

DOI: https://doi.org/10.48009/2_iis_2026_145

Human-centered explainable AI for cybersecurity: A design science research approach

Dr. Steven A. Schilhabel, *University of Wisconsin Oshkosh School of Business*, schilhabels@uwosh.edu

Abstract

AI systems are becoming increasingly common in cybersecurity threat detection; however, their unexplained decision-making processes lead to diminished user trust and slow adoption rates. The research addresses this obstacle by developing a Human-Centered Explainable AI (HC-XAI) dashboard that enhances transparency and user trust in AI threat detection classifications. Grounded in Design Science Research (DSR), the artifact integrates three explanation modalities: explanations in the dashboard are derived from rule-based logic, combined with natural language explanations generated by large language models, and visual heatmaps that display token-level attention. Justificatory knowledge from cognitive psychology and human-AI interaction, along with explanation theory, provides the basis for the design process. The artifact was evaluated through a two-phase strategy. The artifact evaluation consisted of an expert walkthrough using HC-XAI heuristics and a mixed-methods user study with twenty-three cybersecurity students. The findings demonstrate excellent usability levels and perceived trustworthiness, while also showing varied preferences depending on the type of explanation. The study presents a multimodal XAI artifact, along with validated principles for HC-XAI applications in critical systems and investigates user preferences regarding explanation clarity and cognitive load. The study yields new insights into the implementation of human-aligned XAI within practical cybersecurity environments.

Keywords: Explainable Artificial Intelligence (XAI), Human-Centered AI (HC-XAI), Cybersecurity, Design Science Research, Trust in AI

Introduction

Machine learning and deep learning applications have grown increasingly common within cybersecurity operations including areas like phishing detection, intrusion detection, malware analysis, and threat intelligence. While ML and DL models can surpass human accuracy in predictive tasks, analytical models often exhibit low transparency within their decision-making processes. Black-box models are often met with user frustration due to cognitive mistrust when adopted within production environments. Cybersecurity analysts desire explanation and justification surrounding AI-generated predictions and recommendations, especially those that may create impactful organizational decisions.

The goal of Explainable Artificial Intelligence (XAI) is to develop analytics that support stakeholder sensemaking of AI-enabled decisions. Existing XAI approaches often focus explanation design on algorithmic transparency while providing minimal regard for human downstream perception or decision-making. Human-Centered Explainable AI (HC-XAI) considers traditional XAI topics while further incorporating notions of explanation usability, explanation cognitive-accessibility, and trust calibration. Analyst sensemaking within cybersecurity environments requires analysts to make quick decisions with often limited information, relying on AI-system recommendations.

While research interest for HC-XAI continues to grow, limited work has been conducted to understand the efficacy of coupling multiple explanation modalities within a single interface or understand how users perceive variant forms of AI-explanation. While many cybersecurity XAI systems have adopted a single form of explanation such as rule-based representations or visual saliency maps, limited research has been conducted to understand how specific forms of generated explanations impact end-user trust, understanding, and cognitive effort.

To better understand how multimodal explanations can be used to augment user trust, interpretability, and understanding within a cybersecurity context, this project will design and evaluate a multimodal HC-XAI artifact for phishing threat classification. This Design Science Research (DSR) project created an artifact that provides three levels of explanations to users including rule-based logic explanations, natural language explanations generated via a large language model (LLM), and heat-map visualization explanations.

Research Questions

RQ1: How can a multimodal HC-XAI artifact be designed to support transparency, trust, and interpretability within a phishing threat detection context?

RQ2: How do users perceive different explanation modalities in terms of trust, confidence, cognitive effort, and confusion?

Contributions

This research contributes three items. First, we designed and evaluated a multimodal HC-XAI artifact focused on a cybersecurity threat classification problem. Second, we evaluated user perception of different explanation modalities and their impact on end-user trust and understanding of AI-model decisions. Third, we derived design principles that contribute to the emerging body of HC-XAI design knowledge.

Paper Structure

The remainder of this paper is structured as follows. The second section covers background information and motivating literature. The third section details solution objectives. The fourth section showcases artifact design. The fifth section exhibits our implemented dashboard. The sixth section conveys our evaluation and results. The seventh discusses how this research contributes to knowledge and design principles. The eighth section concludes, while the ninth section describes limitations and future work.

Problem Identification and Motivation

The integration of artificial intelligence within cybersecurity frameworks remains challenging due to its opaque nature, which restricts both its reliability and application in critical operational settings. AI-based classifiers demonstrate high accuracy for phishing attack detection and other cyber threats but remain unintelligible "black boxes" that prevent analysts from understanding their decision-making processes (Adadi & Berrada, 2018). Opacity creates distrust and fuels algorithm aversion, which leads users to abandon or underutilize systems when they experience unexplained errors (Dietvorst, Simmons, & Massey, 2015).

The absence of trust becomes particularly serious in the context of cybersecurity operations. Alert response analysts require both strong confidence and accountability; however, current systems fail to provide adequate explanation capabilities (Gunning & Aha, 2019). The negative outcomes span slower response

times and incorrect threat prioritization, which together lead to diminished effectiveness in AI-enhanced workflows (Mohale & Obagbuwa, 2025).

HC-XAI aims to bridge the trust gap between users and AI systems by providing customized explanations tailored to individual needs and cognitive styles within specific decision-making contexts (Ehsan & Riedl, 2020). However, operationalizing HC-XAI in cybersecurity remains underdeveloped. Most current tools produce static rule-based outputs or post-hoc visualizations without incorporating multiple explanation formats into a unified and adaptive user interface (Speith, 2022). The field lacks empirical studies examining how various explanation modalities affect trustworthiness and interpretability while also considering cognitive load in this area.

The study fills existing research gaps by developing and accessing a multimodal explanation dashboard intended for phishing threat classification. The study aims to enhance user trust, alongside their comprehension and decision-making ability, by presenting layered explanations that parallel human cognitive processing patterns. The study advances the HC-XAI literature by providing practical applications and theoretical understanding, utilizing Design Science Research methods to evaluate the artifact through heuristic walkthroughs and user testing.

Human-Centered Explainable Artificial Intelligence (HC-XAI) represents an emerging emphasis on designing AI systems that align with human cognitive frameworks and ethical standards, while supporting human decision-making processes (Kong et al., 2024; Schoonderwoerd et al., 2021). The current approach in XAI research represents a shift away from initial AI explainability methods, which centered on algorithmic transparency and post-hoc explanation mechanisms without substantial user consideration (Arrieta et al., 2020; Ali et al., 2023).

Recent HC-XAI research highlights several critical directions: Current HC-XAI research identifies three main research directions, which include creating context-sensitive interactive explanation systems along with the customization of explanations for various user expertise levels while integrating cognitive theories to evaluate mental workload and trust-building effectiveness during explanation delivery (Rong et al., 2023; Kim et al., 2024; Liao & Varshney, 2022). The field of cybersecurity continues to receive inadequate attention within published research studies. Research primarily targets healthcare and finance domains, which require different explanation methods than those needed for immediate threat management in high-risk situations.

Explanation interfaces in cybersecurity continue to operate in a static mode with logic-based processes and saliency visualization methods (Speith, 2022). Without multimodal and adaptive interfaces, security analysts struggle to calibrate trust during time-sensitive evaluations of unclear threat classifications accurately.

This research addresses these shortcomings by creating and evaluating a multimodal HC-XAI dashboard specifically designed for phishing threat classification. The artifact advances explainability research by integrating cognitive and interaction design theories with Design Science Research (DSR) methodology, as outlined by Gregor & Hevner (2013) and Hevner et al. (2004).

Objectives of a Solution

The study focuses on creating and analyzing a Human-Centered Explainable AI (HC-XAI) tool to enhance transparency and trust in AI cybersecurity systems, while improving user comprehension and alignment in phishing threat detection contexts.

The overarching objective is to develop and evaluate a multimodal explanation dashboard that can present three distinct types of explanations for AI-based classifications. The dashboard displays a triad of explanation types, including rule-based phishing indicator logic, natural language explanations from LLMs, and attention-focused visual heatmaps.

The solution follows guidance from justification and design principles originating from DSR methodology (Gregor & Hevner, 2013; Peffers et al., 2007). This approach responds to the XAI literature's demands for explanation systems that advance beyond technical accuracy to meet human cognitive processing needs and decision-making requirements (Ehsan & Riedl, 2020; Wang et al., 2020).

The proposed solution aims to achieve several defined design objectives.

Deliver Actionable, Interpretable Explanations

Phishing classification rationales on the dashboard should be displayed in a way that combines technical precision with cognitive accessibility for non-experts, thereby bridging the AI-human understanding divide (Adadi & Berrada, 2018).

Enable Trust Calibration and Confidence Building

To establish suitable trust levels in the system, users require several types of explanations that cross-validate information at multiple layers. This approach helps reduce algorithm aversion while establishing solid confidence in AI outputs among users (Dietvorst et al., 2015; Logg et al., 2019).

Support Comparative Evaluation Across Modalities

This artifact facilitates empirical comparison of user feedback on clarity perception, cognitive load, and trust across three explanation types to generate design principles for HC-XAI systems (Wang et al., 2020).

Inform Design Knowledge and Theoretical Contribution

This research project aims to advance theoretical knowledge about the operation of multimodal XAI interfaces in critical environments, while developing actionable guidelines for future HC-XAI system deployment (Gregor & Jones, 2007).

The subsequent sections will present the artifact design and evaluation strategy, which builds upon these objectives by adhering to the rigor and relevance principles of the Design Science Research paradigm.

Design and Development

The Explainable AI (XAI) dashboard's design followed an iterative artifact-centered process, as outlined in the Design Science Research (DSR) methodology (Peffers et al., 2007; Hevner et al., 2004). Researchers designed this artifact to enhance human-focused explainability within cybersecurity by focusing on both detecting and providing justification for phishing email threats.

Design Rationale

The artifact employs three explanation modalities, including rule-based logic, natural language reasoning, and visual attention heatmaps, which were specifically chosen to match varied cognitive styles and differing expertise levels. Studies show that users gain advantages from layered or redundant explanations when working in high-risk domains according to research findings by Wang et al. (2020) and Ehsan et al. (2021). Multiple explanation formats help users calibrate trust by making it possible to cross-validate outputs between various modalities.

The artifact utilizes justificatory knowledge from cognitive load theory, along with principles of explanation fidelity and decision support system usability, to inform its design. The HC-XAI evaluation framework informed the design heuristics, focusing on clarity and justification quality, as well as interactive control, as noted by Liao and Varshney (2022).

Technical Implementation

The system creates human-aligned phishing detection explanations by incorporating a large language model (LLM) through the OpenAI GPT API within the dashboard. This module generated natural language rationales that worked in combination with rule-based logic and visual heatmaps to deliver multimodal explanations.

Prompt Design

The prompt design applied human-centered explainable AI (HC-XAI) principles to achieve clarity and relevance while minimizing cognitive load, as per Liao & Varshney (2022). The system used prompts that directed the model to create specific and brief non-technical explanations for emails classified as phishing. To improve interpretability, the prompt highlighted four key indicators commonly associated with phishing emails: Phishing emails show urgency along with deceptive URLs and fake sender details while they manipulate emotions to deceive recipients.

The standard prompt structure used was: Provide a non-technical cybersecurity analyst with apparent reasons why the following email should be classified as phishing. Pay attention to indicators such as urgent requests, questionable links, disguised sender details and emotionally charged phrasing. Email content: [USER_INPUT]”

To streamline decision-making processes and avoid excessive wordiness experts limited the response length to no more than one hundred tokens.

Output Stability and Control

We established the temperature parameter at 0.3 to maintain output consistency and reduce response variability which led to deterministic and less creative results. The decision to manipulate the output generation process played a crucial role in standardizing explanations for comparable phishing samples while preserving interpretability. The output system implemented truncation to prevent both excessive generation and irrelevant content.

Validation and Quality Assurance

The artifact’s prototyping phase included a manual review of every explanation produced by the LLM. The explanations received independent evaluations from five expert panelists who assessed clarity, accuracy, and domain alignment through a 5-point Likert scale. Explanations that received scores below 3 in any dimension were identified as unacceptable and removed while the prompt underwent revision. The iterative review process continued until more than 90% of the output achieved ratings of four or higher.

Theoretical Justification

We chose a generative LLM because it offers more flexibility and context understanding than static templates or rule-based formatting while delivering natural language rationales for human users. Our method adheres to the HC-XAI principle of creating reflective narratives that prioritize users over mechanistic justifications (Ehsan et al., 2021). The utilization of GPT-based explanations enables future developments, including user profiling and personalization, while supporting longitudinal adaptation, which serves as fundamental objectives for next-generation human-AI collaboration systems (Rong et al., 2023; Kim et al., 2024).

Development Iterations

Experts used an HC-XAI checklist for a heuristic evaluation of the artifact's first iteration, which resulted in interface layout changes, as well as adjustments to the framing of explanations and response speed. Based on expert feedback, modifications included:

- Improved heatmap accessibility through the adjustment of color gradients.
- Rewording LLM prompts for more concise explanations.
- Simplifying terminology in the rule-based output display.

The refined version for the user testing phase, described in Section 6, integrated all these changes.

Demonstration

The Explainable AI (XAI) dashboard was developed as a web-based application through Streamlit deployment, enabling interactive functionality. The demonstration events showcased the artifact's essential features by utilizing controlled scenarios that replicated real-world cybersecurity operations. The primary goal of this phase was to confirm that the artifact achieved its design goals by permitting users to engage with various explanation methods for phishing threat assessment.

Access to the dashboard was provided through a secure internal network connection or via a publicly accessible academic portal. The system required participants to enter example text that reproduced potential phishing emails immediately after their login. Three explanation modes were made available: The system presented three explanation modes: (1) a rule-based logic that identified known phishing features in the text, (2) natural language explanations produced using a large language model (LLM), and (3) visual token-level attention heatmaps that highlighted suspect language patterns.

The demonstration fulfilled two essential functions throughout the DSR process. The demonstration first verified that the artifact operated effectively and maintained technical stability when used in expected scenarios. The demonstration provided evaluators with an opportunity to investigate the artifact's explanatory affordances and usability through a heuristic walkthrough before formal empirical study assessment. Participants assessed the artifact's interpretability and decision-support potential by interacting with realistic threats while toggling between various explanation modes.

The demonstration verified that the artifact reached the necessary development stage for formal evaluation and served as an effective demonstration of HC-XAI principles within cybersecurity. Users provided meaningful feedback through operational exposure, which served to refine iterations and validate design principles.

Explainable AI Dashboard for Cybersecurity Threats

Simulate detection explanations for suspicious emails using multiple modes.

 Enter suspicious email content:

 Select explanation type:

- ☒ Rule-based
- ☐ LLM (GPT)
- ☐ Visual Heatmap

 Analyze

Figure 1. Interface.

Participants executed the structured testing protocol outlined in Appendix C by performing each explanation task consecutively. During the rule-based explanation task, participants determined which rules got activated by the input text. The LLM explanation evaluation involved checking both semantic coherence and clarity of the generated justification. The color-coded heatmap indicated to users which words the system identified as suspicious, based on the model's attention weights.

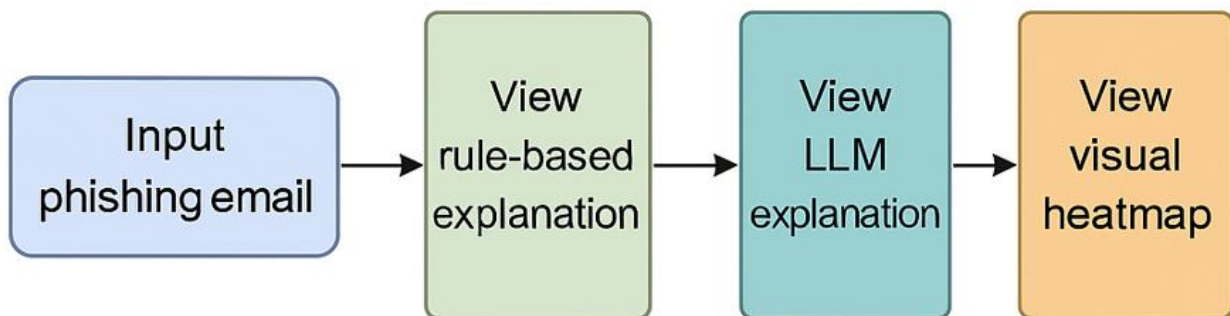


Figure 2. Interaction Sequence

Evaluation

The artifact's effectiveness assessment involved a two-phase evaluation that followed the Design Science Research (DSR) guidelines, as outlined by Hevner et al. (2004) and Peffers et al. (2007). We evaluated the

multimodal Human-Centered Explainable AI (HC-XAI) dashboard to assess its effectiveness in terms of user comprehension, trust, and usability for phishing threat classification.

To evaluate users' experiences with each explanation modality, we collected Likert-scale ratings and qualitative feedback on four cognitive constructs: trust, confidence, effort, and confusion.

Table 1 presents descriptive statistics for all participants, illustrating initial perceptual trends through the evaluated constructs. Selected open-ended comments from participants included in Table 2 help reveal the reasoning and emotional responses that underpin their numeric ratings.

Phase 1: Heuristic Walkthrough

During Phase 1, expert evaluators conducted a heuristic walkthrough of the prototype using a customized HC-XAI evaluation checklist outlined in Appendix C. The evaluation checklist used for the HC-XAI prototype incorporates findings from studies about human–AI interaction and XAI usability heuristics (Wang et al., 2020; Gunning & Aha, 2019). The walkthrough confirmed essential explanation attributes, such as interpretability, alongside consistency, while measuring cognitive load and trust calibration.

The research team found that users could easily navigate through the dashboard and that various explanation methods integrated seamlessly. Noted areas for improvement included: Experts suggested improving user experience by clarifying rule-based terms and optimizing heatmap colors for better accessibility while incorporating new user onboarding instructions. The insights led to direct changes in the artifact development during Phase 2.

Phase 2: User Study

To evaluate how users cognitively engaged with the different explanation types, we collected both quantitative Likert-scale ratings and qualitative open-ended responses across four key constructs: trust, confidence, cognitive effort, and confusion.

The primary descriptive statistics combined with participant feedback showed distinct variations in user experience among the three explanation methods: LLM, rule-based, and heatmap. The disparities between modalities suggest various levels of clarity perception and mental workload among users.

We employed inferential statistical tests, including the Friedman test and post-hoc Wilcoxon signed-rank tests, to determine if the observed patterns were statistically significant instead of occurring by chance. Subsequent sections combine description and inference findings to demonstrate the explanation design effects on interpretability.

The cognitive process that users engage in when interacting with different explanation types plays a crucial role in determining the interpretability outcomes for human-centered explainable AI systems. Participants evaluated mental effort and explanation clarity for natural language (LLM), rule-based logic, and heatmap modes through subjective Likert ratings and open-ended qualitative feedback.

LLM explanations achieved top average clarity ratings, with rule-based logic following closely behind; however, heatmaps lagged both in terms of perceived clarity and the mental effort required.

Table 1. Descriptive stats (mean/SD by modality)

Survey Item	Mean	Std. Dev.	Interpretation
The dashboard was easy to use	3.76	1.64	Usable, though user experience varied
It was clear how to select different explanation modes	3.81	1.66	Interface navigation was clear for most users
The interface allowed quick application of explanations	4.69	0.48	Users found interaction fast and efficient
Rule-based explanations were helpful and clear	3.24	1.55	Mixed feedback on clarity of rule-based output
The logic behind rule-based detection was clear	2.90	1.58	Several users had difficulty understanding the logic
The AI-generated explanation was easy to follow	4.38	0.81	Strong clarity in natural language explanations
Explanation matched user expectations	4.38	1.02	Explanations aligned well with user reasoning
Heatmap highlighted key terms clearly	3.81	1.11	Visual highlights were mostly effective
Visual display aided understanding	3.75	1.18	Display supported understanding for most users
I trust the system's classification	3.19	1.47	Trust levels varied, with moderate overall response
Explanations increased confidence in the system	4.38	0.62	Most users felt explanations improved confidence
Explanations required significant mental effort	2.00	0.89	Explanations were easy to process cognitively
Dashboard explanations were confusing	1.69	0.60	Very few users found the explanations confusing

Participants reported that heatmaps appeared “visually dense,” “unclear,” and “hard to read,” especially when color gradients lacked explanations or when multiple tokens received a highlight. Users reported that LLM explanations demonstrated naturalness and conciseness, which mimicked human speech, while qualitative feedback showed visual-only explanations generated cognitive friction for users.

Table 1. Sample open-ended quotes by explanation type

Theme	Summary Insight	Sample Participant Quote
Clarity of Explanations	Users appreciated multimodal explanations but noted that terminology was dense	“The rules were fine, but some language could be clearer for non-experts.”
Visual Understanding	Heatmaps were liked, though precision in highlighting varied	“I liked the red words, but not all of them seemed relevant.”
Trust Enhancement	Layered explanations (rules, AI, visuals) improved trust and perceived accuracy	“Having multiple types of explanations made me feel more confident.”
Suggestions for Improvement	Participants requested onboarding, examples, and guidance for interpretation	“A short tutorial or guide would make the tool easier to adopt.”
Use in a Professional Setting	Several were open to using it with stronger threat rationale and secure handling	“Yes, but I’d need assurance about accuracy and explanation reliability.”

Users experienced varying perceptions across different cognitive dimensions. The analysis revealed that heatmaps generated higher mental effort ratings, although they did not always confuse. The participants observed that interpreting visual explanations required more cognitive processing but remained understandable when supplemented by email context. The relationship between effortful tasks and cognitive

load theory shows that such tasks do not become misleading when users possess sufficient knowledge about the domain.

Inferential analysis reinforces this point. The Friedman test demonstrated statistically significant differences in user perceptions across the four constructs of trust, confidence, effort, and confusion, with a $\chi^2(3)$ value of 35.69 and a p-value less than 0.001, while showing a strong effect size, indicated by Kendall's W at 0.74. Wilcoxon signed-rank tests demonstrated that users experienced greater confidence than both effort and confusion ($p = 0.0009$ and $p = 0.0005$, respectively), while trust surpassed confusion significantly ($p = 0.0002$). The lack of significant difference between effort and confusion scores ($p = 0.0956$) demonstrates that perceived workload was not consistently associated with misunderstanding.

These results underscore a key HC-XAI insight: Explanation effectiveness requires cognitive usability rather than merely saliency or attention alignment. Design scaffolding through interpretive legends, tooltips, and hybrid visual-text formats should facilitate future visual modality enhancements, thereby minimizing user effort.

Evaluation Summary

According to the evaluation, the HC-XAI dashboard proved to enhance user understanding and trust but demonstrated variable effectiveness across different explanation modalities. Users continuously achieved high usability scores and moderate-to-strong trust levels in each modality but displayed significant differences in interpretability and cognitive workload across those modalities. The analysis demonstrates the artifact's adherence to HC-XAI principles but identifies areas for improvement, particularly in the domain of visual explanations.

Table 2. Friedman Test Results

Construct	Median Rating	SD
Trust	3.00	1.47
Confidence	5.00	0.62
Cognitive Effort	2.00	0.89
Confusion	2.00	0.60

Table 3. Wilcoxon Pairwise Comparison Results

Comparison	p-value	Significant
Trust vs Confidence	0.0047	Yes
Trust vs Effort	0.0019	Yes
Trust vs Confusion	0.0002	Yes
Confidence vs Effort	0.0009	Yes
Confidence vs Confusion	0.0005	Yes
Effort vs Confusion	0.0956	No

The Friedman test results showed significant statistical differences in user perceptions between the constructs of trust, confidence, cognitive effort, and confusion as evidenced by $\chi^2(3) = 35.69$ and $p < .001$ along with a strong effect size of Kendall's W = 0.74. The application of Bonferroni-adjusted Wilcoxon signed-rank tests ($\alpha = 0.0083$) confirmed that trust levels surpassed those of confusion ($p = .0002$). In contrast, confidence levels were higher than both effort ($p = .0009$) and confusion ($p = .0005$). The data

show no meaningful statistical difference when comparing trust to confidence ($p = 0.0047$) and effort to confusion ($p = 0.0956$), indicating subtle distinctions in participants' perceptions. The limited number of participants ($n = 23$) requires careful interpretation as it constrains the study's generalizability and statistical power. The alignment between quantitative and qualitative results strengthens the credibility of the identified trends.

Natural language explanations and rule-based logic received the best clarity and support ratings from participants, but heatmaps scored the lowest, particularly in terms of interpretability and perceived mental effort. Heatmaps appeared visually dense to users, who sometimes found them unclear, but still appreciated their visual appeal.

The results align with visual explanation studies, which show that saliency maps often produce interpretability problems, especially when they lack contextual anchoring or semantic framing. The studies by Liao and Varshney (2022) and Kim et al. (2024) demonstrate that saliency mismatches, together with semantic grounding deficiencies, reduce explanation clarity. Research by Kim et al. (2024) demonstrates that explanation clarity suffers when saliency maps create mismatches with images and lack meaningful semantic connections. The results demonstrate that users require more cognitive effort when using heatmaps but did not consistently report confusion because heatmaps presented challenges in understanding without being completely inaccurate.

Users struggled to understand the heatmap because it lacked interactive features, including tooltips, token-level explanations, and threshold adjustment capabilities. Research indicates that interactive visualizations help decrease cognitive load and enhance comprehension by directing attention toward what users require (Kim et al., 2024). Subsequent dashboard updates should incorporate interactive elements, such as hover text and confidence thresholds, to enhance interpretability.

Through its design implementations, the artifact successfully achieved its primary goal of delivering explainable outcomes that accommodate various cognitive user preferences. Heatmap visualizations require additional development work, focusing on design scaffolding and interactive elements, to align with human-centered cognitive models and eliminate interpretive ambiguity.

The research outcomes support HC-XAI design principles, which focus on progressive disclosure, along with user-controlled transparency, as key to delivering effective explanations (Liao & Varshney, 2022). Interactivity through token-level tooltips and explanation toggles, along with adjustable confidence thresholds, minimizes interpretation resistance by tailoring explanations to both cognitive contexts and decision stages of users. The broader HC-XAI objective of active user engagement with explanations is demonstrated using interactive design elements (Kim et al., 2024).

Interpreting Heatmap Explanations and Cognitive Effort

A surprising result from the analysis was that participants tended to rate heatmap-based explanations lower than natural language and rule-based explanations on various subjective measures. While most participants found visual explanations helpful for finding red flags within emails, heatmaps still received higher cognitive effort ratings and lower clarity ratings on average.

Qualitative feedback from participants suggests that this may be due to the cognitive effort required to interpret attention-based visualizations. Highlighted tokens show users which features contributed to a model's classification, but do not indicate why those features mattered. Participants must both reason about the semantics of highlighted tokens and translate those visualizations into the decision to label an email as

phishing. This mental translation requires extra work on behalf of users, even if the visualization correctly depicts model attention.

These results support previous HC-XAI work regarding the lack of fidelity many saliency-based explanations have when provided standalone (Liao & Varshney, 2022; Kim et al., 2024). Visual attention can easily show where a model is looking, but not necessarily why it is looking there. Users may understand what the visual explanation is showing them but lack context to grasp its full meaning.

Some participants noted that heatmaps felt cluttered or were hard to read. Color scales were not always intuitive, and emails with many highlighted tokens made it hard for users to identify which words were phishing indicators. While most participants were able to effectively use heatmap explanations, this could be problematic in a practical cybersecurity setting.

These results suggest that, from an HC-XAI perspective, visual explanations may require additional scaffolding to support interpretation. Tooltips on hover, confidence scores, adjustment sliders, comment ability on a token-level, and natural language justifications for why certain tokens were highlighted are all ways to decrease cognitive effort by providing semantic meaning to low-level visual cues.

These findings indicate that explanations must be both accurate and understandable to effectively support user decision-making. Visibility into a model's internal processes is valuable; however, explanations that require substantial cognitive effort may limit their practical usefulness. Hybrid explanations that allow users to better interpret model behavior should be implemented into the dashboard.

Contributions to Knowledge

The study advances practical applications and theoretical frameworks for Human-Centered Explainable AI (HC-XAI) in cybersecurity by creating and assessing a new multimodal explanation dashboard. The research adheres to Design Science Research principles by creating actionable design knowledge and validating artifact performance, while generating new theoretical insights (Gregor & Hevner, 2013; Hevner et al., 2004).

Design Theory Instantiation for Multimodal HC-XAI

Researchers have developed a comprehensive design theory for multimodal Explainable AI systems within cybersecurity through this study. By combining rule-based logic with explanations from large language models and visual heatmaps, the artifact fills existing system gaps while providing interpretability for users with different cognitive styles. The study implements DSR components, which include constructs and principles of form and function, mutability, alongside testable propositions defined by Gregor & Jones (2007), and delivers an applicable HC-XAI artifact example.

Empirical Foundations for Interface Design Preferences

Empirical data on the effects of different explanation types on user trust, clarity, and cognitive load were collected from a study that employed mixed-method evaluation techniques, including heuristic walkthroughs and user testing. Research shows that different explanation modalities affect how users make decisions (Wang et al., 2020) while also providing practical guidance for designing user interfaces that accommodate diverse explanation needs.

Practical Heuristics for HC-XAI Development

The study yielded validated design and evaluation heuristics for creating explainable interfaces in high-risk environments. These include:

- Users benefit from multiple explanation types being cross-validated, such as combining LLM and rule-based systems.
- Non-expert users require effective onboarding and simplified language to understand complex interfaces.
- Attention-based explanations require visual accessibility features to support user understanding.

These heuristics offer general applicability to next-generation HC-XAI systems, which operate in security-sensitive areas where interpretability, trust calibration, and response speed are essential features.

Design Principles for Human-Centered Explainable AI in Cybersecurity

One goal of Design Science Research is to produce reusable design knowledge rather than simply instantiate a single artifact (Gregor & Hevner, 2013). Although HC-XAI dashboard serves as one instantiation of design tailored to phishing threat detection, evaluation results offer generalizable knowledge on how to design explainable AI systems that best support human decision-making behavior in cybersecurity settings. Four design principles were derived from the artifact evaluation findings and participant feedback.

DP1: Support Trust Calibration Through Multimodal Explanations. Participants indicated they felt more confident in their decisions when they could access multiple explanation modalities at once. Combined rule-, natural language-, and attention-based explanations enabled users to compare across forms of explanation and validate AI results. HC-XAI systems should not lean on one form of explanation but provide supplementary layers of information that users can utilize to check AI recommendations.

DP2: Prioritize Natural Language Explanations as the Primary Explanation Layer. Natural language explanations generated by the AI were rated highest for understandability across the board. Users described them as easy and intuitive to understand when compared to rules and visual explanations. Cybersecurity settings often demand quick decision-making; therefore, natural language explanations should be prioritized as the first explanation layer to reduce cognitive burden while maintaining elevated levels of interpretability. More complex explanation modalities can be used as supplemental evidence.

DP3: Augment Visual Explanations with Contextual Scaffolding. Participants noted that while heatmaps had the potential to be useful, they required more effort to understand. Simply knowing where a model's attention is focused does not provide users with enough information to understand why certain features contributed to the model's decision. Visual explanations should be supplemented with contextual information to provide users with enough knowledge to make sense of what they are seeing. Tooltips, legend keys, explanatory annotations, or links to text can serve as scaffolding to make visual explanations more intuitive.

DP4: Employ Progressive Disclosure to Support Novice and Expert Users. Different users have various levels of expertise and need explanations. While some users only wanted to know whether a URI was dangerous or not, other users wanted to know more about how the AI made its decisions. HC-XAI interfaces should use progressive disclosure to start with a high-level explanation before giving users the option to delve deeper. This strategy prevents unnecessary details from taxing users' cognitive load while accommodating both novice and expert users.

These principles build off previous HC-XAI work to produce design knowledge that can be leveraged in future work. This includes a shift away from placing exclusive focus on creating transparent algorithms

towards consideration of human cognitive factors like trust calibration and information processing. Therefore, these design principles help populate the space of HC-XAI design knowledge and offer tangible recommendations for future work.

Advancing HC-XAI Research within DSR

This study demonstrates how Design Science Research can be leveraged to develop and evaluate explainable cybersecurity solutions that address both technical and human-centered requirements. To date, much of the explainable AI literature has focused on algorithmic performance and technical explanation generation mechanisms. Little explainable AI research has considered how decisions about the form that an explanation takes impact user trust, interpretability, and cognitive workload during operational decision-making tasks.

Viewed through the lens of DSR contributions, however, this study offers more than the creation of a usable artifact. This research created and validated reusable design knowledge about how to design explanations in cybersecurity settings. The iterative development, expert validation, and user evaluation of this artifact showed that effective human-AI explanation for cybersecurity requires explanations that can be presented in multiple modalities, are accessible via natural language, provide context for visual explanations, and can be progressively shown.

According to the knowledge contribution types outlined by Gregor and Hevner (2013), the contributions of this work can be considered an invention contribution. While explainability and trust have been recognized problems in cybersecurity for some time there has been little work considering multimodal HC-XAI interfaces that can reason about decisions via rules, explain those rules via natural language, and show visual attention over tested cyber surfaces all in one system. While the artifact alone could be considered an innovative work around a well-defined problem, it was created in conjunction with design knowledge that can be applied to future HC-XAI systems.

More generally, this research also validates some previously identified important concepts in HC-XAI literature. It adds to growing evidence that algorithmic transparency alone is not enough. Explanations must also be cognitively accessible to their target users. Study participants preferred explanations that were easy to understand while still providing enough information for them to calibrate trust.

This research also showed that DSR can create end-to-end AI solutions with meaningful contributions to both practice and theory. Previous DSR works in AI have often focused on either the creation of an artifact or theory generation. By unifying all three stages of knowledge discovery this work was able to provide technical takeaways for cybersecurity professionals while also developing design theory that can help guide future HC-XAI work.

Conclusion

The research introduced and assessed a Human-Centered Explainable AI dashboard designed to address interpretability and trust issues in phishing threat detection using AI within the cybersecurity field. The artifact integrated rule-based explanations, natural language descriptions, and visual elements into a single interface, following the Design Science Research methodology, to assist users in understanding and making informed decisions.

The evaluation confirmed that the artifact functioned well and successfully improved trust and interpretability for non-expert cybersecurity users. The multimodal explanation framework received

positive feedback from users; however, qualitative feedback highlighted the need for improved onboarding support and clearer visual explanations. Research results demonstrate that user expectations and experience affect how individuals perceive and utilize various explanation types.

This research offers a theory-based and actionable solution to a pressing challenge in explainable cybersecurity AI, achieved through iterative design cycles and empirical testing in conjunction with design theory principles. This research provides direction for upcoming HC-XAI systems while demonstrating how to combine technical elements with human-centered aspects inside a DSR framework.

Limitations and Future Research

Although this research contributes to the expanding field of Human-Centered Explainable AI (HC-XAI) applications in cybersecurity, it is essential to acknowledge several limitations.

Sample Size and Generalizability

One limitation of this study is that the evaluation was conducted with twenty-three cybersecurity students rather than professional cybersecurity analysts. Although this limits the direct generalizability of the findings to operational security environments, having students meet our participant requirements was suitable for our preliminary DSR Evaluate assessment as participants had basic cybersecurity knowledge and were able to act as proxies for end users of explainable security solutions in the future.

Participant expertise has been shown to play a significant role in Human-Centered XAI work as users' explanation preferences can vary based on background knowledge, domain expertise, and problem-solving heuristics. Cybersecurity analysts will have experience assessing potential risks using technical artifacts, threat intelligence, and previous experiences with phishing attacks which can influence how an analyst evaluates AI-powered recommendations. These individuals may also perceive explanation modalities differently than students or non-expert users.

Analysts might benefit more from rule-based explanations as they could easily map to an investigation process or even mirror existing security rules. Non-expert users might prefer natural language explanations because they can help to decrease the technical jargon required to understand a solution. In addition, cybersecurity experts who are familiar with reviewing screens of emails and logs to identify threats may better understand heatmap visualizations because they are familiar with parsing copious amounts of technical data.

Lastly, professionals in security settings may have different trust calibration behaviors than students. Cybersecurity professionals often operate under stressful environments and make decisions on alerts that could be false positives, cause alert fatigue, or hold them accountable for security breaches. As such, professionals in operational settings might perceive and use explanations differently than students. The evaluation setting did not replicate these organizational pressures and provided participants with greater time and flexibility to make decisions.

These limitations highlight areas of research that can be performed in the future. This evaluation provides an important initial step toward understanding how multimodal explanations influence user trust, confidence, and interpretability of solutions in cybersecurity. The evaluation should be repeated with professional cybersecurity analysts, incident responders, and security operations center employees to determine whether users' explanation preferences and trust behaviors differ.

Explanation of Design Constraints

Static explanation modes, such as rule-based systems, LLM-generated outputs, and visual heatmaps, are currently supported by the artifact without adaptive personalization capabilities. Recent studies in HC-XAI literature by Rong (2023) and Kong (2024) demonstrate that effective explanations require customization for different users. Users have varying needs for explanation types, which depend on their cognitive style, expertise level, and contextual requirements.

Subsequent versions should incorporate user profiling to tailor explanations to individual user needs, such as using visual signals for experts while providing narrative descriptions for novices. Machine learning models that utilize training data from usage patterns, combined with trust signals and interaction history, can create adaptive systems.

Evaluation Methodology

Trust, confidence, mental effort, confusion, and usability measures were used primarily based on post-task survey questions and qualitative participant feedback. These types of perceptual measures are emphasized outcome variables in HC-XAI literature because the utility of an explanation is subjective to how it is perceived, comprehended, and acted upon by the end user.

Objective performance-based metrics are needed to determine if explanations resulted in more effective operator decision-making. Users may perceive that they trust or have more confidence in the system, but this does not always correlate to higher levels of classification accuracy or effective cybersecurity decisions. Although some explanations can be clear and understandable to users, these may not have any practical performance benefits.

The current evaluation focused on assessing human-centered aspects of the artifact during an early DSR evaluation cycle. Many HC-XAI studies similarly emphasize interpretability, trust calibration, and usability as foundational requirements for effective human-AI collaboration. With that said, the current results should only be considered as user perceptions of explanation effectiveness.

Objective performance-based measures should be included in future iterations to determine if explanations had any effect on cybersecurity operator performance. This can include accuracy in phishing identification, reaction time to make decisions, reducing false-positives and false negatives, and long-term calibration of trust. Comparative experiments can also be used to evaluate if one modality of explanation leads to better performance by analysts in operational settings.

Using a balance of perceptual and objective performance measures can help better understand HC-XAI and provide stronger evidence on how multimodal explanations can affect cybersecurity decision-making.

Technical Constraints

The present dashboard's dependency on cloud-hosted APIs such as OpenAI limits its functionality within restricted or air-gapped cybersecurity settings. Subsequent iterations of the dashboard should implement on-device inference and edge-computing solutions, along with encryption methods, to address data sensitivity issues.

Multimodal Interaction Strategy

The research supported cross-modal evaluations but failed to examine how explanations could be used together or in sequence (such as combining LLM with heatmap visualizations). Research experiments are needed to determine the effects of stacking explanations instead of separating them on user understanding, trust calibration, and cognitive load capacity.

Usage Sequencing

Finally, the study did not track how users sequenced through different explanation types. Our research design allowed participants to experience various explanation formats in shuffled sequences. However, it did not include an analysis of navigation patterns between modalities or the dependency on specific explanations during decision-making processes. The area of explanation usage sequencing represents an important opportunity for future research investigations. Analyzing user patterns in sequencing and prioritizing explanations provides a deeper understanding of interpretability requirements and cognitive strategies within high-stakes decision-making environments.

Appendices

Appendix A. Design Theory Table

Table 4. Design Theory Table (Gregor & Jones, 2007)

Component	Description
Purpose and Scope	The XAI interface design needs to enable users to understand and trust AI-generated cybersecurity threat alerts.
Constructs	The explanation modes — rule-based, LLM, and heatmap — affect users' perceived trust and clarity, while influencing their cognitive effort.
Principles of Form and Function	The artifact should enable users to switch between different explanation types, ensuring that explanations align with their cognitive models and decision requirements.
Artifact Mutability	The system can be adjusted to support new methods of explanations and emerging threat domains.
Testable Propositions	Multiple explanation modalities enhance users' understanding and trust levels, while non-expert individuals tend to favor explanations provided by large language models (LLMs).
Justificatory Knowledge	Research by Gregor and Benbasat (1999), together with Gunning and Aha (2019) and Wang et al. (2020), informs the development of cognitive psychology theories. (2020), cognitive psychology theories.
Principles of Implementation	Developed through Python and Streamlit technology, with separate modules designed for each explanation method.
Expository Instantiation	Developed through Python and Streamlit technology, with separate modules designed for each explanation method.

Appendix B. Knowledge Contribution Matrix

Table 5. Knowledge Contribution Matrix (Gregor & Hevner, 2013)

Problem Domain Maturity	Solution Domain Maturity	Contribution Type	Description
Established: Research has extensively demonstrated the requirement for AI explainability and trustworthiness in cybersecurity applications.	Nascent: The field of cybersecurity lacks sufficient research and development into multimodal XAI dashboards.	Invention	The artifact introduces a unique system that merges three different explanation methods into one adaptive interface to address a missing component in HC-XAI for cybersecurity.

Appendix C. Survey Instrument (Post-IRB)

After completing their session with the HC-XAI dashboard, participants took the completed post-intervention survey. The University of Wisconsin–Oshkosh IRB authorized the survey tool that contains Likert-scale questions and open-ended queries to assess usability quality alongside explanation clarity, trustworthiness, cognitive workload, and professional utility.

Post-Intervention Survey

- Consent and IRB compliance section
- Demographics: Age, gender, experience with cybersecurity
- Dashboard Usability
- Explanation Evaluation: Rule-Based, AI-Generated, Heatmap
- Trust and Cognitive Effort
- Open-ended survey items ask participants to evaluate the dashboard in terms of clarity and trustworthiness while assessing its professional utility.

A complete item list can be reviewed in supplemental materials and through a request process.

Table 6. Post Test Survey

Construct	Survey Item
Dashboard Usability	The dashboard was easy to use.
	It was clear how to select different explanation modes.
	I felt comfortable navigating the interface.
Rule-Based Explanation	The rule-based explanation helped me understand the classification.
	The explanation was clear and concise.
	I would feel confident relying on the rule-based explanation.
AI-Generated Explanation	The AI-generated explanation helped me understand the classification.
	The explanation was clear and concise.
	I would feel confident relying on the AI-generated explanation.
Heatmap Explanation	The heatmap helped me understand the classification.
	The explanation was clear and concise.
	I would feel confident relying on the heatmap explanation.
Trust & Effort	The explanations increased my trust in the classification.
	I had to think hard to interpret the explanations.

Note: Following the Likert-based items, participants were presented with open-ended questions to provide qualitative feedback on the dashboard's clarity, usefulness, and potential areas for improvement.

Instructions for Use: Experts should complete this checklist after interacting with the artifact. Experts can evaluate items by providing scores and comments, which will assist in progressive refinement.

Appendix D. Suspicious Email (generated by ChatGPT)

Subject: Urgent: Account Suspension Notification

Body:

Dear User,

We have detected unusual activity in your account that violates our terms of service. As a result, your access has been suspended.

To restore access, please verify your credentials immediately by clicking the secure link below:
<http://account-verification-secure.com/login>

Failure to act within 24 hours will result in permanent suspension of your account.

Thank you for your prompt attention to this matter.

Sincerely,

Account Security Team

Appendix E. Rule-based Model Response to Suspicious Email

Explainable AI Dashboard for Cybersecurity Threats

Simulate detection explanations for suspicious emails using multiple modes.

☒ Enter suspicious email content:

Subject: Urgent: Account Suspension Notification

Body:

Dear User,

We have detected unusual activity in your account that violates our terms of service. As a result, your

☒ Select explanation type:

☒ Rule-based

☐ LLM (GPT)

☐ Visual Heatmap

 Analyze

Explanation

 Financial keywords detected (bank/account/login)

 Urgent language detected

 Suspicious link present

Figure 3. Rule-based Response to Suspicious Email

Appendix F. LLM (GPT) Model Response to Suspicious Email

Explainable AI Dashboard for Cybersecurity Threats

Simulate detection explanations for suspicious emails using multiple modes.

Enter suspicious email content:

Subject: Urgent: Account Suspension Notification
Body:
Dear User,

We have detected unusual activity in your account that violates our terms of service. As a result, your

Select explanation type:

- ☐ Rule-based
☒ LLM (GPT)
☐ Visual Heatmap

Analyze

Explanation

GPT Explanation:

There are several red flags in this email that indicate it may be a phishing attempt:

1. **Urgency:** The subject line and the warning of "urgent" in the email body are common tactics used by phishers to create a sense of panic or urgency in the recipient, prompting them to act without thinking critically.
2. **Unusual Activity Claim:** The email claims that there has been unusual activity in your account, but it does not provide any specific details or context about what that activity might be. Legitimate service providers usually provide more specific information to help users understand the issue.
3. **Request for Immediate Action:** The email urges you to verify your credentials immediately by clicking on a link. Phishers often use this tactic to trick users into clicking on malicious links that lead to fake login pages designed to steal their credentials.
4. **Suspicious Link:** The link provided in the email (<http://account-verification-secure.com/login>) does not appear to be a legitimate domain for the company mentioned. Phishers often use deceptive URLs to make their emails appear legitimate, but careful inspection can reveal inconsistencies.
5. **Consequences of Inaction:** The email threatens that failure to act within 24 hours will result in permanent account suspension. This is another common tactic used by phishers to pressure recipients into taking immediate action without questioning the legitimacy of the request.
6. **Generic Greeting:** The email addresses the recipient as "Dear User," which is a generic greeting that suggests the sender does not have specific information about the recipient's account.

In conclusion, this email exhibits several characteristics commonly associated with phishing attempts, including urgency, vague claims of account issues, requests for immediate action, suspicious links, threats of consequences, and generic greetings. It is advisable to be cautious and verify the legitimacy of such emails by contacting the company directly through official channels before taking any action.

Figure 4. LLM (GPT) Response to Suspicious Email

Appendix G. Visual Heatmap Model Response to Suspicious Email

Explainable AI Dashboard for Cybersecurity Threats

Simulate detection explanations for suspicious emails using multiple modes.

☒ Enter suspicious email content:

Subject: Urgent: Account Suspension Notification

Body:

Dear User,

We have detected unusual activity in your account that violates our terms of service. As a result, your

☒ Select explanation type:

☐ Rule-based

☐ LLM (GPT)

☒ Visual Heatmap

 Analyze

Explanation

Simulated Attention Heatmap

Subject: Urgent: **Account** Suspension Notification Body: Dear User, We have detected unusual activity in your **account** that violates our terms of service. As a result, your access has been temporarily **suspended**. To restore access, please **verify** your **credentials immediately** by clicking the secure link below: <http://account-verification-secure.com/login> Failure to act within 24 hours will result in permanent suspension of your **account**. Thank you for your prompt attention to this matter. Sincerely, **Account** Security Team

 Highlighted words may indicate urgency, spoofing, or phishing.

Figure 5. Visual Heatmap Model Response to Suspicious Email

References

- Adadi, A., & Berrada, M. (2018). Peeking inside the black box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- Ali, S., Shamsuddin, S. M., Salah, K., & Jayaraman, R. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy AI. *Information Fusion*, 99, 101805. <https://doi.org/10.1016/j.inffus.2023.101805>
- Alqaraawi, A., Schuessler, M., Weiß, P., & Kulesza, T. (2020). Evaluating saliency map explanations for convolutional neural networks: A user study. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). <https://doi.org/10.1145/3313831.3376218>
- De Silva, C., Halloluwa, T., & Vyas, D. (2025). A multilayered research framework for human-centered AI: Defining the path to explainability and trust. *arXiv Preprint*. <https://doi.org/10.48550/arXiv.XXXX.XXXX> (Replace with actual DOI once published)

- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126. <https://doi.org/10.1037/xge0000033>
- Ehsan, U., & Riedl, M. O. (2020). Human-centered explainable AI: Toward a reflective sociotechnical approach. *arXiv Preprint*. <https://doi.org/10.48550/arXiv.2001.00093>
- Ehsan, U., Rai, A., & Riedl, M. O. (2021). Explainability in human–AI interaction: Toward a reflective sociotechnical approach. In *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). <https://doi.org/10.1145/3411764.3445188>
- Gunning, D., & Aha, D. W. (2019). DARPA’s Explainable Artificial Intelligence (XAI) program. *AI Magazine*, 40(2), 44–58. <https://doi.org/10.1609/aimag.v40i2.2850>
- Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 37(2), 337–355. <https://doi.org/10.25300/MISQ/2013/37.2.01>
- Gregor, S., & Jones, D. (2007). The anatomy of a design theory. *Journal of the Association for Information Systems*, 8(5), 312–335. <https://doi.org/10.17705/1jais.00129>
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105. <https://doi.org/10.2307/25148625>
- Kim, J., Maathuis, H., & Sent, D. (2024). Human-centered evaluation of explainable AI applications: A systematic review. *Frontiers in Artificial Intelligence*, 7, 1456486. <https://doi.org/10.3389/frai.2024.1456486>
- Kong, X., Liu, S., & Zhu, L. (2024). Toward human-centered XAI in practice: A survey. *Machine Intelligence Research*, 1–31. <https://doi.org/10.1007/s11633-023-1403-2>
- Liao, Q. V., & Varshney, K. R. (2022). Human-centered explainable AI (XAI): From algorithms to user experiences. *ACM Transactions on Interactive Intelligent Systems*, 12(2), 1–45. <https://doi.org/10.1145/3495244>
- Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90–103. <https://doi.org/10.1016/j.obhdp.2018.12.005>