You Know Why, but Still Rely: The Impact of Explainable AI on Trust, Task Load, and Performance in Cybersecurity Decision-Making

Neele Roch
ETH Zurich

Hannah Sievers
ETH Zurich

Noé Zufferey
ETH Zurich

Verena Zimmermann
ETH Zurich

Abstract

With the increasing digitisation of institutions, the demand for effective cybersecurity measures is rising rapidly. Simultaneously, the complexity and volume of cybersecurity tasks are outpacing the capacity of available practitioners. Leveraging AI to augment human cybersecurity expertise has the potential to reduce complexities and cognitive overload. Transparent and human-understandable insights into AI decisions are not only demanded by governance authorities, such as the EU, but also by practitioners themselves when collaborating with AI in high-risk contexts. We report on a between-subjects study (N = 139) that investigated the effects of explainable AI (XAI) explanations on trust, usability, perceived task load, and collaborative task performance among users with cybersecurity domain knowledge in the context of malicious domain blocking. The provision of explanations in this context did not foster trust; in fact, users with domain knowledge reported lower trust after interaction with XAI. Qualitative results suggest that they apply their own decision-making criteria, and that exposing AI decision boundaries may introduce ambiguity and foster mistrust. Although the inclusion of XAI did not increase perceived task load, it also failed to improve performance. These findings raise important questions about the effectiveness of current XAI approaches in knowledge-centric, decision-making settings and underscore the need for more context-sensitive, user-aligned explanation strategies in cybersecurity.

1 Introduction

With the rising threat and costs of cyberattacks, cybersecurity operations are pivotal for organisations [37,62]. This rising demand is also reflected in an increasing need for cybersecurity practitioners [66]. However, trained and knowledgeable candidates are scarce. Researchers have suggested that augmenting existing cybersecurity practitioners with AI capabilities could improve efficacy, support practitioners, and help address the workforce gap [27,30,31]. However, in high-risk cybersecurity scenarios and applications, collaboration with AI raises concerns regarding transparency and accountability. Governance authorities, such as the European Union [24], along with cybersecurity practitioners [58] emphasised the need for AI transparency for responsible integration into cybersecurity contexts. It has been proposed that transparency via explainable AI (XAI) could support trust building and potentially improve human-AI joint performance [15]. Simultaneously, these explanations may support users of AI tools to recognise when AI errs, and override its recommendation.

While the technical implementation of AI into cybersecurity tasks and contexts has been well-explored, its effectiveness widely depends on the human factor. Understanding how factors, such as transparency and the related provision of explanations, impact human-AI collaboration is crucial for the deployment of trustworthy, usable, and successful AI tools in cybersecurity contexts.

Despite the potential benefits of augmenting cybersecurity tasks with AI tools, prior research has shown that domain-knowledgeable users (DKU) are prone to falling back to manual control, reflecting tendencies for algorithm aversion. Interestingly, explanations increase the trusting intention for DKUs, suggesting that explainability may support overcoming aversion tendencies [10] and facilitate successful collaboration.

Trust shapes behavioural intention, however, actual reliance on automated systems, such as AI tools, is also influenced by contextual factors such as workload, self-confidence, risk, and time pressure, underscoring the need to consider task load as an indicator for the task's mental, physical and temporal demands alongside measures of trust and reliance [42].

Previous work showed that explanations can increase task load due to the effort required to interpret them, sometimes prompting over-reliance on AI to reduce cognitive strain [15], yet explanations can also mitigate under-reliance for self-confident users [60]. Research has shown that in the case of imperfect XAI, expertise influences reliance and performance, emphasising expertise and domain knowledge as critical moderators [48]. Personal factors beyond domain knowledge

also predict trust, especially in early adoption phases [50], and perceived usefulness drives reliance among experienced users [10].

Nevertheless, findings related to performance gains from explanations are inconsistent, and effects from proxy tasks may not transfer to real-world settings, stressing the need for context-specific evaluations of explainability methods [15].

Despite prior findings on the effect of explanations on DKUs, it remains unclear whether the use of XAI fosters trust and complementary team performance. Literature on human-AI collaboration showed contrasting results, with some claiming that explanations can enhance joint performance in human-AI collaboration [15], whereas others conclude that explanations seldom lead to performance gains [30].

Consequently, the evaluation of trust in this context requires a theoretically founded investigation, entailing trust in automation and reliance behaviour. Previous research has found relationships between trust and usability, as well as performance, trust, and task load. However, the composite of these variables together remains yet to be studied and may further explain contradicting results.

Therefore, our research systematically investigates whether, in the high-risk context of cybersecurity, the provision of explanations for the AI's decision can support the collaboration of cybersecurity practitioners and AI. As a use case for our investigation, we selected the relevant and critical task of malicious domain blocking (MDB) through a two-step evaluation of various cybersecurity tasks.

This study addresses key gaps in the literature by focusing on users with domain knowledge and evaluating XAI explanations in a domain-specific cybersecurity classification task. Building on Lee and See's theoretical framework of trust in automation, we examine the holistic effects of XAI on trust, task load, performance, and usability. We thereby distinguish between attitudinal trust as expressed through personal reports and opinions, and behavioural reliance as reflected in the actual use and the decisions taken, supported by the AI tool.

Through a between-subjects experiment, we investigated the following four research questions: (RQ1) How does the use of an XAI tool for malicious domain blocking influence security-knowledgeable users' trust in the system and their reliance behaviour?, (RQ2) How does the integration of an XAI tool affect the combined performance of security-knowledgeable users and AI in detecting malicious domains?, (RQ3) How does the use of an XAI tool for malicious domain blocking influence security-knowledgeable users' perceived task load?, and (RQ4) How does an XAI tool for malicious domain blocking impact security-knowledgeable users' perceptions of the tool's usability?

We found that DKUs who received explanations had lower trust in automation scores after collaborating, which may stem from a wider range of discrepancies between the DKUs' and the AI's decision-making. Reliance behaviour, however, remained consistently high, independently of explanations being provided. Our findings suggest that AI explanations in this scenario did not offer significant utility, as DKUs were not able to rectify AI errors better when explanations were available, and no improvement was noted in the overall performance. We contribute to the field in several important ways: (1) focus on human-AI collaboration, with knowledgeable users assessing true and false AI classifications, capturing real-world decision contexts, (2) distinguish (self-reported) attitudinal trust from behavioural reliance, and use theoretically grounded constructs rather than proxy metrics, and (3) provide an integrated view of how feature importance explanations affect multiple user outcomes and their interplay, something fragmented or overlooked in previous research.

2 Related Work

This section first introduces and motivates the MDB task. It then summarises the previous inconclusive findings related to the effect of explanations on trust, performance, task load, and perceived usability in human-AI collaboration.

2.1 Malicious Domain Blocking

Malicious domains represent a significant aspect of cyberattacks, endangering internet users by facilitating the distribution of malicious services, compromising their security and privacy [43]. A broad body of research is concerned with technically detecting such domains [1, 2, 35, 63]; however, constantly evolving manipulation tactics make it difficult for technical solutions alone to succeed [3, 4]. Therefore, employing human expertise alongside AI can be crucial to the success of detecting malicious domains. Integrating XAI into cybersecurity-specific tasks is intended to support system trust and enhance users' understanding of AI decisions in this highly critical domain, where transparency and understanding are essential for mitigating risks and evaluating security decisions [23]. The choice of XAI method is particularly sensitive in cybersecurity, as inappropriate methods can inadvertently expose security vulnerabilities. Warnecke et al. [70] evaluated various XAI methods for deep learning in cybersecurity and identified Integrated Gradients (IG) and Layer-wise Relevance Propagation (LRP) as the most effective across criteria such as completeness and stability. These methods access model weights and gradients directly, offering faithful and security-aligned explanations, making them especially suitable for decision-making tasks in cybersecurity.

While technically robust explanations are essential for AI in cybersecurity, it is equally important to consider how users perceive and approach the MDB task when using technical support. Althobaiti et al. [4] evaluated various features of URLs and their usefulness for classification of URLs into *malicious* or *benign* for humans or technical systems. Extending

this work, they designed and evaluated a "URL feature report" for lay users that provided relevant information to participants by aggregating and highlighting important information and supported them in improving their MDB capabilities [3]. Further research has explored lay users' ability to detect malicious or benign URLs correctly and found that only 75% of domains were correctly classified [56], while falsely classifying 66.34% as benign, and only 33.66% falsely as malicious.

Although existing work has developed AI tools for MDB, it has largely focused on lay users, despite such systems often being used by DKUs. To address this gap, we integrate XAI to support DKU-AI collaboration in cybersecurity contexts. Building on prior findings, we adopt an instance-based explanation method suited to the MDB task and empirically examine how XAI affects DKU performance, trust, task load, and usability.

2.2 Explanations in Human-AI Decision-Making

Practitioners across domains, including cybersecurity, have expressed the need for transparent AI to enhance trust [58,65]. Cultivating appropriate trust can be crucial for technology acceptance and effective collaboration. In their foundational model of trust in automation (TiA), Lee and See (2004) defined TiA as "*the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability*" [42, p. 54]. Trust exceeding a system's capabilities is referred to as over-trust, and trust falling short of capabilities and leading to disuse is under-trust [42]. Lee and See further posit that while trust shapes behavioural intentions, actual reliance is influenced by contextual factors such as workload, self-confidence, risk, and time pressure [42]. This highlights the need to account for task load alongside measures of attitudinal trust and behavioural reliance, as they are inherently intertwined [42]. On the one hand, AI explanations may point users to the relevant parts of the AI decision, reducing cognitive effort [32]. On the other hand, explanations can also increase task load by requiring additional analysis and contextualisation [15,46], causing users to over-rely on the AI's recommendation to avoid cognitive effort [15].
Domain-oriented arguments are crucial for DKUs' adherence to AI advice, as explanations are primarily used to resolve discrepancies between users' own judgment and the system's recommendation, and to verify AI decisions [6,28].
Users that are familiar with a task tend to base their use of a system on its perceived usefulness [10], suggesting that usability is a critical, yet frequently overlooked, factor in studies of TiA. Related work further shows that perceived usefulness and ease of use can be pivotal factors of user trust across contexts [73]. AI explanations have been shown to reduce under-reliance, with self-confidence serving as a mediating factor [60]. However, findings on domain familiarity are mixed. Higher familiarity resulted in higher than average trust [59] in one study; yet showed no significant correlation with trust or performance in another [46]. Even though some studies report an increase in combined performance when providing explanations [15,40], others, including systematic reviews of existing literature, concluded that explanations seldom lead to performance gains [30,74]. For DKUs, Paleja et al. [52] found that collaborative performance decreased, whereas novices showed improved performance on the same task. This contrast highlights knowledge and familiarity as influential factors for improved performance in human-AI collaboration.
Current findings suggest that the effects of XAI on trust, performance, and task load observed in proxy tasks may not reliably transfer to real-world tasks [15]. This may help explain contradictory findings while underscoring the need for context-specific evaluations of XAI methods.
These findings show that research in this domain has not yet come to a conclusive result regarding whether AI explanations enhance or reduce trust, affect performance, alter task load, or interact with usability [15,51,52,59,61,69]. These inconsistencies may stem from the use of diverse measurement scales, variations in study designs, and a lack of theoretical clarity distinguishing trust from reliance. Some studies examined trust across different explanation types [15], while others investigated the provision of explanations versus their absence [51]. To address gaps from inconsistent prior findings, the present study systematically examines how XAI affects DKUs' trust, task performance, cognitive workload, and perceived usability in a real-world cybersecurity decision-making task, explicitly involving domain-knowledgeable users rather than lay participants.

3 Use Case Selection and Technical Prototype for MDB Task

This section describes the use case selection and technical development of the prototype used for the experimental task, integrating technical feasibility and real-world applicability.

Use Case Selection. To identify a relevant and practically applicable cybersecurity use case, two researchers reviewed all use cases reported in a literature review on AI in cybersecurity [36]. The evaluation considered (a) the feasibility and usefulness of providing transparency, (b) the independence of organisational strategy (to avoid DKUs or practitioners having to act contrary to their usual strategy), and (c) the availability of data to develop the (X)AI tool. This yielded nine use cases that were suitable for implementation in both a transparent and non-transparent condition.

These use cases were then presented to a small group of cybersecurity experts, who evaluated them based on their relevance (*high, medium, or low*) and their perceived criticality (*high or low*) via a short survey on Qualtrics. Malicious do-

main blocking and phishing detection were assessed as most critical and relevant. Because decision-making around malicious URLs, specifically allowing or blocking them, is a common task for DKUs and is also widely examined in phishing research, we selected malicious domain blocking as our use case due to its broader applicability.

Shallow Neural Network and Integrated Gradients. The shallow neural network (SNN) that classifies the URLs as malicious or benign follows Senanayake et al. [63] and was developed using PyTorch (version 2.4.0) [53], and scikit-learn (version 1.6.1) [14]. It uses features such as special character count, and number of name servers. We used IG, an XAI technique for attributing the prediction of a deep network to its input features [64], following Warnecke et al. [70] as a security-relevant explainability method with well-maintained PyTorch implementations, ensuring reproducibility and comparability. We implemented IG using the Captum library [17]. For a detailed description of the model, training parameters, and technical evaluation, see the Zenodo Repository of the MDB Prototype. To enable the empirical assessment of the technological prototype, we developed a dashboard that allowed DKUs to inspect various features of the URLs. The dashboard design was informed by Althobaiti et al. [3] to ensure alignment with previously evaluated URL-classification dashboards. The dashboard was implemented with a Python backend (version 3.11.11) [54] and a React frontend. A detailed description of the interface is provided in subsection 4.1.

4 Method

The independent variable in this between-subjects design was whether the AI provided an *explanation* of the features that contributed to its recommendation or offered *no explanation*, representing a standard opaque baseline. For clarity, we refer to the explainable condition as *XAI* and the non-explainable condition as *AI-only* throughout the remainder of the paper. In both conditions, participants were shown the AI's classification for each URL, which we refer to as the *AI classification*.

Our experiment assessed multiple dimensions of human-AI interaction through several dependent variables. *Performance* reflects the accuracy of the human-AI team, with decisions considered correct if they matched the ground truth for the URL. *Trust in automation* captures participants' belief that an AI agent will help achieve their goals under uncertainty and vulnerability [42], measured using the 19-item *Trust in Automation (TiA)* scale [39]. The TiA scale, based on Lee and See's model of trust in automation [42], covers perceived reliability, familiarity, trust, understanding, and developer intentions on a five-point Likert scale. *Reliance* measures how frequently participants' decisions aligned with the AI's recommendations, serving as a behavioural component for trust and complementing the self-reported TiA from the post-task survey, commonly done in XAI evaluations [38, 40]. *Perceived task load* reflects participants' subjective mental effort, measured using the Rating Scale Mental Effort (RSME) [75], rated after each task, and the *NASA Raw Task Load Index (NASA-RTLX)* [29] (without the physical demand dimension). Finally, usability of the AI dashboard was measured with the *System Usability Scale (SUS)* [13], capturing effectiveness, efficiency, and satisfaction on a five-point Likert scale [34]. Explanation-specific usability in the XAI condition was additionally assessed with the *Explanation Satisfaction Scale (ESS)* [33].

The pre-screening and the main study surveys were administered using ETH Zurich's Qualtrics instance [55], and participants were recruited through the online panel provider Prolific.

4.1 Procedure

This section describes the pre-screening and main study procedure for both conditions (illustrated in Figure 1).

Pre-Screening. The study began with a pre-screening (see OSF Project for details). This enabled targeting DKUs in cybersecurity, verified through a pre-screening on education, work experience, and current relation to cybersecurity. The DKU phrasing aims to reflect participants' familiarity with the cybersecurity domain without overstating their expertise related to the selected MDB task. While some participants may have been practising cybersecurity professionals, we adopt this more conservative terminology to ensure that claims about generalizability remain well-grounded.

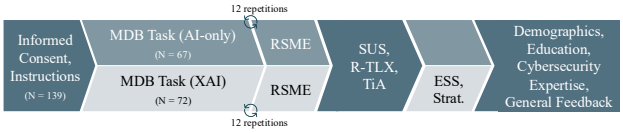

Figure 1: Main study procedure for both the XAI and AI-only condition.

Informed Consent and Instructions. All participants received a written explanation of the study and were asked for their informed consent. Following this, participants received instructions on the MDB task. Participants were tasked with deciding whether to allow or block 12 URLs based on provided information on the dashboard. They were randomly assigned to either the control (AI-only) or experimental (XAI) condition and forwarded to the respective version of the dashboard.

Repeated Tasks. The MDB task asked DKUs to classify a domain (e.g., `http://malicious.domain`) as malicious or benign using the dashboard. Both conditions received different versions of the dashboard. The dashboard structure consisted of four main components: *(A) the AI classification*,

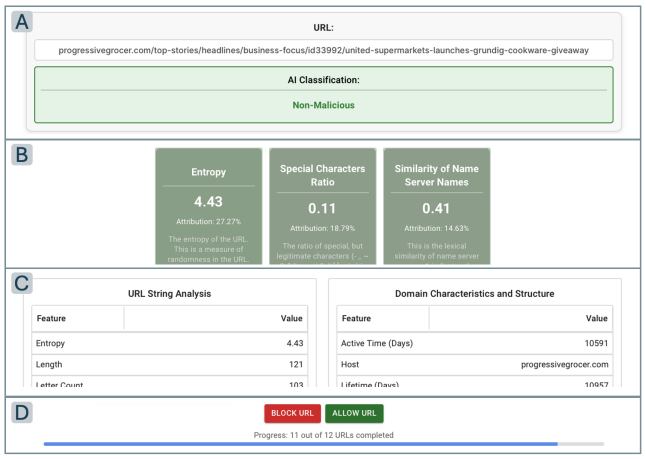

Figure 2: Dashboard screenshot, showing the (A) AI classification, (B) AI explanation, (C) data and (D) decision component.

(B) the explanation (only shown to the XAI condition), *(C) the data*, and *(D) the decision components*. The components of the dashboard are shown in Figure 2, and screenshots of the full dashboards are shown in the OSF project.

The *AI classification component (A)* contained the analysed URL as text, and the SNN's classification of that URL (i.e., malicious or non-malicious), colour coded (red or green), depending on the classification.

Below this was the *explanation component (B)*, which was only shown to the XAI condition. The IG attributions for each URL were used to dynamically display the three most relevant features for the SNN's classification as cards in the dashboard. To make the attributions of each feature understandable to the user, we standardised the outputs of the method. Each card consisted of the feature name, its value, and a percentage-wise attribution of that feature to the classification. We also provided a brief textual explanation of each feature and its typical behaviour for malicious or non-malicious URLs, based on recent work [63]. The cards colour intensity varied respective to their standardised attribution.

The *data component (C)* was shown to both conditions and contained the full report of the extracted features from the URL in a tabular format. Six tables contained clustered information: URL string analysis, domain characteristics and structure, DNS and network information, encryption and HTTP response, webpage content and structure, and geographical and hosting information.

The *decision component (D)* at the bottom of the page consisted of two buttons ("Block" and "Allow"), and a progress bar to indicate how many task repetitions were left.

The task was repeated 12 times: four correctly classified malicious, four correctly classified non-malicious, two falsely classified malicious, and two falsely classified non-malicious URLs. The false cases were included to examine whether explanations help DKUs correct AI errors. The order of URLs was randomized for each participant.

For each task repetition, we collected the participant's decision, mental effort, and decision time, allowing us to derive participants' correctness (alignment with ground truth) and reliance (alignment with AI decision).

Post-Task Survey. Following the 12 task repetitions, the participants were redirected back to the Qualtrics survey where they first answered the SUS, then the NASA-RTLX followed by the TiA scale. Participants in the XAI condition were asked to describe how they used the explanations in their decisions and completed the ESS. All participants were asked to provide demographic information, including gender and educational background. We repeated the questions from the pre-screening to ensure consistency with these answers. Additionally, we asked participants about their familiarity with AI and their use of AI in their current organisation. Lastly, we asked all participants for voluntary general feedback, giving participants the opportunity for additional feedback. Details of all scales and questions can be found in the OSF project.

4.2 Participants

A Priori Power analysis. We determined the required sample size through an a priori power analysis using G*Power [26]. For our t-test analysis to have 95% power to detect an effect of 80%, with a significance level α of .05, at least 42 participants were required in each of the two conditions. For the MANOVA analysis to have 95% power to detect a medium effect ($f_2 = 0.2$), with a significance level α of .05 and two response variables, at least 108 participants were required. We estimated an additional 10% for dropouts and outliers, resulting in at least 120 participants.

Recruitment. Participants were recruited online through Prolific and the pre-screening for eligibility for our study was conducted on the 2nd of April 2025. The main study was open for eligible participants from April 3rd through April 7th, 2025. Compensation was £0.30 for the pre-screening for an expected completion time of 2 mins, and £3 for the main study with an estimated completion time of 20 mins.

Inclusion Criteria. Through the pre-screening, we selected participants who had at least two years of experience in cybersecurity and were currently (at least partially) working in a cybersecurity role. Additionally, participants were filtered through a knowledge test, asking them a general cybersecurity-related question about best practice for encryption of data at rest, which should be easily answered by knowledgeable users in the cybersecurity domain. Participants who fulfilled these criteria were admitted to the main study. The screening questions were again included in the main study, and the same criteria were applied. Participants who did not fulfil the criteria were excluded

from the study. Additionally, the main study contained two attention checks that participants needed to pass. The pre-screening was completed by n=750 participants, of which n=220 were eligible and admitted for the main study. The main study was completed by n=195 participants. After excluding participants based on the previously described criteria, the final sample consisted of N=139 complete responses.

Participant Characteristics. The final sample (N=139) for the main study included 67 participants in the AI-only condition and 72 participants in the XAI condition. 53 participants had a cybersecurity role, and 86 participants viewed cybersecurity as part of their role. Qualitative analysis of all participants' role descriptions showed that most participants (87%) held technical positions such as cybersecurity roles, software development, IT management or support, which naturally involve security-related responsibilities, e.g., managing access controls, or applying secure coding practices, reflecting practical cybersecurity knowledge [7]. The remaining participants worked in related fields, namely research, finance, product management, or data and analytics roles, indicating that cybersecurity responsibilities were part of their role. Table 1 lists all participants' role categories, including examples and number of occurrences.

Code	Examples	No. occ.
IT Management & Support	IT Support, Systems Administrator, Database Manager	41
Executive & Leadership	IT Project Lead, Director of Technology, Chief Technology Officer	30
Cybersecurity	Security Operations Analyst, Digital Forensics Officer, Cloud Security Engineer	29
Software Development & Engineering	Software Developer, Web Developer, Senior Systems Engineer	22
Data & Analytics	Business Analyst, Data Analyst	11
Product Management	Procurement, Innovation Lead	5
Finance	Financial Analyst, Accounts Manager	5
Research	Scientist	1

Table 1: Codes, example descriptions, and no. of occurrences for participants' reported roles within their organizations. Participants sometimes listed multiple roles, resulting in code frequencies exceeding the total number of participants.

The majority of participants (57%) had 3 to 5 years of experience in the field of cybersecurity, and 23 participants had more than 10 years of experience. The participants mostly held academic degrees, with 72 having completed a bachelor's degree, 48 a master's degree, 16 a PhD, 2 having visited a university but not received their degree, and only one person having completed an apprenticeship. Ages were grouped as follows: 18-24 (6%), 25-34 (34%), 35-44 (27%), 45-54 (18%), 55-64 (9%), and 65+ (6%). The majority of participants had an educational background in computer science, engineering, mathematics, or a combination of these. Almost 90% of participants indicated being extremely (40%) or very (48%) familiar with AI. For details, see Appendix A.

5 Results

In the following, we first present the results in line with our research questions and hypotheses from the pre-registration. Then, we report on additional exploratory analyses. The data and R scripts used for the analysis are provided on the OSF project as part of open science. All quantitative data analysis was conducted in R version 4.3.2 [57]. For our exploratory data analysis, we calculated regression models using the *lme4* package [9].

We set the significance threshold at $\alpha = .05$ for all tests. Confidence intervals (CIs) are reported for most analyses, except for the pre-registered MANOVAs on trust, performance, and task load, due to their multivariate nature. CIs provide a range of likely values for the population parameter, comprising the point estimate and margin of error, and offer insight into the potential effect size. Importantly, a CI that includes zero corresponds to a non-significant p-value.

For the pre-registered MANOVAs, we used Pillai's Trace, which is robust to violations of normality and homogeneity of variance [8]. Usability differences between conditions were evaluated with a t-test, and all univariate results were corrected for multiple comparisons using the Benjamini-Hochberg method [11].

5.1 Observing Attitudinal Trust and Behavioural Reliance

Concerning trust, the MANOVA indicated a significant multivariate effect of the experimental conditions, i.e., AI-only as compared to XAI, on trust in automation and reliance, $F(2, 136) = 3.70$, $p = .027$ ($\eta^2_{TiA} = 0.05$, $\eta^2_{reliance} < 0.001$). η^2 suggests a small effect on the attitudinal TiA, and a negligible effect on the participants' reliance [21]. Participants in the AI-only condition reported higher TiA scores ($M = 71.18$, $SD = 8.07$) compared to those who received explanations in the XAI condition ($M = 67.04$, $SD = 9.86$) post-task. Participants in the AI-only condition relied on the AI's recommendation for an average of 9.75 ($SD = 1.69$) decisions, while those receiving additional explanations averaged 9.69 ($SD = 2.12$). This finding suggests that, in our study context, providing explanations for the AI's classifications was associated with lower trust among DKUs.

Following the multivariate tests, we conducted exploratory independent-samples t-tests on individual measures of TiA and reliance. Participants in the AI-only condition reported significantly higher TiA than those in the XAI condition, $t(137) = 2.70$, 95% CI [1.10, 7.17], $d = 0.46$. In contrast, reliance on the AI's recommendations did not differ between conditions, $t(137) = 0.16$, 95% CI [-1.00, <0.01]. This pattern

suggests a mismatch between post-task attitudinal trust and behavioural reliance, whereby participants' reliance behaviour remained consistent despite differences in self-reported trust. As we were interested in how the trust scores could be explained, we fitted a linear model (estimated using OLS) to examine the effect of usability, AI familiarity, task load (effort and performance), expertise, and demographic variables (age, gender, and highest education) on TiA. Prior literature has found these variables to be relevant to explaining trust in human-AI collaboration [72]. The model was significant, $F(21, 117) = 7.10$, $p < .001$, and explained around 48% of the variance in TiA ($R^2 = .56$, $adj.R^2 = .48$). Detailed results can be found in Table 2.

Variable	β	SE	95% CI	p
Intercept	**41.68**	**4.10**	**[33.56,49.79]**	**<.001**
Task Load Effort	-0.05	0.03	[-0.10,0.00]	.066
Task Load Perf.	**0.11**	**0.04**	**[0.03,0.18]**	**.008**
AI Familiarity	**8.84**	**3.47**	**[1.97,15.71]**	**.012**
Usability	**0.24**	**0.04**	**[0.15,0.32]**	**<.001**
Full-Time Practitioner	1.89	1.09	[-0.26,4.05]	.084
Experience	1.81	1.42	[-1.00,4.62]	.204
Gender [Male]	1.52	1.29	[-1.03,4.07]	.241
Age	-1.77	2.45	[-6.61,3.08]	.472
Highest Ed. [Master]	**2.76**	**1.39**	**[0.00,5.52]**	**.050**
Highest Ed. [PhD or higher]	0.58	2.03	[-3.44,4.59]	.777
Highest Ed. [None]	-3.25	5.01	[-13.17,6.66]	.517
Highest Ed. [Apprenticeship]	-8.26	7.05	[-22.22,5.71]	.244
Full-Time Pract. x Experience	**4.84**	**1.69**	**[1.37,8.31]**	**.007**
Observations	139			
R^2 (adj.)	.56 (.48)			

Note: AI Familiarity, Full-Time Practitioner, Experience, Age, and Highest Education were ordinal; Gender was a factor.

Table 2: Linear regression for TiA predictors [72]. Reported: β, SE, 95% CI, and Wald t p-values. Bold indicates 95% CI significance.

The results indicate that TiA can be explained through a variety of factors. The SUS score, as well as the task load performance dimension, had a significant positive effect, indicating that higher perceived usability and better perceived task-related performance are associated with increased trust. Familiarity with AI was also positively associated with TiA, whereas cybersecurity experience did not have a significant effect on trust in general. In other words, participants with more cybersecurity experience did not report higher or lower trust than those with less experience.

Key Findings (RQ1 - Trust & Reliance)

- Our findings suggest that providing explanations coincided with lower reported trust in automation among DKUs.
- There existed a gap between reliance and post-task attitudinal trust of participants, where their behaviour differed from their general trusting attitude.
- Higher usability and better perceived task-related performance were associated with increased trust.

5.2 The Effects of Engagement and Reliance on Performance

Regarding performance, the MANOVA indicated no significant multivariate effect on the average decision time and the total correctness score, $F(2, 136) = 1.30$, $p = .276$. The findings indicate that the effect of AI explanations on the combined measures of decision time and correctness was not statistically significant. This suggests that participants in both conditions required an equivalent amount of time to make a decision and demonstrated a comparable level of accuracy. As a univariate follow-up, we tested whether the individual measures differed between the conditions. As our data violated the normality assumption, we fell back to the more robust non-parametric Wilcoxon rank sum test instead of the t-test. The Wilcoxon rank sum test indicated that there was no significant difference between correctness for either condition, $W = 2735$, 95% CI $[< 0.00, 1.00]$, suggesting that neither condition was able to make more correct decisions than the other. The AI-only condition made an average of $M = 7.78$ ($SD = 1.22$) correct decisions, and the XAI condition an average of $M = 7.42$ ($SD = 1.44$) correct decisions. Figure 3 shows that participants' decisions closely followed the AI's correct or incorrect classifications, reflecting comparable behaviour across conditions and substantial reliance on the AI's recommendations.

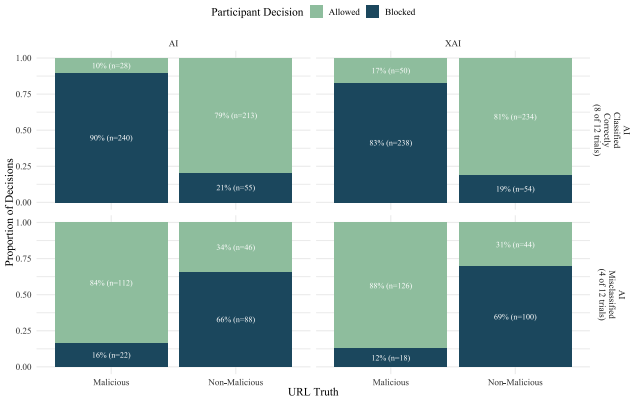

Figure 3: Participant decisions (Allow vs. Block) by URL truth (Malicious vs. Non-malicious), AI accuracy (Correct vs. Incorrect classification), and system type (AI-only vs. XAI).

We further analysed participants' decision times to evaluate their engagement with the explanations and information. Participants in the AI-only condition spent on average 37.50 seconds per decision (SD=25.87), while participants in the XAI condition spent 35.88 (SD=23.85). We additionally fitted a linear mixed model (estimated using REML and nloptwrap optimiser) to investigate whether time spent on trials differed by condition and changed over time. The model included condition, task position, and their interaction as fixed effects, and a random intercept for participants. Results indicated a

significant effect of position (β = -2.64, *SE* = 0.38, *t* = -6.86), suggesting that participants of both conditions spent less time on later tasks. There was no significant effect of condition (β = 6.30, *SE* = 5.46, *t* = 1.15), indicating that explanations did not affect decision-making time. Participants in both conditions dedicated a similar amount of time to each decision. However, a significant negative interaction between condition and position, i.e., task repetition, was found (β = -1.22, *SE* = 0.53, *t* = -2.28). This suggests that the rate of change differed by condition (detailed in Appendix B, Table 6), and that participants in the XAI condition reduced their decision-making time faster over time compared to the AI-only condition. However, upon visual inspection, shown in Figure 4, this effect seems minor.

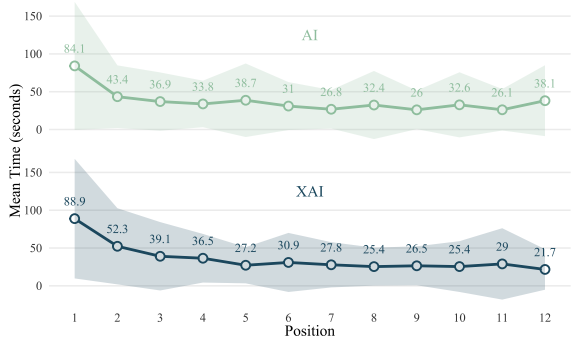

Figure 4: Interaction plot showing average time spent per decision across task positions (1-12), separately for AI-only and XAI conditions. Points represent mean decision times, with error bands indicating ±1 standard deviation.

To deepen our understanding of whether participants in either condition were able to correct the AI when providing false classifications, we performed an exploratory logistic mixed effects analysis of the relationship between correctness and condition and true or false AI classification, i.e., AI correctness. As fixed effects, we entered condition and AI correctness (with an interaction term) into the model, and included random intercepts for participants. The results (shown in Table 3) indicate an effect of AI correctness, but not of condition (AI-only vs. XAI) on the participants' correctness. Participants in the XAI group did show higher correctness; however, participants were significantly more likely to be incorrect when the AI's classification was also incorrect. This can also be observed in Figure 3, when comparing the top and bottom row. These findings indicate that DKUs in our study over-relied on the AI's classification and underscores the strong association between AI errors and participant performance, with false AI recommendations being associated with a higher likelihood of incorrect human decisions. On average, 83.19% of our participants' decisions were correct if the AI made a correct classification, and on average, only 23.38% of the participants' decisions were correct when the

Variable	β	*SE*	95% CI	p
Intercept	**1.70**	**0.12**	**[-1.93, -1.46]**	**<.001**
Condition [XAI]	-0.18	0.16	[-0.50, 0.13]	.253
AI Cor. [Inc.]	**-2.78**	**0.18**	**[-3.14, -2.41]**	**<.001**
Cond. [XAI] x AI Cor. [Inc.]	-0.03	0.26	[-0.53, 0.47]	.907
σ^2 (τ_{00})	3.29 (0.00)			
N	139			
Observations	1668			
R^2 (margi. / cond.)	.346 / .346			

Note: Condition and AI Correct were contrast coded. "Cor." stands for "Correct", and "Inc." for "Incorrect".

Table 3: Logistic mixed-effects regression of system condition (AI-only vs. XAI) and AI classification accuracy (correct vs. incorrect) on participant decision correctness. Reported are unstandardized coefficients (β), SEs, 95% CIs, and Wald t p-values; bold coefficients indicate 95% CI significance..

AI classification was incorrect. To examine the relationship between reliance and correctness overall, we performed an exploratory logistic mixed effects regression analysis of the relationship between the repeated measures (per decision) correctness and reliance, and condition. As fixed effects, we entered reliance and condition (with an interaction term) into the model. As random effects, we had intercepts for participants. The results (shown in Table 4) suggest a significant positive effect of the reliance but not of condition (AI-only vs XAI), nor of their interaction on the participants' correctness. The participants who relied on the AI were significantly more likely to be correct compared to those who did not ($\beta = -1.02$, SE = 0.18, z = -5.52, 95% CI $[0.66, 1.38]$). This suggests that, in general, reliance on the AI's recommendation was associated with increased accuracy.

Variable	β	*SE*	95% CI	p
Intercept	-0.20	0.16	[-0.12., 0.52]	.223
Reli. [True]	**1.02**	**0.18**	**[0.66, 1.38]**	**<.001**
Cond. [XAI]	-0.32	0.23	[-0.77, 0.13]	.165
Cond. [XAI] x Reli. [True]	0.24	0.26	[-0.27, 0.74]	.358
σ^2 (τ_{00})	3.29 (0.00)			
N	139			
Observations	1668			
Marginal R^2	.06			

Note: Condition (Cond.) and Reliance (Reli.) were contrast coded.

Table 4: Logistic mixed-effects regression of system condition (AI-only vs. XAI) and reliance on participant decision correctness. Reported are unstandardized coefficients (β), SEs, 95% CIs, and Wald t p-values; bold coefficients indicate 95% CI significance.

Key Findings (RQ1 - Reliance)

- Whether the human makes a correct decision largely depends on whether the AI correctly classifies or misclassifies URLs, indicating that participants relied on the AI's recommendation during decision-making.
- DKUs in our study over-relied on the AI's classification, underscoring the strong association between AI errors and participant performance, with false AI recommendations being associated with a higher likelihood of incorrect human decisions.
- Reliance on the AI's recommendation was associated with increased accuracy.

Key Findings (RQ2 - Performance)

- Participants in both conditions required an equivalent amount of time to make a decision and demonstrated a comparable level of accuracy.
- Participants in the XAI condition reduced their decision-making time faster over time compared to the AI-only condition.

5.3 Interrelations Among Trust, Task Load, and Usability

Examining the task load, the MANOVA indicated no significant multivariate effect of condition on the task load dimensions (mental effort, mental demand, effort, frustration, and performance) and the averaged repeated measure of mental effort, $F(6, 132) = 1.92$, $p = .082$. We included all dimensions of the NASA-RTLX as separate variables in the MANOVA, as the aggregation of these dimensions is not meaningful for analysis [12]. The results suggest that the participants in both conditions experienced similar levels of task load and average mental effort during the tasks. Explanations, therefore, neither increased nor decreased the task load and mental effort of participants during the decision task. Individual univariate analysis confirmed this finding (details in Appendix C, Table 7).
An independent-samples t-test was pre-registered and conducted to compare usability ratings between the AI-only and XAI conditions. Participants in the AI condition (M = 76.53, SD = 14.40) rated the usability of the MDB tool significantly higher than participants in the XAI condition ($M = 70.24$, $SD = 17.90$), $t(137) = 2.30$, 95% CI $[0.80, 11.76]$, with a moderate effect size (Cohen's $d = 0.38$) [21]. Participants using the AI dashboard that provided explanations, therefore found the tool less usable.

As literature suggests that aspects of usability, i.e., perceived usefulness and ease of use, can be pivotal factors of user trust in different contexts [73], we conducted further exploratory analyses on the relationship of these factors. Both TiA and usability scores from participants were higher when no explanations were provided. Additionally, it can be posited that the task load may have a direct bearing on perceived usability. As the present study contradicts the conclusions of Longo [45], who found that the task load does not impact the usability, we were interested in exploring the relationship between the aforementioned variables in this specific high-risk cybersecurity context, and consequently conducted a correlation analysis (see Appendix D, Figure 6). The Pearson correlation showed that TiA was significantly positively correlated with usability, $r(137) = 0.62$, 95% CI [.50, .71], indicating that participants who trusted the automated system also tended to find it more usable. Additionally, there was a strong negative correlation, $r(137) = -0.59$, 95% CI [-.69, -0.47], between the usability and the frustration dimension of the task load measure. Participants who found the system harder to use felt more frustrated, irritated, or stressed, and higher frustration levels went hand-in-hand with lower usability ratings, r(137) = -.50, 95% CI [-.61, -.36]. Self-perceived task performance was positively correlated with both TiA, r(137) = .42, 95% CI [.28, .55], and usability, r(137) = .38, 95% CI [.23, .52]. Participants who perceived themselves as performing better on the task were more likely to trust the AI tool and to rate the system as easier to use. Performance was also negatively associated with frustration, r(137) = -.44, 95% CI [-.57, -.30]. In short, self-perceived performance, trust, and usability increased together, while frustration negatively impacted all three.

Participants that had seen the explanation component were asked about their satisfaction with the explanations on several dimensions, visualised in Figure 5. Participants found explanations easy to use ($M = 4.00$, $SD = 0.89$), rated them as useful ($M = 4.07$, $SD = 0.83$) and found it enhances their understanding of the AI tool ($M = 4.17$, $SD = 0.84$). Most dimensions have a mean close to 4, indicating satisfaction with the explanations on the captured dimensions.

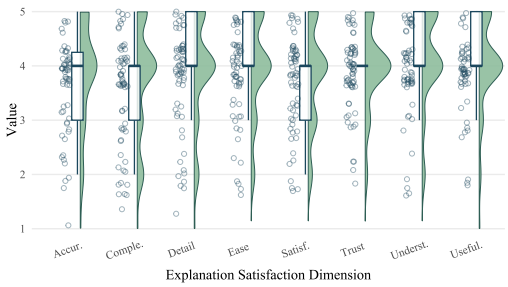

Figure 5: Distribution of participants' ratings of the explanation across eight dimensions (1-5 scale): accuracy, completeness, detail, ease of understanding, satisfaction, trust, understanding, and usefulness. Violin plots show response density, boxplots show central tendency, and points represent individual ratings.

Key Findings (RQ3 - Task Load)

- Explanations neither increased nor decreased the task load and mental effort of participants during the decision-making task.

Key Findings (RQ4 - Usability)

- DKUs using the AI tool that provided explanations found the tool less usable.
- Self-perceived performance, trust, and usability increased together, while frustration negatively impacted all three.

5.4 Decision Strategies and Domain-Knowledge Verification

The qualitative data was analysed by two researchers using MAXQDA24 (version 24.9.1) [67], following Clarke and Braun's [20] approach for thematic analysis. One researcher inductively generated codes from participants' responses regarding how they used the explanations and the AI's recommendations in their decision-making. The second researcher, then, verified the fit of the codes to the coded segments and flagged discrepancies. These were then discussed until resolved, which resulted in the final codebook and coded segments. The final codebook, incl. number of code occurrences, can be found in Appendix E, Table 8.

Qualitative analysis revealed that participants in the XAI condition employed varying strategies to make their decision on allowing or blocking a URL. Many DKUs considered the explanations that were provided during their decision-making as a starting point for their decision "*I took it as a starting point, and sanity checked against the URL itself*" (P01), or "*to guide my choice in allowing or blocking a URL by reviewing the factors it highlighted as influential.*" (P02).

However, often participants additionally had their own indicators that they examined further before making a decision, such as "*main signals were uptime of the host, country and SSL cert*" (P03), "*special characters, entropy, and mail records*" (P04), or "*domain age too, the longer it has been active, the more trustworthy*" (P05). They cross-checked their decision with the additional tabular data provided to them in the dashboard, thereby frequently relying on their own intrinsic knowledge "*I ended up relying on the raw data to inform my decision - namely whether it used SSL and the lifetime of the site.*" (P06)

Others employed a protective strategy for recommended malicious domains and decided to always protectively block a URL that the AI had flagged as malicious, since "*If it was malicious I would already prefer to preemptively block the URL, it is better to be cautious in that case*" (P07). This strategy differs from behaviour observed among lay users, who classified a domain as malicious only when they were confident it was indeed malicious [56].

A few DKUs found the AI's explanations to be inconsistent or not align with their indicators for making a decision and instead decided not to or only rarely use the AI or the explanation components, and the recommendation provided to them in the dashboard, e.g., "*I mostly relied on my own experience. As soon as I saw how unreliable the tool was, and what poor metrics it was using (e.g., existence of MX records is a very poor way to judge the validity of a website domain) I started to ignore it*" (P08).

A small number of DKUs said they "*trusted the AI*" (P09) recommendation and at times relied on its recommendation to make their decision "*I followed the AI explanation judgement to make mine*" (P10). One participant explained that they "*overall trust the AI because the information provided was simple and easy to understand*" (P11).

To further verify cybersecurity knowledge of our participants we analysed their decision strategies more deeply from qualitative responses obtained from the XAI condition. More than half of the participants applied either feature-based (e.g., encryption, lifetime of domain) or heuristic reasoning strategies, e.g., "*With these red flags, blocking it seems like the safest choice*" (P12) for their final decision. Lambe et al. [41] found that as experts accumulate domain knowledge, they also develop domain-specific heuristics, which can lead to faulty decisions even in high-stakes contexts. Additionally, analysis revealed that participants could frequently causally explain their decision-making, e.g., "*I looked at factors like low entropy, the absence of unusual characters, and the lack of suspicious mail records, which suggested the URL is likely safe*" (P13), further underscoring domain understanding.

6 Discussion

In this section, we summarise our main findings, discuss their implications for research and then their implications for XAI design, and conclude with limitations of our study and avenues for future work.

6.1 Research Implications

This section addresses contributions to understanding of collaboration between humans and AI, methodology, and future research directions.

Impact of Explanations on DKU Trust in Cybersecurity XAI-Supported Decision-Making. Aiming to answer our first research question on how providing explanations for AI's decisions may influence the DKUs' trust, we found that DKUs interacting with XAI exhibited lower trust in automation post-task than DKUs interacting with AI-only. This aligns with Bayer et al.'s [10] finding that domain-specific expertise negatively affects trust in AI-supported decisions related to this domain. However, to leverage the potential of human-AI collaboration and have practitioners use AI tools long-term, trust is vital for acceptance of the technology [42].

Drawing on the decision-making strategies from our qualitative data, a reason for the lower TiA in the XAI group may be that the explanations open up more potential for discrepancies between the DKUs' expectations and the AI's actual behaviour. Chen et al. [19] found that, the process to override an AI prediction in AI-assisted decision-making is initiated by a disagreement regarding the outcome of the decision, which then leads the human to identify evidence in the AI explanations that discredits its prediction. In our case, in the AI-only group, DKUs might compare their own decision to the AI's classification, and if they align, this could foster trust. In contrast, if the AI's classification and the DKUs' decisions do not align, this could lead to under-trust and -reliance. DKUs who interacted with the XAI, however, not only compare their decision to the AI's classification, but additionally compare the indicators for their own decision with the indicators that are provided as an explanation for a specific decision. This may reveal additional discrepancies between their decision-making and the AI's, thereby fostering additional distrust. These findings are also supported by Wirtz et al. [71] who similarly found that a mismatch between a user's expectation and the reality of AI applications can negatively affect trust. Wang et al. [69] further found that although feature contribution explanations improved users' subjective understanding of the AI model, these explanations were not able to foster appropriate trust.

Future work could explore the decision-making strategies of cybersecurity practitioners when collaborating with AI, and how these strategies influence TiA. Such insights could help to guide the selection of XAI methods. Although we selected IG thoughtfully, it may not have been the most suitable approach for this context, and alternative methods, such as decision trees or example-based explanations, may better align with practitioners' reasoning processes. The alignment of these XAI methods with the practitioner's decision-making in cybersecurity should then be evaluated through an observational study. Simultaneously, however, the technical evaluation of XAI methods beyond the feature contribution explanations evaluated by Warnecke et al. [70] for their fit in cybersecurity contexts arises as a crucial future research contribution.

Overall, we conclude that the provision of feature contribution explanations did not support building trust in this cybersecurity decision-making task, causing DKUs in our study to trust the AI less if it provided explanations about its decision-making.

The Mismatch of Attitudinal Trust and Reliance Behaviour. Our findings showed that DKUs in the XAI group had lower TiA than the AI-only group; however, they relied on the AI recommendation equally in both conditions. Although related literature suggests that users over-rely on AI to avoid cognitive effort, we did not observe an increase in cognitive effort in our data and therefore cannot attribute the observed over-reliance to this mechanism [15].
Founded in Lee and See's theory on TiA, we can observe a difference in the attitudinal aspect, which affects the trusting intention and then subsequently the reliance action [42]. Nevertheless, the change in attitude was not reflected in the reliance behaviour of DKUs. A longitudinal study on the effects of introducing an AI system in an organisation found initial evidence for a positive relation of trust with reliance, which diminished after ten months [50]. These findings suggest that trust and reliance may be related initially but can decouple as users gain experience. Our exploratory linear regression suggests that familiarity with AI, as well as domain knowledge, played an important role in explaining the DKUs' trust. The more familiar the DKUs were with AI, the higher their declared trust in automation, confirming prior work [72]. Works examining the introduction of an automated intelligent system also found that personal factors were the strongest predictors for trust in the early stages of interaction, while system characteristics become more important in later stages [50]. Chancey et al. [18] found that if people trust a system, they might, especially in high-stakes scenarios, still follow its recommendations even if it gives false recommendations.

Considering our findings alongside the related literature, we conjecture that the observed decline in attitudinal trust in automation may eventually be reflected in how DKUs rely on explainable interfaces. This would mean that users of XAI will rely on the tool less over time, given that they trust the explanations less. Longitudinal research could explore this carry-on effect of the attitude into the trusting intention and then the reliance behaviour on automation in presence of explanations.

The Role of Knowledge and Familiarity in AI Reliance. We also consider reasons for our DKUs over-reliance, when literature has frequently found that knowledgeable users often under-trust and under-rely. For example, research on algorithm appreciation has found that the experts' tendency to adhere to their prior judgments and their failure to use AI for decision support, i.e. under-reliance, led to a lowered accuracy compared to the lay user sample [44].
One plausible explanation lies in the characteristics of our participant population. Cybersecurity practitioners and DKUs work in a highly technical domain and often have a technical background, also evident in our sample (more than 70% have an educational background in computer science, see Table 5). Hence, they might be able to more intuitively understand or even have advanced knowledge about automation and AI. Additionally, we see that over 85% of our participants indicated being at least "very familiar" with AI, which indicates frequent usage and high exposure to this technology. Their educational background, the technical field they are working in, and the reported high familiarity

with AI possibly makes the participants more willing, at least initially, to trust AI-based technologies that they interact with, which could have led the participants of this study to over-rely. Theoretical work further suggests that initial trust is shaped by previous experiences, and individual characteristics, such as propensity to trust [42], which may have contributed to the high levels of reliance observed in our study. Complementary qualitative research on the effects of AI literacy and metaknowledge revealed that high metaknowledge, i.e., being acutely aware of one's own knowledge boundaries, better equips users to critically evaluate the outputs of GenAI [22]. The DKUs experience in this technical field and their educational backgrounds might make them perceptive to their knowledge boundaries in this domain. This may, over time and through feedback loops, then be able to improve reliance and performance in human-AI collaboration.

Another, and potentially complementary explanation for participants' over-reliance relates to their level of domain knowledge and expertise. Although participants were screened for general cybersecurity knowledge, not all of them were experts and therefore may have lacked the knowledge depth, skills, and intuition that experts typically possess, contributing to their increased reliance on the AI system. This relative lack of domain-specific competence may have increased their reliance on the AI system. Furthermore, because we recruited based on general cybersecurity rather than task-specific knowledge (e.g., MDB), it is possible that some participants relied heavily on the AI due to uncertainty or gaps in their own knowledge rather than misattributed trust.
Future work could explore how technical affinity and true domain knowledge interact and impact trust in automation to enable building appropriate trust in domains such as cybersecurity. Further investigating how trust changes over a longer period of time and whether DKUs and practitioners in technical domains are able to adjust their initial over-reliance could further advance human-AI collaboration.

6.2 Design Implications for XAI

This section outlines how our findings can inform the development of XAI systems.

Managing Behavioural Over-Reliance for Human-AI Collaboration in Cybersecurity. Merely one-fourth of participants were able to make a correct decision when the AI did not already provide a correct classification. As clearly visible in Figure 3, DKUs over-relied on the AI in both conditions. Through our exploratory regression analysis, we found that decision correctness was largely related to the correctness of the AI's classification; DKUs were more likely to be correct if they relied on the AI's recommendation. In both conditions, the DKUs' reliance frequently extended to incorrect AI decisions. The provision of explanations, as a typical means to reduce information asymmetry between humans and AI [68], thus did not avoid over-reliance and misuse [42]. DKUs were not able to use explanations to identify cases in which the AI would make unreliable decisions and then correct the AI. Consistent with this, participants in both the AI-only and XAI conditions spent comparable amounts of time on the task.

As already discussed earlier, the over-reliance of DKUs in our study could not be attributed to the increased cognitive effort imposed by explanations [15], as mental effort and task load were similar in both conditions. At the same time, DKUs in both conditions falsely disagreed with the AI's "malicious" recommendation in around 16% of cases when the classification was actually correct, resulting in false blocks of benign URLs. While those errors are comparably harmless, the more concerning observation is that DKUs in both conditions allowed malicious URLs in more than 80% of the decisions if the AI misclassified it as non-malicious (see Figure 3). This behaviour reflects over-trust in the AI's recommendations [38, 69]; participants often trusted the AI when it erred and were unable to correct that recommendation based on their own indicators. This finding is especially critical given the severe implications of permitting malicious URLs. Especially considering their strategy of protectively, and thereby also somewhat blindly, blocking URLs the AI classifies as malicious, this behaviour raises important implications for designing AI in cybersecurity and the introduction of XAI. Overall, these results suggest that AI systems in cybersecurity contexts should ensure rather conservative decision boundaries, as even DKUs tend to, at least behaviourally, over-rely on the AI's recommendation. When over-reliance is likely, AI systems should be designed to accommodate and compensate for this behaviour. Considering the practical utility of explanations, as well as the DKUs potential time engagement with explanations, raises questions of whether standard XAI approaches are effective. This aligns with work by Buçinca et al. [16], who show that cognitive motivation moderates the effectiveness of XAI, based on the assumption that users must be motivated to meaningfully engage with explanations. Approaches that actively prompt deeper cognitive engagement, such as cognitive forcing functions or evaluative AI [47], may therefore be more suitable in these cybersecurity environments to accommodate the need for improved performance.

Concretely, this means that when one outcome clearly poses lower risk, as in MDB, AI systems should prioritise classifying borderline cases into that lower-risk category as part of a protective strategy. Developers and organisations should focus on minimising misclassifications of high-risk outcomes when AI and practitioners collaborate. This approach reduces the likelihood of critical errors while tolerating a higher rate of false positives for malicious classifications. This protective strategy might also generalise

to other use cases in cybersecurity where practitioners and AI collaborate on tasks, such as vulnerability scanning. In these contexts, it might be recommendable to pre-emptively flag potential vulnerabilities in software development, i.e., conservatively classify, as the practitioner might otherwise miss indicators, due to the unstructured nature of the task and risk vulnerabilities in software products.

Strategies for Under-Trust. Lastly, our findings also provide indications of strategies that can be employed when the DKUs tend to under-trust during collaboration on a joint task with AI. Our correlation analysis revealed that trust, performance, and usability are strongly interconnected (see Appendix D, Figure 6 for more details). These results suggest that when DKUs perceived that they performed well on a task, which can also be viewed as a proxy for their perceived domain knowledge and self-confidence, this was related to higher trust in automation, less frustration during the task, and a higher perceived usability of the AI tool.

Therefore, if the user has higher domain knowledge, AI can support them in increasing their perceived performance, e.g., through providing respective feedback, which in turn may foster trust between users and AI. Further, when the users' and the AI's decisions align, this may act as a reinforcing factor on the TiA and thereby also the behavioural component of reliance between the practitioner and the AI. Related literature confirms that people's trust in AI is driven by its perceived quality, such as ease of use, and performance, and further depends on performance and effort expectancy [25,49]. A usable AI tool that takes into account the users' domain knowledge and supports them on that specific task with AI may enable better adjustment to appropriate TiA and reliance. Therefore, iteratively designing XAI tools that thoroughly consider their target users and their knowledge level, enabling them to feel they successfully complete their task, is crucial for successful collaboration and acceptance of the AI system.

7 Limitations

We acknowledge several limitations, beginning with the limited scope of our experiment. As we were aiming to isolate the effect of explanations on DKUs' decision-making in cybersecurity, we selected the method of feature contribution provided through IG, and selected the task of malicious domain blocking as a relevant exemplary use case in cybersecurity. Drawing on prior literature and input from cybersecurity practitioners, we selected the MDB task that captures cognitive demands of cybersecurity work (i.e., evaluating uncertain evidence, interpreting system explanations, and balancing trust and reliance on automation), is familiar to practitioners while enabling experimental control and reproducibility. We consulted the literature to choose the type (feature contribution) of explanation and the explainability method (IG), and appropriate deep learning models. However, this setup inadvertently limits the generalisability of our findings to other methods. Our findings shine light on binary decision-making tasks with IG, however, future work could explore and compare further XAI methods to deepen the understanding of XAI-supported decision-making in cybersecurity, also on more complex collaboration tasks. Additionally, we were required to select URLs for the DKUs to investigate. DKUs might have already encountered some of the URLs that they were asked to investigate and may have already developed an intuition about their malicious nature. Nevertheless, we aimed to have a balanced set of URLs. There are two limitations regarding the domain knowledge of our sample. We designed a recruiting strategy through pre-screeners to identify participants with knowledge and expertise in the field of cybersecurity, however did not specifically target domain knowledge for the task of malicious domain blocking. We additionally re-verified the characteristics of participants from the pre-screening in the main study, and conducted qualitative analysis to further verify domain knowledge. Nevertheless, we cannot definitively verify their level of knowledge due to the online setup of the experiment. It is possible that despite our efforts some of our participants did not have expert knowledge of cybersecurity or the MDB task.

8 Conclusion

Seeking to explore the applicability of transparent and human-understandable insights into AI decisions to support cybersecurity DKUs, we investigated the potential of providing XAI explanations in the context of the cybersecurity-specific high-criticality task of malicious domain blocking. In a between-subject study with (N = 139) cybersecurity DKUs, we compared the effects of providing XAI explanations vs. AI-only on participants' trust, performance, perceived task load, and AI tool usability. We found that DKUs interacting with XAI reported lower trust in automation, which might be partially explained by a disclosed discrepancy between the DKUs' own decision strategies and the XAI's explanations, as implied by accompanying qualitative insights. Additionally, while neither the DKUs' performance nor the task load differed between the conditions, DKUs interacting with the XAI on the task rated the usability as lower than DKUs interacting with the AI-only. Interestingly, DKUs in both conditions relied on the AI's classifications even in misclassified cases – indicating behavioural over-reliance on the AI's decisions. Future research should examine the gap between attitudinal trust and behavioural reliance more closely, taking into consideration the effect of familiarity and expertise on DKUs' trust. Our findings point to multiple strategies to accommodate these observations, such as the use of conservative decision boundaries in cybersecurity decision-making contexts.

Ethical Considerations

This research evaluated human-AI collaboration in the context of malicious domain blocking using XAI. The study design followed established ethical standards for psychological research involving humans [5] and received approval from our IRB. In line with the principles of the Menlo Report, we ensured that the design, data collection, analysis processes and dissemination aligned with established best practices for ethical Information and Communication Technologies and psychological research. We discuss ethical concerns for survey participants, end users of AI-based security systems, malicious actors, human-AI collaboration and policy research communities, AI system and cybersecurity developers, general public and society and the involved researchers.

Respect for Persons. The primary risks for participants included stress from cybersecurity scenarios or interacting with AI, and inadvertent exposure of responses. To mitigate these, participants received detailed study information, provided informed consent, could withdraw or request data deletion at any time, and were explicitly informed when interacting with AI. Only non-identifying demographic data were collected to reduce privacy risks. Participants may also have benefited by learning about AI in cybersecurity and contributing to research. All researchers involved in this study completed training on human-subjects research ethics, and data were securely handled throughout. The researchers also collaboratively ensured participants' autonomy, safeguarding their integrity and supporting credible scientific contributions.

Beneficence. We carefully considered the potential risks and benefits associated with both the study design and the use of XAI in cybersecurity contexts. The primary stakeholders include study participants, end users of AI-based cybersecurity tools, the human-AI collaboration and policy research communities, developers of AI and cybersecurity tools, malicious actors, and the general public. Participants faced minimal risk of inadvertent disclosure of their perspectives, which we mitigated through strict anonymization and reporting results only in aggregated form. End users may over-rely on AI recommendations, potentially experiencing harm through misclassified domains, false positives, or false negatives affecting services or access; we address this risk by discussing design strategies to accommodate harmful behaviour and reduce over-reliance. The research and policy communities risk misinterpreting or overstating findings when extrapolating beyond the studied population. To mitigate this, we explicitly highlight study limitations, emphasize the need for evaluation across tasks and explanation methods, and use cautious phrasing to avoid presenting results as definitive, acknowledging that results should be interpreted cautiously and within the context of this specific study. Developers of AI tools in cybersecurity face risks from exposure of XAI limitations or misuse of findings. We mitigated this by relying on XAI techniques previously evaluated in cybersecurity research and using an open-access, non-vendor-specific prototype that provides high-level insights without revealing sensitive security mechanisms. Similarly, while malicious actors could attempt to exploit insights into XAI-assisted tools, we avoided disclosing task-specific, operational, or vendor-sensitive details that could enable exploitation. For the general public, risks include over-trusting AI security systems, service disruptions due to incorrect blocking, and the normalization of AI-driven decision-making. We mitigate these risks by emphasizing contextual interpretation of results, improving transparency about AI decision-making, encouraging human-centred and iterative system design that accounts for real user behaviour, and highlighting the need for explanation approaches beyond feature attribution, such as cognitive forcing functions and evaluative AI. These benefits from our reserach for various stakeholders include advancing understanding of how XAI can support domain-specific knowledge, informing the design of cybersecurity and URL-classification tools to better support practitioners, contributing to safer browsing for end users, and providing empirical insights to guide future research and evidence-based AI governance. The decision to conduct this research was guided by the potential to advance understanding of human-AI collaboration in high-stakes cybersecurity tasks, while carefully weighing ethical considerations and potential risks. The decision to publish the findings was reached after applying mitigation strategies to minimize risks, ensuring participant anonymity, reporting aggregated results, and highlighting study limitations, thereby allowing the benefits of disseminating insights for research, tool development, and evidence-based policy to outweigh potential harms.

Justice. Our recruitment strategy aimed to involve participants whose background and expertise were appropriate to the study goals while avoiding the inclusion of vulnerable groups. For example, we included only adult participants. Because comprehension of cybersecurity-related task instructions was essential, we recruited English-speaking participants from the US and the UK. All participants received equivalent compensation that aligned with Prolific's fair payment guidelines (£3 for approximately 20 minutes). To promote equitable access to our findings and research process, we have made our research prototype, study plan, and analysis pipeline publicly available.

Respect for Law and Public Interest. The study complied with all applicable legal and institutional requirements. IRB approval was obtained prior to data collection, and no compliance issues or other concerns arose over the course of the research. By promoting transparency in the design and evaluation of high-risk AI applications, this work aligns with emerging regulatory expectations, including those articulated in the EU AI Act [24].

Open Science

This research, including the hypotheses, the data analysis plan, and the power analysis, has been pre-registered on OSF to enhance transparency and replicability of the work (Pre-registration). The source code for the prototype system is hosted on a Zenodo Repository for long-term access, while the full R analysis scripts and processed anonymized datasets (including performance, task load, usability, and trust measures) are archived on OSF (OSF Project incl. materials). To protect participant confidentiality, free-text responses and detailed demographic information are not released publicly but can be provided upon reasonable request, subject to institutional review of ethics.

References

[1] ALHOGAIL, A., AND AL-TURAIKI, I. Improved detection of malicious domain names using gradient boosted machines and feature engineering. *Information Technology and Control 51*, 2 (2022), 313–331. https://doi.org/10.5755/j01.itc.51.2.30380.

[2] ALSAEDI, M., GHALEB, F., SAEED, F., AHMAD, J., AND ALASLI, M. Multi-modal features representation-based convolutional neural network model for malicious website detection. *IEEE Access 12* (2024), 7271–7284. https://doi.org/10.1109/ACCESS.2023.3348071.

[3] ALTHOBAITI, K., MENG, N., AND VANIEA, K. I don't need an expert! making url phishing features human comprehensible. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)* (2021), pp. 1–17. https://doi.org/10.1145/3411764.3445574.

[4] ALTHOBAITI, K., RUMMANI, G., AND VANIEA, K. A review of human- and computer-facing url phishing features. In *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* (2019), pp. 182–191. https://doi.org/10.1109/EuroSPW.2019.00027.

[5] AMERICAN PSYCHOLOGICAL ASSOCIATION. Ethical principles of psychologists and code of conduct (2002, amended effective june 1, 2010, and january 1, 2017), 2017. https://www.apa.org/ethics/code.

[6] ARNOLD, V., CLARK, N., COLLIER, P. A., LEECH, S. A., AND SUTTON, S. G. The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions. *MIS Quarterly 30*, 1 (2006), 79–97. http://www.jstor.org/stable/25148718.

[7] ASSAL, H., AND CHIASSON, S. 'Think secure from the beginning': A survey with software developers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)* (2019), p. 1–13. https://doi.org/10.1145/3290605.3300519.

[8] ATEŞ, C., KAYMAZ, Ö., KALE, H. E., AND TEKINDAL, M. A. Comparison of test statistics of nonnormal and unbalanced samples for multivariate analysis of variance in terms of type-i error rates. *Computational and Mathematical Methods in Medicine 2019* (2019), 2173638. https://doi.org/10.1155/2019/2173638.

[9] BATES, D., MÄCHLER, M., BOLKER, B., AND WALKER, S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*, 1 (2015), 1–48. https://doi.org/10.18637/jss.v067.i01.

[10] BAYER, S., GIMPEL, H., AND MARKGRAF, M. The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems 32*, 1 (2022), 110–138. https://doi.org/10.1080/12460125.2021.1958505.

[11] BENJAMINI, Y., AND HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological) 57*, 1 (1995), 289–300. https://www.jstor.org/stable/2346101.

[12] BOLTON, M. L., BILTEKOFF, E., AND HUMPHREY, L. The mathematical meaninglessness of the NASA task load index: A level of measurement analysis. *IEEE Transactions on Human-Machine Systems 53*, 3 (2023), 590–599. https://doi.org/10.1109/THMS.2023.3263482.

[13] BROOKE, J. Sus: A "quick and dirty" usability scale. *Usability Evaluation in Industry* (1996), 189–194. https://www.taylorfrancis.com/chapters/edit/10.1201/9781498710411-35/sus-quick-dirty-usability-scale-john-brooke.

[14] BUITINCK, L., LOUPPE, G., BLONDEL, M., PEDREGOSA, F., MUELLER, A., GRISEL, O., NICULAE, V., PRETTENHOFER, P., GRAMFORT, A., GROBLER, J., LAYTON, R., VANDERPLAS, J., JOLY, A., HOLT, B., AND VAROQUAUX, G. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning* (2013), pp. 108–122. https://inria.hal.science/hal-00856511v1.

[15] BUÇINCA, Z., LIN, P., GAJOS, K. Z., AND GLASSMAN, E. L. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)* (2020), pp. 454–464. https://doi.org/10.1145/3377325.3377498.

[16] BUÇINCA, Z., MALAYA, M. B., AND GAJOS, K. Z. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *In Proceedings of the ACM on Human-Computer Interaction 5*, CSCW1 (2021), 1–21. https://doi.org/10.1145/3449287.

[17] CAPTUM.AI. Integrated Gradients — Captum Documentation, 2026. https://captum.ai/docs/extension/integrated_gradients.

[18] CHANCEY, E. T., BLISS, J. P., YAMANI, Y., AND HANDLEY, H. A. H. Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human Factors: The Journal of the Human Factors and Ergonomics Society 59*, 3 (2017), 333–345. https://doi.org/10.1177/0018720816682648.

[19] CHEN, V., LIAO, Q. V., WORTMAN VAUGHAN, J., AND BANSAL, G. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-Computer Interaction 7*, CSCW2 (2023), 1–32. https://dl.acm.org/doi/10.1145/3610219.

[20] CLARKE, V., AND BRAUN, V. Thematic analysis. *The Journal of Positive Psychology 12*, 3 (2017), 297–298. https://doi.org/10.1080/17439760.2016.1262613.

[21] COHEN, J. *Statistical power analysis for the behavioral sciences*. Routledge, 2013. https://doi.org/10.4324/9780203771587.

[22] COLVILLE, S., AND OSTERN, N. K. Trust and distrust in GAI applications: The role of AI literacy and metaknowledge. In *Proceedings of the International Conference on Information Systems (ICIS)* (2024). https://aisel.aisnet.org/icis2024/user_behav/user_behav/1.

[23] DESAI, B., PATIL, K., MEHTA, I., AND PATIL, A. Explainable AI in cybersecurity: A comprehensive framework for enhancing transparency, trust, and human-AI collaboration. In *2024 International Seminar on Application for Technology of Information and Communication (iSemantic)* (2024), pp. 135–150. https://doi.org/10.1109/iSemantic63362.2024.10762690.

[24] EUROPEAN PARLIAMENT AND COUNCIL OF THE EUROPEAN UNION. Regulation (EU) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending certain union legislative acts (artificial intelligence act), 2024. http://data.europa.eu/eli/reg/2024/1689/oj.

[25] Fan, W., Liu, J., Zhu, S., and Pardalos, P. M. Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). *Annals of Operations Research 294*, 1 (2020), 567–592. https://doi.org/10.1007/s10479-018-2818-y.

[26] Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods 39*, 2 (2007), 175–191. https://doi.org/10.3758/BF03193146.

[27] Fügener, A., Grahl, J., Gupta, A., and Ketter, W. Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information Systems Research 33*, 2 (2021), 678–696. https://doi.org/10.1287/isre.2021.1079.

[28] Gregor, S., and Benbasat, I. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly 23*, 4 (1999), 497–530. http://www.jstor.org/stable/249487.

[29] Hart, S. G. NASA task load index (TLX): Computerized version - volume 1.0. NASA Ames Research Center, 1986. https://humansystems.arc.nasa.gov/groups/tlx/.

[30] Hemmer, P., Schemmer, M., Vössing, M., and Kühl, N. Human-AI complementarity in hybrid intelligence systems: A structured literature review. In *Proceedings of the 25th Pacific Asia Conference on Information Systems (PACIS 2021)* (2021), pp. 1–14. https://aisel.aisnet.org/pacis2021/78.

[31] Hemmer, P., Westphal, M., Schemmer, M., Vetter, S., Vössing, M., and Satzger, G. Human-AI collaboration: The effect of AI delegation on human task performance and task satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)* (2023), p. 453–463. https://doi.org/10.1145/3581641.3584052.

[32] Herm, L.-V. Impact of explainable AI on cognitive load: Insights from an empirical study. In *Proceedings of the 31st European Conference on Information Systems (ECIS 2023)* (2023). https://aisel.aisnet.org/ecis2023_rp/269.

[33] Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science 5* (2023). https://doi.org/10.3389/fcomp.2023.1096257.

[34] International Organization for Standardization. ISO 9241: Ergonomics of human-system interaction. Tech. Rep. ISO 9241, International Organization for Standardization, Geneva, Switzerland, 2010. https://www.iso.org/standard/77520.html.

[35] Iwahana, K., Takemura, T., Cheng, J., Ashizawa, N., Umeda, N., Sato, K., Kawakami, R., Shimizu, R., Chinen, Y., and Yanai, N. Madmax: Browser-based malicious domain detection through extreme learning machine. *IEEE Access 9* (2021), 78293–78314. https://doi.org/10.1109/ACCESS.2021.3080456.

[36] Kaur, R., Gabrijelčič, D., and Klobučar, T. Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion 97* (2023), 101804. https://doi.org/10.1016/j.inffus.2023.101804.

[37] Kerner, S. M. 35 cybersecurity statistics to lose sleep over in 2024, 2024. https://www.techtarget.com/whatis/34-Cybersecurity-Statistics-to-Lose-Sleep-Over-in-2020.

[38] Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., and Shaw, T. H. Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology 12* (2021). https://doi.org/10.3389/fpsyg.2021.604977.

[39] Körber, M. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)* (2019), pp. 13–30. https://doi.org/10.1007/978-3-319-96074-6_2.

[40] Lai, V., and Tan, C. 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 (FAT* '19)* (2019), pp. 29–38. https://doi.org/10.1145/3287560.3287590.

[41] Lambe, K. A., O'Reilly, G., Kelly, B. D., and Curristan, S. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ Quality & Safety 25*, 10 (2016), 808–820. https://doi.org/10.1136/bmjqs-2015-004417.

[42] Lee, J. D., and See, K. A. Trust in automation: Designing for appropriate reliance. *Human Factors 46*, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392.

[43] Li, K., Yu, X., and Wang, J. A review: How to detect malicious domains. *Communications in Computer and Information Science 1424* (2021), 152–162. https://doi.org/10.1007/978-3-030-78621-2_12.

[44] Logg, J. M., Minson, J. A., and Moore, D. A. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes 151* (2019), 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005.

[45] Longo, L. Experienced mental workload, perception of usability, their interaction and impact on task performance. *PLOS ONE 13*, 8 (2018), e0199661. https://doi.org/10.1371/journal.pone.0199661.

[46] Lotfalian Saremi, M., Ziv, I., Asan, O., and Bayrak, A. E. Trust, workload, and performance in human–artificial intelligence partnering: The role of artificial intelligence attributes in solving classification problems. *Journal of Mechanical Design 147*, 011702 (2024). https://doi.org/10.1115/1.4065916.

[47] Miller, T. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)* (2023), p. 333–342. https://doi.org/10.1145/3593013.3594001.

[48] Morris, J., Newman, S., Palaniappan, K., Fan, J., and Lin, D. "Do you know you are tracked by photos that you didn't take": Large-scale location-aware multi-party image privacy protection. *IEEE Transactions on Dependable and Secure Computing 20*, 1 (2023), 301–312. https://doi.org/10.1109/TDSC.2021.3132230.

[49] Mostafa, R. B., and Kasamani, T. Antecedents and consequences of chatbot initial trust. *European Journal of Marketing 56*, 6 (Oct. 2021), 1748–1771. https://doi.org/10.1108/EJM-02-2020-0084.

[50] Müller, L. S., Nohe, C., Reiners, S., Becker, J., and Hertel, G. Adopting information systems at work: a longitudinal examination of trust dynamics, antecedents, and outcomes. *Behaviour & Information Technology 43*, 6 (2024), 1096–1128. https://doi.org/10.1080/0144929X.2023.2196598.

[51] Ooge, J., Kato, S., and Verbert, K. Explaining recommendations in e-learning: Effects on adolescents' trust. In *Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22)* (2022), pp. 93–105. https://doi.org/10.1145/3490099.3511140.

[52] Paleja, R., Ghuy, M., Ranawaka Arachchige, N., Jensen, R., and Gombolay, M. The utility of explainable AI in ad hoc human-machine teaming. In *Advances in Neural Information Processing Systems* (2021), vol. 34, pp. 610–623. https://proceedings.neurips.cc/paper/2021/file/05d74c48b5b30514d8e9bd60320fc8f6-Paper.pdf.

[53] PASZKE, A., GROSS, S., MIRZA, S., ZHA, L., YANG, Z., GLOROT, X., ANTIGA, L., DESMAISON, A., KOPF, A., BREVDO, E., CHANAN, G., CAI, T., STEPHENS, B., ZHANG, L., V., M. A., AND CHINTALA, S. Pytorch: An imperative style, high-performance deep learning library, 2019. https://pytorch.org.

[54] PYTHON SOFTWARE FOUNDATION. *Python Programming Language, version 3.11.11*, 2024. https://www.python.org/downloads/release/python-31111/.

[55] QUALTRICS. Qualtrics – Experience Management Software, 2020. https://www.qualtrics.com/.

[56] QUINKERT, F., DEGELING, M., BLYTHE, J., AND HOLZ, T. Be the phisher – understanding users' perception of malicious domains. In *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security (ASIA CCS '20)* (2020), p. 263–276. https://doi.org/10.1145/3320269.3384765.

[57] R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2023. https://www.R-project.org/.

[58] ROCH, N., SIEVERS, H., SCHÖNI, L., AND ZIMMERMANN, V. Navigating autonomy: Unveiling security experts' perspectives on augmented intelligence in cybersecurity. In *Twentieth Symposium on Usable Privacy and Security (SOUPS 2024)* (2024), pp. 41–60. https://www.usenix.org/conference/soups2024/presentation/roch.

[59] SCHAFFER, J., O'DONOVAN, J., MICHAELIS, J., RAGLIN, A., AND HÖLLERER, T. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)* (2019), pp. 240–251. https://doi.org/10.1145/3301275.3302308.

[60] SCHEMMER, M., KUEHL, N., BENZ, C., BARTOS, A., AND SATZGER, G. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (2023), pp. 410–422. https://doi.org/10.1145/3581641.3584066.

[61] SCHOEFFER, J., KUEHL, N., AND MACHOWSKI, Y. "There is not enough information": On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)* (2022), pp. 1616–1628. https://doi.org/10.1145/3531146.3533218.

[62] SEGURA TEAM. 32 cybersecurity stats you can't ignore in 2025, 2025. https://segura.security/post/cybersecurity-stats-you-cant-ignore/.

[63] SENANAYAKE, J., RAJAPAKSHA, S., YANAI, N., KOMIYA, C., AND KALUTARAGE, H. Madonna: Browser-based malicious domain detection through optimized neural network with feature analysis. In *IFIP Advances in Information and Communication Technology (IFIP AICT), Volume 679* (2024), pp. 279–292. https://doi.org/10.1007/978-3-031-56326-3_20.

[64] SUNDARARAJAN, M., TALY, A., AND YAN, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17)* (2017), pp. 3319–3328. https://doi.org/10.48550/arXiv.1703.01365.

[65] TAYLOR, H. M., JAY, C., LENNOX, B., CANGELOSI, A., AND DENNIS, L. Should AI systems in nuclear facilities explain decisions the way humans do? an interview study. In *31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (2022), pp. 956–962. https://doi.org/10.1109/RO-MAN53752.2022.9900852.

[66] U.S. BUREAU OF LABOR STATISTICS. Occupational outlook handbook: Information security analysts, 2025. https://www.bls.gov/ooh/computer-and-information-technology/information-security-analysts.htm.

[67] VERBI SOFTWARE. MAXQDA, version 24.9.1, 2024. https://www.maxqda.com.

[68] VÖSSING, M., KÜHL, N., LIND, M., AND SATZGER, G. Designing transparency for effective human-AI collaboration. *Information Systems Frontiers 24*, 3 (2022), 877–895. https://doi.org/10.1007/s10796-022-10284-3.

[69] WANG, X., AND YIN, M. Are explanations helpful? a comparative study of the effects of explanations in AI-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI '21)* (2021), pp. 318–328. https://doi.org/10.1145/3397481.3450650.

[70] WARNECKE, A., ARP, D., WRESSNEGGER, C., AND RIECK, K. Evaluating explanation methods for deep learning in security. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)* (2020), pp. 158–174. https://doi.org/10.1109/EuroSP48549.2020.00018.

[71] WIRTZ, B. W., WEYERER, J. C., AND AND, C. G. Artificial intelligence and the public sector—applications and challenges. *International Journal of Public Administration 42*, 7 (2019), 596–615. https://doi.org/10.1080/01900692.2018.1498103.

[72] YANG, R., AND WIBOWO, S. User trust in artificial intelligence: A comprehensive conceptual framework. *Electronic Markets 32*, 4 (2022), 2053–2077. https://doi.org/10.1007/s12525-022-00592-6.

[73] ZHANG, S., MENG, Z., CHEN, B., YANG, X., AND ZHAO, X. Motivation, social emotion, and the acceptance of artificial intelligence virtual assistants—trust-based mediating effects. *Frontiers in Psychology Volume 12* (2021). https://doi.org/10.3389/fpsyg.2021.728495.

[74] ZHANG, Y., LIAO, Q. V., AND BELLAMY, R. K. E. 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), pp. 295–305. https://doi.org/10.1145/3351095.3372852.

[75] ZIJLSTRA, F. R. H., AND VAN DOORN, L. The construction of a scale to measure subjective effort. *Ergonomics 43*, 10 (1985), 124–139. https://www.researchgate.net/publication/266392097_The_Construction_of_a_Scale_to_Measure_Perceived_Effort.

Appendix A: Participants Characteristics

Participants were able to select multiple educational backgrounds to accommodate interdisciplinary studies.

Field	No. Responses
Computer Science	108
Engineering	36
Mathematics	35
Business & Economics	28
Social Sciences	9
Psychology	4
Other	7
AI Familiarity	**No. Responses**
Extremely Fam.	55
Very Fam.	67
Moderately Fam.	15
Slightly Fam.	2

Table 5: Participants' Educational Backgrounds and Self-Reported AI Familiarity

Appendix B: Linear mixed-effects regression of condition and task repetition on average decision time

Variable	β	*SE*	95% CI	p
Intercept	**54.65**	**3.93**	**[46.93, 62.36]**	**<.001**
Cond. [XAI]	6.30	5.46	[-4.41, 17.02]	.249
Position	**-2.64**	**0.38**	**[-3.39, -1.88]**	**<.001**
Cond. [XAI] x Pos.	**-1.22**	**0.53**	**[-2.27, -0.17]**	**.023**
σ^2 (τ_{00})	1417.50 (499.12)			
N	139			
Observations	1668			
R^2 (marg. / cond.)	.07 / .31			

Note: Condition was contrast coded; Position is trial number.

Table 6: Linear mixed-effects regression of condition (AI-only vs. XAI) and task repetition on average decision time. Reports include unstandardized coefficients (β), SEs, 95% CIs, and Wald t p-values; bold coefficients indicate 95% CI significance.

Appendix C: Wilcoxon rank-sum test results for task load and mental effort across conditions

	AI-only		XAI			
Variable	*M*	*SD*	*M*	*SD*	W	95% CI
Task Load						
Effort	56.94	26.40	64.60	23.31	2061.5	$[-16.00, 2.00]$
Mental Demand	53.63	27.21	59.51	22.69	2221.5	$[-11.00, 4.00]$
Temporal Demand	28.94	21.21	33.44	20.96	2072	$[-12.00, 2.00]$
Performance	84.13	16.66	79.19	17.31	2865	$[< 0.01, 0.1]$
Frustration	16.42	19.19	21.40	23.51	2168	$[-9.00, < 0.01]$
Average Mental Effort	60.42	31.11	58.23	30.44	2491.5	$[-9.42, 13.08]$

Table 7: Wilcoxon rank-sum test results comparing average ratings and standard deviations for task load and mental effort across conditions (AI-only vs. XAI). Reported statistics include means (M), standard deviations (SD), Wilcoxon test statistic (W), and 95% CIs.

Appendix D: Pearson correlation matrix for trust in automation (TiA), perceived usability (SUS), and task load measures

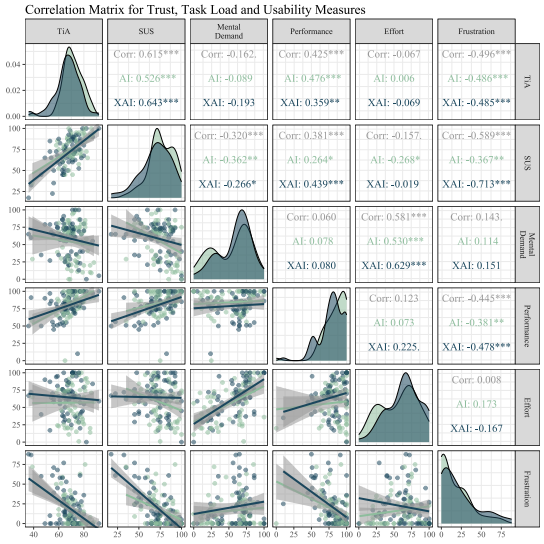

Figure 6: Pearson correlation matrix illustrating relationships among trust in automation (TiA), perceived usability (SUS), and task load measures (NASA-RTLX subscales: Mental Demand, Performance, Effort, and Frustration). Correlation coefficients are annotated with significance levels: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Appendix E: Code Book for Qualitative Analysis

Code	Description	Example	No. occurances
Use of explanation for decision-making			
Own indicators for decision	Participants have their own indicators that they review and base their decision on.	"*I would expect it to be mentioning, e.g., SSL encryption status prominently but instead it seemed to focus on characteristics of the URL.*"	21
Background knowledge	Participant explicitly mentions having some own knowledge, and thereby expectations they can rely on for this task.	"*I mostly relied on my own experience.*"	12
Human control	Participant specifically mentions being the one making the final decision.	"*In the end, I will make the final decision.*"	2
Consider AI classification	The participant uses the classification in some way, by looking or reviewing it for their decision.	"*Based on the automated review of the URL I could see a summary on why a URL is malicious or not malicious and then decide based on it.*"	6
Trust AI classification	Participant's description of decision-making reflects reliance on the AI's malicious/non-malicious classification.	"*[I] relied a lot on the AI classification of either malicious or non-malicious.*"	8
Consider AI explanations	The participant uses the explanation in some way, by looking or reviewing it for their decision.	"*By considering all the features highlighted by the AI and applying common patterns of malicious and non-malicious URLs.*"	41
Trust AI explanations	Participant's description of decision-making reflects reliance on the AI's explanation.	"*I relied on what the system provided to make a decision.*"	3
Inconsistent explanations	Participant expresses frustration due to inconsistent or unreliable explanations.	"*In the end, I really couldn't use its judgment one way or the other. It seemed incredibly inconsistent, using the same metrics to show some URLs as safe and others as malicious. The results seemed almost random.*"	3
Feature descriptions	Participants mention the descriptions of each feature, and the min and max anchor points provided.	"*The explanation gave indication of what is a bad value and what is good. This helped me weigh up the different properties and determine whether the values of those properties would indicate maliciousness or not.*"	3
Rely on statistics	Participants mention mainly relying on the descriptive tabular data summarised below the AI explanations.	"*I glanced at this then looked in more detail at some key metrics below that I tend to base decisions on.*"	5
Little use	The participant makes minimal use of the AI features in the dashboard.	"*I used it very little.*"	3
Domain-specific knowledge analysis			
Feature-based reasoning	Participants focused on specific or characteristics of the explanation to make sense of the maliciousness of the URL.	"*main signals were uptime of the host, country and SSL cert.*"	29
Heuristic reasoning	Participants used experience-based shortcuts or rules of thumb to make decisions about a URL.	"*With these red flags, blocking it seems like the safest choice.*"	25
Causal explanation & decision justification	Participants explained why something justified their choices and why specific indicators made the suspicious of or trust the URL.	"*To complement it - I placed greater weight on the length of time the domain had been in exisitence. [sic]*"	44
Risk articulation	Participants identified possible risks or things that could go wrong based on their decision.	"*I didn't like how it didn't seem to weigh the fact the site was using SSL high enough.*"	8

Table 8: Codes, descriptions, and example quotes of XAI participants' (n=72) decision-making strategies.