Center for Management Science Research
ISSN: 3006-5291 3006-5283 Volume 4, Issue 1, 2026

EVALUATING PUBLIC TRUST IN ARTIFICIAL INTELLIGENCE-DRIVEN CYBERSECURITY SYSTEMS: A CASE STUDY OF UNDERGRADUATE COMPUTER SCIENCE STUDENTS

Abdul Wasay Sial[1], Muhammad Faseeh Anjum[2], Hassan Raza[3], Muhammad Ahsan Ali[4], Ehtesham Ul Haque Ijaz[5], Khadija Ahsan[6], Dr. Muhammad Arfan Lodhi[*7]

[1,2,3,4,5]*BS Computer Science, Air University Multan Campus, Multan, Pakistan*
[6]*BS Physics, NFC Institute of Engineering and Technology (NFC-IET), Multan, Pakistan*
[*7]*Higher Education Department, Punjab, Pakistan*

[1]wasaysial30@gmail.com, [2]faseehj11@gmail.com, [3]hvsvnnn@gmail.com, [4]aliahsan6971@gmail.com, [5]ehteshamulhaqueijaz@gmail.com, [6]khadijaahsan1618@gmail.com, [*7]samaritan_as@hotmail.com

DOI: https://doi.org/10.5281/zenodo.20567269

Keywords
AI trust, cybersecurity, explainability, transparency, Pakistan, BSCS students, public trust

Article History
Received: 02 December 2025
Accepted: 12 January 2026
Published: 27 January 2026

Corresponding Author: *
Dr. Muhammad Arfan Lodhi

Abstract

This study aims to investigate the trust of the BSCS undergraduate students with the AI based cybersecurity tools. Understanding whether a person of the field trusts these AI systems is of highly importance. This study has the major question of on what factors dose the trust of the people depends upon. The findings show that the general public holds a relatively neutral opinion on the matter meaning they don't fully trust the AI nor they are against it. The factors of the perceived explainability have a high effect on whether the system is to be trusted or not, for a system to be trusted, the system should give highly detailed explanations of its actions. In addition to the perceived explainability, the factor of privacy concern acts as a independent factor on the overall trust in AI based cyber security systems. This study provides an insight of trust in AI in a developing country and gives factual recommendations to policy makers based on these insights to build a trustworthy AI cybersecurity infrastructure in Pakistan.

1. Introduction

The progress of AI in today's world has changed the landscape of digital protection which has raised important questions about the technical performance and also how people respond to these technologies and whether they trust these technologies or not. In the technical aspect the AI systems have evolved significantly, but their social and psychological aspects are still unstudied especially in developing countries like Pakistan where the AI frameworks are still relatively new. This paper investigates the trust of AI cybersecurity systems among the BSCS students of Pakistani universities. The reason for selecting this population is because they are at the intersection of the technical literacy and emerging professional responsibility. The study has been organized into following sections: Section 1 establishes the background, research questions and the significance of the study. Section 2 reviews the literature and previous work done related to this field. Section 3 describes the research methodology. Section 4 presents the data analysis and its findings. Section 5 tells the implications of the study. Section 6 and 7 are conclusion and recommendations respectively.

1.1 Background of the Study

The new digital landscape faces far more complex cyber security related threats than the traditional rule based security systems. Consequently, Artificial Intelligence (AI) and Machine Learning (ML) are driving a new generation of adaptive, autonomous cyber defense mechanisms (Kaloudi & Li, 2020; Ofusori et al., 2024). The public trust in these systems lags far behind their technical capabilities. Trust is of paramount importance in the cybersecurity field, otherwise the users

may bypass the critical protocols. This trust deficit is compounded by the "black box" opacity of complex AI models, where a lack of explainability undermines user confidence and raises ethical and legal accountability concerns (Capuano et al., 2022; Cheong, 2024). Addressing this issue requires an approach to foster the trust (Lahusen et al., 2024), especially because modern risk frameworks neglect the factors that are important in AI trust (Polemi et al., 2024). With respect to the psychology, the sustained trust on AI systems depends upon the factor of perceived competence and the consideration of the human values, which is very important in this case as the data that these systems monitor is highly sensitive (Li et al., 2024). Different frameworks try to implement these metrics across the ethical and the security dimensions (Mylrea & Robinson, 2023). This is because the issues like privacy concerns and the data stealing continue to threaten the confidence of the public in these tools. Factors like these contribute to the lower trust rate of general public in these systems. While public trust in the field of cybersecurity is of very great importance, the studies that measure the trust of the general public in these systems, particularly in developing countries like Pakistan, remains critically low and scarce (Azizi et al., 2025).

1.2 Statement of the Problem

The adoptability of AI based cyber security systems in different institutes like governmental corporate sector and the educational institutions has reached an all time high. These systems deal with highly sensitive data and, often make important automated decisions about data access, threat recognition and user behavioural monitoring. These systems make these decisions with minimal human intervention and limited transparency. The BSCS students of the universities of Pakistan, who will design and deploy these systems are frequently required to interact with these systems, and implicitly will rely on these AI based cybersecurity tools, but the factors that affect this kind of trust have not been extensively studied. Broad AI trust research identifies the trust of the public in the general sense but does not study the particular trust in the cyber security based AI tools. The technical cybersecurity literature studies the system performance monitoring, threat hierarchy, and adversarial AI dimensions but it very rarely studies the perceptions of the end users and the social factors of trust in these AI systems (Cheong, 2024; Lahusen et al., 2024; Raman et al., 2024).

Different studies on AI transparency acknowledge trust implications but the public trust is typically not measured through the primary survey data rather demonstrate the importance of the XAI for user confidence. They also do not provide population specific or survey based results of the trust level of the public. As a result there is quite a significant and practically consequential gap of research. Moreover, there is very limited survey based evidence dealing with the level of trust that BSCS students of Pakistani universities hold towards the AI based cybersecurity tools and systems. There is insufficient quantitative data which identifies and explain different factors such as perceived transparency, perceived competence and privacy concerns that contributes to the overall trust in these systems. This gap is, in particular, more evident in the Pakistani context where AI based systems are just starting out, public awareness of such tools is uneven and the extent of research on this kind of topic is quite limited. It is of very high importance for the policy makers of Pakistan. If the policy makers do not understand the factors that affects or undermines the public trust in the AI based cybersecurity tools, they could design and deploy the systems that, although, are technically capable but are socially mistreated. Thus could lead to the failure of the necessary cooperation that is very much important for effective and long lasting digital security.

1.3 Research Questions

This study is guided by the following three research questions:

1. What is the level of public trust among BSCS students across Pakistani universities in AI-based cybersecurity systems?
2. What factors specifically perceived transparency, perceived competence, privacy concerns, and explainability significantly influence public trust in AI-

based cybersecurity systems among BSCS students?

3. Is there a significant positive relationship between the perceived transparency of AI-based cybersecurity systems and the level of public trust among BSCS students?

1.4 Significance of the Study

The significance of this study includes important contributions at the theoretical, practical and policy levels. In a theoretical aspect, this study extends the already established frameworks of trust in AI based cybersecurity tools including Mayer et al.'s (1995) integrative trust model and the three dimensional human-AI trust framework and the AI-TMM (Mylrea & Robinson, 2023). It specifically extends the AI-TMM into the new-found domain of the cybersecurity. The empirical and statistical testing of the different constructs of the study such as the transparency, competence, privacy concern and the explainability shows how these factors operate in this domain. The study adds viable and important domain specific evidence to the broader AI trust literature. It also reflects on the need of more targeted, context-sensitive research (Scharowski et al., 2024; Afroogh et al., 2024). In a practical aspect, this study will provide real and actionable guidance to the cybersecurity experts, system designers and developers. This study helps in the understanding of the specific factors that strongly predict the trust in AI systems. This will enable professionals to make products, systems or apps that are most likely be backed by the trust of the public, and to sustain those systems. From a policy perspective, this study will prove to be very beneficial and is directly related to the emerging AI landscape of Pakistan. Drawing on the regulatory frameworks by Lahusen et al. (2024) and Cheong (2024), this study will prove to be quite beneficial in the development and deployment of the national standards for AI transparency and accountability. This will prove to be beneficial in contributing a technically sound and secure as well as socially accepted digital ecosystem.

2 Review of the Related Literature

The literature informing this study spans three principal intellectual traditions: the psychology and sociology of trust in artificial intelligence; the governance, transparency, and accountability of AI systems; and the technical and social dimensions of AI-driven cybersecurity. This review synthesizes recent scholarship across these traditions, organized according to theoretical foundations, empirical studies, and the conceptual framework underpinning the present investigation.

2.1 Theoretical Foundations

2.1.1 Mayer, Davis, and Schoorman's (1995) Integrative Trust Model

The integrative model of organizational trust proposed by Mayer et al. (1995) remains among the most cited and empirically validated frameworks in trust scholarship. The model identifies three antecedents of trust in a trustee: ability (perceived competence in a specific domain), benevolence (perceived motivation to act in the trustor's interest), and integrity (perceived adherence to acceptable ethical and professional principles). Applied to AI systems, these dimensions map onto AI competence and reliability, system safety and user aligned design, and ethical and transparent operation, respectively. Afroogh et al. (2024) explicitly build on this tradition in their taxonomy of AI trustworthiness, distinguishing between technical trustworthiness metrics — including safety, accuracy, and robustness and axiological metrics encompassing ethical compliance, legal accountability, and respect for user autonomy. This tripartite structure directly informs the independent variables selected for the present study.

2.1.2 The Three-Dimensional Human-AI Trust Framework

Li et al. (2024) propose an integrative three-dimensional framework for understanding trust in AI, synthesizing research findings from interpersonal trust, human-automation interaction, and emerging human-AI trust literature. The framework's three dimensions the trustor (human user), the trustee (AI system), and the contextual environmen provide a comprehensive and psychologically grounded account of why trust in AI is both

similar to and qualitatively different from interpersonal trust. The authors identify warmth and competence as the most robust cross-contextual trust formation factors, demonstrating that users evaluate AI systems not only on their technical capabilities but also on their perceived alignment with human values and interests. In cybersecurity contexts, where AI systems perform consequential and privacy-sensitive monitoring tasks, this dual cognitive-moral assessment is particularly critical. The framework directly informs the present study's conceptual model and questionnaire instrument design.

2.1.3AI Trust Framework and Maturity Model (AI-TMM)

Mylrea and Robinson (2023) introduce the AI Trust Framework and Maturity Model (AI-TMM), a domain-specific theoretical and evaluative framework grounded in information theory. By applying an entropy-based lens to AI system design and governance, AI-TMM operationalizes trust as an "agreed-upon understanding between humans and machines about system performance." High entropy corresponding to unpredictability, opacity, and behavioral inconsistency in AI systems is identified as a primary trust-reducing factor, particularly in competitive and uncertain environments such as cybersecurity. The AI-TMM proposes repeatable, measurable evaluation metrics across performance, governance, and ethical dimensions, making it a practical theoretical anchor for empirical trust research in AI cybersecurity contexts and directly informing the study's trust construct operationalization.

2.2 Review of Empirical Studies

The empirical literature relevant to this study is organized into four thematic clusters including general AI trust, AI governance and accountability, AI in cybersecurity, and explainability and trust in AI security systems. This clustering facilitates a systematic identification of converging findings, contradictions, and the research gap the present study addresses.

2.2.1Cluster A: General AI Trust

Afroogh et al. (2024), in a systematic literature review spanning multiple domains of human-machine interaction, demonstrate that trust in AI is a multidimensional construct encompassing technical performance, ethical conduct, legal compliance, and user autonomy. Their proposed taxonomy distinguishes between technical trustworthiness metrics (safety, accuracy, robustness, explainability) and non-technical axiological metrics (ethical compliance, legal accountability, dignity preservation). Critically, they identify major "trust- breakers" including opaque decision-making, threats to user autonomy, lack of accountability, and discriminatory outcomes that are directly relevant to public concerns about AI cybersecurity tools. Scharowski et al. (2024), through a 24-year bibliometric review of empirical AI trust research, corroborate this complexity, highlighting the importance of calibrated trust neither excessive over-trust nor irrational under-trust and identifying cybersecurity as a notably underrepresented domain in empirical trust research, directly validating the gap addressed by the present study. Li et al. (2024) complement these findings from a psychological perspective, demonstrating that perceived competence, behavioral consistency, and value alignment are the most robust predictors of sustained trust across interpersonal, automation, and AI trust contexts. Achuthan et al. (2024), examining current trends and future directions in AI-driven cybersecurity and privacy, identify a growing tension between the operational effectiveness of AI security tools and their social acceptance among end users. Their analysis of emerging research directions highlights that advancing cybersecurity with AI requires simultaneous investment in privacy-preserving architectures and transparent system design two dimensions that directly correspond to the independent variables investigated in the present study. The authors argue that public trust in AI cybersecurity cannot be assumed on the basis of technical performance alone, and that researchers must empirically investigate trust formation processes across different user populations and governance contexts. This calls for contextually grounded empirical trust research directly motivates the present survey-

based investigation of BSCS students in Pakistan, and the study's findings contribute to the research agenda Achuthan et al. identify as a priority for the field.

2.2.2 AI Governance and Accountability

Lahusen et al. (2024) examine trust in AI from an interdisciplinary governance perspective, focusing on the deployment of algorithmic decision-making (ADM) systems by public authorities. Their analysis introduces the concept of "watchful trust" a form of conditional, informed, and actively maintained trust appropriate to high-stakes sociotechnical systems and argues that ensuring the trustworthiness of AI governance requires simultaneously addressing technical properties, institutional practices, and regulatory frameworks. Cheong (2024) extends this analysis through a detailed review of the legal and ethical challenges of AI transparency and accountability, identifying four thematic areas: technical explainability approaches, legal and regulatory frameworks (including the EU AI Act), ethical and societal considerations, and multi-stakeholder governance models. His work demonstrates that transparency is not a singular technical property but a complex, contested concept requiring contextual operationalization a finding that directly informs how perceived transparency is defined and measured in the present study. Raman et al. (2024) further contextualize these challenges in the cybersecurity domain, examining how AI's dual-use potential its simultaneous capacity for defense and offense, surveillance and protection creates acute ethical tensions that must be resolved through transparent and accountable governance if public trust is to be maintained.

2.2.3 Technical AI in Cybersecurity

Kaloudi and Li (2020), in a foundational survey of the AI-based cyber threat landscape, provide a comprehensive taxonomy of both offensive AI applications including automated vulnerability exploitation, adversarial attacks on ML models, and AI-generated social engineering and defensive AI capabilities, establishing the technological context within which public trust must be understood. Ofusori et al. (2024), in a comprehensive bibliometric and systematic review covering AI techniques in cybersecurity from 2014 to 2024, document the rapid evolution of ML, deep learning, and natural language processing applications in security tasks, and identify data privacy as an increasingly prominent user concern in the AI cybersecurity literature. Malatji and Tolah (2024) extend this technical landscape with a comprehensive framework for understanding adversarial and offensive AI dimensions, arguing that public distrust of AI cybersecurity systems may reflect rational concerns about AI's broader dual-use ecosystem rather than simply technical skepticism about specific tools. Most directly relevant to the present study, Azizi et al. (2025) examine the social impacts of AI- based cyber defense systems through a systematic literature review, documenting significant public concerns about mass surveillance, data privacy erosion, and algorithmic bias. Their conclusion that ethical frameworks and public awareness campaigns are necessary preconditions for trustenabling deployment of AI cybersecurity directly motivates the empirical investigation of the present research.

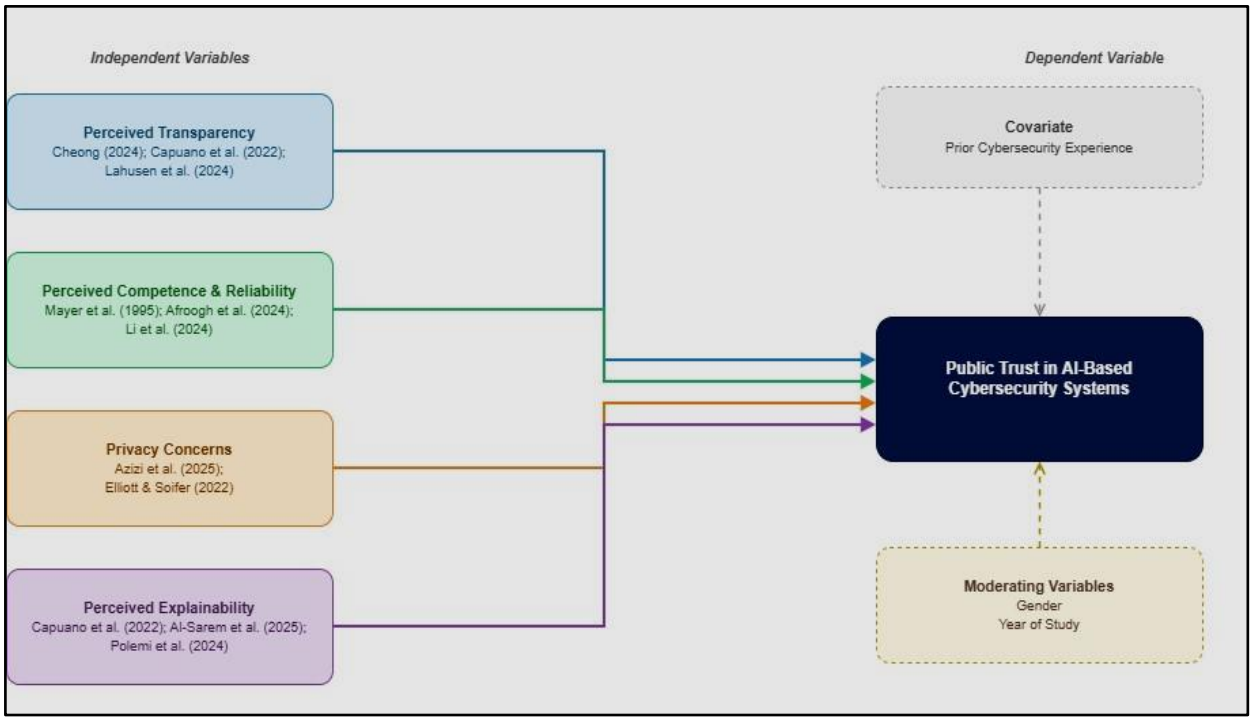
2.2.4 Explainability and Trust in AI Security Systems

Capuano et al. (2022), in a comprehensive survey of explainable AI (XAI) in cybersecurity, review methods such as SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and attention-based mechanisms as applied to intrusion detection, malware classification, and network anomaly detection. Their central argument is that explainability functions as a critical trust bridge between AI system performance and end-user confidence: when users can understand even at a high level why an AI system has raised a security alert, their confidence in the system's judgments and their willingness to act on its recommendations are substantially enhanced. Al-Sarem et al. (2025) extend this finding specifically to AI-powered intrusion detection systems (IDS), demonstrating that XAI integration improves acceptance among both technical security analysts and non-expert organizational stakeholders. Polemi et al. (2024), drawing on NIST AI RMF and ENISA frameworks, identify the neglect of human factors and socially

mediated threats as critical gaps in AI risk management, reinforcing that trustworthiness is a sociotechnical achievement rather than a purely technical property. Finally, Elliott and Soifer (2022) provide a philosophical grounding for the privacy-trust relationship, distinguishing between informational privacy and contextual integrity, and arguing that AI cybersecurity tools frequently violate users' contextual privacy expectations a violation that erodes trust in ways that are not captured by purely technical performance metrics.

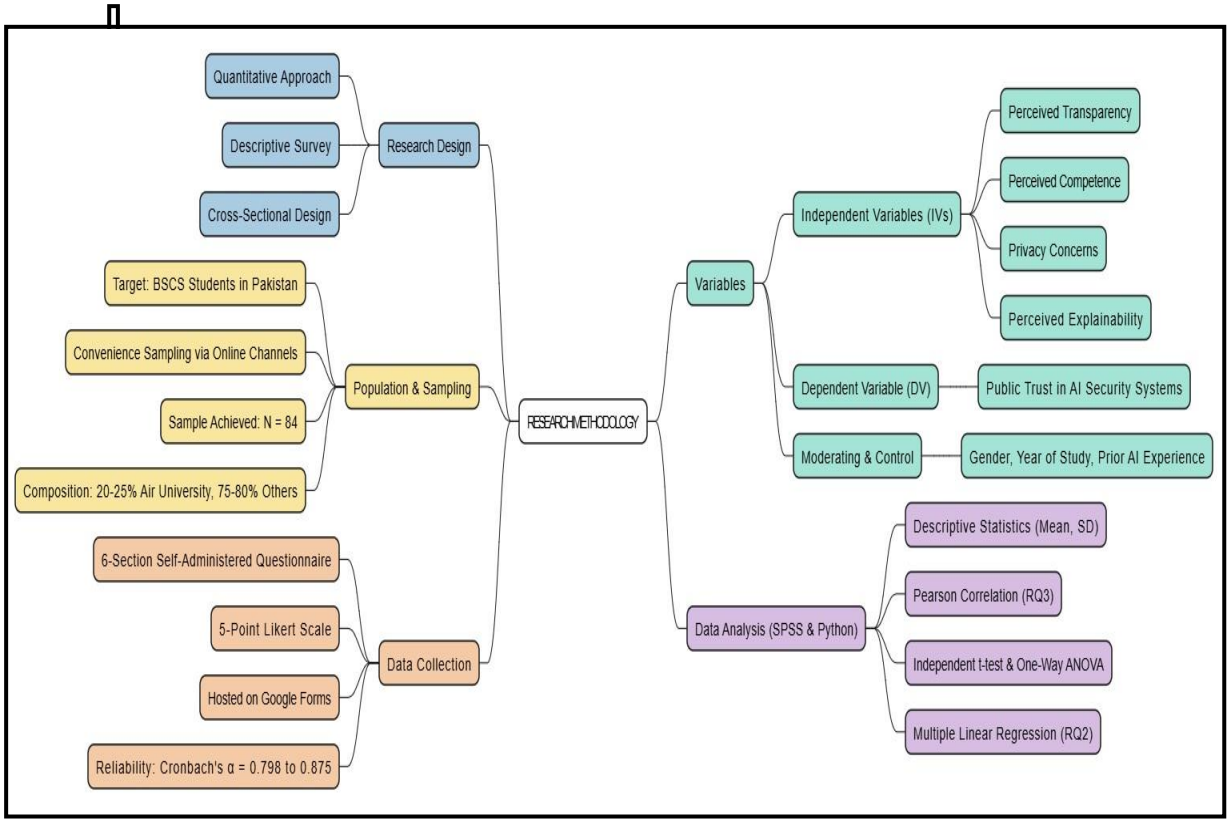
2.3 Conceptual Framework

The conceptual framework of this study proposes that Public Trust in AI-Based Cybersecurity Systems (the dependent variable) is shaped by four theoretically grounded independent variables. Perceived Transparency, Perceived Competence and Reliability, Privacy Concerns, and Perceived Explainability. This framework is synthesized from the theoretical models and empirical findings reviewed in the preceding sections and is designed to guide the operationalization of constructs, the construction of the survey instrument, and the interpretation of findings. Perceived Transparency refers to users' beliefs about the degree to which AI cybersecurity systems openly communicate their operations, decision-making processes, and data handling practices. This construct is grounded in the governance frameworks of Cheong (2024) and Lahusen et al. (2024), both of whom identify transparency as a foundational condition for trustworthy AI, and is further informed by Capuano et al. (2022), who demonstrate that user understanding of AI system behavior is a prerequisite for trust calibration in security contexts. Perceived Competence and Reliability refers to users' assessments of the technical accuracy, effectiveness, and dependability of AI cybersecurity systems.



3. Research Methodology

This section describes the research design, population and sampling strategy, data collection instruments, data collection procedure, data analysis techniques, and delimitations of the present study. The methodology is designed to be systematic, transparent, and appropriate to the study's quantitative, survey-based orientation.



A quantitative approach is appropriate given the study's objective of measuring the level of public trust in AI-based cybersecurity systems and statistically testing the relationships between trust and its proposed antecedents. A descriptive design enables the systematic characterization of these variables as they naturally exist in the study population without experimental manipulation, while a cross-sectional design allows data to be gathered from participants at a single point in time, providing a snapshot of trust attitudes at the time of data collection. This design is consistent with established methodological practices in AI trust and technology acceptance research (Li et al., 2024; Afroogh et al., 2024). The study examines the relationship between four independent variables Perceived Transparency (IV1), Perceived Competence and Reliability (IV2), Privacy Concerns (IV3), and Perceived Explainability (IV4) and one dependent variable: the Level of Public Trust in AI-Based Cybersecurity Systems (DV). The relationship between variables is both correlational testing the degree and direction of association between each IV and the DV and predictive examining through multiple regression analysis which IVs are statistically significant predictors of trust. Gender and year of study function as moderating variables, and prior cybersecurity experience is included as a covariate.

3.2 Population and Sampling

The target population for this study comprises currently enrolled Bachelor of Science in Computer Science (BSCS) students across Pakistani universities. While the survey was initially distributed through channels at Air University Multan Campus, the online distribution method reached BSCS students attending multiple universities across Pakistan, with approximately 20–25% of respondents enrolled at Air University Multan Campus and the remaining 75–80% drawn from other Pakistani higher education institutions. This broader sample composition strengthens the generalizability of findings beyond a single institution and more accurately represents the attitudes of technically educated undergraduate computer science students across Pakistan's higher education landscape. BSCS students across Pakistani universities represent a particularly appropriate target population for this study because they possess foundational knowledge of computing, networking, and cybersecurity concepts, enabling them to form

informed and meaningful judgments about AI-based security systems. They also represent the emerging professional cohort that will design, deploy, and interact with these systems in their careers, making their trust attitudes of significant practical importance regardless of institutional affiliation.

3.3 Data Collection Instruments

Data was collected using a structured, self-administered questionnaire developed specifically for this study, drawing on validated trust constructs from Afroogh et al. (2024) and Mylrea and Robinson (2023). The instrument employs a five-point Likert scale with response options ranging from 1 (Strongly Disagree) to 5 (Strongly Agree), which is the standard measurement approach for attitudinal and perceptual constructs in social science survey research. The questionnaire is organized into six sections: Section A captures demographic and background information (gender, year of study, and prior cybersecurity experience, measured through self-rated familiarity with AI security tools and number of relevant courses completed); Section B measures Perceived Transparency through five items assessing users' beliefs about the understandability and communicative clarity of AI cybersecurity systems (grounded in Cheong, 2024; Capuano et al., 2022); Section C measures Perceived Competence and Reliability through five items assessing users' beliefs about the accuracy, effectiveness, and dependability of AI cybersecurity systems (grounded in Afroogh et al., 2024; Li et al., 2024); Section D measures Privacy Concerns through five items assessing user anxiety about data collection, surveillance, and potential misuse of personal information (grounded in Azizi et al., 2025; Elliott & Soifer, 2022); Section E measures Perceived Explainability through four items assessing the adequacy of AI systems' explanations of their security decisions and alerts (grounded in Capuano et al., 2022; AlSarem et al., 2025); and Section F measures the dependent variable Public Trust in AI-Based Cybersecurity Systems through six items assessing overall trust, willingness to rely on these systems, and confidence in their ethical operation (grounded in Afroogh et al., 2024; Li et al., 2024). Prior to main data collection, the questionnaire will be pilot-tested on a convenience sample of n = 30 BSCS students, with internal consistency reliability assessed using Cronbach's Alpha (acceptable threshold: $\alpha \geq 0.70$ per subscale) and content validity established through faculty expert review.

3.4 Data Collection Procedure

Primary data collection will be conducted via an online, self-administered questionnaire hosted on Google Forms. Following institutional ethical clearance and departmental permission, the survey link will be distributed to eligible BSCS students through official class communication channels including WhatsApp groups and social media platforms reaching BSCS students across multiple Pakistani universities. Participation will be entirely voluntary, and all participants will be informed through a preamble displayed at the beginning of the questionnaire of the study's purpose, the voluntary and anonymous nature of their participation, and their unconditional right to withdraw at any time without consequence. Informed consent will be obtained digitally through a mandatory acknowledgment checkbox on the questionnaire's first page. No personally identifiable information will be collected. Survey responses will be stored securely in Google Forms and exported to a password protected spreadsheet for analysis. The data collection window will span approximately two weeks, with a single reminder distributed through the same channels after the first week to encourage participation among non-respondents. Questionnaires with more than 20% missing data will be excluded from analysis prior to processing.

3.5 Data Analysis Technique

Data analysis will be conducted using SPSS (Statistical Package for Social Sciences) and Python (with Pandas and SciPy libraries), proceeding in three structured phases. In the first phase, descriptive statistics including frequencies, percentages, means, and standard deviations will be computed for all study variables to characterize the sample and profile the distribution of responses across each construct; demographic data will be presented through frequency tables and bar charts. In the

second phase, inferential statistical tests will be applied to address each research question: to answer RQ1, mean scores on the Public Trust subscale will be calculated and interpreted against the scale midpoint (3.0), with scores above 3.0 indicating a positive trust disposition; to address RQ3, Pearson correlation analysis will test the bivariate relationship between perceived transparency and public trust at a significance level of $p < 0.05$; independent samples t-tests will compare trust scores by gender, and one- way Analysis of Variance (ANOVA) with post-hoc Tukey HSD compared trust scores across year-of-study groups.

3.6 Delimitations of the Study

Several delimitations define the scope and boundaries of this investigation. First, the study is limited to BSCS students across Pakistani universities who responded to an online convenience survey. While this broadens the sample beyond a single institution, it remains limited to technically educated undergraduate computer science students and is not representative of the general public, non-computing disciplines, or working industry professionals. Findings should therefore be generalized with appropriate caution. Second, as a cross-sectional study, the research captures trust attitudes at a single point in time and is not designed to examine how trust in AI cybersecurity systems evolves in response to system updates, data breach events, or regulatory changes. Third, the reliance on self-reported perceptual data means the study measures stated attitudes toward AI cybersecurity systems rather than observing actual behavioral reliance on these tools in real-world security scenarios; the gap between stated attitudes and actual behavior remains beyond the scope of this investigation. Fourth, the questionnaire addresses AI-based cybersecurity systems as a broad category rather than evaluating specific commercial products or institutional deployments, which may limit the precision of findings for practitioners concerned with particular system implementation

4 Data Analysis

This section presents the findings of the data analysis conducted on valid responses obtained from 84 BSCS students at Air University Multan Campus. The analysis is organized into four sub-sections: demographic analysis of the sample, item-level analysis of each questionnaire section independently, construct-level descriptive statistics and reliability analysis, and inferential statistical analyses addressing each research question.

4.1 Demographic Analysis

The final study sample comprised N = 84 respondents. Table 1 presents the full demographic profile of the sample across all five background variables collected in Section A of the questionnaire.

Table 1 *Demographic Profile of the Sample (N = 84)*

Variable	Category	n	%
Gender	Male	71	84.5
	Female	13	15.5
Year of Study	Year 1	3	3.6
	Year 2	12	14.3
	Year 3	38	45.2
	Year 4	31	36.9
AI Familiarity	Very Low	2	2.4
	Low	2	2.4
	Moderate	42	50.0
	High	27	32.1
	Very High	11	13.1
Cybersecurity Courses	None indicated	25	29.8
	1–2 courses	43	51.2
	3–4 courses	11	13.1
	5 or more courses	5	6.0

Used AI Cybersecurity Tool	Yes	32	38.1
	No	33	39.3

As shown in Table 1, the sample was predominantly male (n = 71, 84.5%), consistent with the gender composition of the BSCS program. The majority of respondents were in Year 3 (n = 38, 45.2%) and Year 4 (n = 31, 36.9%), indicating that most participants possessed considerable academic exposure to computing and cybersecurity concepts. Half of all respondents (50.0%) rated their AI familiarity as moderate, with a further 45.2% rating it as high or very high, suggesting a generally informed sample capable of providing meaningful perceptual assessments of AI cybersecurity systems. A total of 32 respondents (38.1%) confirmed they had personally used an AI-based cybersecurity tool, while 33 (39.3%) had not. Figures 1 and 2 illustrate the gender and year-of-study distributions visually.

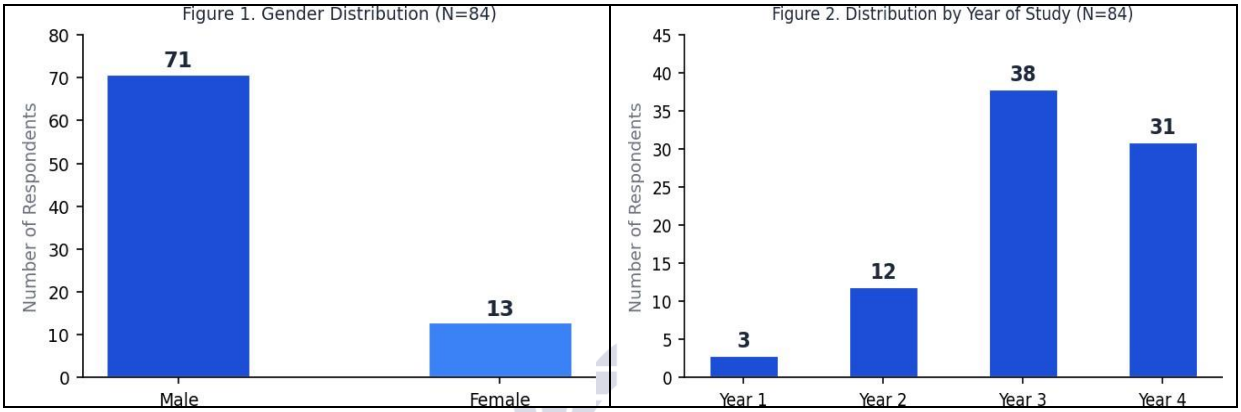


Figure 1. Gender Distribution of Respondents (N = 84)
Figure 2. Distribution of Respondents by Year of Study (N = 84)

4.2 Item-Level Analysis of each Section

In accordance with the professor's requirement for independent section-level analysis, this sub-section presents item-by-item descriptive statistics mean (M), standard deviation (SD), and response frequency distribution for every individual question across Sections B through F. This granular analysis enables identification of specific items that most strongly shaped each construct's overall score and highlights areas of agreement or disagreement within each section. Figure 3 provides a visual overview of item-level mean scores across all five sections.

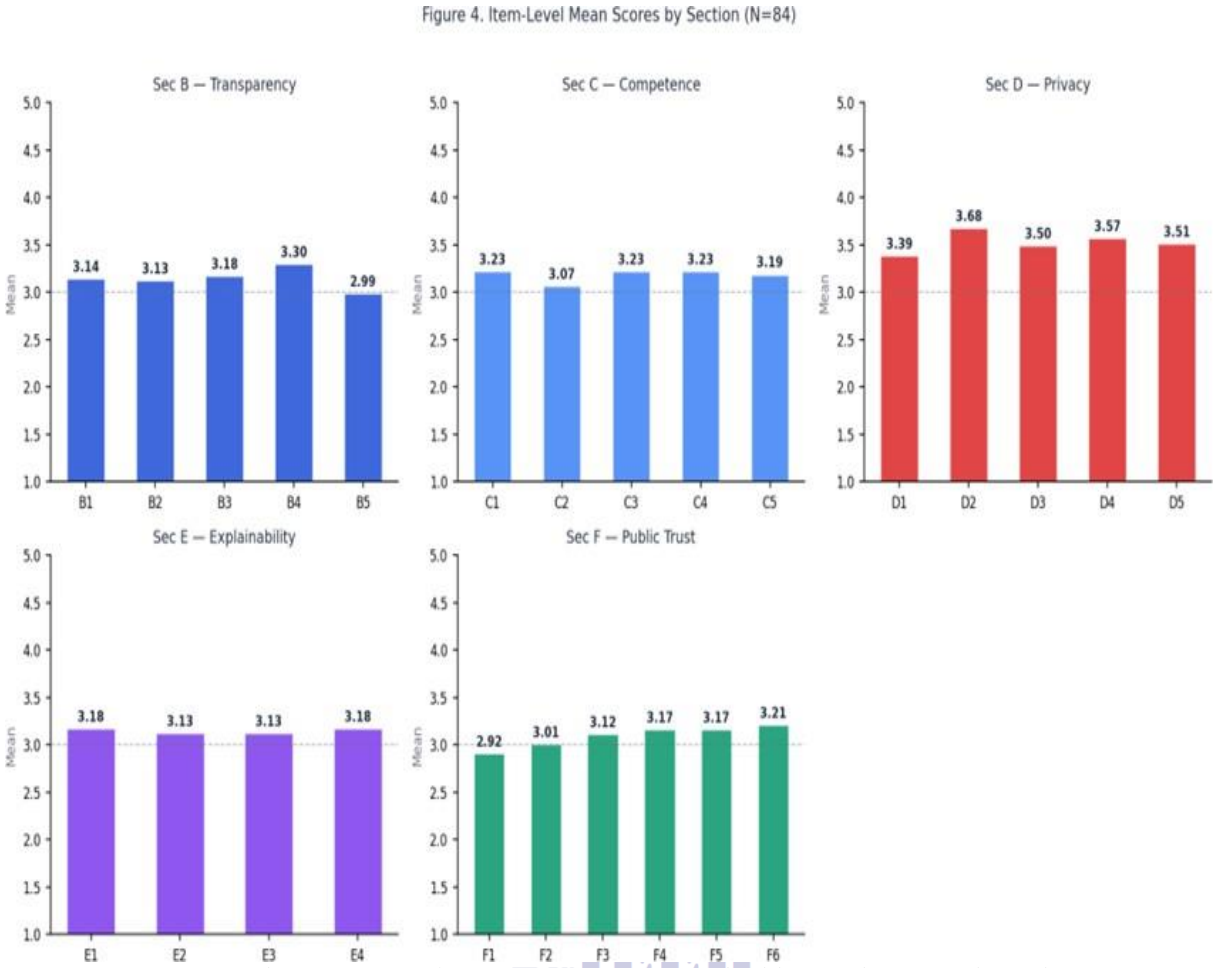


Figure 3. Item-Level Mean Scores by Section (N = 84). Dashed line indicates scale midpoint (3.0).

4.2.1Section B: Perceived Transparency (Items B1–B5)

Table 2 presents item-level descriptives for Section B, which measured respondents' perceptions of how openly and understandably AI-based cybersecurity systems communicate their operations and decisions.

Table 2

Table 2. Item-Level Descriptive Statistics: Section B — Perceived Transparency

item	Statement (abbreviated)	M	SD	SD%	D%	N%	A%	SA%
B1	I understand how AI cybersecurity systems make decisions.	3.14	1.16	13.1	11.9	32.1	33.3	9.5
B2	AI systems clearly communicate reasons behind alerts.	3.13	1.11	9.5	16.7	34.5	29.8	9.5
B3	Operation of AI cybersecurity systems is easy to understand.	3.18	1.10	9.5	14.3	34.5	32.1	9.5
B4	AI systems provide sufficient info on data protection.	3.30	1.00	7.1	10.7	34.5	40.5	7.1
B5	AI systems operate openly without hiding information.	2.99	1.09	14.3	13.1	35.7	33.3	3.6
Section Mean: M = 3.148, SD = 0.876, Cronbach's α = .860								

Note. SD = Strongly Disagree, D = Disagree, N = Neutral, A = Agree, SA = Strongly Agree. All values are percentages (%) except M and SD.

The Perceived Transparency construct recorded an overall mean of M = 3.148 (SD = 0.876), marginally above the scale midpoint. Item B4 "AI systems provide sufficient information

about how they protect users' data" yielded the highest mean in this section (M = 3.30), with 47.6% of respondents agreeing or strongly agreeing, suggesting that data protection communication is perceived slightly more positively than other transparency dimensions. Item B5 "AI systems operate openly without hiding important information" recorded the lowest mean (M = 2.99), the only item in this section to fall below the midpoint, with 27.4% disagreeing or strongly disagreeing. This finding indicates that system openness is the weakest perceived dimension of transparency. Across all items, the largest response category was Neutral (32.1%–35.7%), reflecting widespread ambivalence about the transparency of AI cybersecurity systems among this sample

4.2.2 Section C — Perceived Competence and Reliability (Items C1–C5)

Table 3 presents item-level descriptives for Section C, which assessed respondents' beliefs about the accuracy, effectiveness, and technical reliability of AIbased cybersecurity systems.

Table 3 Item-Level Descriptive Statistics: Section C: Perceived Competence and Reliability

Item	Statement (abbreviated)	M	SD	SD%	D%	N%	A%	SA%
C1	AI systems are accurate in detecting real cyber threats.	3.23	1.00	4.8	16.7	39.3	29.8	9.5
C2	AI systems are reliable enough for critical environments.	3.07	1.08	10.7	15.5	36.9	29.8	7.1
C3	AI systems outperform traditional security systems.	3.23	1.05	9.5	8.3	40.5	33.3	8.3
C4	AI systems consistently protect against evolving threats.	3.23	1.03	7.1	11.9	42.9	27.4	10.7
C5	AI systems are technically sophisticated enough.	3.19	1.01	4.8	17.9	41.7	25.0	10.7
Section Mean: M = 3.188, SD = 0.845, Cronbach's Alpha = .875								

Note. SD = Strongly Disagree, D = Disagree, N = Neutral, A = Agree, SA = Strongly Agree. All values are percentages (%) except M and S

The Perceived Competence construct recorded an overall mean of M = 3.188 (SD = 0.845), reflecting a modestly positive perception of AI cybersecurity system effectiveness. Items C1, C3, and C4 shared the highest mean scores in this section (M = 3.23 each), with approximately 39%–41.5% of respondents selecting Neutral and a meaningful proportion (36.3%–38.1%) agreeing or strongly agreeing that AI systems are accurate and outperform traditional alternatives. Item C2 reliability in critical environments such as hospitals and banks recorded the lowest mean (M = 3.07), reflecting greater hesitancy among respondents about deploying AI cybersecurity in high-stakes settings. The consistency of neutral responses (36.9%–42.9%) across all items indicates that competence perceptions remain uncertain rather than firmly positive or negative within this sample.

4.2.3 Section D: Privacy Concerns (Items D1–D5)

Table 4 presents item-level descriptives for Section D, which assessed the extent of respondents' concerns about personal data handling, surveillance risks, and privacy implications of AI-based cybersecurity systems. Note that higher scores in this section indicate stronger privacy concern rather than more positive perceptions.

Table 4. Item-Level Descriptive Statistics: Section D: Privacy Concerns

Item	Statement (abbreviated)	M	SD	SD%	D%	N%	A%	SA%
D1	AI systems collect more data than necessary.	3.39	1.17	6.0	16.7	31.0	25.0	21.4
D2	AI systems could be used to surveil users without knowledge.	3.68	1.01	3.6	6.0	32.1	35.7	22.6
D3	Personal data could be accessed by unauthorized parties.	3.50	1.04	4.8	8.3	36.9	32.1	17.9
D4	AI systems pose a risk to users' digital privacy rights.	3.57	1.01	3.6	8.3	34.5	34.5	19.0
D5	Uncomfortable with behavioral data analysis by AI systems.	3.51	1.08	7.1	7.1	29.8	39.3	16.7
Section Mean: M = 3.531, SD = 0.862, Cronbach's Alpha = .870								

Note. SD = Strongly Disagree, D = Disagree, N = Neutral, A = Agree, SA = Strongly Agree. All values are percentages (%) except M and SD. Higher scores indicate stronger privacy concern.

Privacy Concerns recorded the highest overall mean of all constructs (M = 3.531, SD = 0.862), indicating that this sample holds elevated concerns about the privacy implications of AI cybersecurity systems. Item D2 "AI systems could be used to monitor or surveil users without their knowledge" yielded the highest mean in this section (M = 3.68) and across all 25 Likert-scale items in the entire questionnaire, with 58.3% of respondents agreeing or strongly agreeing with this concern. Item D4 risk to digital privacy rights showed an equal split between Neutral (34.5%) and Agree (34.5%) responses, with a further 19.0% strongly agreeing, underscoring widespread concern about privacy rights. Even the lowest-scoring item in this section, D1 (M = 3.39), exceeded the scale midpoint substantially, confirming that privacy anxiety is a pervasive and consistent dimension across all five items. These results present the clearest and most unambiguous finding in the dataset: BSCS students at Air University Multan Campus harbour strong and consistent concerns about the privacy practices of AI cybersecurity systems.

4.2.4 Section E: Perceived Explainability (Items E1–E4)

Table 5 presents item-level descriptives for Section E, which assessed the extent to which respondents believed AI cybersecurity systems adequately explain their decisions, alerts, and security recommendations to users.

Table 5. Item-Level Descriptive Statistics: Section E: Perceived Explainability

Item	Statement (abbreviated)	M	SD	SD%	D%	N%	A%	SA%
E1	AI systems explain why they flag certain activities.	3.18	1.03	7.1	15.5	38.1	31.0	8.3
E2	I understand reasoning behind AI security recommendations.	3.13	1.05	6.0	21.4	35.7	27.4	9.5
E3	AI systems make decisions easy for non-expert users.	3.13	1.00	8.3	11.9	45.2	27.4	7.1
E4	Explanations are sufficient to feel confident about decisions.	3.18	0.97	3.6	19.0	42.9	25.0	9.5
Section Mean: M = 3.155, SD = 0.801, Cronbach's Alpha = .798								

Note. SD = Strongly Disagree, D = Disagree, N = Neutral, A = Agree, SA = Strongly Agree. All values are percentages (%) except M and SD.

Perceived Explainability recorded an overall mean of M = 3.155 (SD = 0.801), modestly above the scale midpoint. All four items in this section yielded near-identical mean scores (range: 3.13–3.18), indicating a consistent if moderate level of perceived explainability

across all dimensions measured. Item E3 "AI systems make it easy for non- expert users to understand why a security decision was made" attracted the highest proportion of Neutral responses (45.2%), suggesting that accessibility of explanations for general users is a dimension where opinions are most divided. Item E2 attracted the highest proportion of Disagree responses (21.4%), indicating that understanding the detailed reasoning behind AI security recommendations is the most challenging aspect of explainability for this sample.

4.2.5Section F: Public Trust in AI-Based Cybersecurity Systems (Items F1–F6)

Table 6 presents item-level descriptives for Section F, the dependent variable of this study. This section measured respondents' overall trust in AI-based cybersecurity systems across six facets: protective trust, reliance willingness, user interest alignment, ethical confidence, recommendation intention, and general trust level.

Table 6 Public Trust in AI Cybersecurity Systems (N = 84)

Item	Statement (abbreviated)	M	SD	SD%	D%	N%	A%	SA%
F1	I trust AI cybersecurity systems to protect my data.	2.92	1.07	11.9	19.0	40.5	22.6	6.0
F2	I would willingly rely on AI cybersecurity for my devices.	3.01	1.00	7.1	21.4	40.5	25.0	6.0
F3	AI cybersecurity systems act in the best interest of users.	3.12	1.03	8.3	14.3	42.9	26.2	8.3
F4	I feel confident about the ethical standards of AI systems.	3.17	0.90	4.8	15.5	41.7	34.5	3.6
F5	I would recommend AI cybersecurity systems to others.	3.17	1.11	10.7	11.9	36.9	31.0	9.5
F6	My overall trust in AI cybersecurity systems is high.	3.21	1.14	8.3	14.3	41.7	19.0	16.7
Section Mean: M = 3.099, SD = 0.820, Cronbach's Alpha = .875								

Note. SD = Strongly Disagree, D = Disagree, N = Neutral, A = Agree, SA = Strongly Agree. All values are percentages (%) except M and SD.

Perceived Explainability recorded an overall mean of M = 3.155 (SD = 0.801), modestly above the scale midpoint. All four items in this section yielded near-identical mean scores (range: 3.13–3.18), indicating a consistent if moderate level of perceived explainability across all dimensions measured. Item E3 "AI systems make it easy for non- expert users to understand why a security decision was made" attracted the highest proportion of Neutral responses (45.2%), suggesting that accessibility of explanations for general users is a dimension where opinions are most divided. Item E2 attracted the highest proportion of Disagree responses (21.4%), indicating that understanding the detailed reasoning behind AI security recommendations is the most challenging aspect of explainability for this sample. The relative uniformity of scores across all four items suggests that explainability perceptions are multidimensionally consistent, without any single dimension being clearly better or worse addressed.

4.3 Construct-Level Descriptive Statistics and Reliability

Table 7 presents construct-level descriptive statistics and Cronbach's Alpha reliability coefficients for all five study constructs. Construct scores were computed as the mean of their respective items.

Table 7 *Construct-Level Descriptive Statistics and Reliability (N =84)*

Construct	Items	M	SD	Min	Max	α
Perceived Transparency	5	3.148	0.876	1.00	5.00	.860
Perceived Competence	5	3.188	0.845	1.00	5.00	.875
Privacy Concerns	5	3.531	0.862	1.00	5.00	.870
Perceived Explainability	4	3.155	0.801	1.00	5.00	.798
Public Trust (DV)	6	3.099	0.820	1.00	4.50	.875

Note. M = Mean; SD = Standard Deviation; α = Cronbach's Alpha; DV = Dependent Variable. All scales: 1–5 Likert. Cronbach's Alpha ≥ .70 is the acceptable threshold for internal consistency.

Each of the five Cronbach's Alpha coefficients surpassed .70, varying between .798 for Explainability and .875 shared by Competence and Trust this supports reliable internal structure throughout every scale. At M = 3.531, Privacy Concerns stood out with the highest average, far surpassing remaining variables and revealing notably stronger reactions within the sample. Following close behind were Perceived Competence at 3.188, then Explainability at 3.155, alongside Transparency slightly lower at 3.148 - all positioned just above neutral ground. Sitting at the bottom was Public Trust, registering a mean of 3.099, which hints: even when people view AI features somewhat favorably, such views still fall short of building firm confidence. Displayed visually in Figure 4 are the averages per construct, including error margins based on standard deviations.

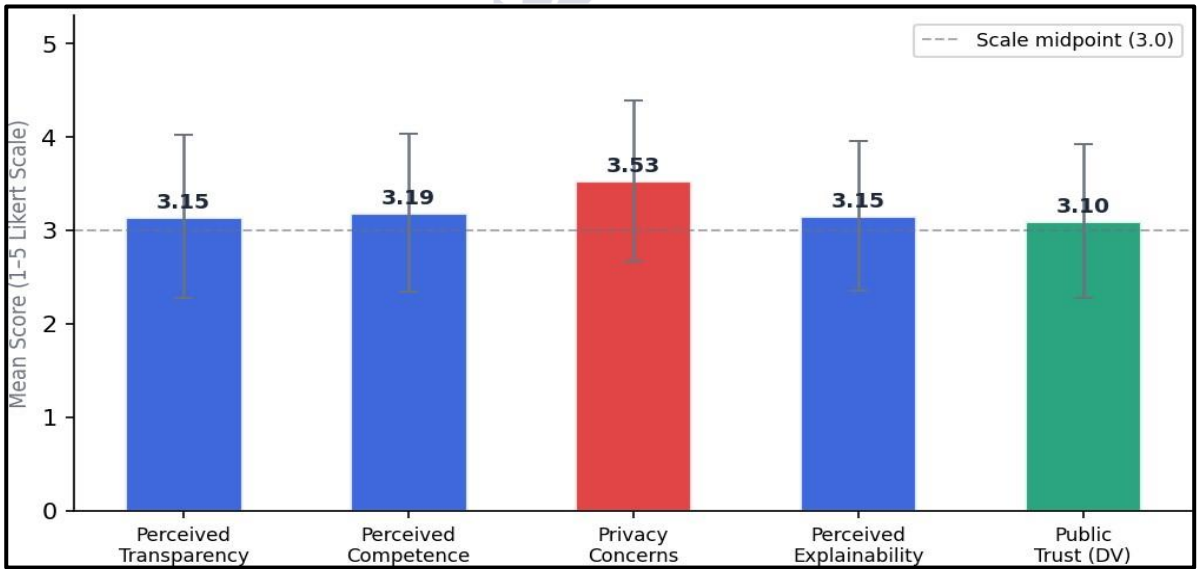


Figure 4. Construct-Level Mean Scores with Standard Deviation Error Bars (N = 84).

4.4 Inferential Statistical Analysis

4.4.1 RQ1: Measuring the Level of Public Trust

To address RQ1 "What is the level of public trust among BSCS students in AI-based cybersecurity systems?" the mean Public Trust score (M = 3.099, SD = 0.820) was computed and a one-sample t-test was conducted to determine whether the observed mean differed significantly from the scale midpoint of 3.0. The test yielded t (83) = 1.109, p = .271, indicating that while the mean trust score was above the midpoint, this difference was not statistically significant at α = .05. These results demonstrate that BSCS students across Pakistani universities hold a marginally positive but essentially neutral disposition toward AI-based cybersecurity systems cautiously leaning

toward trust without reaching a level of clear, confident endorsement. The item-level analysis in Section 4.2.5 further reveals that this neutral overall mean masks notable within-construct variation.

4.4.2 RQ3: Pearson Correlation Analysis

To address RQ3 "Is there a significant positive relationship between perceived transparency and public trust?" and to explore all bivariate relationships between IVs and the DV, Pearson correlation analysis was conducted. Table 8 presents the full correlation matrix

Table 8 Pearson Correlation Matrix for All Study Constructs (N = 84)

Construct	1	2	3	4	5 (Trust)
1. Transparency	—				
2. Competence	.577**	—			
3. Privacy Concerns	.296**	.171	—		
4. Explainability	.653**	.720**	.233*	—	
5. Public Trust	.427**	.574**	-.016	.570**	—

Note. * p < .05 (two-tailed). ** p < .001 (two-tailed).

The correlation results are presented in Table 8 and illustrated in Figure 5. Perceived Explainability demonstrated the strongest positive correlation with Public Trust (r = .570, p< .001), followed closely by Perceived Competence (r = .574, p < .001) and Perceived Transparency (r = .427, p < .001). All three correlations were positive, moderate in magnitude, and highly statistically significant, confirming that higher perceptions of explainability, competence, and transparency are each meaningfully associated with greater public trust in AI cybersecurity systems. Addressing RQ3 directly: Perceived Transparency was significantly and positively correlated with Public Trust (r = .427, p < .001), confirming the hypothesized positive relationship. Privacy Concerns, however, showed a near-zero and non-significant correlation with Public Trust (r = -.016, p = .883), indicating that privacy concern levels did not meaningfully covary with overall trust in this sample.

4.4.3 Gender Differences in Trust: Independent Samples T-Test

An independent samples t-test was conducted to examine whether Public Trust scores differed significantly between male and female respondents. Male respondents (n = 71) recorded a mean trust score of M = 3.042 (SD = 0.818), while female respondents (n = 13) recorded a higher mean of M = 3.410 (SD = 0.789). The t-test yielded t(82) = -1.498, p = .138, indicating no statistically significant gender difference in public trust at the $\alpha = .05$ level. Although female respondents reported higher mean trust than male respondents, this difference did not reach statistical significance in the current sample. This contrasts with the finding in the earlier 71-respondent sample and illustrates the sensitivity of this comparison to sample size, particularly given the small and unequal female subsample (n = 13). This finding should therefore be interpreted cautiously.

4.4.4 One-Way ANOVA

A one-way ANOVA was conducted to examine whether Public Trust scores differed significantly across the four year-of-study groups. Mean trust scores were as follows: Year 1 (n = 3, M = 3.111, SD = 0.585), Year 2 (n = 12, M = 3.222, SD = 0.952), Year 3 (n = 38, M = 3.171, SD = 0.714), and Year 4 (n = 31, M = 2.962, SD = 0.920). The ANOVA yielded F (3, 80) = 0.466, p = .707, indicating no statistically significant difference in public trust across year-of-study groups. Trust attitudes toward AI-based cybersecurity systems appear to remain broadly stable across academic years within this sample, suggesting that increasing academic exposure to computing and cybersecurity does not substantially shift trust perceptions in a consistent directional manner.

4.4.5 RQ 2: Multiple Linear Regression Predictors of Public Trust

To address RQ2 “What factors significantly influence public trust in AI-based cybersecurity

systems?" a multiple linear regression analysis was conducted with Public Trust as the dependent variable and Perceived Transparency, Perceived Competence, Privacy Concerns, and Perceived Explainability as standardized predictors. Table 9 presents the full regression results.

Table 9 Multiple Linear Regression

Predictor	B	SE	β	t	p	VIF
(Intercept)	3.099	0.071	—	43.90	.000	—
Perceived Transparency	0.063	0.097	.063	0.65	.521	1.89
Perceived Competence	0.264	0.104	.264	2.54	.013*	2.17
Privacy Concerns	-0.139	0.074	-.139	-1.88	.064	1.10
Perceived Explainability	0.266	0.112	.266	2.37	.020*	2.53
R2 = .408; Adjusted R2 = .378; F(4, 79) significant at p < .001						

Note. * p < .05. B = unstandardized coefficient; SE = standard error; β = standardized beta coefficient; VIF = Variance Inflation Factor. All VIF < 10, confirming absence of multicollinearity.

The regression model was statistically significant and explained 40.8% of the variance in Public Trust (R2 = .408, Adjusted R2 = .378). This is a substantial proportion of explained variance for a social-science survey study, confirming that the four IVs collectively possess strong predictive utility in accounting for public trust in AI-based cybersecurity systems. All VIF values ranged from 1.10 to 2.53, well below the critical threshold of 10, confirming the absence of problematic multicollinearity.

5. Implications of the Study

The findings of this study carry meaningful implications at theoretical, practical, and policy levels. Theoretically, the confirmation that perceived competence and perceived explainability are the two significant predictors of public trust extends Mayer et al.'s (1995) integrative trust model and Li et al.'s (2024) three-dimensional human-AI trust framework into the specific domain of AI-based cybersecurity, while the dissociation between privacy concerns and trust challenges prevailing assumptions in AI governance scholarship that privacy anxiety is a primary driver of reduced trust suggesting that technically educated users are capable of simultaneously holding strong privacy concerns and maintaining a conditional, neutral trust disposition. Practically, the findings present a clear design priority for cybersecurity system developers: investment in user-facing explainability interfaces and accessible performance disclosure should be treated as core trust-building requirements, not optional features, given that these two factors independently and significantly predicted trust in this sample; the elevated and pervasive privacy concerns documented across the dataset further signal an unstable equilibrium that risks translating into active trust erosion if left unaddressed. From a policy standpoint, the neutral overall trust level among technically educated students underscores that public trust in AI cybersecurity cannot be assumed on the basis of technical deployment alone, and Pakistani regulatory bodies including the NTA and the Personal Data Protection Bill framework should incorporate provisions requiring AI cybersecurity systems to disclose data collection practices, provide meaningful explanations of AI-driven decisions, and submit to independent accountability audits, aligned with the NIST AI Risk Management Framework and international transparency standards.

6 Conclusion

This study set out to determine to how much extent the students of BSCS of Pakistani universities trust AI based cybersecurity systems. After extensive data analysis, this study provided answers to the three research questions. In the response of first question, the data analysis showed that the population sample or the students of BSCS of universities of Pakistan have a almost neutral stand. Even though they are slightly positive but they don't blindly or fully trust these systems. This is evident in the way that one of the questions that asked if students directly trusted AI to

protect the personal information, the response is below neutral. This gives valuable insights for the policy makers of Pakistan. With regard to the second research question, the study found two main factors that affected, and actively built trust of the people in these systems. Those two are the perceived competence and the perceived explainability. This means that if the AI is accurate and can properly explain and back its action through facts, the students will rely on it. This is interesting as the students' reports high privacy concerns but these concerns did not lower their trust. This indicates that the fear or the anxiety is completely separate from the other factors in trusting these systems. All in all, the study shows that AI trust can be increased if the AI properly explains its actions. For the third research question, the data analysis showed that there is a very strong correlation between the transparency and the public trust meaning the more transparent the AI is, the more it is trusted. This further cements the idea that AI needs to be more transparent in its responses to increase its trust.

The current study, as all other studies have, a few limitations that should be acknowledged. The convenience sample size was significantly smaller then the initial proposed sample size due to the lack of time and the fact that the responses are predominantly male which shows the fact that there is a higher concentration of male students in the Pakistani universities. It is also important to note that the survey was collected at a specific point in time. It also asked about the AI in cybersecurity in general rather than a specific product. This limits the extent to which these findings can be applied to individual products. All in all, the trust in AI based cyber security related systems among the BSCS students of Pakistani universities is at a cautious neutral. It depends upon to how much extent the AI system explain its actions and be transparent about its actions and responses. Additionally the transparency in AI is directly related to the public trust in these systems.

7 Recommendations

Based on the public response of this study, the following recommendations are directed at four key points. The Developers and System designer of Cybersecurity domain should ensure the integration of user-facing Explainable AI mechanisms and workings such as plain-language alert explanations and accessible decision-log interfaces as their basic and primary trust-building design requirement, They should integrate the metrices like detection accuracy and false positive rates continuously available to the system users and ensure that it should become a part of their performance matrix rather than just restricting this information to technical documentation.AI cybersecurity tools deployed by the institutions should contain system rollout with structured user education programs covering not only the system capabilities but also the data collection practices and available redress mechanisms, As finding that trust does not increase with academic year points toward that passive exposure to these tools is not enough to develop the required trust. It is necessary to incorporate the specific transparency, privacy disclosure and accountability requirements for AI cybersecurity system into national governance frameworks and in the Higher Education Commission should consider mandating AI ethics and cybersecurity trust in curriculum of first-year BSCS to cultivate the calibrated trust dispositions among the future of AI professionals from the outset of their training. Therefore the Pakistani policymakers and regulatory bodies should consider it as a necessity. Finally, future researchers should pursue longitudinal and mixed-methods studies with larger and more demographically diverse samples to investigate how trust evolves over time, to explore the cognitive mechanisms behind the privacy-trust dissociation identified in this study, and to account for the substantial proportion of trust variance that the present model leaves unexplained.

References

Achuthan, K., Ramanathan, S., Srinivas, S., & Raman, R. (2024). Advancing cybersecurity and privacy with artificial intelligence: current trends and future research directions. *Frontiers in Big Data*, 7, 1497535. https://doi.org/10.3389/fdata.2024.1497535

Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: progress, challenges, and future directions. *Humanities and Social Sciences Communications*, *11*(1). https://doi.org/10.1057/s41599-024-04044-8

Azizi, A., Mohammadi, M. Q., & Samadzai, A. W. (2025). Ai In Cyber Defense: Privacy Risks, Public Trust, And Policy Challenges. *Jurnal Ilmiah Dinamika Sosial*, *9*(1), 103–118. https://doi.org/10.38043/jids.v9i1.6278

Benk, M., Kerstan, S., Von Wangenheim, F., & Ferrario, A. (2024). Twenty-four years of empirical research on trust in AI: a bibliometric review of trends, overlooked issues, and future directions. *AI & Society*, *40*(4), 2083–2106. https://doi.org/10.1007/s00146-02402059-y

Capuano, N., Fenza, G., Loia, V., & Stanzione, C. (2022). Explainable Artificial Intelligence in CyberSecurity: a survey. *IEEE Access*, *10*, 93575–93600. https://doi.org/10.1109/access.2022.3204171

Cheong, B. C. (2024). Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making. Frontiers in Human Dynamics, 6, 1421273. https://doi.org/10.3389/fhumd.2024.1421273

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Elliott, D., & Soifer, E. (2022). AI technologies, privacy, and security. *Frontiers in Artificial Intelligence*, *5*, 826737. https://doi.org/10.3389/frai.2022.826737

Kaloudi, N., & Li, J. (2020). The AI-Based cyber threat landscape. *ACM Computing Surveys*, *53*(1), 1–34. https://doi.org/10.1145/3372823

Lahusen, C., Maggetti, M., & Slavkovik, M. (2024). Trust, trustworthiness and AI governance. *Scientific Reports*, *14*(1), 20752. https://doi.org/10.1038/s41598-024-71761-0

Li, Y., Wu, B., Huang, Y., & Luan, S. (2024). Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust. *Frontiers in Psychology*, *15*, 1382693. https://doi.org/10.3389/fpsyg.2024.1382693

Malatji, M., & Tolah, A. (2024). Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI. *AI And Ethics*, *5*(2), 883–910. https://doi.org/10.1007/s43681-024-00427-4

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. *Frontiers in Artificial Intelligence*, *8*, 1526221. https://doi.org/10.3389/frai.2025.1526221

Mylrea, M., & Robinson, N. (2023). Artificial Intelligence (AI) Trust Framework and Maturity Model: Applying an entropy lens to improve security, privacy, and ethical AI. *Entropy*, *25*(10), 1429. https://doi.org/10.3390/e25101429

Ofusori, L., Bokaba, T., & Mhlongo, S. (2024). Artificial intelligence in cybersecurity: a comprehensive review and future direction. *Applied Artificial Intelligence*, *38*(1). https://doi.org/10.1080/08839514.2024.2439609

Polemi, N., Praça, I., Kioskli, K., & Bécue, A. (2024). Challenges and efforts in managing AI trustworthiness risks: a state of knowledge. *Frontiers in Big Data*, *7*, 1381163. https://doi.org/10.3389/fdata.2024.1381163