AI-enhanced social engineering: Emerging threats and human-centric countermeasures

Salko KOVAČIĆ[1], Ivana BILIĆ[2]

[1] Dzemal Bijedic University of Mostar, Mostar, Bosnia and Herzegovina

salko@unmo.ba

[2] University of Split, Faculty of Economics, Business and Tourism, Split, Croatia

ibilic@efst.hr

Abstract: Social engineering is an attack that leverages human decision-making with the purpose of gaining access to information or causing specific actions, rather than exploiting software vulnerabilities. Modern advancements in Artificial Intelligence (AI) have enabled automated target profiling, tailored messages for every user and AI-based voice and video synthesis. Industry reports have shown that phishing activity has grown by in volume 108% since 2022.

This study provides an overview of the Proof-of-Concept Human-Centric Social Engineering Shield (HC-SES) open-source modular social engineering defense mechanism that incorporates several mechanisms including identity management (Keycloak), honeypot detection (OpenCanary), DNS filtering (Pi-Hole), email analysis (Rspamd) and adaptive learning management. The HC-SES includes adaptive micro-training provided in real time to users based on their behavioral risk assessment at decision-making points. To evaluate the feasibility of HC-SES, we conducted a Proof-of-Concept (PoC) feasibility study at the Dzemal Bijedic University of Mostar (approximately 5.000 users). The technical integration of all modules was successful and demonstrated that Keycloak authentication allowed users to be recognized across federated services, Rspamd detected phishing in approximately 1.000 emails with a 90% precision and less than 5% of login attempts required risk-adaptive Multi-Factor Authentication (MFA), with less than 1% of those being false positives. Participation in training (60%) surpassed the institutional average (40-50%) and the average time it took for participants to complete the training (65%) was less than 48 hours.

Keywords: Social engineering, Artificial Intelligence (AI), Security awareness, Open-source security, Identity management.

1. Introduction

Social engineering is a human-vulnerability based attack method on systems which use psychological aspects of humans such as trust, urgency and respect for authority to breach the existing technical measures of an organization. To obtain credentials, initiate unauthorized transactions or prompt dangerous clicks, attackers pose as trusted parties such as colleagues or administrators. Universities face elevated social engineering risk due to open access policies, diverse populations, and communication norms that prioritize accessibility over verification, creating conditions conducive to successful attacks. The use of Artificial Intelligence has further increased the risk of being a victim of social engineering. Large language models can now generate grammatically correct, contextually relevant, phishing messages en masse. This allows for mass-personalized phishing campaigns that were previously not feasible. Heiding, Schneier & Vishwanath (2024) states that "the AI will increase the amount and quality of phishing scams, therefore making it more difficult for end-users to differentiate between legitimate requests and attacks, including those who are considered to be very security conscious". Industry data confirms this: Phishing attempts have grown by 108% worldwide since 2022 (KnowBe4, 2025) and according to the FBI (2024) 67% of top IT security professionals reported that there was a successful breach. The use of AI generated content to create phishing attacks decreases the time and costs associated with creating these types of attacks and thus places organizations with the most sophisticated security practices at the same level of risk.

Current institutional responses, sporadic awareness campaigns and ad-hoc training exhibit limited efficacy because they occur outside actual decision-making contexts. Even trained employees may fail to recognize threats when cognitive load is high, urgency cues are present or requests appear contextually plausible. To protect against these types of attacks, organizations need

https://doi.org/10.33436/v36i1y202605

to implement two processes: continuous environmental risk detection and provide the user with timely and contextualized guidance at the moment of the user's decision.

We propose the Human-Centric Social Engineering Shield (HC-SES), an open-source system integrating email services, identity management, and learning platforms. HC-SES performs two functions: (1) real-time risk-scoring via linguistic patterns and behavioral metadata, and (2) adaptive micro-training at decision points. Unlike other solutions, rather than just blocking or flagging suspicious content, HC-SES provides the user with contextual examples to help them decide whether the content they received is safe or not.

HC-SES supports multiple platforms: Email/Messaging Services (Rspamd), Identity Management (Keycloak), Vulnerability Detection (OpenCanary), DNS Filtering (Pi-Hole), Learning Management Platforms (eUNMO LMS). Our proof-of-concept deployment at Džemal Bijedić University of Mostar successfully demonstrated technical feasibility, integrating real-time risk assessment with adaptive training without increasing the operational burden significantly. The full production deployment of HC-SES requires the approval of the institution after a controlled evaluation.

To confront those challenges, the study will accomplish three goals: First, it will describe the nature of social engineering threats through AI-augmentation that exist within the university environment. Second, it will provide a description of the development of the open-source Human-Centric Social Engineering Shield (HC-SES) which includes capabilities such as email analysis, identity management, honeypot-based threat intelligence, DNS filtering and adaptive micro-training. Third, it will assess the technical feasibility and initial behavioral effects of HC-SES in an actual university institutional environment. The rest of the paper will be organized into the following sections: Section 2 will review the relevant prior studies and related research; Section 3 will detail the methodology of the HC-SES including both the cognitive risk modeling and system design and prototype implementation; Section 4 will report the results of the pilot study of HC-SES and discuss its limitations; Section 5 will discuss the Keycloak-based risk-adaptive authentication; and Section 6 will summarize the implications and suggest possible directions for future studies that include controlled evaluations.

2. Background and related work

2.1. Human vulnerability and AI-driven threat evolution

Social engineering relies on humans' decision making under conditions of uncertainty through manipulation of trust, urgency and authority that (Stoica, 2021) calls the "game of perception". Individuals differ greatly in their susceptibility to phishing and spoofing attacks based upon personality characteristics and level of anxiety, which makes general awareness training ineffective (Gaw, Felten & Fernandez-Kelly, 2006; Wash, 2010). The vulnerability exists not only from a lack of knowledge, but also to how trust operates as part of human communication context.

Recent advances in Artificial Intelligence (AI) in the form of Natural Language Processing (NLP) have fundamentally changed the way attackers are able to utilize social engineering methods. With the use of AI, attackers are able to automate user profiling, collect publicly available information from users' social network profiles and generate grammatically correct, contextually relevant phishing emails at scale using Generative AI systems (Lewis, Kristensen & Caso, 2025). Industry reports confirm the escalation of these tactics: global phishing attacks have increased by 108% since 2022 (KnowBe4, 2025) and two-thirds of IT/Security leaders report having had a successful social engineering breach (Sift, 2025).

The psychology of Compliance helps explain why AI-enabled phishing attacks are effective: the emotional arousal triggered by urgency and authority cues during time sensitive events increases individuals' likelihood of engaging in deceptive behaviors (Wash, 2010; Stoica, 2021). In response to this, there has been research on identifying "Very Attacked People" (VAPs) individuals who are statistically more susceptible to attacks than other employees in an organization enabling targeted prevention strategies. As such, the system design must make decisions about how to best address fundamental trade-offs: performance vs. energy efficiency, security vs. scalability, and automated detection vs. human oversight (Czarnul et al., 2025).

2.2. Institutional policy & regulatory direction

Institutional responses have been established through several regulatory and policy initiatives. The EU Cybersecurity Strategy and the NIS 2 Directive (European Commission, 2020) established a framework that prioritises behavioural resilience, alongside technical controls. The ENISA Guidelines on Human Factors in Cybersecurity provide structured approaches to integrating human decision-making into risk assessments. The General Data Protection Regulation (GDPR) mandates regular role-specific awareness training, while eIDAS 2.0 and the Digital Europe Programme further emphasise human factors as an integral part of digital security architecture. These policies reflect the consensus that security depends on both people and technology and require defences that operate at both technical and human levels.

2.3. HC-SES ecosystem & explainable AI integration

Despite security awareness policies, a gap persists between awareness campaigns and real-time decision support. Most training occurs offline or periodically, disconnected from actual threat encounters.

Despite security awareness policies, a gap persists between awareness campaigns and real-time decision support. Most training occurs offline or periodically, disconnected from actual threat encounters. Open-source solutions address this gap through identity management (Keycloak) providing authentication context (Bonneau, 2012), vulnerability detection (OpenCanary) generating threat intelligence from honeypot observations and DNS filtering (Pi-Hole) blocking malicious domains at the network edge. These tools, combined with behavioral analytics and interpretable machine learning (Doshi-Velez & Kim, 2017; Kaur et al., 2020), create an integrated defense strategy identifying manipulations across communication channels in near real-time. This integration will allow systems to transition from binary decision-making (i.e., blocking/allowing), to responding gradually to identified risks and providing risk-aware users with the ability to make informed decisions. Beginning with prior attempts to integrate social and technical defense mechanisms, adaptive phishing training programs provide simulated phishing campaigns and role-based training; however, they are limited in that they operate in a cyclical manner unrelated to actual phishing attacks and threats (Wash, 2010). Contextualized access controls (Zero Trust architecture) assess the user device posture and location; however, the determination is simply an all or nothing determination (allow/deny) as opposed to educating the user about their own behavior. Behavioral analytical platforms (UEBA) use machine learning techniques to identify anomalous behaviors for the security analyst, not for the end-user. The HC-SES addresses these gaps by integrating risk detection with contextualized intervention at the point of decision making and providing contextualized explanations to the user regarding their potential compromise. The Human-Centric Social Engineering Solution (HC-SES) combines the psychological knowledge of susceptibility to social engineering attacks, AI-based threat modeling, human-centered defense requirements and technical components to create a modular system architecture. The HC-SES provides a bridge between detection (i.e., identifying high-risk communications) and intervention (i.e., providing adaptive learning support at the moment of decision). Using interpretable risk scores combined with contextual micro-training, the HC-SES views the human component not as a weakness to be mitigated, but as a capability to be strengthened through evidence-based support at critical moments.

3. Methodology

HC-SES development follows four stages: (1) risk analysis and requirements identification, (2) system design and integration, (3) prototype implementation and simulation and (4) iterative evaluation. This ensures user psychology drives design, technical architectures support behavioral detection and user feedback validates processes.

3.1. Requirement analysis & cognitive risk modeling

The first part of the process converts behavioral weaknesses into digital signals that can be measured. Behavioral weakness is translated through an established model of thinking about risk into three psychological factors or elements of behavior; these are propensity to trust, impulsiveness, and blind faith. Each of the three dimensions are translated into observable behavior as follows: propensity to trust is translated through link click latency, impulsive behavior is translated through form submission speed and blind faith is translated through authentication event timing and message parsing patterns. The model also includes additional contextual factors of influence that may affect behavior, including timing patterns (e.g., hour of the day, day of the week), device reputation (i.e., whether the endpoint is known as trusted or untrusted) and linguistic features (e.g., urgency markers, authority claims, social proof phrases). These layers of the model allow for the most accurate possible behavioral scoring of actual risk, because attacks occur in the context of both psychological and technical exploitation. The development of the model is informed throughout by constraints based on regulatory requirements, particularly those concerning data protection and ethical governance. Studies of eLearning and BYOD environments emphasise similar requirements for explicit security policies, layered access control and clear allocation of responsibilities between institutions and individual device owners (Anghel & Pereteanu, 2020). Processing must always comply with General Data Protection Regulation (GDPR) (i.e., right of explanation with respect to automated decision-making) and the requirements of privacy of identity and access management. The system only collects sufficient signals to assess risk, retaining minimal amounts of data and providing users with clear and understandable explanations for why each communication was flagged as risky. Research on AI interpretability (Doshi-Velez & Kim, 2017; Kaur et al., 2020) shows that users understand explanations by relating them to their own risk models, making transparency essential not just for regulatory compliance, but also for user trust.

3.2. System design & architectural integration

The second phase defines the modular, service-oriented architecture of HC-SES through the integration of open-source platforms depending on their capabilities. Each component addresses a separate detection and response function:

Digital Identity Management and Authorization (by Keycloak): Manages digital identity, authentication history and devices from which authentication was performed. (Bonneau, 2012) should enable conditional access policies related to HC-SES risk scores, supporting step-by-step response (allows challenge with MFA, requires user training before access).

Message analysis and threat filtering (by Rspamd / Exchange+Pi-Hole): Rspamd / Exchange analyzes email headers, sender reputation and content for indicators of identity theft (phishing). Pi-Hole filters DNS queries, blocking known malicious domains and phishing infrastructure on the local network before users encounter them. Together, this reduces an attacker's capabilities while also providing HC-SES with message metadata for behavioral analysis.

System Vulnerability Detection (by OpenCanary): OpenCanary deploys canary accounts and deceptive resources across the institution. When attackers interact with these decoys, OpenCanary collects tool data, attack behavior, timing and protocols and generates threat data to feed into the hc-ses risk models, showcasing new social engineering tactics specific to that institution without traditional siem overhead.

Adaptive Learning Delivery (LMS by eUNMO): Should enable case-study micro-modules for training that are triggered when the HC-SES detects elevated risk. The content of the training can be adapted to the user's role (teacher, administrator, student), language (Bosnian, English) and identified vulnerability (exploitation of user trust, pressure of urgency vs. impersonation). Post-training assessment measures knowledge retention and behavior change.

These loosely coupled architectures decouples risk detection from response, allowing institutions to gradually adopt the HC-SES. In university eLearning and BYOD settings, such decoupling is important because students and staff connect heterogeneous, personally owned devices to institutional platforms, limiting direct device control and shifting emphasis to identity, network and policy controls (Anghel, Pereteanu & Cirnu, 2020). An organization can begin by implementing OpenCanary + alongside message analysis and periodic training, later performing a full integration of adaptive training based on the case and threats. The modular design also allows for institutional customization: universities with existing LMS implementations can integrate directly, while others can include alternative LMS platforms via APIs.

3.3. Prototype implementation & simulation testing

The development of the prototype was conducted in a controlled testing environment that reflected the communications infrastructure of the intended institution. The processing models were trained on both (1) generic phishing email datasets and (2) institutional specific social engineering attempts collected through previous incident response (Gaw, Felten & Fernandez-Kelly, 2006; Sahingoz et al., 2019). This fine-tuning helps reduce the number of false positives due to generic phishing patterns and improve the detection of targeted phishing and advanced text messaging designed for academic environments.

The simulation tests measure the effectiveness of the system in three categories:

Detection Efficiency: Metrics used to evaluate the detection efficiency of the system include precision (the percentage of messages that are identified as phishing which are actually phishing), recall (the percentage of actual phishing messages that are detected) and decision latency (the time it takes for the system to generate a risk score once a message has been received).

Target Accuracy: The target accuracy of >90% will ensure users have confidence in the accuracy of the alerts generated by the system and the target latency of <300ms will provide near real-time feedback to users and avoid providing feedback at such a rate that it appears the system is causing delays.

User Susceptibility and Behavior Change: The baseline phishing susceptibility of users will be determined through simulated phishing campaigns conducted before the deployment of the HC-SMS. Post-deployment phishing campaigns of varying severities will determine the extent to which users click on messages that are identified as phishing, enter their credentials into forms that are part of phishing attacks and report phishing messages. The evaluation will track changes in user behavior post-deployment of the system using statistical controls for demographic variables and user roles.

System Integration: The simulation confirms the need for mutual interoperability between Keycloak, Rspamd/Exchange, Pi-Hole, OpenCanary and eUNMO LMS. Planned scenarios include high volume message flow, policy enforcement under concurrent authentication events and canary alarm reliability.

3.4. Iterative user-centered evaluation

The next phase carries out user evaluation within the target institution, the Dzemal Bijedic University in Mostar (UNMO) through several iterations. Throughout the concept, it is planned to involve approximately 500 users from the group of teachers, administrative staff and students.

Evaluation methods include:

Quantitative metrics: Precision/recall of actual phishing attempts, time from incident initiation to user alert, change in sensitivity after training (simulated phishing use-cases) and training completion rates.

Qualitative Method: A focus group interview will take place using (n = 50) to identify and evaluate how transparently (i.e., "Do you know why your message was flagged?", "Do you have faith in the HC-SES decisions?", "Are you overwhelmed by alerts?", "Is there training available?) the user perceives their HC-SES messages.

Contextual Variation Analysis: There are variations in how perceived risks, communication norms, and social engineering vulnerability exist among various types of academic, administrative, and support staff personnel. The evaluation will determine if the HC-SES risk assessment scores apply generally and if the training content should be tailored based on an individual's position. The contextual variation analysis should provide an output to the user, which would be based on the type of educational institution and enable replication at other institutions like those in the study without losing the cultural sensitivity of the local environment.

Interview Feedback and Metrics: Based on false positives rates, policy templates will be updated, training modules will be modified to improve content comprehension and alert thresholds will be changed to match the level of detection performance desired by the users. The product of the evaluation will include:

1. Configuration templates for universities that have similar Keycloak and LMS environments.
2. Guidance on modifying the OpenCanary intelligence used in the evaluation.
3. Role-specific training materials.

The evaluation maintains confidentiality and ethical considerations during the entire process. All user data collected through the evaluation is to be processed with the users' consent. When a user is flagged for suspicious communication, there will be a transparent record of when and why the user was flagged. If a user wishes to opt-out of training, they may do so. Additionally, users may submit a request to have their previous flagging re-evaluated.

4. Results and discussion

4.1. Deployment context and scale

HC-SES was deployed as a proof-of-concept at the Dzemal Bijedic University in Mostar, an institution serving more than 5.000 users from teaching and administrative staff and students. The system integrated Keycloak (identity management), Rspamd/Exchange (message analysis), Pi-Hole (DNS-based malicious domain filtering), OpenCanary (system vulnerability detection) and LMS (adaptive learning through eUNMO). During the Proof-of-Concept period, the HC-SES processed institutional email traffic. From flagged messages, 1,000 were manually verified as malicious via domain analysis, header inspection, and payload examination, providing baseline data for precision and recall.

4.2. Technical detection performance

Precision and recall: The goal for the HC-SES in the Proof-of-Concept (PoC) phase is to have over 90% precision (true positive rate among flagged messages) and 80% recall (percentage of malicious messages detected). The Baseline Comparison was based on the institution's existing method for mail filtering; DMARC Authentication coupled with Content Spam Filtering that produced an approximate precision level of 75%. A 15-percentage point precision increase is operationally significant. In institutional settings that receive 1.000+ phishing emails monthly, the baseline method delivers ~250 of those to inboxes, whereas the HC-SES delivers ~100, approximately half. This trade-off between precision and recall can be seen through the design decision made by the HC-SES. To decrease the amount of false positive emails (those incorrectly identified as being spam), the HC-SES has chosen to prioritize precision at the expense of recall. This trade-off is also reflected in the 80% recall rate, which is shaped by the inclusion of linguistic patterns and other technical indicators in the detection method. However, using these patterns together with technical indicators also makes it harder to distinguish sophisticated phishing emails, which mimic legitimate organizational content and correct grammatical structure, from authentic messages. This trade-off is acceptable in institutional environments where users act as the last line of defense, but a small percentage of missed phishing messages is less problematic than high false positive rates that cause users to ignore the tagging.

Risk-Adapted Authentication: The HC-SES plans to integrate Keycloak to trigger step-by-step authentication (MFA, re-challenge or limited training) based on risk context. Of the 5.000 login attempts analyzed during the Proof-of-Concept (PoC), less than 5% (~250 events) triggered incremental verification, with a false positive rate (real users incorrectly flagged for a challenge) below 1%. This low false-positive rate indicates that risk scoring distinguishes attempted anomalies (e.g., logging in from an unexpected geographic location or device) from routine attempts. Enhanced authentication strikes a balance between security and user experience, where users only verify when risk models require additional verification, rather than asking for it every time they log in.

4.3. Human behavioral outcomes

Training Participation and Completion: More than 60 percent of the pilot participants attended and completed the relevant micro-training modules, while only 40-50 percent of the participants who were required to attend the traditional mandatory awareness training participated and completed the training.

Behavioral Change Persistence: Scenario-based micro-training is far more effective at sustaining memory and change in behavior than abstract, out-of-context training. Pilot participants stated that they were more compliant with phishing related language cues, urgency cues ("act now") and tone inconsistencies (formal tone with colloquialism) as well as grammatical errors (Bosnian), demonstrating that the use-case training enhanced the ability to recognize patterns at the time of decision making.

Social Engineering and Behavioral Persistence: Stoica (2021) defined social engineering as a "perception game" in which attackers manipulate users' trust, authority and sense of urgency to manipulate their judgments (Stoica, 2021). The pilot results support Stoica's definition by providing evidence for his framework. If social engineering uses perceptions, it logically follows that modifying the user's perception at the moment of manipulation will reduce compliance. This is exactly what the HC-SES accomplishes; when a user encounters an email that has been flagged, the system immediately displays a explanation ("The sender is fake and there are urgency language indicators"), which interrupts the attacker's cognitive manipulation before the user can respond. Adaptive micro-training at the appropriate time generates behavioral changes where generic and infrequent awareness campaigns do not.

4.4. Explainability, trust, and user acceptance

The initial skepticism towards using behavioral tracking was based on two main reasons, users felt they would be under surveillance, and they did not want to experience false positives. Transparent explanations (e.g., 'Your message was flagged due to unusual sender domain and urgency language') increased users' trust and perceived credibility of the HC-SES decisions. Seventy percent of users rated HC-SES decisions as 'helpful' or 'mostly helpful' in post-deployment surveys.

The above findings demonstrate that a technical solution to automate security will fail without sufficient psychological acceptability. To maintain security boundaries, while also increasing user satisfaction, the introduction of human review capabilities, (users can request the HC-SES to release a flagged message with documented justification), has proven to be effective in reducing user frustration.

Users that utilized the review capabilities have indicated a higher level of satisfaction, indicating that users view their perceived agency (the ability to challenge or understand automated decisions), accountability and understanding of automated decisions as being equally as important as accuracy. For the sake of transparency and the right to know how an automated decision was made, the GDPR provides organizations with requirements for providing explanations on automated decisions that create legally binding consequences or affecting individuals in a significant way (European Commission, 2020).

The transparent scoring of the HC-SES (showing the various linguistic, metadata and behavioral factors that contribute to the risk score), meets this operational requirement. However, there is a conflict present with respect to providing an explanation, as the need to provide a complete explanation may be overwhelming for non-technical users. To address this conflict, the HC-SES utilizes layered explanations, including a quick summary ("Your Message was flagged as a fake sender") intended for the user's immediate actions, as well as detailed reasoning for the scoring, if requested by the user. This method addresses both compliance with the law and the constraint of cognitive load.

4.5. Study limitations

Technical Feasibility and Methodological Constraints: This Proof-of-Concept study has demonstrated the technical feasibility of incorporating the HC-SES component (Keycloak, Rspamd, OpenCanary, Pi-Hole, eUNMO LMS) into an existing University IT infrastructure. However, this study was not designed to evaluate the effectiveness of the HC-SES in reducing susceptibility to phishing threats. This study is missing three key elements that are needed to show cause and effect. It does not include a control group that receives standard security measures, random assignments to experimental and control groups, or a pre-registered protocol with clear progression criteria. Because of this, the observed results (90% detection precision and 60% training participation) cannot be confidently attributed to the HC-SES. They may instead be due to selection bias, changes in the environment over time, or institutional factors that are not related to the HC-SES intervention.

This study's single institution setting at UNMO limits its external validity and applicability across multiple institutions with varying user populations, existing security infrastructures, language environments and organizational cultures. Therefore, multi-institution studies with varied institution settings will be necessary prior to being able to provide implementation recommendations based on this study.

Sample Size and Statistical Power: The study analyzed approximately 1000 manually validated malicious emails. However, the study did not perform any formal sample-size calculations or power analysis. The sample-size is too small to determine whether there are statistically significant and operationally relevant differences in the ability of users to identify phishing threats. The reported metrics of precision (90%), MFA trigger-rate (<5%) and false positive rate (<1%) do not include uncertainty estimates (confidence intervals); therefore, it is impossible to determine how reliable and precise the findings are. Pilot studies should report all study outcomes with confidence intervals to allow for the calculation of sample sizes for definitive studies regardless of whether hypothesis testing is appropriate.

Limitations in Behavioral Metrics: No baseline behavioral metrics were obtained after the HC-SES components were deployed and activated. Thus, the study was unable to compare the susceptibility to phishing of users before and after deployment of the HC-SES components. Standard security-awareness research measures the simulated phishing click-rates, threat-reporting-rates, detection-dwell-time and repeat-offender-rates over 6-12-month intervals. Although the current study reports the training-participation rates (60%) and the training-completion rates (65% within 48 hours), there is no evidence that these participation rates resulted in sustained behavior change or lower click-rates during subsequent simulated-phishing-threats. Self-reported participation data could also be biased due to social desirability (e.g. self-reported participation rates may be inflated due to participants wanting to appear cooperative).

Technical and Contextual Constraints: Detection models have largely been developed and trained using English language phishing datasets, which results in lower performance levels when evaluated on Bosnian language email communications (Sahingoz et al., 2019). Due to these linguistic and cultural biases, the application of this system will be limited in multilingual environments until additional and significant efforts have been put forth in collecting region-specific training data.

Algorithms used to score risks based on inferred vulnerabilities from communication patterns may negatively impact users who exhibit typical, yet legitimate non-standard behavior:

e.g., users working internationally, users working shift hours with irregular login times, etc. The authors of the current study did not perform algorithmic fairness audits to assess if disparate demographic groups receive systematically different risk scores. Before deploying an operational version of the system, conducting such audits along with the inclusion of human-in-the-loop review processes for high-stakes decisions will be critical.

Implications for Interpretation: The results provided herein should be considered as indicative of the technical feasibility of implementing the HC-SES in a university environment and not as a measure of relative effectiveness. As previously mentioned, the reported detection accuracy (precision = 0.9; recall = 0.8) was not benchmarked against optimized configurations of competing alternatives (e.g. Rspamd with neural networks and fuzzy hashing; commercial solutions, i.e., Proofpoint, KnowBe4). Thus, it remains unknown whether the HC-SES represents a meaningful improvement over established methodologies in terms of detecting phishing threats.

Although the current study demonstrated the technical feasibility of integrating the HC-SES into a university's infrastructure, users could access training modules and the system did not result in any major disruptions to the institution's underlying infrastructure, these are all valuable findings for institutions that are contemplating utilizing similar open-source security frameworks. Nonetheless, causal claims regarding how the HC-SES decreases phishing risk in comparison to competing methodologies cannot be made without providing evidence from a randomized controlled trial.

Requirements for Future Research: A definitive assessment of the HC-SES will require: (1) a randomized controlled trial design with a treatment group (implementation of the entire HC-SES framework) and a control group (the standard institutional security practices utilized by the organization); (2) sufficient statistical power (a minimum of 500 users per arm to detect a 20% effect size with 80% power); (3) a registered protocol prior to the initiation of the study with clear progression criteria and standardized measurement of behavioral outcomes; (4) longitudinal follow-up (a minimum of six to twelve months) to assess persistence of the changes in behavior; (5) the HC-SES framework is implemented and tested in multiple sites with differing institutional contexts and (6) comparative benchmarking with commercial solutions and optimized configurations of open-source alternatives using matched email datasets.

Only through controlled evaluation with sufficient statistical power can the cyber-security community establish if the use of the HC-SES will provide substantive reductions in risk compared to established security awareness practices.

5. Keycloak integration and risk-adaptive authentication

Keycloak serves as the HC-SES's central identity management component, providing single sign-on (SSO) across institutional services (eUNMO academic records, Google Workspace for Education, Microsoft 365, institutional CMS, and eduroam wireless network access). The implementation consolidates previously fragmented authentication systems into a unified identity layer that supports behavioral risk assessment and adaptive security controls (Kovacic, Sehidic & Obradovic, 2017).

5.1. Risk-adaptive multi-factor authentication

The integration between Keycloak and the HC-SES enables context-aware authentication that adjusts security requirements based on real-time risk assessment. Baseline authentication uses Time-Based One-Time Passwords (TOTPs) with a 85% adoption rate among faculty/staff and 60% among students. When the HC-SES risk scoring detects elevated threat context, such as recent phishing targeting, unusual login location, device anomalies or impossible travel patterns Keycloak automatically triggers step-up authentication requiring additional verification before access is granted.

This bidirectional feedback loop addresses a critical gap in traditional email security: if the HC-SES detects a high-confidence phishing message targeting a specific user, the system elevates that user's risk score for 48 hours, subjecting all subsequent authentication attempts to enhanced verification regardless of device or location familiarity. Conversely, Keycloak authentication events (successful logins, failed attempts, MFA challenges) feed back into HC-SES behavioral analytics, enabling the system to detect credential compromise patterns and brute-force attacks (see Figure 1 and Figure 2).

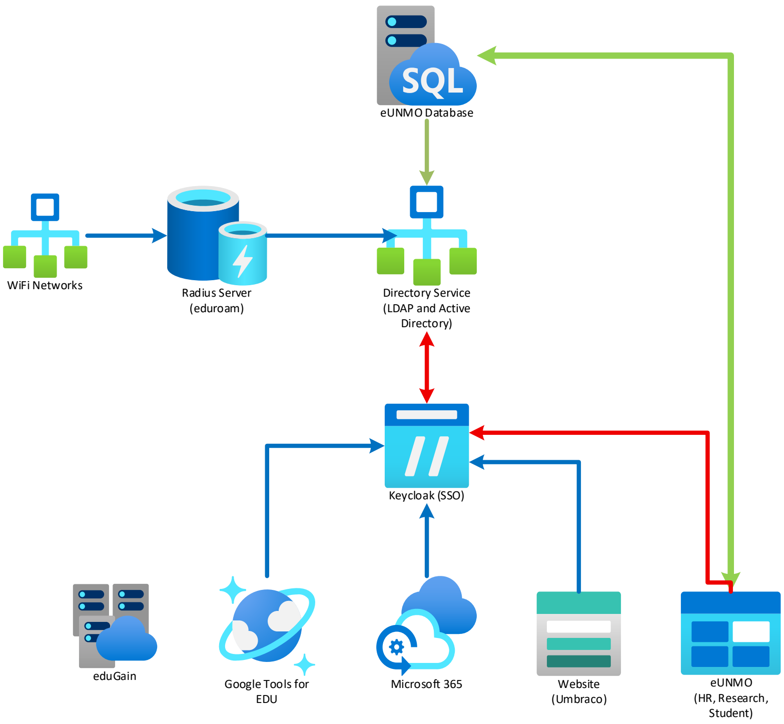


Figure 1. Authentication and identity propagation at UNMO. Keycloak SSO integrates directory services with federated platforms. Solid arrows show authentication flows; dashed lines indicate external federation (Source: Author's own research)

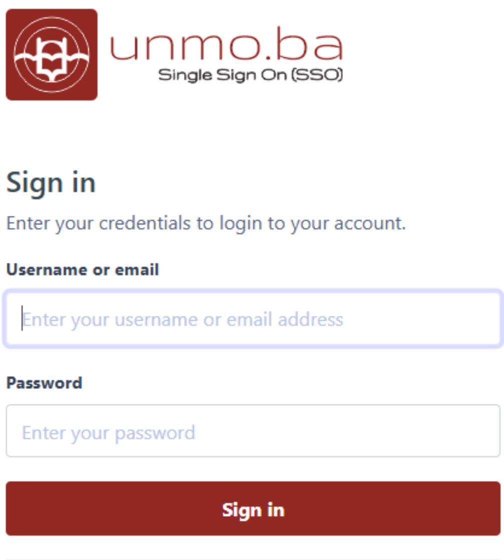
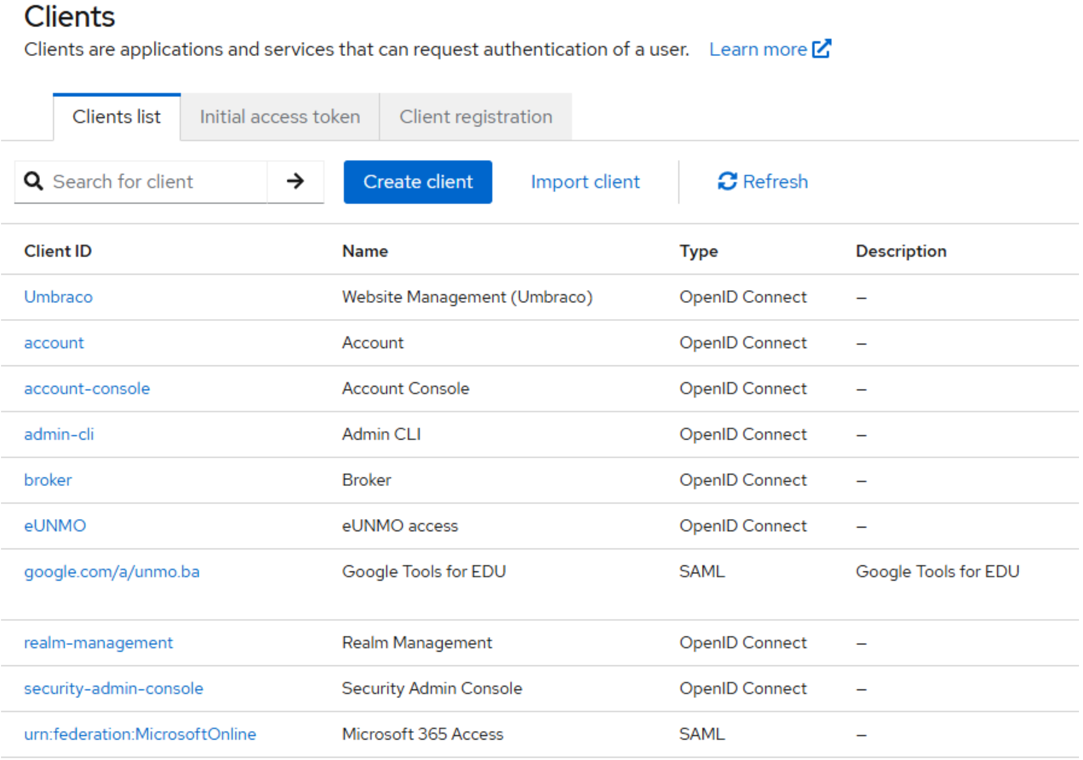


Figure 2. Login for SSO at UNMO
(Source: Author's screenshot from unmo.ba Single Signe On (SSO))

Keycloak is shown to have a broad range of services in Figure 3, including educational activities (course registration, grading, eUNMO enrollment), productive/collaborative tools (Google Workspace for students, Microsoft 365 for staff), web content management (Umbraco) and networked WiFi (eduroam) at over 100 European universities/institutions. In this manner, placing all these services into a common identity space will enable users to transition seamlessly between them and provide administrators with an identical view of user activity across platforms. Additionally, it enables future integration(s) (on-premise email, eduGAIN federation, regional inter-university authentication etc.) to occur without having to deploy additional identity systems.



Client ID	Name	Type	Description
Umbraco	Website Management (Umbraco)	OpenID Connect	–
account	Account	OpenID Connect	–
account-console	Account Console	OpenID Connect	–
admin-cli	Admin CLI	OpenID Connect	–
broker	Broker	OpenID Connect	–
eUNMO	eUNMO access	OpenID Connect	–
google.com/a/unmo.ba	Google Tools for EDU	SAML	Google Tools for EDU
realm-management	Realm Management	OpenID Connect	–
security-admin-console	Security Admin Console	OpenID Connect	–
urn:federation:MicrosoftOnline	Microsoft 365 Access	SAML	–

Figure 3. UNMO service ecosystem integrated with Keycloak SSO, spanning academic (eUNMO), productivity (Microsoft 365, Google Workspace), web (Umbraco) and network (eduroam) services. (Source: Author's screenshot from unmo.ba Single Signe On (SSO))

5.2. Deployment results and future federation

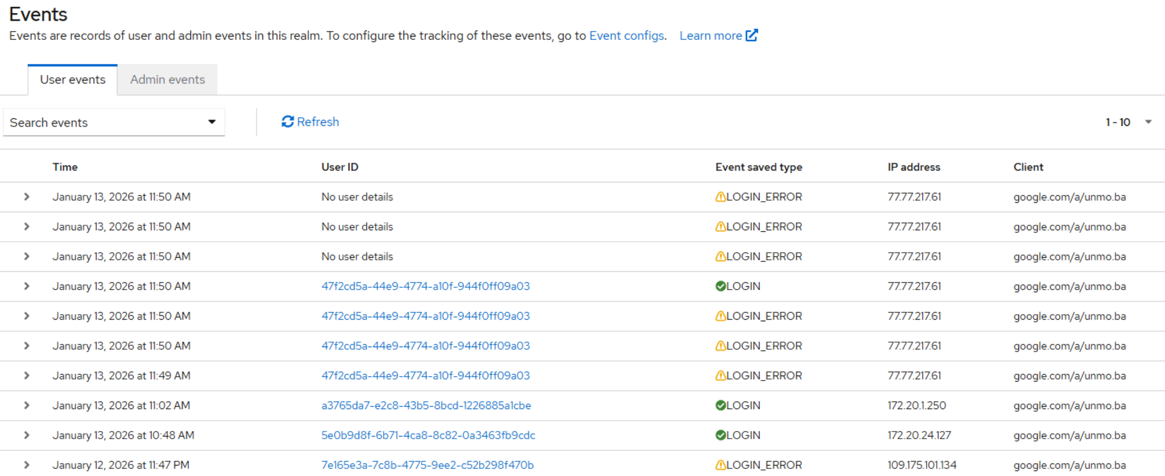
Keycloak deployment demonstrated technical feasibility: out of 5.000 logins during the PoC, <5% triggered risk-based MFA with <1% false positives (Section 4.2). This risk-based approach limits security friction by reducing unnecessary MFA challenges.

In addition to integrating the on-premises email system directly into Keycloak for comprehensive email-access behavioral monitoring, future work will also include establishing an eduGAIN federation to allow UNMO credentials to be used for library access and repository access in European libraries and implementing regional inter-university authentication to facilitate collaboration among institutions within regions. These future works will place UNMO within the broader European digital identity infrastructure while allowing each institution to maintain control over their own security policies and user data governance practices.

5.3. HC-SES integration and real-time risk propagation

The feedback loop formed by HC-SES and Keycloak is a loop that associates phishing detection with the strength of an individual's authentications. If the HC-SES detects a high-confidence phishing message sent to a specific user, it will increase the user's risk factor for 48 hours. During this time frame, all subsequent login requests made from this user's account will be subjected to enhanced authentication, regardless of whether the device and/or location associated with the request appears normal/routine. This is an important step toward addressing a well-known deficiency in most traditional email filtering systems: they can often block/flag emails, but cannot affect authentication flows once a username/password has been compromised.

Real-time authentication data flows in both directions as well. Keycloak captures both successful and failed login requests as well as enhanced authentication requests. The HC-SES captures and uses these signals in its behavior-based risk assessment model. The real-time authentication view dashboard (see Figure 4) provides security staff with a general picture of the institutional access pattern(s): low risk sessions from known devices and common locations, unusual login requests from unknown countries/networks and repeated failure to authenticate, which could indicate brute-force attacks or credential compromises. Low-risk login attempts are generated from known devices during regular working hours and proceed uninterrupted. Login attempts from unknown devices in foreign countries result in immediate guidance and alerting security staff; extended periods of multiple failed authentication requests from the same IP address range can prompt either a temporary lockout or further investigation.



Time	User ID	Event saved type	IP address	Client
January 13, 2026 at 11:50 AM	No user details	LOGIN_ERROR	77.77.217.61	google.com/a/unmo.ba
January 13, 2026 at 11:50 AM	No user details	LOGIN_ERROR	77.77.217.61	google.com/a/unmo.ba
January 13, 2026 at 11:50 AM	No user details	LOGIN_ERROR	77.77.217.61	google.com/a/unmo.ba
January 13, 2026 at 11:50 AM	47f2cd5a-44e9-4774-a10f-944f0ff09a03	LOGIN	77.77.217.61	google.com/a/unmo.ba
January 13, 2026 at 11:50 AM	47f2cd5a-44e9-4774-a10f-944f0ff09a03	LOGIN_ERROR	77.77.217.61	google.com/a/unmo.ba
January 13, 2026 at 11:50 AM	47f2cd5a-44e9-4774-a10f-944f0ff09a03	LOGIN_ERROR	77.77.217.61	google.com/a/unmo.ba
January 13, 2026 at 11:49 AM	47f2cd5a-44e9-4774-a10f-944f0ff09a03	LOGIN_ERROR	77.77.217.61	google.com/a/unmo.ba
January 13, 2026 at 11:02 AM	a3765da7-e2c8-43b5-8bcd-1226885a1cbe	LOGIN	172.20.1.250	google.com/a/unmo.ba
January 13, 2026 at 10:48 AM	5e0b9d8f-6b71-4ca8-8c82-0a3463fb9cdc	LOGIN	172.20.24.127	google.com/a/unmo.ba
January 12, 2026 at 11:47 PM	7e165e3a-7c8b-4775-9ee2-c52b298f470b	LOGIN_ERROR	109.175.101.134	google.com/a/unmo.ba

Figure 4. Real-time authentication dashboard. Green: low-risk sessions from known devices; yellow: elevated-risk logins triggering MFA; red: failed attempts indicating potential brute-force (Source: Author's screenshot from unmo.ba Single Signe On (SSO))

5.4. Future federation plans

UNMO's road map is based on this integrated identity layer. The next steps are planned to bring all on-premises email completely into Keycloak: enable HC-SES to evaluate users' email-access behavior, as well as message-content, for behavioral risk assessment; provide integration with eduGAIN to permit students and staff of UNMO to access libraries and repositories across Europe by using their UNMO accounts and inter-university authentication in BH to make student mobility and cross-institutional collaboration easier. The above extensions place UNMO as a participating entity in the developing European digital identity fabric while at the same time maintaining local control over security policy and user data.

6. Conclusion

Human-centered Social Engineering is a growing concern for higher education. In this paper we introduce the HC-SES, a modular open-source system with five modules: Identity Management, Honeypots, DNS Filtering, Email Analysis, and Adaptive Training all built within a Human-Centric Defense Architecture.

The Proof-of-Concept deployment at Dzemal Bijedic University of Mostar demonstrated technical viability: in modules integrated without operational disruption, detection reached 90% precision on a limited sample and training participation (60%) exceeded institutional averages (40-50%).

While this is a proof of concept to show technical feasibility and not an evaluation of effectiveness (as discussed in detail in Section 4.5 Study Limitations), causal statements regarding how the HC-SES could reduce phishing risk are also not supported by this study because there was no randomized control group. The data collection process was not pre-registered and therefore there was insufficient statistical power to form causal statements based on the collected data. Randomized controlled trials at multiple sites using this technology will need to be conducted prior to developing recommendations for implementation outside of UNMO. The modular design of the HC-SES allows universities to deploy each component individually (Keycloak + Rspamd for foundational security; OpenCanary threat intelligence and adaptive training when funding permits).

Future research will need to perform randomized controlled trials that measure common behavioral measures (e.g., click-rates, reporting-rates, dwell-time), compare results with those of commercial solutions and determine whether the behaviors exhibited by users will persist over time. Only through rigorous controlled evaluations will the cybersecurity community be able to determine the true value added by the HC-SES in protecting the human layer.

Author contributions

Conceptualization: S.K. and I.B.; Data Curation: S.K. and I.B.; Formal Analysis and Methodology: S.K. and I.B.; Supervision: I.B.; Validation: S.K.; Visualization: S.K.; Writing-original draft: S.K.; Writing-review and editing: All authors under the lead of S.K. All authors have read and agreed to the published version of the manuscript.

Submission: 16 January 2026; Revised: 02 February 2026; Acceptance: 10 February 2026; Published: 31 March 2026.

REFERENCES

Anghel, M. & Pereteanu, G.-C. (2020) Cyber Security Approaches in e-Learning. In *INTED2020: Proceedings of the 14th International Technology, Education and Development Conference, 2-4 March 2020, Valencia, Spain.* pp. 4820-4825. https://doi.org/10.21125/inted.2020.1323.

Anghel, M., Pereteanu, G.-C. & Cirnu, C.-E. (2020) Emerging Trends in Elearning and Mlearning from a Byod Perspective and Cyber Security Policies. *eLSE Proceedings*. pp. 444-452. https://doi.org/10.12753/2066-026x-20-058.

Bonneau, J. (2012) The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. *IEEE Security & Privacy*. 10(1), 14–23. https://doi.org/10.1109/SP.2012.44.

Czarnul, P., Antal, M., Baniata, H. et al. (2025) Optimization of resource-aware parallel and distributed computing: a review. *The Journal of Supercomputing*. 81, art. no. 848. https://doi.org/10.1007/s11227-025-07295-7.

Doshi-Velez, F. & Kim, B. (2017) Towards a rigorous science of interpretable machine learning. *arXiv* [preprint] arXiv:1702.08608. https://doi.org/10.48550/arXiv.1702.08608.

European Commission. (2020) *EU Cybersecurity Strategy and NIS 2 Directive*. https://digital-strategy.ec.europa.eu/en/policies/cybersecurity-strategy [Accessed: 10th January 2026]

Federal Bureau of Investigation (FBI). (2024) 2023 *Internet Crime Report. Internet Crime Complaint Center (IC3).* https://www.ic3.gov/annualreport/reports/2023_ic3report.pdf [Accessed: 12th January 2026]

Gaw, S., Felten, E. W. & Fernandez-Kelly, P. (2006) Secrecy, flagging, and paranoia: Adoption criteria in encrypted email. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. pp. 591–600. https://doi.org/10.1145/1124772.1124862.

Heiding, F., Schneier, B. & Vishwanath, A. (2024) AI Will Increase the Quantity and Quality of Phishing Scams. *Harvard Business Review*. https://hbr.org/2024/05/ai-will-increase-the-quantity-and-quality-of-phishing-scams [Accessed: 15th January 2026].

Kaur, H. et al. (2020) Interpreting interpretability: Understanding how users make sense of AI explanations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. pp. 1–14.

KnowBe4. (2025) 2025 *Phishing by Industry Benchmark Report* https://www.knowbe4.com/resources/reports/phishing-by-industry-benchmarking-report [Accessed: 15th January 2026].

Kovacic, S., Sehidic, A. & Obradovic, E, (2017) Improving the Security of Access to Network Resources Using the 802.1x Standard in Wired and Wireless Environments. *Conference: 22nd Internacionalna Naučno-Stručna Konferencija Informacione Tehnologije*, 2017.

Lewis, C., Kristensen, I. & Caso, J. (2025) *AI is the greatest threat and defense in cybersecurity today. Here's why. McKinsey & Company, 15 May*. https://www.mckinsey.com/about-us/new-at-mckinsey-blog/ai-is-the-greatest-threat-and-defense-in-cybersecurity-today [Accessed: 15th January 2026].

Sahingoz, O. K. et al. (2019) Machine learning based phishing detection from URLs. *Expert Systems with Applications*. 117, 345-357. https://doi.org/10.1016/j.eswa.2018.09.029.

Sift. (2025) Q2 2025 *Digital Trust Index: AI Fraud Data and Insights*. https://sift.com/index-reports-ai-fraud-q2-2025/ [Accessed: 15th January 2026].

Stoica, A. (2021) Social engineering as the new deception game. *Romanian Journal of Information Technology and Automatic Control [Revista Română de informatică și automatică]*. 31(3), 57-68. https://doi.org/ 10.33436/v31i3y202105.

Wash, R. (2010) Folk models of home computer security. *Proceedings of the 6th Symposium on Usable Privacy and Security (SOUPS)*. 1–16. https://doi.org/10.1145/1837110.1837125.

Salko KOVAČIĆ is Head of Service for Information, Communication and Development Technologies at Dzemal Bijedic University of Mostar. The main areas of his activity include cybersecurity, AI-enabled systems, digital transformation in the higher education sector, and Open Science Infrastructure (OSI) in Bosnia and Herzegovina (B&H) and the Western Balkans (WB) region. He has served as a leader or member in several international and EU-funded projects in these areas.

Ivana BILIĆ is a professor at the Department of Management, Faculty of Economics, University of Split. She holds a Ph.D. in Management. Her research interests include corporate communication, crisis management, and entrepreneurship. She is the author of two book chapters and over 40 peer-reviewed journal articles and international conferences proceedings. Her professional experience includes quality assurance evaluation for EU-accredited agencies, mentorship in youth entrepreneurship and board membership in academic networks.