Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Explainable artificial intelligence in interface design influences user outcomes and cybersecurity literacy

**Ahmed Al-sa'di**

ahmed.al-sa'di@weltec.ac.nz

Wellington Institute of Technology

**Mahmoud Al-Sarayreh**

German Jordanian University

**Esraa Ahmad**

Swinburne University of Technology

**Research Article**

**Keywords:** explainable artificial intelligence, interface design, human-AI interaction, privacy literacy, cybersecurity literacy, user autonomy

**Posted Date:** September 25th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10798262/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Abstract

Explainable artificial intelligence (XAI) exposes information that users can examine when assessing AI-generated recommendations. However, limited evidence is available on whether explanation-supported interaction develops competencies relevant to privacy and cybersecurity literacy.

This study examined whether an XAI-enhanced design tool affects creative confidence, perceived learning, reported trust, perceived autonomy, and usability during a UI design task.

We conducted a controlled between-groups experiment with 128 participants. Participants were randomly assigned to an XAI condition (*n = 64*) or a standard AI condition (*n = 64*). Both groups completed the same UI design task. The XAI tool presented rationales and trade-offs with each suggestion, whereas the standard tool presented suggestions alone. Results show that the XAI condition produced a larger creative-confidence gain (*d = 0.68*). Post-task perceived learning (*d = 1.17*) reported trust (*d = 0.97*), and perceived autonomy (*d = 0.42*; all *p<.05*) were also higher. Usability was higher under XAI (*d = 1.78*). Logged interaction counts were higher in XAI condition (*d = 2.66*). Across the full sample, these counts correlated with perceived learning, reported trust, usability, and task quality (all *r≥.35*).

Thus, XAI was associated with stronger self-reported outcomes and higher usability during AI-assisted design. The findings support explanations as an interface feature that may promote critical engagement with AI recommendations. However, trust calibration, privacy literacy, dark-pattern recognition, and resistance to deceptive interfaces were not measured directly. The results indicate potential relevance to privacy and cybersecurity literacy, not demonstrated behavioural resilience. Direct transfer should be evaluated using realistic privacy-decision tasks.


# 1 Introduction

Dark patterns are interface designs that steer users toward choices that may conflict with their interests [14]. They can encourage disclosure of personal data, intrusive permissions, or recurring payments. These designs do not compromise technical infrastructure. Instead, they exploit defaults, visual hierarchy, and interaction costs to influence decisions [1, 2]. The User interface therefore becomes part of the privacy and cybersecurity risk. The issue is pertinent to usable security, because in order for consent, authentication, and access controls to be effective, users must understand what the options presented to them mean.

There are numerous dark patterns at play. Many large-scale and empirical studies have documented such findings on shopping websites, mobile apps, and social media [3–5]. According to [6], LLM-based GUI agents performing actions on behalf of users can also be manipulated. Furthermore, Article 25 of the EU Digital Services Act deals with interface practices limiting free and informed ones [7]. While regulation may limit some impositions, it cannot eliminate them totally. During the interaction, the user and the AI agent must both evaluate the interface choices.

Privacy along with cybersecurity literacy provides a useful basis for this evaluation. At the interface level, users need to understand requests for their data, assess system recommendations, and challenge inappropriate defaults. It is harder to understand the basis of an output if it doesn't explain it. XAI can reveal information that aids in evaluating the recommendation [18–20, 31]. Simply being transparent does not indicate better literacy or protective behaviour.

Most user-centred XAI research examines explanation quality, decision support, trust, and acceptance. Evidence on learning and user agency remains more limited. Dark patterns rely on attentional constraints, default effects, and rapid decisions [1, 2]. Explanations may encourage users to pause and examine a recommendation. Yet this potential connection has rarely been tested in an educational design task. Its transfer to privacy decisions is also unknown.

This study examined user outcomes associated with an XAI-enhanced UI design tool. We conducted a controlled between-subjects experiment and compared it with a standard AI tool. The measured outcomes were creative-confidence gain, post-task perceived learning, reported trust, perceived autonomy, usability, task quality, and logged interactions. Thus, the study evaluates explanation-supported interaction during UI design. It also considers how the observed outcomes may inform future privacy and cybersecurity literacy interventions.

The study did not measure privacy literacy, dark-pattern recognition, avoidance behaviour, or trust calibration. Consequently, the findings should not be interpreted as evidence of resistance to deceptive interfaces. Direct transfer requires separate testing with realistic privacy decisions. Section 2 reviews the theoretical background. Section 3 describes the method, Section 4 presents the results, and Section 5 discusses the findings, limitations, and future work.

## 2 Theoretical Background

# 2.1 Dark Patterns as Interface-Level Privacy Threats

Dark-pattern research has developed from descriptive examples to systematic classification. Mathur et al. [3] characterised these designs as asymmetric, covert, deceptive, information-hiding, or restrictive. Gray et al. [8] proposed an ontology that includes obstruction, sneaking, interface interference, forced action, and social engineering. Although the schemes differ in scope, both show how interface design can constrain informed choice. Article 25 of the Digital Services Act similarly addresses interface practices that deceive, manipulate, or materially impair free and informed decisions [7].

Consequences are visible in privacy decisions. Nouwens et al. [9] found that the design of consent banners influenced users' consent choices. The California Privacy Rights Act also addresses interfaces that subvert or impair user autonomy and decision-making [10]. Similar risks may extend to automated agents. Tang et al. [6] reported that LLM-based GUI agents were susceptible to manipulative interface designs, which reinforces the need for human oversight.

Awareness does not always prevent manipulation. In a controlled study, Bongard-Blanchy et al. [11] found that users could recognise dark patterns yet remain unable to resist their influence. One explanation is that deceptive designs take advantage of rapid, low-effort processing [12]. Therefore, factual knowledge may be insufficient when the interface makes compliance easy or refusal difficult. Support for deliberate evaluation may be needed alongside awareness.

## 2.2 Privacy Literacy and Interface Decision-Making

Privacy literacy encompasses more than simply knowing organisations collect personal data. The Online Privacy Literacy Scale includes knowledge of institutional practices and technical protection and privacy laws and policies [13]. Cognitive biases and interface design also affect applied decisions. Waldman [14] suggests that these factors may help bridge the gap between the expressed concerns about privacy and protective action. As a result, privacy interventions should manage knowledge and decision conditions.

There are various ways to assist privacy and security choices. According to Acquisti et al. [15], designing nudges to assist users is possible but effects depend on the designers and context. Experiential learning offers a relevant educational foundation because knowledge is generated through task execution and reflection [16]. According to Bandura's self-efficacy framework, success experience is a source of perceived capability [17]. This justifies the need for interactive tools. Nonetheless, they do not demonstrate that XAI influences privacy literacy, and the current study did not assess that construct.

## 2.3 Cognitive Load and Decision Fatigue

Dark patterns often exploit limited attention and rapid decision-making [12]. Automated recommendations may add another source of pressure when users defer to system output without adequate verification. Parasuraman and Manzey [21] describe automation bias and complacency as attention-related problems in human use of automation. Explanations could support closer examination, but they may also increase information demands. Cognitive load and decision fatigue were not manipulated or measured in the present experiment. They are therefore treated as theoretical considerations only.

## 2.4 XAI, Reported Trust, and Automation Bias

User-centred XAI questions whether an explanation aids people in assessing and utilizing an AI output. Poursabzi-Sangdeh and his associates demonstrated that improved model interpretability does not enhance human decision-making. How users use and interact with information, is important. Liao et al. [19] found similar explanation needs; for example, understanding the output and system limitations and possible actions. Consequently, it ought to assist with assessment, not merely contribute more information to the interface. Trust has a nuanced definition in AI assisted decision-making. User's subjective evaluation is reported trust. Conversely, trust calibration questions whether the justification for their reliance is correct with respect to system operation. The effect of confidence information and explanations on decision accuracy and calibration measures was examined by Zhang et al [20].

Automation bias poses a correlated risk, as users may accept automatic advice without sufficient checks. XAI may aid users in verifying a recommendation.

Nonetheless, explanations may not always curtail bias. Calibration should therefore be measured directly. It is governance requirement to be transparent. Under Article 13 of the EU AI Act, information must be provided to enable the deployer to interpret the output of a high-risk AI system and use the system properly. The connection between explanation, oversight and user agency is acknowledged. Users can access information through education. This does not evaluate the security performance and regulatory compliance. Still, the current experiment assesses trust, perceived autonomy, and other design-task outcomes.

## 2.5 Creative Confidence and Perceived Autonomy

Creativity confidence is simply the notion one has that they can create ideas and act on them. The belief in the presence of AI does change the task because it might depend upon the user's retaining directional autonomy over the task. Research suggests generative AI has both positive and negative effects on the creativity and creative self-efficacy of students [24, 25]. In a design process, AI suggestion transparency might thus matter. Nonetheless, creative confidence is independent from the quality on the task itself and needs to be tested as separate out-come.

Self-efficacy gives a more extended account of perceived capability [17]. Users that believe they are in the loop may find it easier to question the suggestions that AI makes and maintain control over final decisions. Survey research about higher education suggests perceived benefits and concerns associated with generative AI [26]. Such observations offer a conceptual connection among confidence, autonomy, and critical engagement. Gains in creative confidence do not elicit privacy-protective behaviour. This requires direct tests.

# 3 Methods

## 3.1 Study Design

A between-subjects design was used with two conditions; explainable AI (XAI) and standard AI (STD). The XAI interface displayed an explanation panel alongside each AI-generated design suggestion, whereas this panel was absent from the standard interface. The task, tool functions, and study procedure were otherwise kept the same across conditions. Participants were allocated using stratified randomisation, with prior AI familiarity and UI design experience used as stratification factors to support balance between groups. The study tested whether providing explanations with AI-generated suggestions would improve participants' creative confidence, perceived learning, trust, autonomy, and engagement with AI reasoning compared with the standard AI condition. Institutional ethical approval was obtained. Participants provided digital informed consent before taking part and were debriefed after completing the study.

## 3.2 Participants and Recruitment

Participants were recruited through Prolific [27]. The platform is widely used for online behavioural research. Eligibility required participants to be at least 18 years old and fluent in English. They also had to reside in an English-speaking country, complete at least ten prior Prolific studies, and maintain a minimum approval rating of 95%.

A total of 141 participants were recruited across a generated-task subset and an in-the-wild task subset. Each design task was completed by multiple participants to account for performance variability. Participants were excluded for failing the embedded attention check (*n = 8*) or reporting technical problems that prevented task completion (*n = 5*). The final analytic sample contained 128 participants, with 64 in each condition. Compensation was USD 15 per hour. Mean session duration was approximately 45 minutes, including onboarding, task completion, and post-task questionnaires.

The final sample included 59 men (46.1%) and 57 women (44.5%). It also included 9 non-binary participants (7.0%) and 3 participants who preferred not to disclose their gender (2.3%). Mean age was 24.66 years in the XAI group (*SD = 4.09*). The corresponding value for the standard group was 25.27 years (*SD = 4.08*). Prior AI familiarity averaged 3.70 (*SD = 1.14*) for XAI and 3.56 (*SD = 1.10*) for standard AI. UI design experience was 10.06 months in the XAI group (*SD = 5.10*). The standard group reported a mean of 11.45 months (*SD = 5.44*). All baseline differences were reported as non-significant (all *p>.10*).

Participants accessed the tool through a browser. Standard mouse and keyboard input was used for scrolling, clicking, and entering text into design fields. Each participant completed one condition. The task and questionnaire were administered in a single uninterrupted session.

## 3.3 Materials and Design Task

The design scenario concerned health-research administration. Participants were asked to improve a dashboard used by a research officer to monitor recruitment, deadlines, and reporting across several concurrent studies, including a dementia caregiver study and a digital health study. The task did not involve clinical content, patient data, or health-related decision-making, and no clinical knowledge was required. The scenario therefore provided an information-management context rather than an assessment of health literacy.

Both groups used the same AI-assisted UI design tool and completed the same design task under the same functional requirements. Participants created interface components for the specified user scenario within a 30-minute task period. Pilot testing was conducted before data collection to confirm that the task could be completed within the allocated time while retaining sufficient variation in design quality and participant engagement.

The AI-assisted tool generated design suggestions in response to participants' interactions with the interface. The underlying suggestions were presented consistently across conditions; the principal experimental difference concerned the explanatory information accompanying them. In the XAI condition, each suggestion was accompanied by an explanation panel describing the relevant design

principle, contextual considerations, possible trade-offs, and a confidence indicator. In the standard AI condition, participants received the design suggestions without the dedicated explanation panel.

The AI-assisted tool was implemented in Python using VScode (V 1.133.0 ) and used OpenAI's GPT-5.5 model to generate design suggestions. In the XAI condition, the model output was supplemented with an explanation panel presenting the design rationale, relevant contextual considerations, potential trade-offs, and a confidence indicator. The standard AI condition presented the corresponding design suggestions without the dedicated explanation panel.

Interface events were logged automatically in both conditions. In the XAI condition, the recorded interaction count reflected participants' use of the explanation panel. In the standard AI condition, the count referred to interactions with adjacent interface elements, including tooltips, suggestion labels, and secondary information icons. These events represented different forms of interaction and were therefore not interpreted as equivalent measures of explanation engagement.

## 3.4 Outcome Measures

### 3.4.1 Creative Confidence

Creative confidence was measured before and after the design task using a three-item, 1–7 Likert scale. Internal consistency was $\alpha = .88$ before the task and $\alpha = .92$ after the task. A composite score was calculated at each time point. Creative-confidence gain was defined as the post-task composite minus the pre-task composite. Thus, this variable represented within-participant change.

### 3.4.2 Reported Trust, Perceived Autonomy, and Perceived Learning

Reported trust in the AI tool, perceived autonomy, and perceived learning were measured after the task. Each construct was assessed using three items on a 1–7 Likert scale. Internal consistency was $\alpha = .83$ for reported trust, $\alpha = .82$ for perceived autonomy, and $\alpha = .86$ for perceived learning. Item means formed the composite scores. Accordingly, these variables were post-task ratings and did not represent pre-to-post change.

### 3.4.3 System Usability Scale

System usability was assessed with the 10-item System Usability Scale (SUS) [28]. The instrument yields a score from 0 to 100. For descriptive interpretation, scores above 68 were classified as above average. Scores from 50 to 68 were treated as marginal, and scores below 50 as unacceptable. The same standardised measure was used in both conditions.

### 3.4.4 Task Quality and Logged Interactions

Two trained raters, blinded to condition assignment, scored task quality on a 0-100 scale. Inter-rater reliability was reported as $ICC=.87$ before data collection. Scores differing by more than 10 points were

resolved through discussion. Interface interactions were logged automatically for both conditions. As described in Section 3.3, these counts reflected different actions in the XAI and standard conditions.

## 3.5 Statistical Analysis

Independent-samples t-tests examined between-condition differences. Paired t-tests assessed pre-to-post change within each condition. Cohen's d was used to report effect sizes, and Pearson correlations examined associations among the outcome variables. Two hierarchical regression models evaluated predictors of creative-confidence gain. Condition was entered in the first step. Reported trust, perceived autonomy, perceived learning, SUS, logged interactions, and prior AI familiarity were added in the second step. All analyses were conducted in Python using pandas and SciPy. The two-tailed significance threshold was $\alpha = .05$. Cohen's benchmarks were used: small ($d=.20$), medium ($d=.50$), and large ($d=.80$) [29].

# 4 Results

## 4.1 Baseline Characteristics and Task Duration

Table 1 summarizes the participant characteristics and between-condition comparisons. The reported tests did not identify statistically significant differences in age, prior AI familiarity, or UI design experience. Mean age was 24.66 years in the XAI condition ($SD = 4.09$) and 25.27 years in the standard AI condition ($SD = 4.08$). The reported comparison was $t(126) = -0.84$, $p=.401$, $d = 0.15$. No statistically significant difference was reported for prior AI familiarity, $t(126) = 0.71$, $p=.478$. The same result was reported for UI design experience, $t(126) = 1.49$, $p=.138$.

Mean task completion time was 30.63 minutes in the XAI condition ($SD = 5.26$). The corresponding mean for standard AI was 28.93 minutes ($SD = 5.96$). This difference was not statistically significant, $t(126) = 1.71$, $p=.090$, $d = 0.30$.

Table 1

Descriptive Statistics and Between-Condition Comparisons ($N = 128$)

| Variable | XAI ( $\varvec{n} = 64$ ) | Std AI ( $\varvec{n} = 64$ ) | $\varvec{t}\,(126)$ | $\varvec{p}$ | $\varvec{d}$ |
|---|---|---|---|---|---|
| | $M$ (SD) | $M$ (SD) | | | |
| Age (years) | 24.66 (4.09) | 25.27 (4.08) | $-0.84$ | .401 | 0.15 |
| Prior AI familiarity (1–7) | 3.70 (1.14) | 3.56 (1.10) | 0.71 | .478 | 0.13 |
| UI experience (months) | 10.06 (5.10) | 11.45 (5.44) | 1.49 | .138 | 0.26 |
| Task time (min) | 30.63 (5.26) | 28.93 (5.96) | 1.71 | .090 | 0.30 |
| Task quality (0–100) | 82.80 (7.81) | 75.54 (7.11) | 5.50 | $<$.001*** | 0.97 |
| Explanation interactions | 8.08 (2.82) | 1.91 (1.68) | 15.05 | $<$.001*** | 2.66 |
| CC pre | 4.21 (1.18) | 3.95 (1.15) | 1.25 | .215 | 0.22 |
| CC post | 5.22 (1.19) | 4.52 (1.17) | 3.36 | .001** | 0.60 |
| CC gain | 1.02 (0.55) | 0.57 (0.76) | 3.82 | $<$.001*** | 0.68 |
| Trust score (1–7) | 5.64 (0.72) | 4.88 (0.85) | 5.51 | $<$.001*** | 0.97 |
| Autonomy score (1–7) | 5.14 (0.79) | 4.79 (0.87) | 2.36 | .020* | 0.42 |
| Perceived learning (1–7) | 5.57 (0.74) | 4.62 (0.89) | 6.62 | $<$.001*** | 1.17 |
| SUS score (0–100) | 71.52 (7.99) | 58.36 (6.79) | 10.05 | $<$.001*** | 1.78 |

# 4.2 Creative Confidence

Figure 1 shows pre- and post-task creative confidence. Both conditions showed significant gains. In the XAI condition, the mean increased from 4.21 (*SD = 1.18*) to 5.22 (*SD = 1.19*). The reported test was *t(63) = 14.70*, *p<.001*, *d = 1.84*. The standard condition increased from 3.95 (*SD = 1.15*) to 4.52 (*SD = 1.17*). Here, the test was *t(63) = 6.00*, *p<.001*, *d = 0.75*. Gain scores were also higher in the XAI condition (*M = 1.02* vs. 0.57). The between-condition test gave *t(126) = 3.82*, *p<.001*, *d = 0.68*. Thus, the XAI condition showed the larger creative-confidence gain.

# 4.3 Effect Sizes for Study Outcomes

Figure 2 summarizes Cohen's d values for the primary outcomes. Perceived autonomy had the smallest reported effect (*d = 0.42*). All other effects were medium or larger. The largest standardized difference was observed for logged interface interactions (*d = 2.66*). However, these counts represented different actions in the two conditions and are therefore interpreted descriptively.

# 4.4 Reported Trust, Perceived Autonomy, and Perceived Learning

Reported trust was higher in the XAI condition (*M = 5.64*, *SD = 0.72*) than in the standard AI condition (*M = 4.88*, *SD = 0.85*). The reported effect size was *d = 0.97*. Item means ranged from 5.61 to 5.66 under XAI and from 4.83 to 4.91 under standard AI. Perceived autonomy was also higher under XAI (*M = 5.14* vs. 4.79; *d = 0.42*). Perceived learning showed the largest effect among these measures (*M = 5.57* vs. 4.62; *d = 1.17*). Its item means ranged from 5.52 to 5.66 under XAI and from 4.50 to 4.73 under standard AI. These findings represent post-task differences; they do not indicate pre-to-post gains.

# 4.5 System Usability

Figure 3 shows the SUS score distribution. In the XAI group, 39 of 64 participants (61%) scored 68 or above, and none scored below 50. In the standard AI group, 4 of 64 participants (6%) scored 68 or above. A further 55 participants (86%) scored from 50 to below 68, while 5 (8%) scored below 50.

Mean SUS was higher in the XAI condition (*M = 71.52*, *SD = 7.99*) than in the standard AI condition (*M = 58.36*, *SD = 6.79*). The reported comparison was *t(126) = 10.05*, *p<.001*, *d = 1.78*.

# 4.6 Logged Interface Interactions

Figure 4 presents the distribution of logged interface interactions. The XAI group recorded a mean of 8.08 interactions per session (*SD = 2.82*). Its median was 8 (*IQR = 4.25*; range 2 to 15). The corresponding mean in the standard group was 1.91 (*SD = 1.68*). Its median was 2 (*IQR = 2*; range 0 to 8). In that group, fifteen participants recorded no interactions. Counts of one to three were observed for 38 participants, while 11 recorded four or more.

The event definitions differed between conditions. XAI participants could open a dedicated explanation panel. The standard interface did not provide this panel. Its counts referred to clicks on suggestion-label tooltips and secondary information icons. These elements displayed brief metadata, such as a confidence percentage or category tag. Therefore, the counts should not be interpreted as a common measure of explanation engagement. Within the standard group, interaction count was not significantly correlated with any outcome (all $r<.19$, all $p>.13$). This analysis did not detect an association between the peripheral clicks and the measured outcomes.

# 4.7 Descriptive Outcomes by Prior AI Familiarity

Figure 5 reports creative-confidence gain by prior AI familiarity. The XAI condition had the higher mean gain at every level. The low-familiarity means were $M = 1.13$ versus 0.63. At medium familiarity, the means were $M = 1.00$ versus 0.49. The high-familiarity means were $M = 1.00$ versus 0.82.

For the low-familiarity subgroup ($n = 17$), mean reported trust was 5.67 under XAI and 4.67 under standard AI. The corresponding perceived-learning means were 5.29 and 4.92, while SUS means were 69.38 and 58.61. These comparisons are descriptive. They should be interpreted cautiously because of the small subgroup sizes.

# 4.8 Correlations and Hierarchical Regression

Figure 6 and Table 2 report Pearson correlations across the full sample. Logged interaction count was positively associated with perceived learning ($r=.42$), SUS ($r=.52$), and reported trust ($r=.37$). Positive correlations were also reported for task quality ($r=.35$), perceived autonomy ($r=.24$), and creative-confidence gain ($r=.21$). The strongest correlations were with SUS and perceived learning. These pooled associations do not establish direction or causation. Moreover, because event definitions and several outcomes differed by condition, the correlations may partly reflect between-condition separation.

Table 2

Pearson Correlations Among Primary Outcome Variables ($N = 128$)

| Variable | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1. CC Gain | – | | | | | | |
| 2. Trust | .178* | – | | | | | |
| 3. Autonomy | .115 | −.050 | – | | | | |
| 4. Perceived Learning | .177* | .277** | .076 | – | | | |
| 5. SUS Score | .231** | .251** | .018 | .273** | – | | |
| 6. Expl. Interactions | .208* | .373** | .243** | .419** | .521** | – | |
| 7. Task Quality | .177* | .278** | .142 | .148 | .268** | .351** | – |

Hierarchical regression examined predictors of creative-confidence gain. Model 1 included condition only and accounted for 10.4% of the variance (XAI *β = 0.45*). Model 2 added reported trust, perceived autonomy, perceived learning, SUS, logged interactions, and prior AI familiarity. It accounted for 12.0% of the variance (full-model XAI *β = 0.50*). Condition remained the strongest reported predictor in both models. However, the modest increase in explained variance does not demonstrate a mediating pathway or causal mechanism.

# 5 Discussion

# 5.1 Interpretation and Scope of the Findings

The findings should be interpreted in relation to the measured outcomes. The XAI condition showed a larger creative-confidence gain. It also produced higher post-task scores for perceived learning, reported trust, perceived autonomy, and usability. Thus, explanation-supported design affected user responses during the task. The results do not show that participants acquired privacy literacy or resilience to deceptive interfaces. No dark-pattern exposure or avoidance task was included.

This distinction defines both the contribution and its boundary. Prior research shows that awareness may not be sufficient to prevent dark-pattern influence [11]. Work on explanations also indicates that confidence information can affect trust and reliance [20]. Self-efficacy provides a conceptual basis for perceived capability [17]. Together, these studies motivate further testing of explanation-supported learning. However, the present experiment did not test a pathway from XAI interaction to dark-pattern avoidance. Behavioural transfer must therefore remain a research question.

# 5.2 Logged Interactions and Outcome Associations

Figure 6 shows that logged interaction counts were correlated with perceived learning, reported trust, usability, task quality, and creative-confidence gain. However, the correlations were calculated across the full sample. Condition affected both the meaning of the logged events and several outcomes. Thus, the associations may partly reflect between-condition differences. They support a relationship between interaction counts and outcomes, not an explanation-engagement mechanism. This proposed mechanism was not tested.

The condition-specific events should be interpreted separately. In the standard group, counts represented clicks on tooltips and metadata icons. These counts were not significantly correlated with the measured outcomes (all $r<.19$, all $p>.13$). This result does not establish that explanation engagement caused the XAI outcomes. It only indicates that the logged events in the standard interface were not associated with them. Feature-importance scores and confidence percentages are also forms of explanatory information, although their usefulness depends on design and context [31]. Explanation quality should therefore be evaluated directly rather than inferred from interaction count.

Mean SUS was 71.52 in the XAI condition, which was above the descriptive threshold of 68. The standard condition mean was 58.36 and fell below that threshold. Therefore, the XAI interface received the higher usability rating. The result does not establish that transparency reduced hesitation or rework because these outcomes were not measured. The difference may reflect explanation design, interface organisation, or other features of the condition. Accordingly, usability should be evaluated alongside explanation quality. It should not be treated as evidence of improved efficiency.

## 5.3 Prior AI Familiarity and Generalisability

The recruitment criteria required a Prolific approval rating of 95% and at least ten completed studies. These requirements may select participants who are familiar with online research. However, digital engagement was not measured directly. The sample should therefore be described as Prolific-eligible rather than digitally advanced. This distinction limits generalisation. The familiarity results were also descriptive; no moderation test was reported.

Descriptive results favoured XAI across the three prior AI-familiarity levels. In the low-familiarity subgroup ($n = 17$), reported trust means were $M = 5.67$ under XAI and $M = 4.67$ under standard AI. The corresponding SUS means were $M = 69.38$ and $M = 58.61$. These observations are exploratory. Low prior AI familiarity is not equivalent to low digital literacy. Moreover, the study did not measure digital literacy or cognitive load. The results cannot determine whether explanation panels impose additional demands on lower-literacy users.

The current evidence is most applicable to participants who met the Prolific criteria and completed this UI design task. It does not establish XAI as a privacy-literacy intervention for digitally engaged or lower-literacy populations. Future research could test simpler language, layered presentation, and human support. Such adaptations must be evaluated directly. Their effect on the measured outcomes cannot be inferred from the present sample.

## 5.4 Reported Trust and Trust Calibration

Reported trust was higher in the XAI condition than in the standard AI condition (*d = 0.97*). This result does not demonstrate trust calibration. Calibration requires trust to vary appropriately with system accuracy or reliability. Neither factor was manipulated in this experiment. Therefore, the finding is limited to a post-task difference in reported trust.

Reported trust and appropriate reliance are different constructs. High trust can be problematic when an automated recommendation is incorrect or unsuitable [21]. Explanations and confidence information may influence trust and decision-making [20]. However, their effect depends on system performance and user interpretation. Thus, higher trust should not be described as beneficial without a measure of reliance accuracy.

Participants in the XAI condition saw information about the stated reasoning behind each recommendation. This information may have contributed to perceived transparency or understanding. However, neither construct was measured as an explanatory variable. They cannot therefore account for the between-condition difference in reported trust. The data also do not show whether this trust was warranted or calibrated.

Explanation design can shape user reliance. Greene et al. [30] caution that an explanation may be used persuasively rather than informatively. Selective evidence, excessive certainty, or biased framing could increase confidence without improving judgement. Thus, explanation quality should be assessed in relation to appropriate reliance, not transparency alone.

Article 13 of the EU AI Act requires transparency that supports interpretation and appropriate use of high-risk AI systems [22]. The present study is relevant to this broader objective because it examines user responses to explanations. It did not test a high-risk system, regulatory compliance, or appropriate reliance. Future studies should vary system accuracy or explanation quality and then measure whether users adjust their reliance. Such designs are required to assess calibration directly.

## 5.5 Potential Relevance to Human-AI Oversight

Tang et al. [6] identified attentional tunnelling during human oversight of GUI agents. Supervisors may focus on the agent's selected path and overlook manipulative elements elsewhere in the interface. The present task did not involve autonomous agents or measure interface-wide attention. Therefore, the results cannot show that explanation-supported interaction reduces this failure. Future work could test whether explanation-supported design affects oversight of agent actions and attention to surrounding interface elements.

Workspace-integrated AI assistants create a practical need for human oversight. However, the present study did not examine these deployment contexts or measure AI-reasoning literacy. The findings may inform a hypothesis about explanation-supported oversight. They do not demonstrate transfer beyond the design task.

## 5.6 Implications for Privacy and Cybersecurity Literacy

The findings support further examination of experience-based approaches alongside knowledge-based privacy education. Prior research suggests that awareness does not always lead to avoidance [11, 12]. In the present study, explanations were associated with user outcomes during a design task. Privacy resilience was not measured. Therefore, a competency-based privacy-literacy model should be presented as a framework for future testing rather than an outcome demonstrated by this experiment.

The potential cybersecurity relevance lies at the interface level. Users often evaluate permissions, credentials, and data-sharing requests. Yet the study did not include phishing, access-control, consent, or other security tasks. It also did not compare experiential learning with declarative instruction. Thus, the results cannot determine whether explanation-supported practice improves cybersecurity literacy or transfers across threat categories. Such transfer requires direct evaluation with security-relevant tasks.

The results suggest that explanations can encourage closer inspection of AI recommendations during design. They do not show that repeated XAI engagement creates a general habit of questioning interface choices. Dark-pattern knowledge, privacy decisions, and behavioural avoidance were not measured. Thus, whether explanation-supported practice affects resistance to deceptive interfaces remains open.

## 5.7 Potential Relevance to Digital Health

The scenario concerned health-research administration, but the task contained no clinical content, patient data, consent flow, or health decision. Therefore, the findings should not be generalised to digital or public health practice. Cognitive load, trust calibration, and transfer to health interfaces were also not measured. The factors discussed in Section 2.3 remain theoretical in relation to this experiment. Future work may test explanation-supported interaction in health settings, but no clinical design recommendation follows from the present results.

# 6 Limitations and Future Directions

The primary limitation is the absence of a direct measure of dark-pattern recognition or avoidance. The experiment cannot establish a pathway from XAI interaction to privacy literacy or behavioural resistance. Future studies should use standardised deceptive-interface tasks before and after an XAI intervention. Such designs could test recognition accuracy and avoidance behaviour directly. Until then, resilience remains a theoretical implication rather than a demonstrated outcome.

All outcomes were collected during a single session of approximately 45 minutes. Only creative confidence was measured before and after the task. Reported trust, perceived learning, autonomy, and usability were post-task measures. Their persistence cannot be assessed from this design. Longitudinal studies with follow-up at one, three, and six months could examine durability and transfer across interface domains. Such evidence is required before concluding that XAI produces lasting cognitive change.

The Prolific recruitment criteria may limit generalisability, as discussed in Section 5.3. Digital literacy was not measured. Therefore, above-average digital engagement should not be inferred from eligibility alone. Older adults, users in low-resource settings, second-language users, and people with limited AI experience may respond differently. Dedicated studies should examine these populations using explanation designs suited to their needs.

The experiment examined one explanation-panel implementation. The accuracy, depth, and quality of its reasoning were not evaluated independently. Real-world XAI systems vary from detailed accounts to feature-importance displays [31]. Therefore, the observed effects should not be generalised to all forms of explanation. Future work should vary explanation depth and quality while measuring the same outcomes.

Greene et al. [30] warn that commercial incentives may produce persuasive or distorted explanations. Such designs could increase reliance without improving judgement. The present study did not vary explanation quality. It also did not measure automation bias or trust calibration. Future experiments should test whether different explanation designs improve or reduce appropriate reliance. This question is relevant to both implementation practice and explainability standards.

The task was not health-specific. Although the scenario involved health-research administration, no clinical information or patient decision was presented. The findings therefore cannot support recommendations for health-interface design. Replication could use patient portals, data-sharing flows, symptom-checker tools, or caregiver systems. Such studies should directly measure cognitive load and decision fatigue, as discussed in Section 2.3. Until this transfer is tested, the relevance to health settings remains unknown.

# 7 Conclusion

Dark patterns influence decisions through defaults, visual prominence, and interaction costs. Users may recognise a manipulative design without resisting it. This provides a rationale for studying interventions that support closer evaluation of interface recommendations. The present experiment examined explanations within an AI-assisted design task. It did not test dark-pattern recognition or avoidance.

This study compared XAI and standard AI design tools under a controlled task. The XAI condition showed the larger creative-confidence gain. Perceived learning also showed a large between-condition difference ($d = 1.17$), while reported trust and usability were higher. Logged interaction counts differed markedly between conditions, but the events did not represent the same action. Their pooled correlations with the outcomes do not establish a behavioural mechanism.

The results do not show whether explanation-supported design affects responses to deceptive interfaces. Direct testing requires dark-pattern recognition and avoidance tasks. The present session identified a change in creative confidence and post-task differences in the other outcomes. It did not establish long-term retention or transfer. Longitudinal studies are therefore needed.

The contribution is limited to the measured outcomes and the design-task context. The findings do not demonstrate privacy resilience, cybersecurity literacy, or durable changes in user reasoning. Their value lies in showing that an explanation-supported interface can affect user responses during a controlled task. Future work should test transfer, durability, diverse participants, and realistic privacy or security decisions. The transparency objectives of the EU AI Act make meaningful user engagement important. However, the present study does not establish a regulatory effect or a cybersecurity requirement.

## Declarations

**Funding:** This research received no specific funding from public, commercial, or not-for-profit sectors.

**Conflict of interest / Competing interests:** The authors declare no conflict of interest.

**Ethics approval :** This study was approved by the Whitireia and WelTec Research and Ethics Committee (RP 332–2022) and was carried out in accordance with the guidelines and regulations set out by that committee. All participants provided informed consent prior to participating in the research.

**Acknowledgements**: The authors express gratitude to all participants who generously contributed their time and professional insights to this research. The willingness of participants to share their experiences across four national contexts was invaluable to the study.

**Consent for publication:** Not applicable.

**Data availability:** The dataset supporting these findings is available from the corresponding author on reasonable request.

**Materials availability:** Not applicable.

**Code availability:** Not applicable.

**Author contribution:** All authors contributed to study conception, analysis, and manuscript preparation.

**Clinical trial number:** Not applicable.

# Author Contribution

A.A. led the study conception and design, developed the research methodology, conducted the main analysis, and prepared the initial draft of the manuscript. M.S. made substantial contributions to the study design, methodological development, analysis, interpretation of findings, and manuscript preparation. E.A. contributed to the health and public-health framing of the study, reviewed the alignment of the research scope with privacy, cybersecurity, and digital-health implications, and contributed to manuscript writing and critical revision. A.A. led the overall development and revision of the manuscript. All authors reviewed and approved the final manuscript.

# References


1. Acquisti A, Brandimarte L, Loewenstein G (2015) Privacy and human behaviour in the age of information. Science 347(6221):509-514. https://doi.org/10.1126/science.aaa1465
2. Solove DJ (2013) Introduction: Privacy self-management and the consent dilemma. Harv Law Rev 126:1880-1903
3. Mathur A, Acar G, Friedman MJ, Lucherini E, Mayer J, Chetty M, Narayanan A (2019) Dark patterns at scale: Findings from a crawl of 11K shopping websites. Proc ACM Hum-Comput Interact 3(CSCW):81. https://doi.org/10.1145/3359183
4. Di Geronimo L, Braz L, Fregnan E, Palomba F, Bacchelli A (2020) UI dark patterns and where to find them: A study on mobile applications and user perception. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). ACM, New York, pp 1-14. https://doi.org/10.1145/3313831.3376600
5. Mildner T, Freye M, Savino GL, Doyle PR, Cowan BR, Malaka R (2023) Defending against the dark arts: Recognising dark patterns in social media. In: Proceedings of the 2023 ACM Designing Interactive Systems Conference (DIS '23). ACM, New York, pp 2362-2374. https://doi.org/10.1145/3563657.3595964
6. Tang J, Chen C, Li J, Zhang Z, Guo B, Khalilov I, Gebreegziabher SA, Yao B, Wang D, Ye Y, Li T, Xiao Z, Yao Y, Li TJ (2026) Dark patterns meet GUI agents: LLM agent susceptibility to manipulative interfaces and the role of human oversight. In: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26). ACM, New York, Article 403. https://doi.org/10.1145/3772318.3791568
7. European Parliament, Council of the European Union (2022) Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services (Digital Services Act). Off J Eur Union L 277:1–102. https://eur-lex.europa.eu/eli/reg/2022/2065/oj
8. Gray CM, Santos CT, Bielova N, Mildner T (2024) An ontology of dark patterns knowledge: Foundations, definitions, and a pathway for shared knowledge-building. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. ACM, New York, Article 289. https://doi.org/10.1145/3613904.3642436
9. Nouwens M, Liccardi I, Veale M, Karger D, Kagal L (2020) Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). ACM, New York, pp 1-13. https://doi.org/10.1145/3313831.3376321
10. State of California (2020) California Privacy Rights Act of 2020 (Proposition 24). https://oag.ca.gov/privacy/ccpa
11. Bongard-Blanchy K, Rossi A, Rivas S, Doublet S, Koenig V, Lenzini G (2021) "I am definitely manipulated, even when I am aware of it. It's ridiculous!" Dark patterns from the end-user perspective. In: Proceedings of the 2021 ACM Designing Interactive Systems Conference. ACM, New York, pp 763–776. https://doi.org/10.1145/3461778.3462086

12. Kahneman D (2011) Thinking, fast and slow. Farrar, Straus and Giroux, New York
13. Trepte S et al (2015) Do people know about privacy and data protection strategies? Towards the "Online Privacy Literacy Scale" (OPLIS). In: Gutwirth S, Leenes R, de Hert P (eds) Reforming European Data Protection Law. Law, Governance and Technology Series, vol 20. Springer, Dordrecht. https://doi.org/10.1007/978-94-017-9385-8_14
14. Waldman AE (2020) Cognitive biases, dark patterns, and the 'privacy paradox'. Curr Opin Psychol 31:105-109. https://doi.org/10.1016/j.copsyc.2019.08.025
15. Acquisti A, Adjerid I, Balebako R, Brandimarte L, Cranor LF, Komanduri S, Leon PG, Sadeh N, Schaub F, Sleeper M, Wang Y, Wilson S (2017) Nudges for privacy and security: Understanding and assisting users' choices online. ACM Comput Surv 50(3), Article 44. https://doi.org/10.1145/3054926
16. Kolb DA (1984) Experiential learning: Experience as the source of learning and development. Prentice Hall, Englewood Cliffs
17. Bandura A (1997) Self-efficacy: The exercise of control. W.H. Freeman, New York
18. Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Vaughan JW, Wallach H (2021) Manipulating and measuring model interpretability. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). ACM, New York, Article 237. https://doi.org/10.1145/3411764.3445315
19. Liao QV, Gruen D, Miller S (2020) Questioning the AI: Informing design practices for explainable AI user experiences. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). ACM, New York, pp 1-15. https://doi.org/10.1145/3313831.3376590
20. Zhang Y, Liao QV, Bellamy RKE (2020) 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). ACM, New York, pp 295-305. https://doi.org/10.1145/3351095.3372852
21. Parasuraman R, Manzey DH (2010) Complacency and bias in human use of automation: An attentional integration. Hum Factors 52(3):381-410. https://doi.org/10.1177/0018720810376055
22. European Parliament, Council of the European Union (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Off J Eur Union L 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj
23. Kelley T, Kelley D (2013) Creative confidence: Unleashing the creative potential within us all. Crown Business, New York
24. Mou T-Y (2026) Artificial intelligence and student creativity: An exploratory study of students' experiences with AI tools. Comput Hum Behav Rep 21:100988. https://doi.org/10.1016/j.chbr.2026.100988
25. Habib S, Gardiner Z (2026) Generative AI and creative self-efficacy: A systematic review. Think Skills Creat 62(Part A):102234. https://doi.org/10.1016/j.tsc.2026.102234

26. Chan CKY, Hu W (2023) Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. Int J Educ Technol High Educ 20:43. https://doi.org/10.1186/s41239-023-00411-8
27. Palan S, Schitter C (2018) Prolific.ac – A subject pool for online experiments. J Behav Exp Finance 17:22-27. https://doi.org/10.1016/j.jbef.2017.12.004
28. Brooke J (1996) SUS: A quick and dirty usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland AL (eds) Usability Evaluation in Industry. Taylor and Francis, London, pp 189-194
29. Cohen J (2013) Statistical power analysis for the behavioral sciences. Routledge, New York
30. Greene T, Goethals S, Martens D, Shmueli G (2025) Monetization could corrupt algorithmic explanations. AI Soc 40:6291–6308. https://doi.org/10.1007/s00146-025-02352-4
31. Arrieta AB, Diaz-Rodriguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F (2020) Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82-115. https://doi.org/10.1016/j.inffus.2019.12.012

# Figures

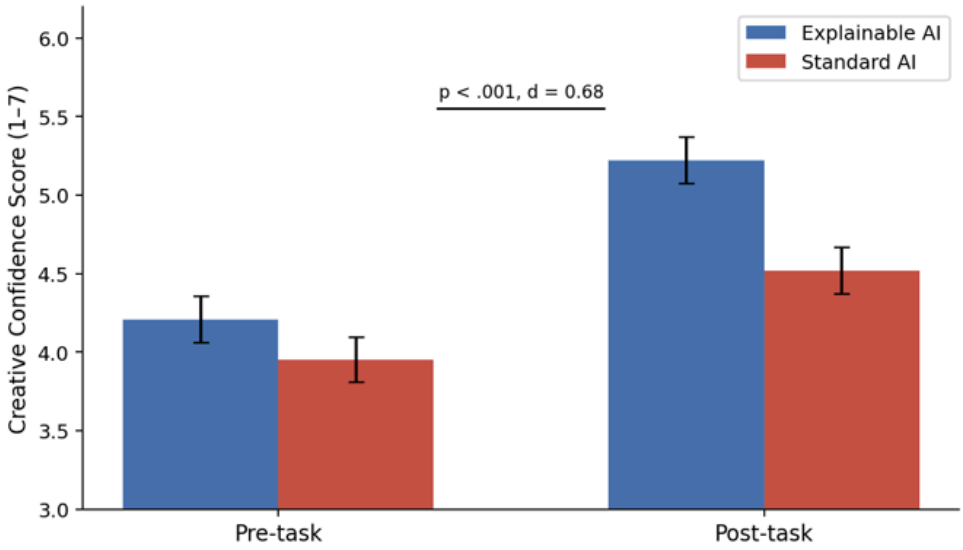



***Fig. 1*** *Creative confidence before and after task completion by condition (error bars = ±1 SE)*

**Figure 1**

See image above for figure legend.

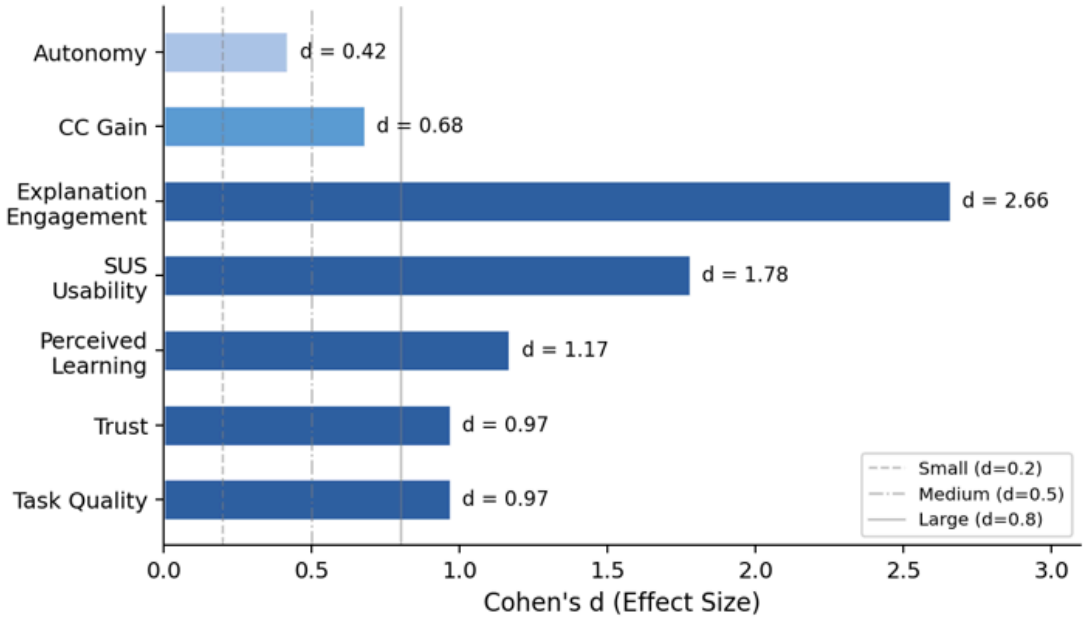



***Fig. 2*** *Cohen's $d$ effect sizes for all primary outcomes. Dashed vertical lines mark small, medium, and large thresholds*

**Figure 2**

See image above for figure legend.

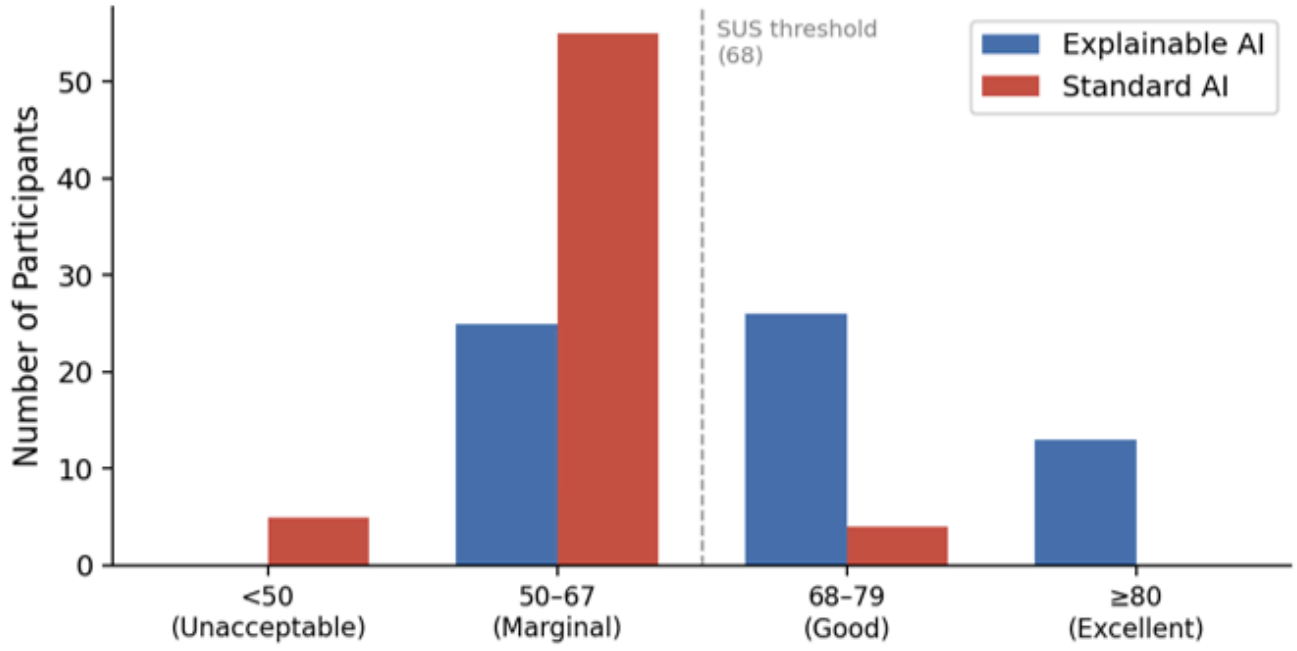



**Figure 3**

Distribution of SUS usability scores by condition. Dashed line marks the above-average usability threshold (SUS = 68)

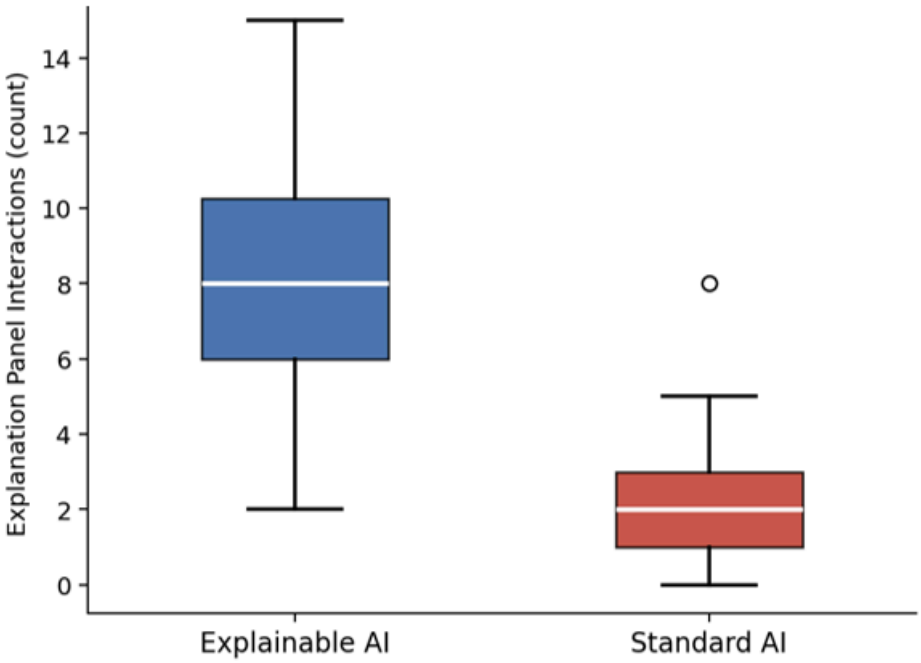



***Fig. 4*** *Distribution of logged interface interactions by condition*($d = 2.66$, $p < .001$)

**Figure 4**

See image above for figure legend.

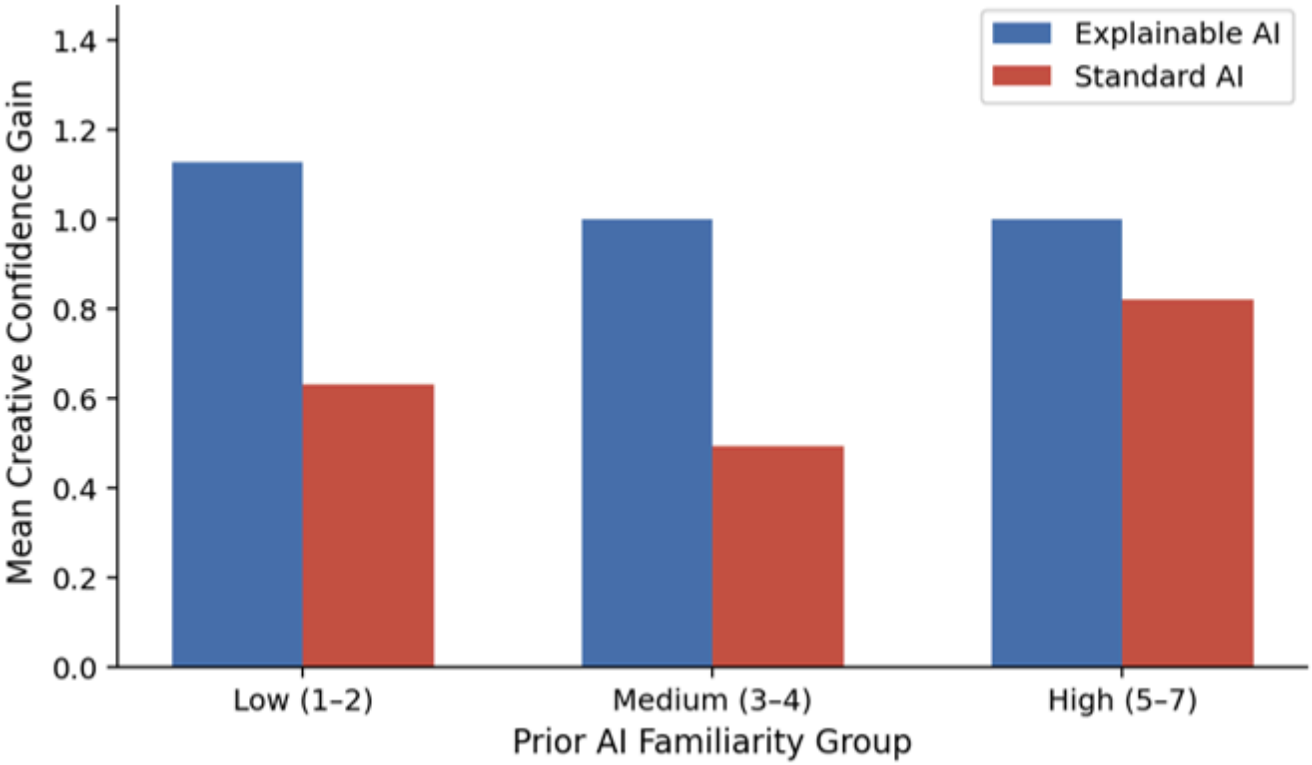



**Figure 5**

Creative confidence gains by prior AI familiarity level and condition

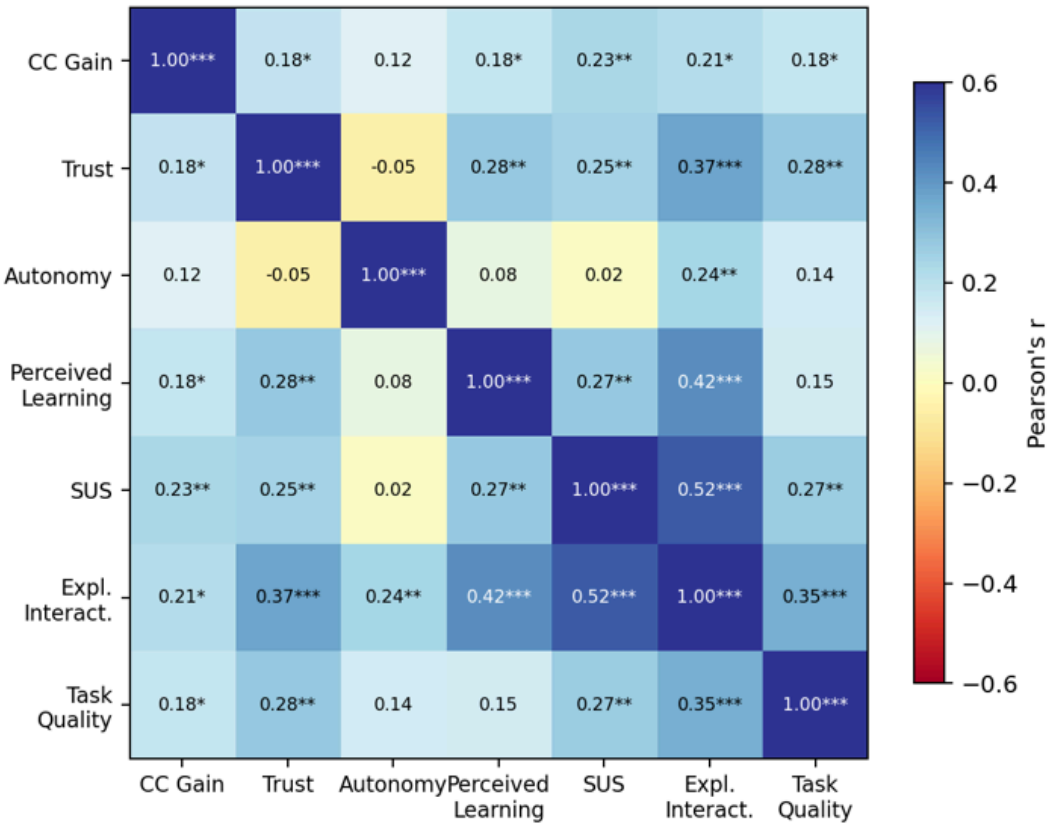



***Fig. 6*** *Correlation matrix of primary outcome variables* ($N = 128$). $^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$

**Figure 6**

See image above for figure legend.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- XAISupplementaryDatav2.xlsx