segmentation mask) ground truth evaluations. I use the same subset of images and saliency map explanations from Grad-CAM technique similar to the previous section for the purposes of this human-subjects study. Figure 6.3-(Top) shows examples of heatmap overlays from the Grad-CAM technique used in the user study.

I designed a simple interface to collect user feedback about the quality of heatmap overlays from the Grad-CAM saliency explanation technique. The participants started by reading task instructions followed by a series of images for review and rate. Given an image from the test set,

the target classification, and a heatmap overlay, participants were instructed to "review and rate the heatmaps which explain what parts the AI used to make it's classification decision" and were asked to rate the "goodness" of the AI decision on the scale of 1 to 10. A total of 200 unique participants' were recruited from Amazon Mechanical Turk and paid $0.20 per HIT to review and rate 14 images. The first four image ratings (identical images were used for all participants) were used as training and attention check examples; these were disregarded for data collection.

Figure 6.2-b shows a scatterplot of the evaluation scores ($1.0 - \text{MAE}$) between human judgment ratings and ground truth scores from objects' segmentation masks and human attention masks. The two regression lines for human-attention ground truth (in orange) and segmentation mask (in blue) show both baselines produce different evaluation scores from the user rating scores. To test for the statistical significance of observed differences, I first normalize user ratings across participants by subtracting each participant's mean rating. Then, I use a Pearson's correlation test and linear regression test to compare the human judgment rating scores and the two ground truth scores. The user ratings show a moderate-strength correlation with object segmentation mask baseline ($r = -0.121$, $p = 0.002$) and small correlation with human-attention mask baseline ($r = -0.306$, $p < 0.001$). I also observe signs of user bias, noting that none of the participants rated any of the saliency map instances in the test set below 3-stars even though there are multiple examples with scores below 0.3 for both ground truth evaluation types. These cases were specifically from the examples with multiple occurrences of the target object in which the saliency map was only pointing to one of the target objects. This could potentially indicate a side effect of lower user attention in reviewing cases with incomplete saliency explanations.

To compare measurements between evaluation approaches, I run a linear regression analysis and find that the segmentation mask scores fit with a slope of $w = 0.313$ and intercept of $b = 0.268$ (Figure 6.2-b, blue trend line), and the fit for human-attention mask scores has a slope of $w = 0.428$ and intercept of $b = 0.210$ (Figure 6.2-b, orange trend line). Note that the difference between the two linear regression models' slopes with the ideal slope of 1.0 is higher with the segmentation-mask baseline. To examine the statistical significance difference between the measures, I use

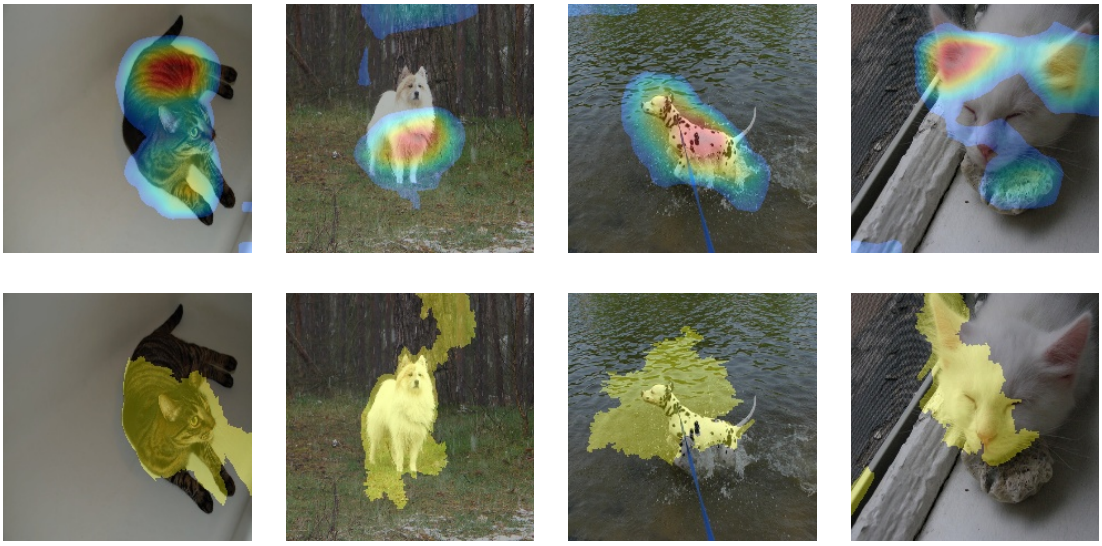


Figure 6.3: Examples of heat-map overlay of saliency maps using the Grad-cam **(Top)** and LIME **(Bottom)**.

ANCOVA with a custom model to test for homogeneity of the regression slopes between the two regression models as well as between the calculated regression model and the ideal of slope 1.0 with zero bias. The homogeneity test fails with a significant difference of $p < 0.001$ between the two regression models and the ideal line. The analysis indicates the subjective measurement of explanations goodness produces significantly different results from both objective ground truth measures.

6.5 Discussion

In this section, I review and discuss the evaluation experiments and open problems around model explanation evaluation. The evaluation experiment results showed that the human-attention benchmark has allowed for a higher level of granularity in the evaluation of saliency maps and reflected human attention to the features in comparison to the single-layer object's segmentation map. As compared to the human judgment rating evaluations, I observed signs of participants' bias in their ratings.

6.5.1 Implications of Results

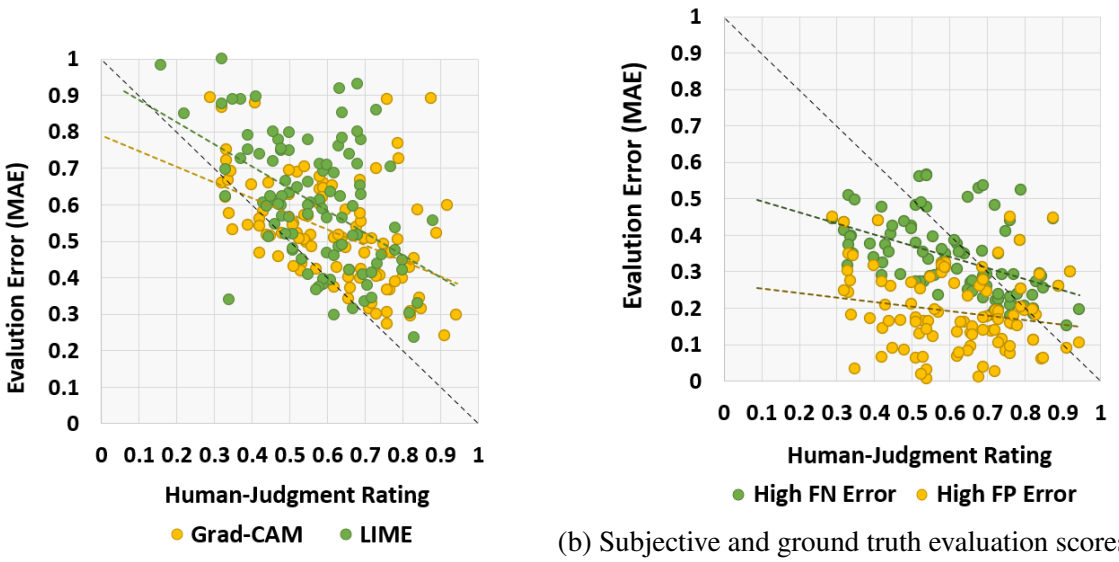
I ran human-subject experiments to understand the differences between the subjective and objective evaluation of saliency explanations. Although the evaluation results from the three methods had positive correlations, the experimental results showed significant differences among all evaluation scores. The difference in scores was mainly due to the clear non-uniform distribution of weights in human attention masks while the segmentation mask weights are uniformly distributed for all features (e.g., pixels, words).

While segmentation mask benchmarks are mainly used for object segmentation evaluation and weakly supervised object localization [92, 293], the human-attention baseline reflects human factors in feature attribution. For example, in annotations of living things, users were more likely to select facial features as important features while the segmentation mask offers a uniformly weighted single-layer mask. This is reflected in the evaluation results with human judgment with participants' ratings of explanations being closer to the human-attention baseline rather than the segmentation mask baseline. Due to the same effect, evaluation results with the human-attention baseline could be extended to better anticipate user acceptance and trust in model explanations when putting on different applications.

6.5.2 User Biases in Rating

I explored the human judgment evaluation results to find other possible external or internal factors that could affect participants' subjective ratings. For example, human judgment ratings may include user biases toward visual appearance or completeness of saliency maps resulting in incorrect ratings. I reviewed and compared the results from human judgment for Grad-CAM and LIME explanations to identify possible biases. Also, I reviewed the results to assess possible participants' biases toward model explanation FP and FN error types.

To evaluate the effect of visual appearance of saliency explanations, I compared participants' rating of saliency map explanations from LIME [86] technique to Grad-CAM explanations on the same subset of images and the same classifier. The saliency explanations from the LIME technique



(a) Subjective and ground truth evaluation scores from LIME and Grad-CAM Explanations.

(b) Subjective and ground truth evaluation scores for samples with high FP and High FN explanation MAE error.

Figure 6.4: Discrepancies between averaged human judgment rating of saliency explanations and human-attention baseline evaluation. Evaluation scores are not normalized and the black dashed lines shows the ideal regression line with the slope equal to -1.0 and intercept of zero. **(a)** Participants evaluate saliency explanations from LIME and Grad-CAM differently. **(b)** Participants evaluate saliency explanations' FP error (model looking at background pixels) differently than FN errors (model not looking at target pixels).

(Figure 6.3-(Bottom)) are visually more chunky and pixelated (mainly due to use of super pixels in LIME's pipeline) compared to smooth concept activation maps from Grad-CAM technique (Figure 6.3-(Top)). I analyze results after running a new user study to collect participants' subjective ratings of LIME explanations.

I used two linear regression models to compare participants' ratings between the two groups, see 6.4-(a). I find the slope of $w = -0.428$ and intercept of $b = 0.789$ for the user ratings on samples with LIME saliency map (Figure 6.4-(a) green trend line) and slope of $w = -0.607$ and intercept of $b = 0.947$ for samples with Grad-CAM saliency map (Figure 6.4-(a), yellow trend line). I would have expected to see the similar regression slopes between the two groups if the users were evaluating both saliency map explanation types similarly. However, the test for homogeneity between the two regression slopes shows a significant difference ($p < 0.001$) between the two model error types. This indicates that users rated the saliency maps differently, although ground

truth evaluation score (Figure 6.4-(a), y axis) for both sets of samples.

I then analyze participants' rating behavior with respect to different explanation error types. I first divided the samples for the test set into two groups with high FP (when the model is looking at background pixels) explanation error and high FN explanation errors (when the model is missing foreground pixels). Using linear regression models, I find the slope of $w = -0.121$ and intercept of $b = 0.265$ for the samples with FP explanation error score (Figure 6.4-(b) yellow trend line) and slope of $w = -0.306$ and intercept of $b = 0.525$ for samples with high FN explanations error score (Figure 6.4-(b), green trend line). I would have expected to see the similar regression slopes between the two groups if the users were evaluating both saliency error types similarly. However, the test for homogeneity between the two regression slopes shows a significant difference ($p < 0.001$) between the two explanation error types. This indicates that users pay less attention to FP explanation errors and in turn, are more critical for FN explanation errors. Looking at image samples from the user study, these images included several examples in which the target object was on a smaller scale and the model saliency map was largely exceeded to the background pixels.

6.5.3 Reproducibility and Objectivity Trade-off

One way to categorize different evaluation measures is by their objectivity and reproducibility of results. As implemented in this paper, users' subjective rating of explanations could collect results for correctness and goodness of model generated explanations. Ribeiro et al. [86] presented a case for correction of model explanation in which users reject wrong features and add new features for quantitative evaluation of model explanations. A different method is to collect user feedback through the direct comparison of explanations from multiple interpretability techniques. For example, users could review several options to choose the best machine-generated explanation and provide justifications for their choices. However, although these methods can provide detailed insights, subjective user feedback is not reusable for new models and interpretability techniques. This limitation indeed exists in studies for evaluating XAI systems in different applications and domains [112], including tasks and scenarios concerned with the fairness of the decision-making system.

On the other hand, objective evaluation that utilizes ground truth, provides quantitative and reproducible results, yet lacks the guidance of human correctness and goodness scores that show which improvements would be most significant. My benchmark bridge the trade-off between objectivity and subjectivity of a baseline to satisfy both evaluation aspects.

6.6 Findings and Conclusion

I proposed a human-attention baseline for direct evaluation of machine learning saliency explanations. Based on the human-attention baseline, I created evaluation benchmarks for four public datasets that can significantly reduce evaluation time and costs over design cycles. The proposed baseline and four benchmarks contribute to the XAI framework by improving computational methods for direct evaluation of model explanations (Layer 3, Guideline 7, Section 3.4.4).

In a series of experiments, I compared the (1) proposed baseline with (2) binary feature mask baseline evaluation and (3) subjective user rating evaluations. Although the evaluation results from these three methods had positive correlations, the experimental results showed significant differences among all evaluation scores. The comparison between the binary feature mask and human-attention mask revealed that the human-attention baseline can better reflect the human factors in feature attribution whereas binary feature mask offers a uniformly weighted single-layer mask. Additionally, I observed that human judgment ratings may include user biases toward visual appearance and completeness of saliency maps resulting in incorrect ratings. For example, in the experiments, I identified significantly different user rating between human judgment for Grad-CAM and LIME explanations. Similarly, participants' had biases toward different errors types in model saliency explanations which under mines the evaluation quality as compared to baseline. Results suggest human-attention baseline for XAI evaluation can be a better substitute for user subjective rating specially in safety and fairness critical domains and applications such as medical, law and autonomous systems.

7. DISCUSSION AND CONCLUSION

This dissertation explored problems and proposed solutions for building effective XAI systems to improve human-AI interactions. In this section, I review a summary of my contributions and a discussion on existing open problems worth exploring in future work.

7.1 Summary

I organized my research contributions in the four following parts:

C1: A Design and Evaluation Framework for Explainable AI Systems. I reviewed over 200 XAI-related research papers to organize different XAI design goals and evaluation measures across disciplines. Table 3.1 presents a list of selected papers and my categorization of XAI design and evaluation methods that organizes literature along two main dimensions of: *design goals* and *evaluation methods*, and an auxiliary dimension of *targeted users* for the XAI system. From my review, I provide summarized ready-to-use tables of evaluation methods and recommendations for different goals in XAI research. This categorization revealed the necessity of an interdisciplinary effort for designing and evaluating XAI systems. Additionally, I want to draw attention to the resources in the social sciences field that can facilitate the extent of social and cognitive aspects of explanations.

As a product of my review, I proposed a design and evaluation framework that connects design goals and evaluation methods for end-to-end XAI systems design in multidisciplinary teams. The proposed framework provides step-by-step design guidelines paired with evaluation methods to close the iterative design and evaluation loops in system design. I hope my framework drives further discussion about the interplay between different design goals and evaluation outcomes in XAI systems. Although the presented framework is intended to give guidance on what evaluation measures are appropriate to use at which design stage to build XAI systems, it is not meant to offer all detailed aspects of interface and interaction design and development of interpretable machine learning techniques. Therefore, the framework is to benefit from other design guidelines for additional details and considerations in each design step.

C2: Case Study and Examples for XAI Framework. I presented a case study to demonstrate the generative function of the framework during a system design process. In our case study, a multi-disciplinary team of researchers with machine learning and HCI backgrounds collaborate on the design and development of a XAI system for fake news detection. The case study presents a practical example of how-to-use of my framework for XAI system design. I reviewed system design steps and different between-layer and within-layer framework guidelines that were used in this case study. In the end, I presented a thorough review and analysis of system evaluation results. Results showed that users' interaction with the AI and XAI assistants affected their performance, mental model, and trust. However, model explanations in our studies did not improve task performance or increase user trust and mental model. Instead, explanations helped users' to build an appropriate mental model of intelligent assistants and adjust their trust accordingly, given the limitations of the models.

In addition to the XAI system design case study, I analyze two existing XAI systems to demonstrate the framework's descriptive functionality to describe their design process workflow (between-layers) and design and evaluation choices (within each layer). Both analyses are aiming to find insights from their work and intended to suggest future design iterations.

In conclusion, my case study and examples revealed multiple challenges and open problems in designing effective XAI systems. The challenges for XAI designers like aligning design goals for machine learning algorithms and users interactions components. The open problems like the dissonance between the AI reasoning and human sense-making. Additionally, the main case study led to the identification of two limitations in existing XAI evaluation methods which addressed them in this dissertation. The new evaluation methods (contributions C3 and C4) also contribute to the XAI framework by improving evaluation of XAI systems during system design cycles.

C3: User Trust Dynamics in Explainable AI. The first contribution to improve the proposed XAI framework is a study to demonstrate the importance of dynamics of user behavior with XAI systems. This study contributes to the XAI framework by introducing important aspects of Human-XAI interactions and the value of recurrent user behavior measurements in XAI systems. My study

showed that users' trust and reliance on complex XAI systems change overtime and studying dynamics of user behavior is essential for accurate evaluation to improve the XAI system during design cycles. Specifically, I analyzed the effects of interpretability on dynamics of user behavior and trust over time in a human-XAI collaborative setup in the fake news detection case study.

My study results indicate that users working with the same intelligent system can perceive the system accuracy differently depending on how the model and its decision making is explained. Also, the study results show model explanations effect on user trust level as well as how it morphs over time by analyzing study measurements over time. Recurring measurements of user reliance revealed whether model explanations are persuasive (resulting in an increase of user overtrust) or implausible (resulting in a decrease of user trust) to the user. However, my findings suggest the dynamics of self-reported subjective performance measures were not aligned with the objective behavioral measures. This could be an indicator of possible lead or lag in reflections of trust between my two measurements of trust. This latency between users' exposure to the system, adapting their behavior, and coming to their conclusions have also been reported in previous research [10]. I conclude that the recurring measurements of user trust in complex systems (e.g., AI-based systems) is invaluable to understand the dynamics of user behavior and complement the limitations of self-report measurements.

C4: A Human-Attention Benchmark. My final contribution to the XAI framework is a human-attention baseline for quantitative evaluation of model saliency explanations. This human-attention baseline contributes to the inner-layer of the XAI framework by proposing a new baseline for direct evaluation of machine learning saliency explanations. My publicly available human-attention benchmark enables fast, replicable, and objective execution of evaluation experiments for saliency explanations. The human-attention evaluation benchmark covers a subset of four major public datasets in image and text domains. This human-grounded benchmark resolves the main limitations of user review and feedback in controlled user studies such as study costs, imprecision, and subjectivity to the task.

`https://github.com/SinaMohseni/ML-Interpretability-Evaluation-Benchmark`

I demonstrated the benchmark's utility for quantitative evaluation of model explanations to foster the interest of the machine learning community. In a series of experiments, I studied the relationships and trade-offs between my benchmark and two common evaluation approaches: (1) binary annotation mask and (2) human subjective review and feedback. The study results indicated the significant difference between evaluation with a human-attention baseline as compared to two previous methods. My experiments also revealed user biases in their subjective rating when exposed to different visual appearance and error types of saliency explanations. I conclude that human-attention baseline is the most accurate ground-truth for direct evaluation (i.e. feature-level) of model saliency explanations when compared to binary segmentation mask and human subjective review.

7.2 Open Problems

In the following I am reviewing research limitations and future opportunities to extend my research.

L1: Limitations and opportunities in the Framework. My framework provides a basis for XAI system design in interdisciplinary teamwork and I presented a case study and two examples to demonstrate, validate, and improve the framework. However, no framework is perfect or entirely comprehensive. I acknowledge that the validity and usefulness of a framework are to be proven in practice with further case studies from the community. The presented case study and examples serve as a practical examples of using my framework in a multidisciplinary collaborative XAI design and development effort. The lessons learned and pitfalls in our end-to-end implementation case study are incorporated in the framework guidelines as well as added to this dissertation through contributions C3 (studying dynamic of user trust) and C4 (proposing a human-attention benchmark). Without doubt, future work is needed to examine practicality and usefulness of this framework in various domains and setups.

Moreover, this framework has a common limitation of many multidisciplinary design frameworks of being light on specific details at each step. Rather than contributing detailed guidelines for each framework layer, the framework is intended to pave the path for efficient collaboration

among and within different teams, which is essential for XAI system design given the inherently interdisciplinary nature of the area. The diversity of design goals and evaluation methods at each layer can help maintain the balance of attention from the design team to different aspects of XAI system. This higher level of freedom allows for extendability with other design guidelines (see the discussion in Section 3.5.5) to integrate with more tailored approaches for specific domains. Therefor, I identify a possible direction to continue this framework as to be adopting and merging other human-AI interaction guidelines with it so to achieve more detailed and tailored design framework.

L2: Limitations and opportunities in the Case Study. In our case study research, we designed and implemented model explanations from multiple models as part of an ensemble approach for fake news detection. The case study served well to demonstrate how to use the XAI framework and validate its guidelines. Plus, this approach allowed us to study how different types of explanations affect users in fake news detection and analysis of study results showed the value of using recurring measurements in XAI system evaluation. However, I recognize a few limitations in our studies and analysis that could become more clear in future work. Firstly, the fake news detection domain tackles a complicated problem. Though the presented study used the same curated list of news and articles for all conditions, it is not clear how participants' prior knowledge might have influenced the results.

Second, for the analysis of user trust dynamics, I did not observe dependency or correlation between dynamics of user trust and their reliance over time. In addition to the latency between user learning and experience (reflected in the behavioral reliance measurements) and their coming to a conclusion about the system (reflected in the self-report trust measurements), other factors such as potential cognitive biases or users' lack of conscious awareness of behavior could have been affected the measurements which require further investigation.

Considering the complicated nature of the news review and fake news detection task, future research is needed to assess the effectiveness of other types of explanations, such as knowledge graphs and multi-modal evidence retrieval on users with XAI fake news detection assistants. Fur-

thermore, study results from dynamic measurement of user behavior motivated potential design approaches such as adaptive, interactive, and personalized explanations to prevent users from overtrusting and undertrusting in the intelligent agent. Future work is needed to continue studying personalized explanations as a trust calibration mechanism to help users in building a justified trust relative to the system performance.

L3: Limitations and opportunities in the Human-Attention Benchmark. The proposed human-attention benchmark allows for direct evaluation of model saliency explanations and contributed to the inner-level of the framework as a computational evaluation technique. Evaluation experiments showed the proposed benchmark is more efficient and accurate baseline compared to the binary baseline and subjective human rating. A limitation of creating this human-attention benchmark is the annotation cost for multi-level human explanation masks. However, annotation cost for an open-sourced benchmark could be justified when compared to repeated novel rounds of evaluations for subjective human judgments. As typically, the iterative design and evaluation of machine learning based systems require multiple rounds of training and test. My human attention benchmark can significantly reduce evaluation costs over design cycles.

In my future work, I plan to study annotators' opinion when annotating objects in different size and pose to learn general patterns in human attention. This could potentially help to optimize the number of annotators for each sample.

Another direction for this benchmark would be extend it to domains with fairness, safety, or legality concerns to assessing model trustworthiness with expert-grounded baseline. Specifically, creating a subset of edge case instances or scenarios with expert annotated explanations to assure that the model is learned features and patterns that are aligned with design specifications.

Lastly, I am interested in examining the use case of the human-attention benchmark for tuning models to improve prediction rationale and its effects on explanation quality.

7.3 Conclusions

This dissertation proposed a XAI design and evaluation framework that provides step-by-step design guidelines paired with evaluation methods for end-to-end XAI system design in multidis-

ciplinary teams. This framework was the product of an in-depth literature review and analysis to organize diverse XAI design goals and evaluation measures across three disciplines of machine learning, human-computer interactions, and visualization. I presented a case study to demonstrate the generative function of the framework with a practical example for XAI system design. I reviewed system design steps and different between-layer and within-layer framework guidelines that were used in this case study. Two additional analysis of existing XAI systems were presented to demonstrate the framework's descriptive functionality for analyzing their design process workflow (between-layers decision points) and design and evaluation choices (within-layer decision points).

The main case study led to the identification of two limitations in existing XAI evaluation methods which I addressed them in two follow up studies. The new evaluations for (C3) studying dynamics of user trust and (C4) direct evaluation of saliency explanations contributed to the XAI framework by improving the toolbox of evaluation methods of XAI systems in system design cycles.

My case study revealed multiple challenges and open problems in designing effective XAI systems. The challenges arising from multidisciplinary nature of XAI systems (e.g., aligning design goals for different XAI system components) and inherent limitations of machine learning algorithms (e.g., the dissonance between the AI reasoning and human sense-making). I hope my framework drives further discussion about the interplay between different design goals and evaluation outcomes in XAI systems. However, this framework is not intended to offer detailed guidelines for interface and interaction design or development of interpretable machine learning techniques. Hence, it can benefit from adapting design guidelines from related AI-based system design frameworks for additional details and considerations.

The proposed framework will have potentially impacts in designing transparency demanding applications of AI, such as domains concerned with ethical, legal, and safety aspects of intelligent systems. This framework will primarily translate and connect design goals among teams with multidisciplinary focus where complex XAI system outcomes like morality, fairness, and legality

are expected to be delivered. Also, my additions to the framework have promising future impacts on XAI design for critical domains. On one hand, algorithm development and testing using expert-grounded benchmarks (C4) for direct evaluation of model trustworthiness are valuable to improve model evaluation cycles. On the other hand, studying users' understanding of explanations and their trust dynamic (C3) — especially when experiencing persuasive or deceptive explanations — in critical domains can lead to the creation of new design considerations and user trust calibration mechanisms in XAI systems. Future work is needed to examine practicality and usefulness of this framework in such domains and setups.

REFERENCES

[1] M. Nourani, C. Roy, T. Rahman, E. D. Ragan, N. Ruozzi, and V. Gogate, "Don't explain without verifying veracity: An evaluation of explainable ai with video activity recognition," *arXiv preprint arXiv:2005.02335*, 2020.

[2] M. Hamidi-Haines, Z. Qi, A. Fern, F. Li, and P. Tadepalli, "Interactive naming for explaining deep neural networks: A formative study," *arXiv preprint arXiv:1812.07150*, 2018.

[3] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable ai systems," *arXiv preprint arXiv:1811.11839*, 2019.

[4] S. C. Woolley, "Automating power: Social bot interference in global politics," *First Monday*, vol. 21, no. 4, 2016.

[5] C. O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

[6] D. Gunning, "Explainable artificial intelligence (xai)," *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2017.

[7] B. Poulin, R. Eisner, D. Szafron, P. Lu, R. Greiner, D. S. Wishart, A. Fyshe, B. Pearcy, C. MacDonell, and J. Anvik, "Visual explanation of evidence with additive classifiers," in *Proceedings Of The National Conference On Artificial Intelligence*, vol. 21, p. 1822, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[8] S. Gregor and I. Benbasat, "Explanations from intelligent systems: Theoretical foundations and implications for practice," *MIS Quarterly*, pp. 497–530, 1999.

[9] B. Goodman and S. Flaxman, "Eu regulations on algorithmic decision-making and a "right to explanation". arxiv preprint," *arXiv preprint arXiv:1606.08813*, 2016.

[10] S. Carpendale, "Evaluating information visualizations," in *Information Visualization*, pp. 19–45, Springer, 2008.

[11] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.

[12] A. Datta, M. C. Tschantz, and A. Datta, "Automated experiments on ad privacy settings," *Proceedings on Privacy Enhancing Technologies*, vol. 2015, no. 1, pp. 92–112, 2015.

[13] L. Sweeney, "Discrimination in online ad delivery," *Queue*, vol. 11, no. 3, p. 10, 2013.

[14] E. Bozdag and J. van den Hoven, "Breaking the filter bubble: democracy and design," *Ethics and Information Technology*, vol. 17, no. 4, pp. 249–265, 2015.

[15] J. Heer, "Agency plus automation: Designing artificial intelligence into interactive systems," *Proceedings of the National Academy of Sciences*, vol. 116, no. 6, pp. 1844–1850, 2019.

[16] V. Bellotti and K. Edwards, "Intelligibility and accountability: human considerations in context-aware systems," *Human–Computer Interaction*, vol. 16, no. 2-4, pp. 193–212, 2001.

[17] T. Zarsky, "The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making," *Science, Technology, & Human Values*, vol. 41, no. 1, pp. 118–132, 2016.

[18] M. Ananny and K. Crawford, "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability," *New Media & Society*, vol. 20, no. 3, pp. 973–989, 2018.

[19] N. Diakopoulos, "Algorithmic-accountability: the investigation of black boxes," *Tow Center for Digital Journalism*, 2014.

[20] B. Y. Lim and A. K. Dey, "Assessing demand for intelligibility in context-aware applications," in *Proceedings of the 11th international conference on Ubiquitous computing*, pp. 195–204, ACM, 2009.

[21] N. Diakopoulos, "Enabling accountability of algorithmic media: Transparency as a constructive and critical lens," in *Transparent Data Mining for Big and Small Data*, pp. 25–43, Springer, 2017.

[22] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, S. Schieber, J. Waldo, D. Weinberger, and A. Wood, "Accountability of ai under the law: The role of explanation," *arXiv preprint arXiv:1711.01134*, 2017.

[23] B. Mittelstadt, "Automation, algorithms, and politics| auditing for transparency in content personalization systems," *International Journal of Communication*, vol. 10, p. 12, 2016.

[24] M. Turilli and L. Floridi, "The ethics of information transparency," *Ethics and Information Technology*, vol. 11, no. 2, pp. 105–112, 2009.

[25] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort, "Auditing algorithms: Research methods for detecting discrimination on internet platforms," pp. 1–23, 2014.

[26] M. Eslami, A. Rickman, K. Vaccaro, A. Aleyasen, A. Vuong, K. Karahalios, K. Hamilton, and C. Sandvig, "I always assumed that i wasn't really that close to [her]: Reasoning about invisible algorithms in news feeds," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 153–162, ACM, 2015.

[27] A. Hannak, P. Sapiezynski, A. Molavi Kakhki, B. Krishnamurthy, D. Lazer, A. Mislove, and C. Wilson, "Measuring personalization of web search," in *Proceedings of the 22nd International Conference on World Wide Web*, pp. 527–538, ACM, 2013.

[28] M. Eslami, K. Vaccaro, K. Karahalios, and K. Hamilton, ""' be careful; things can be worse than they appear": Understanding biased algorithms and users' behavior around them in rating platforms.," in *ICWSM*, pp. 62–71, 2017.

[29] J. Tang, H. Gao, H. Liu, and A. Das Sarma, "etrust: Understanding trust evolution in an online world," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 253–261, ACM, 2012.

[30] N. Tintarev and J. Masthoff, "Designing and evaluating explanations for recommender systems," in *Recommender Systems Handbook*, pp. 479–510, Springer, 2011.

[31] M. K. Lee, D. Kusbit, E. Metsky, and L. Dabbish, "Working with machines: The impact of algorithmic and data-driven management on human workers," in *Proceedings of the 33rd*

Annual ACM Conference on Human Factors in Computing Systems, pp. 1603–1612, ACM, 2015.

[32] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, "'it's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 377, ACM, 2018.

[33] E. Rader, K. Cotter, and J. Cho, "Explanations as mechanisms for supporting algorithmic transparency," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 103, ACM, 2018.

[34] S. Stumpf, V. Rajaram, L. Li, W.-K. Wong, M. Burnett, T. Dietterich, E. Sullivan, and J. Herlocker, "Interacting meaningfully with machine learning systems: Three experiments," *International Journal of Human-Computer Studies*, vol. 67, no. 8, pp. 639–662, 2009.

[35] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[36] D. S. Weld and G. Bansal, "The challenge of crafting intelligible intelligence," *Commun. ACM*, vol. 62, p. 70–79, May 2019.

[37] M.-A. Clinciu and H. Hastie, "A survey of explainable ai terminology," in *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, pp. 8–13, 2019.

[38] R. Tomsett, D. Braines, D. Harborne, A. Preece, and S. Chakraborty, "Interpretable to whom? a role-based model for analyzing interpretable machine learning systems," *arXiv preprint arXiv:1806.07552*, 2018.

[39] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 582, ACM, 2018.

[40] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[41] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016.

[42] S. Mohseni, M. Pitale, V. Singh, and Z. Wang, "Practical solutions for machine learning safety in autonomous vehicles," in *The AAAI Workshop on Artificial Intelligence Safety (Safe AI)*, 2020.

[43] Z. C. Lipton, "The mythos of model interpretability," *arXiv preprint arXiv:1606.03490*, 2016.

[44] C. Molnar, *Interpretable machine learning*. Lulu. com, 2019.

[45] M. Eiband, H. Schneider, M. Bilandzic, J. Fazekas-Con, M. Haug, and H. Hussmann, "Bringing transparency design into practice," in *23rd International Conference on Intelligent User Interfaces*, IUI '18, (New York, NY, USA), pp. 211–223, ACM, 2018.

[46] B. Lim, "Improving understanding, trust, and control with intelligibility in context-aware applications," *Human-Computer Interaction*, 2011.

[47] R. R. Hoffman, M. Johnson, J. M. Bradshaw, and A. Underbrink, "Trust in automation," *IEEE Intelligent Systems*, vol. 28, no. 1, pp. 84–88, 2013.

[48] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.

[49] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing theory-driven user-centric explainable ai," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, (New York, NY, USA), pp. 601:1–601:15, ACM, 2019.

[50] B. Lepri, N. Oliver, E. Letouzé, A. Pentland, and P. Vinck, "Fair, transparent, and accountable algorithmic decision-making processes," *Philosophy & Technology*, pp. 1–17, 2017.

[51] B. D. Horne, D. Nevo, J. O'Donovan, J.-H. Cho, and S. Adalı, "Rating reliability and bias in news articles: Does ai assistance help everyone?," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, pp. 247–256, 2019.

[52] M. Yin, J. Wortman Vaughan, and H. Wallach, "Understanding the effect of accuracy on trust in machine learning models," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019.

[53] M. Nourani, S. Kabir, S. Mohseni, and E. D. Ragan, "The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, pp. 97–105, 2019.

[54] R. Kocielnik, S. Amershi, and P. N. Bennett, "Will you accept an imperfect ai?: Exploring designs for adjusting end-user expectations of ai systems," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, p. 411, ACM, 2019.

[55] X. J. Yang, V. V. Unhelkar, K. Li, and J. A. Shah, "Evaluating effects of user experience and system transparency on trust in automation," in *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pp. 408–416, IEEE, 2017.

[56] M. Eiband, D. Buschek, A. Kremer, and H. Hussmann, "The impact of placebic explanations on trust in intelligent systems," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, p. LBW0243, ACM, 2019.

[57] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: Comparing automatically generated explanations," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 109–116, IEEE, 2016.

[58] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," *arXiv preprint arXiv:1802.07810*, 2018.

[59] A. Papenmeier, G. Englebienne, and C. Seifert, "How model accuracy and explanation fidelity influence user trust," *IJCAI Workshop on Explainable Artificial Intelligence (XAI) 2019*, 2019.

[60] G. Bansal, B. Nushi, E. Kamar, D. S. Weld, W. S. Lasecki, and E. Horvitz, "Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 2429–2437, 2019.

[61] A. Ray, Y. Yao, R. Kumar, A. Divakaran, and G. Burachas, "Can you explain that? lucid explanations help human-ai collaborative image retrieval," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, pp. 153–161, 2019.

[62] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz, "Beyond accuracy: The role of mental models in human-ai team performance," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, pp. 2–11, 2019.

[63] V. Lai and C. Tan, "On human predictions with explanations and predictions of machine learning models: A case study on deception detection," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 29–38, 2019.

[64] J. R. Goodall, E. D. Ragan, C. A. Steed, J. W. Reed, G. D. Richardson, K. M. Huffer, R. A. Bridges, and J. A. Laska, "Situ: Identifying and explaining suspicious behavior in networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 204–214, 2018.

[65] D. M. Best, A. Endert, and D. Kidwell, "7 key challenges for visualization in cyber network defense," in *Proceedings of the Eleventh Workshop on Visualization for Cyber Security*, pp. 33–40, ACM, 2014.

[66] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings*

of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1721–1730, ACM, 2015.

[67] J. Krause, A. Perer, and K. Ng, "Interacting with predictions: Visual inspection of black-box machine learning models," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5686–5697, ACM, 2016.

[68] S. Liu, X. Wang, J. Chen, J. Zhu, and B. Guo, "Topicpanorama: A full picture of relevant topics," in *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pp. 183–192, IEEE, 2014.

[69] M. Liu, S. Liu, X. Zhu, Q. Liao, F. Wei, and S. Pan, "An uncertainty-aware approach for exploratory microblog retrieval," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 250–259, 2016.

[70] C. Robinson, F. Hohman, and B. Dilkina, "A deep learning approach for population estimation from satellite imagery," in *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, pp. 47–54, ACM, 2017.

[71] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim, "The role of uncertainty, awareness, and trust in visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 240–249, 2016.

[72] Y. Ahn and Y.-R. Lin, "Fairsight: Visual analytics for fairness in decision making," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 1086–1095, 2019.

[73] M. Tory and T. Moller, "Evaluating visualizations: do expert reviews work?," *IEEE Computer Graphics and Applications*, vol. 25, no. 5, pp. 8–11, 2005.

[74] M. Liu, J. Shi, K. Cao, J. Zhu, and S. Liu, "Analyzing the training processes of deep generative models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 77–87, 2018.

[75] N. Pezzotti, T. Höllt, J. Van Gemert, B. P. Lelieveldt, E. Eisemann, and A. Vilanova, "Deepeyes: Progressive visual analytics for designing deep neural networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 98–108, 2018.

[76] M. Kahng, P. Y. Andrews, A. Kalro, and D. H. P. Chau, "A cti v is: Visual exploration of industry-scale deep neural network models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 88–97, 2018.

[77] Q. Wang, J. Yuan, S. Chen, H. Su, H. Qu, and S. Liu, "Visual genealogy of deep neural networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 11, pp. 3340–3352, 2020.

[78] F. Hohman, H. Park, C. Robinson, and D. H. P. Chau, "Summit: scaling deep learning interpretability by visualizing activation and attribution summarizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 1096–1106, 2019.

[79] H. Strobelt, S. Gehrmann, H. Pfister, and A. M. Rush, "Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 667–676, 2018.

[80] Y. Ming, S. Cao, R. Zhang, Z. Li, Y. Chen, Y. Song, and H. Qu, "Understanding hidden memories of recurrent neural networks," *arXiv preprint arXiv:1710.10777*, 2017.

[81] W. Zhong, C. Xie, Y. Zhong, Y. Wang, W. Xu, S. Cheng, and K. Mueller, "Evolutionary visual analysis of deep neural networks," in *ICML Workshop on Visualization for Deep Learning*, 2017.

[82] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Communications of the ACM*, vol. 63, no. 1, pp. 68–77, 2019.

[83] M. Denil, A. Demiraj, and N. de Freitas, "Extraction of salient sentences from labelled documents," *International Conference on Learning Representations (ICLR)*, 2015.

[84] N. Poerner, H. Schütze, and B. Roth, "Evaluating neural network explanation methods using hybrid documents and morphological prediction," in *56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

[85] A. Kádár, G. Chrupała, and A. Alishahi, "Representation of linguistic form and function in recurrent neural networks," *Computational Linguistics*, vol. 43, no. 4, pp. 761–780, 2017.

[86] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i you? explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, ACM, 2016.

[87] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *AAAI Conference on Artificial Intelligence*, 2018.

[88] W. J. Murdoch, P. J. Liu, and B. Yu, "Beyond word importance: Contextual decomposition to extract interactions from lstms," *ICLR*, 2018.

[89] M. Du, N. Liu, F. Yang, S. Ji, and X. Hu, "On attribution of recurrent neural network predictions via additive decomposition," in *The World Wide Web Conference*, pp. 383–393, 2019.

[90] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, pp. 98–136, Jan. 2015.

[91] A. Osman, L. Arras, and W. Samek, "Towards ground truth evaluation of visual explanations," *arXiv preprint arXiv:2003.07258*, 2020.

[92] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9215–9223, 2018.

[93] M. Du, N. Liu, Q. Song, and X. Hu, "Towards explanation of dnn-based prediction with guided feature inversion," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1358–1367, 2018.

[94] O. Zaidan, J. Eisner, and C. Piatko, "Using "annotator rationales" to improve machine learning for text categorization," in *Human Language Technologies*, pp. 260–267, 2007.

[95] T. Lei, R. Barzilay, and T. S. Jaakkola, "Rationalizing neural predictions," in *EMNLP*, 2016.

[96] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. J. Gershman, and F. Doshi-Velez, "Human evaluation of models built for interpretability," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, pp. 59–67, 2019.

[97] P. Schmidt and F. Biessmann, "Quantifying interpretability and trust in machine learning systems," *arXiv preprint arXiv:1901.08558*, 2019.

[98] J. Schneider, J. Handali, M. Vlachos, and C. Meske, "Deceptive ai explanations: Creation and detection," *arXiv preprint arXiv:2001.07641*, 2020.

[99] H. Lakkaraju and O. Bastani, """ how do i fool you?": Manipulating user trust via misleading black box explanations," *arXiv preprint arXiv:1911.06473*, 2019.

[100] L. Chu, X. Hu, J. Hu, L. Wang, and J. Pei, "Exact and consistent interpretation for piecewise linear neural networks: A closed form solution," *arXiv preprint arXiv:1802.06259*, 2018.

[101] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3145–3153, JMLR. org, 2017.

[102] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.

[103] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 2662–2670, 2017.

[104] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 11, pp. 2660–2673, 2017.

[105] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," in *Advances in Neural Information Processing Systems*, pp. 9734–9745, 2019.

[106] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[107] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, pp. 818–833, Springer, 2014.

[108] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[109] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, "The (un) reliability of saliency methods," *arXiv preprint arXiv:1711.00867*, 2017.

[110] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems*, p. 9505–9515, Curran Associates, Inc., 2018.

[111] B. Herman, "The promise and peril of human evaluation for model interpretability," *arXiv preprint arXiv:1711.07414*, 2017.

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

[113] R. R. Hoffman and G. Klein, "Explaining explanation, part 1: theoretical foundations," *IEEE Intelligent Systems*, vol. 32, no. 3, pp. 68–73, 2017.

[114] R. R. Hoffman, S. T. Mueller, and G. Klein, "Explaining explanation, part 2: empirical foundations," *IEEE Intelligent Systems*, vol. 32, no. 4, pp. 78–86, 2017.

[115] G. Klein, "Explaining explanation, part 3: The causal landscape," *IEEE Intelligent Systems*, vol. 33, no. 2, pp. 83–88, 2018.

[116] R. Hoffman, T. Miller, S. T. Mueller, G. Klein, and W. J. Clancey, "Explaining explanation, part 4: a deep dive on deep nets," *IEEE Intelligent Systems*, vol. 33, no. 3, pp. 87–95, 2018.

[117] T. Miller, "Explanation in artificial intelligence: insights from the social sciences," *arXiv preprint arXiv:1706.07269*, 2017.

[118] A. Weller, "Challenges for transparency," *arXiv preprint arXiv:1708.01870*, 2017.

[119] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," *AI Magazine*, vol. 35, no. 4, pp. 105–120, 2014.

[120] K. Höök, "Steps to take before intelligent user interfaces become real," *Interacting with Computers*, vol. 12, no. 4, pp. 409–426, 2000.

[121] J. Zhu, A. Liapis, S. Risi, R. Bidarra, and G. M. Youngblood, "Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation," in *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, pp. 1–8, IEEE, 2018.

[122] D. Sacha, M. Sedlmair, L. Zhang, J. A. Lee, D. Weiskopf, S. North, and D. Keim, "Human-centered machine learning through interactive visualization," ESANN, 2016.

[123] A. Endert, W. Ribarsky, C. Turkay, B. Wong, I. Nabney, I. D. Blanco, and F. Rossi, "The state of the art in integrating machine learning into visual analytics," in *Computer Graphics Forum*, vol. 36, pp. 458–486, Wiley Online Library, 2017.

[124] J. Choo and S. Liu, "Visual analytics for explainable deep learning," *IEEE Computer Graphics and Applications*, vol. 38, no. 4, pp. 84–92, 2018.

[125] F. M. Hohman, M. Kahng, R. Pienta, and D. H. Chau, "Visual analytics in deep learning: An interrogative survey for the next frontiers," *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[126] T. Spinner, U. Schlegel, H. Schäfer, and M. El-Assady, "explainer: A visual analytics framework for interactive and explainable machine learning," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 1064–1074, 2020.

[127] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, p. 93, 2018.

[128] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[129] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau, "Visual analytics in deep learning: An interrogative survey for the next frontiers," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 8, pp. 2674–2693, 2019.

[130] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 80–89, IEEE, 2018.

[131] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," in *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, pp. 241–250, ACM, 2000.

[132] T. Kulesza, S. Stumpf, M. Burnett, and I. Kwan, "Tell me more?: The effects of mental model soundness on personalizing an intelligent agent," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, (New York, NY, USA), pp. 1–10, ACM, 2012.

[133] S. Stumpf, S. Skrebe, G. Aymer, and J. Hobson, "Explaining smart heating systems to discourage fiddling with optimized behavior," in *CEUR Workshop Proceedings*, vol. 2068, 2018.

[134] M. Bilgic and R. J. Mooney, "Explaining recommendations: Satisfaction vs. promotion," in *Beyond Personalization Workshop, IUI*, vol. 5, p. 153, 2005.

[135] A. Bunt, M. Lount, and C. Lauzon, "Are explanations always important?: a study of deployed, low-cost intelligent interactive systems," in *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pp. 169–178, ACM, 2012.

[136] F. Gedikli, D. Jannach, and M. Ge, "How should i explain? a comparison of different explanation types for recommender systems," *International Journal of Human-Computer Studies*, vol. 72, no. 4, pp. 367–382, 2014.

[137] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong, "Too much, too little, or just right? ways explanations impact end users' mental models," in *Visual Languages and Human-Centric Computing (VL/HCC), 2013 IEEE Symposium on*, pp. 3–10, IEEE, 2013.

[138] B. Y. Lim, A. K. Dey, and D. Avrahami, "Why and why not explanations improve the intelligibility of context-aware intelligent systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2119–2128, ACM, 2009.

[139] U. Schmid, C. Zeller, T. Besold, A. Tamaddoni-Nezhad, and S. Muggleton, "How does predicate invention affect human comprehensibility?," in *International Conference on Inductive Logic Programming*, pp. 52–67, Springer, 2016.

[140] S. Berkovsky, R. Taib, and D. Conway, "How to recommend?: User trust factors in movie recommender systems," in *Proceedings of the 22Nd International Conference on Intelligent User Interfaces*, IUI '17, (New York, NY, USA), pp. 287–300, ACM, 2017.

[141] A. Glass, D. L. McGuinness, and M. Wolverton, "Toward establishing trust in adaptive agents," in *Proceedings of the 13th International Conference on Intelligent User Interfaces*, pp. 227–236, ACM, 2008.

[142] S. R. Haynes, M. A. Cohen, and F. E. Ritter, "Designs for explaining intelligent agents," *International Journal of Human-Computer Studies*, vol. 67, no. 1, pp. 90–110, 2009.

[143] D. Holliday, S. Wilson, and S. Stumpf, "User trust in intelligent systems: A journey over time," in *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pp. 164–168, ACM, 2016.

[144] F. Nothdurft, F. Richter, and W. Minker, "Probabilistic human-computer trust handling," in *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp. 51–59, 2014.

[145] P. Pu and L. Chen, "Trust building with explanation interfaces," in *Proceedings of the 11th international conference on Intelligent user interfaces*, pp. 93–100, ACM, 2006.

[146] A. Bussone, S. Stumpf, and D. O'Sullivan, "The role of explanations on trust and reliance in clinical decision support systems," in *Healthcare Informatics (ICHI), 2015 International Conference on*, pp. 160–169, IEEE, 2015.

[147] A. Groce, T. Kulesza, C. Zhang, S. Shamasunder, M. Burnett, W.-K. Wong, S. Stumpf, S. Das, A. Shinsel, F. Bice, *et al.*, "You are the only possible oracle: Effective test selection for end users of interactive machine learning systems," *IEEE Transactions on Software Engineering*, vol. 40, no. 3, pp. 307–323, 2014.

[148] B. A. Myers, D. A. Weitzman, A. J. Ko, and D. H. Chau, "Answering why and why not questions in user interfaces," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pp. 397–406, ACM, 2006.

[149] M. K. Lee, A. Jain, H. J. Cha, S. Ojha, and D. Kusbit, "Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, p. 182, 2019.

[150] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of explanatory debugging to personalize interactive machine learning," in *Proceedings of the 20th international conference on intelligent user interfaces*, pp. 126–137, ACM, 2015.

[151] T. Kulesza, S. Stumpf, M. Burnett, W.-K. Wong, Y. Riche, T. Moore, I. Oberst, A. Shinsel, and K. McIntosh, "Explanatory debugging: Supporting end-user debugging of machine-learned programs," in *Visual Languages and Human-Centric Computing (VL/HCC), 2010 IEEE Symposium on*, pp. 41–48, IEEE, 2010.

[152] J. Krause, A. Dasgupta, J. Swartz, Y. Aphinyanaphongs, and E. Bertini, "A workflow for visual diagnostics of binary classifiers using instance-level explanations," in *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 162–172, IEEE, 2017.

[153] A. S. Ross and F. Doshi-Velez, "Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients," *arXiv preprint arXiv:1711.09404*, 2017.

[154] H. Lakkaraju, S. H. Bach, and J. Leskovec, "Interpretable decision sets: A joint framework for description and prediction," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1675–1684, ACM, 2016.

[155] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu, "Towards better analysis of deep convolutional neural networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 91–100, 2017.

[156] J. J. Dudley and P. O. Kristensson, "A review of user interface design for interactive machine learning," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 8, no. 2, p. 8, 2018.

[157] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, *et al.*, "Guidelines for human-ai interaction," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, p. 3, ACM, 2019.

[158] T. Munzner, "A nested process model for visualization design and validation," *IEEE Transactions on Visualization & Computer Graphics*, no. 6, pp. 921–928, 2009.

[159] M. Beaudouin-Lafon, "Designing interaction, not interfaces," in *Proceedings of the Working Conference on Advanced Visual Interfaces*, pp. 15–22, 2004.

[160] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *International Journal of Man-Machine Studies*, vol. 27, no. 5-6, pp. 527–539, 1987.

[161] G. Wiegand, M. Schmidmaier, T. Weber, Y. Liu, and H. Hussmann, "I drive-you trust: Explaining driving behavior of autonomous cars," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, p. LBW0163, ACM, 2019.

[162] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *arXiv preprint arXiv:1908.09635*, 2019.

[163] J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell, "Fairness under unawareness: Assessing disparity when protected class is unobserved," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 339–348, ACM, 2019.

[164] D. Smilkov, S. Carter, D. Sculley, F. B. Viégas, and M. Wattenberg, "Direct-manipulation visualization of deep networks," *arXiv preprint arXiv:1708.03788*, 2017.

[165] A. P. Norton and Y. Qi, "Adversarial-playground: A visualization suite showing how adversarial examples fool deep learning," in *Visualization for Cyber Security (VizSec), 2017 IEEE Symposium on*, pp. 1–4, IEEE, 2017.

[166] Y. Ming, H. Qu, and E. Bertini, "Rulematrix: Visualizing and understanding classifiers with rules," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 342–352, 2018.

[167] R. Yu and L. Shi, "A user-based taxonomy for deep learning visualization," *Visual Informatics*, vol. 2, no. 3, pp. 147–154, 2018.

[168] E. Alexander and M. Gleicher, "Task-driven comparison of topic models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 320–329, 2015.

[169] F. Du, C. Plaisant, N. Spring, K. Crowley, and B. Shneiderman, "Eventaction: A visual analytics approach to explainable recommendation for event sequences," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 9, no. 4, pp. 1–31, 2019.

[170] K. Wongsuphasawat, D. Smilkov, J. Wexler, J. Wilson, D. Mane, D. Fritz, D. Krishnan, F. B. Viégas, and M. Wattenberg, "Visualizing dataflow graphs of deep learning models in

tensorflow," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 1–12, 2017.

[171] F. Wang and C. Rudin, "Falling rule lists," in *Artificial Intelligence and Statistics*, pp. 1013–1022, 2015.

[172] M. Wu, M. C. Hughes, S. Parbhoo, M. Zazzi, V. Roth, and F. Doshi-Velez, "Beyond sparsity: Tree regularization of deep models for interpretability," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[173] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *ICML Deep Learning Workshop 2015*, 2015.

[174] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?," *Computer Vision and Image Understanding*, vol. 163, pp. 90–100, 2017.

[175] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev, "The building blocks of interpretability," *Distill*, 2018. https://distill.pub/2018/building-blocks.

[176] Q. Zhang, W. Wang, and S.-C. Zhu, "Examining cnn representations with respect to dataset bias," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[177] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3429–3437, 2017.

[178] S. Mohseni, A. Jagadeesh, and Z. Wang, "Predicting model failure using saliency maps in autonomous driving systems," *ICML Workshop on Uncertainty & Robustness in Deep Learning*, 2019.

[179] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755*, 2014.

[180] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Attention bridging network for knowledge transfer," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5198–5207, 2019.

[181] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, "How to explain individual classification decisions," *Journal of Machine Learning Research*, vol. 11, no. Jun, pp. 1803–1831, 2010.

[182] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[183] T. Lombrozo, "Explanation and categorization: How "why?" informs "what?"," *Cognition*, vol. 110, no. 2, pp. 248–253, 2009.

[184] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, IEEE, 2017.

[185] J. Vermeulen, G. Vanderhulst, K. Luyten, and K. Coninx, "Pervasivecrystal: Asking and answering why and why not questions about pervasive computing applications," in *Intelligent Environments (IE), 2010 Sixth International Conference on*, pp. 271–276, IEEE, 2010.

[186] C. J. Cai, J. Jongejan, and J. Holbrook, "The effects of example-based explanations in a machine learning interface," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. 258–262, ACM, 2019.

[187] B. Y. Lim, Q. Yang, A. M. Abdul, and D. Wang, "Why these explanations? selecting intelligibility types for explanation goals," in *IUI Workshops*, 2019.

[188] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harvard Journal of Law & Technology*, vol. 31, p. 841, 2017.

[189] C. J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, F. Viegas, G. S. Corrado, M. C. Stumpe, *et al.*, "Human-centered tools for coping with imperfect algorithms

during medical decision-making," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2019.

[190] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," in *Advances in Neural Information Processing Systems*, pp. 2280–2288, 2016.

[191] A. Lex, M. Streit, H.-J. Schulz, C. Partl, D. Schmalstieg, P. J. Park, and N. Gehlenborg, "Stratomex: Visual analysis of large-scale heterogeneous genomics data for cancer subtype characterization," in *Computer graphics forum*, vol. 31, pp. 1175–1184, Wiley Online Library, 2012.

[192] C. Tikkinen-Piri, A. Rohunen, and J. Markkula, "Eu general data protection regulation: Changes and implications for personal data collecting companies," *Computer Law & Security Review*, vol. 34, no. 1, pp. 134–153, 2018.

[193] S. Mennicken, J. Vermeulen, and E. M. Huang, "From today's augmented houses to tomorrow's smart homes: new directions for home automation research," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 105–115, ACM, 2014.

[194] S. Mohseni, E. Ragan, and X. Hu, "Open issues in combating fake news: Interpretability as an opportunity," *arXiv preprint arXiv:1904.03016*, 2019.

[195] M. Madsen and S. Gregor, "Measuring human-computer trust," in *11th Australasian Conference on Information Systems*, vol. 53, pp. 6–8, Citeseer, 2000.

[196] D. Meyerson, K. E. Weick, and R. M. Kramer, "Swift trust and temporary groups," *Trust in Organizations: Frontiers of Theory and Research*, vol. 166, p. 195, 1996.

[197] S. M. Merritt, H. Heimbaugh, J. LaChapell, and D. Lee, "I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system," *Human Factors*, vol. 55, no. 3, pp. 520–534, 2013.

[198] P. Bobko, A. J. Barelka, and L. M. Hirshfield, "The construct of state-level suspicion: A model and research agenda for automated and information technology (it) contexts," *Human Factors*, vol. 56, no. 3, pp. 489–508, 2014.

[199] R. R. Hoffman, J. K. Hawley, and J. M. Bradshaw, "Myths of automation, part 2: Some very human consequences," *IEEE Intelligent Systems*, vol. 29, no. 2, pp. 82–85, 2014.

[200] B. Cahour and J.-F. Forzy, "Does projection into use improve trust and exploration? an example with a cruise control system," *Safety Science*, vol. 47, no. 9, pp. 1260–1270, 2009.

[201] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *International Journal of Cognitive Ergonomics*, vol. 4, no. 1, pp. 53–71, 2000.

[202] S. Antifakos, N. Kern, B. Schiele, and A. Schwaninger, "Towards improving trust in context-aware systems by displaying system confidence," in *Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices & Services*, pp. 9–14, ACM, 2005.

[203] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson, "When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5092–5103, ACM, 2016.

[204] E. J. Langer, A. Blank, and B. Chanowitz, "The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction.," *Journal of Personality and Social Psychology*, vol. 36, no. 6, p. 635, 1978.

[205] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow, "Visualizing the non-visual: Spatial analysis and interaction with information from text documents," in *Information Visualization, 1995. Proceedings.*, pp. 51–58, IEEE, 1995.

[206] Y. Hu, J. Boyd-Graber, B. Satinoff, and A. Smith, "Interactive topic modeling," *Machine Learning*, vol. 95, no. 3, pp. 423–469, 2014.

[207] J. Choo, H. Lee, J. Kihm, and H. Park, "ivisclassifier: An interactive visual analytics system for classification based on supervised dimension reduction," in *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pp. 27–34, IEEE, 2010.

[208] N. Bryan and G. Mysore, "An efficient posterior regularized latent variable model for interactive sound source separation," in *International Conference on Machine Learning*, pp. 208–216, 2013.

[209] S. Rudolph, A. Savikhin, and D. S. Ebert, "Finvis: Applied visual analytics for personal financial planning," in *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pp. 195–202, Citeseer, 2009.

[210] S. Rosenthal, S. P. Selvaraj, and M. M. Veloso, "Verbalization: Narration of autonomous robot experience.," in *IJCAI*, pp. 862–868, 2016.

[211] F. Hohman, A. Srinivasan, and S. M. Drucker, "Telegam: combining visualization and verbalization for interpretable machine learning," *IEEE Visualization Conference (VIS)*, 2019.

[212] M. Chromik, M. Eiband, S. T. Völkel, and D. Buschek, "Dark patterns of explainability, transparency, and user control for intelligent systems.," in *IUI Workshops*, 2019.

[213] C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, and A. L. Toombs, "The dark (patterns) side of ux design," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 534, ACM, 2018.

[214] S. Jhaver, Y. Karpfen, and J. Antin, "Algorithmic anxiety and coping strategies of airbnb hosts," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 421, ACM, 2018.

[215] A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, and D. H. Chau, "Fairvis: visual analytics for discovering intersectional bias in machine learning," *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2019.

[216] M. Meyer, M. Sedlmair, P. S. Quinan, and T. Munzner, "The nested blocks and guidelines model," *Information Visualization*, vol. 14, no. 3, pp. 234–249, 2015.

[217] R. R. Hoffman, "Theory, concepts, measures but policies, metrics," in *Macrocognition Metrics and Scenarios*, pp. 35–42, CRC Press, 2017.

[218] S. Coppers, J. Van den Bergh, K. Luyten, K. Coninx, I. Van der Lek-Ciudin, T. Vanallemeersch, and V. Vandeghinste, "Intellingo: An intelligible translation environment," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 524, ACM, 2018.

[219] J. Krause, A. Perer, and E. Bertini, "Infuse: interactive feature selection for predictive modeling of high dimensional data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1614–1623, 2014.

[220] W. Curran, T. Moore, T. Kulesza, W.-K. Wong, S. Todorovic, S. Stumpf, R. White, and M. Burnett, "Towards recognizing cool: can end users help computer vision recognize subjective attributes of objects in images?," in *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, pp. 285–288, ACM, 2012.

[221] E. Costanza, J. E. Fischer, J. A. Colley, T. Rodden, S. D. Ramchurn, and N. R. Jennings, "Doing the laundry with agents: a field trial of a future smart energy system in the home," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 813–822, ACM, 2014.

[222] J. Dodge, S. Penney, A. Anderson, and M. M. Burnett, "What should be in an xai explanation? what ift reveals.," in *IUI Workshops*, 2018.

[223] S. Penney, J. Dodge, C. Hilderbrand, A. Anderson, L. Simpson, and M. Burnett, "Toward foraging for understanding of starcraft agents: An empirical study," in *23rd International Conference on Intelligent User Interfaces*, IUI '18, (New York, NY, USA), pp. 225–237, ACM, 2018.

[224] E. Rader and R. Gray, "Understanding user beliefs about algorithmic curation in the facebook news feed," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 173–182, ACM, 2015.

[225] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International Conference on Machine Learning*, pp. 2673–2682, 2018.

[226] B. Nushi, E. Kamar, and E. Horvitz, "Towards accountable ai: Hybrid human-machine analyses for characterizing system failure," in *Sixth AAAI Conference on Human Computation and Crowdsourcing*, 2018.

[227] F. C. Keil, "Explanation and understanding," *Annu. Rev. Psychol.*, vol. 57, pp. 227–254, 2006.

[228] T. Lombrozo, "The structure and function of explanations," *Trends in Vognitive Sciences*, vol. 10, no. 10, pp. 464–470, 2006.

[229] B. Saket, A. Srinivasan, E. D. Ragan, and A. Endert, "Evaluating interactive graphical encodings for data visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 3, pp. 1316–1330, 2017.

[230] B. Letham, C. Rudin, T. H. McCormick, D. Madigan, *et al.*, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," *The Annals of Applied Statistics*, vol. 9, no. 3, pp. 1350–1371, 2015.

[231] J. Kim and J. Seo, "Human understandable explanation extraction for black-box classification models based on matrix factorization," *arXiv preprint arXiv:1709.06201*, 2017.

[232] Z. Zhang, J. Singh, U. Gadiraju, and A. Anand, "Dissonance between human and machine understanding," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–23, 2019.

[233] M. Robnik-Šikonja and M. Bohanec, "Perturbation-based explanations of prediction models," in *Human and Machine Learning*, pp. 159–175, Springer, 2018.

[234] T. Zahavy, N. Ben-Zrihem, and S. Mannor, "Graying the black box: Understanding dqns," in *International Conference on Machine Learning*, pp. 1899–1908, 2016.

[235] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *arXiv preprint arXiv:1702.04595*, 2017.

[236] S. Mohseni and E. D. Ragan, "A human-grounded evaluation benchmark for local explanations of machine learning," *arXiv preprint arXiv:1801.05075*, 2018.

[237] P. Lertvittayakumjorn and F. Toni, "Human-grounded evaluations of explanation methods for text classification," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5198–5208, 2019.

[238] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[239] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," in *Advances in Neural Information Processing Systems*, pp. 9273–9282, 2019.

[240] R. Florez-Lopez and J. M. Ramon-Jeronimo, "Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. a correlated-adjusted decision forest proposal," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5737–5753, 2015.

[241] M. Bojarski, A. Choromanska, K. Choromanski, B. Firner, L. J. Ackel, U. Muller, P. Yeres, and K. Zieba, "Visualbackprop: Efficient visualization of cnns for autonomous driving," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8, IEEE, 2018.

[242] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach, "Women also snowboard: Overcoming bias in captioning models," in *European Conference on Computer Vision*, pp. 793–811, Springer, 2018.

[243] M. Du, N. Liu, F. Yang, and X. Hu, "Learning credible deep neural networks with rationale regularization," in *2019 IEEE International Conference on Data Mining (ICDM)*, pp. 150–159, 2019.

[244] J. K. Doyle, M. J. Radzicki, and W. S. Trees, "Measuring change in mental models of complex dynamic systems," in *Complex Decision Making*, pp. 269–294, Springer, 2008.

[245] M. Schaffernicht and S. N. Groesser, "A comprehensive method for comparing mental models of dynamic systems," *European Journal of Operational Research*, vol. 210, no. 1, pp. 57–67, 2011.

[246] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 2020.

[247] M. Nourani, D. Honeycutt, J. Block, C. Roy, T. Rahman, E. D. Ragan, and V. Gogate, "Investigating the importance of first impressions and explainable ai with interactive video analysis," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, ACM, 2020.

[248] S. Mohseni, F. Yang, S. Pentyala, M. Du, Y. Liu, N. Lupfer, X. Hu, S. Ji, and E. Ragan, "Trust evolution over time in explainable ai for fake news detection," *Fair & Responsible AI Workshop at CHI 2020*, 2020.

[249] J. Kraus, D. Scholz, D. Stiegemeier, and M. Baumann, "The more you know: trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency," *Human Factors*, vol. 62, no. 5, pp. 718–736, 2020.

[250] S. T. Mueller and G. Klein, "Improving users' mental models of intelligent software tools," *IEEE Intelligent Systems*, vol. 26, no. 2, pp. 77–83, 2011.

[251] M. Eslami, S. R. Krishna Kumaran, C. Sandvig, and K. Karahalios, "Communicating algorithmic process in online behavioral advertising," in *CHI*, ACM, 2018.

[252] M. ter Hoeve, M. Heruer, D. Odijk, A. Schuth, and M. de Rijke, "Do news consumers want explanations for personalized news rankings," in *FATREC Workshop on Responsible Recommendation Proceedings*, 2017.

[253] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[254] X. Zhou and R. Zafarani, "Fake news: A survey of research, detection methods, and opportunities," *arXiv preprint arXiv:1812.00315*, 2018.

[255] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 2, p. 32, 2018.

[256] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu, "Combating fake news: A survey on identification and mitigation techniques," *arXiv preprint arXiv:1901.06437*, 2019.

[257] S. Afroz, M. Brennan, and R. Greenstadt, "Detecting hoaxes, frauds, and deception in writing style online," in *Security and Privacy (SP), 2012 IEEE Symposium on*, pp. 461–475, IEEE, 2012.

[258] V. L. Rubin and T. Lukoianova, "Truth and deception at the rhetorical structure level," *Journal of the Association for Information Science and Technology*, vol. 66, no. 5, pp. 905–917, 2015.

[259] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A stylometric inquiry into hyperpartisan and fake news," *arXiv preprint arXiv:1702.05638*, 2017.

[260] M. Potthast, S. Köpsel, B. Stein, and M. Hagen, "Clickbait detection," in *European Conference on Information Retrieval*, pp. 810–817, Springer, 2016.

[261] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," *arXiv preprint arXiv:1809.00888*, 2018.

[262] A. Bharati, D. Moreira, J. Brogan, P. Hale, K. W. Bowyer, P. J. Flynn, A. Rocha, and W. J. Scheirer, "Beyond pixels: Image provenance analysis leveraging metadata," *arXiv preprint arXiv:1807.03376*, 2018.

[263] R. Linder, A. M. Stacy, N. Lupfer, A. Kerne, and E. D. Ragan, "Pop the feed filter bubble: Making Reddit social media a VR cityscape," in *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 619–620, IEEE, 2018.

[264] D. Geschke, J. Lorenz, and P. Holtz, "The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers," *British Journal of Social Psychology*, vol. 58, no. 1, pp. 129–149, 2019.

[265] D. Valcarce, A. Bellogín, J. Parapar, and P. Castells, "On the robustness and discriminative power of information retrieval metrics for top-n recommendation," in *Proceedings of the 12th ACM Conference on Recommender Systems*, pp. 260–268, ACM, 2018.

[266] M. D. Ekstrand, M. Tian, I. M. Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill, and M. S. Pera, "All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness," in *Conference on Fairness, Accountability and Transparency*, pp. 172–186, 2018.

[267] J. Kulshrestha, M. Eslami, J. Messias, M. B. Zafar, S. Ghosh, K. P. Gummadi, and K. Karahalios, "Quantifying search bias: Investigating sources of bias for political searches in social media," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 417–432, ACM, 2017.

[268] Z. Jin, J. Cao, Y. Zhang, and J. Luo, "News verification by exploiting conflicting social viewpoints in microblogs.," in *AAAI*, pp. 2972–2978, 2016.

[269] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *2013 IEEE 13th International Conference on Data Mining*, pp. 1103–1108, IEEE, 2013.

[270] C. Castillo, M. Mendoza, and B. Poblete, "Predicting information credibility in time-sensitive social media," *Internet Research*, vol. 23, no. 5, pp. 560–588, 2013.

[271] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro, "Some like it hoax: Automated fake news detection in social networks," *arXiv preprint arXiv:1704.07506*, 2017.

[272] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks.," in *IJCAI*, pp. 3818–3824, 2016.

[273] K. Shu, S. Wang, and H. Liu, "Exploiting tri-relationship for fake news detection," *arXiv preprint arXiv:1712.07709*, 2017.

[274] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media," *arXiv preprint arXiv:1809.01286*, 2018.

[275] K. Popat, S. Mukherjee, A. Yates, and G. Weikum, "DeClarE: Debunking fake news and false claims using evidence-aware deep learning," in *EMNLP*, ACL, 2018.

[276] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "defend: Explainable fake news detection," in *KDD*, ACM, 2019.

[277] F. Yang, S. K. Pentyala, S. Mohseni, M. Du, H. Yuan, R. Linder, E. D. Ragan, S. Ji, and X. B. Hu, "Xfake: Explainable fake news detector with visualizations," in *The World Wide Web Conference*, pp. 383–393, 2019.

[278] A. T. Nguyen, A. Kharosekar, S. Krishnan, S. Krishnan, E. Tate, B. C. Wallace, and M. Lease, "Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking," in *UIST*, ACM, 2018.

[279] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[280] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013.

[281] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *NAACL-HLT*, pp. 1480–1489, 2016.

[282] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[283] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *EMNLP*, ACL, 2014.

[284] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *KDD*, ACM, 2016.

[285] A. Kumar and R. Rastogi, "Attentional recurrent neural networks for sentence classification," in *Innovations in Infrastructure*, Springer, 2019.

[286] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 25–36, 2013.

[287] Z. Qi, S. Khorram, and F. Li, "Embedding deep networks into visual explanations," *arXiv preprint arXiv:1709.05360*, 2017.

[288] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1084–1102, 2018.

[289] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[290] C. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval," *Natural Language Engineering*, vol. 16, no. 1, pp. 100–103, 2010.

[291] R. R. Hoffman, "A taxonomy of emergent trusting in the human–machine relationship," *Cognitive Systems Engineering*, pp. 137–163, 2017.

[292] R. F. Kizilcec, "How much information?: Effects of transparency on trust in an algorithmic interface," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 2390–2395, ACM, 2016.

[293] J. Choe, S. J. Oh, S. Lee, S. Chun, Z. Akata, and H. Shim, "Evaluating weakly supervised object localization methods right," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3133–3142, 2020.

[294] A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, and N. Berthouze, "Evaluating saliency map explanations for convolutional neural networks: A user study," *arXiv preprint arXiv:2002.00772*, 2020.

[295] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*, IEEE, 2009.

[296] K. Lang, "Newsweeder: Learning to filter netnews," in *ICML*, pp. 331–339, 1995.

[297] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, p. 271, Association for Computational Linguistics, 2004.

[298] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.