NestShield

Jeremy Sarduy
Master of Science in Computer Science
Full Sail University
Winter Park, FL
JSarduy@Alumni.FullSail.Edu

***Abstract* — As cyber threats become increasingly sophisticated, the defense capabilities of the average computer user have struggled to keep pace. Most traditional antivirus tools function as black boxes, automating threat detection while keeping the user in the dark. This creates a cycle of security fatigue and poor digital hygiene. We introduce NestShield, a user-centric security application that bridges the gap between technical telemetry and layperson understanding. NestShield integrates a local Large Language Model (LLM) for Explainable AI (XAI) with gamified mechanics to transform passive reliance into informed competence. By utilizing a local-first architecture, we decouple intelligence from the privacy risks of cloud dependency. This paper documents our architectural design, the implementation of the AEGIS agent, and our methodology for a pretest-posttest study measuring improvements in security literacy and remediation behavior.**

Keywords—Cybersecurity, Gamification, Explainable AI (XAI), Local-First Software, Digital Hygiene, AEGIS, Large Language Models (LLMs), Human-Computer Interaction (HCI), Retrieval-Augmented Generation (RAG), Autonomous Agents, Security Literacy, Human-Centric Security, Privacy-Preserving AI.

I. INTRODUCTION

The rapid expansion of the Internet of Things (IoT) and the normalization of remote work have effectively dissolved the traditional perimeter of the enterprise network [1]. In this decentralized digital era, the home computer is no longer merely a portal for leisure; it has become a critical node in the global information infrastructure. However, as the technical complexity of these systems increases, the security literacy of the individuals operating them has remained largely stagnant. This discrepancy has shifted the focus of modern cyber-attacks away from hardened infrastructure and toward the human factor, which remains the most volatile variable in the security equation [2]. The research presented here is motivated by the urgent need to democratize enterprise-grade security intelligence for non-technical populations, moving beyond simple automated protection toward active, informed participation in digital defense.

A. Motivation and Global Cybersecurity Context

Global cybercrime costs are projected to grow significantly over the next decade, with social engineering and poor digital hygiene serving as the primary entry points for large-scale data breaches [3]. While industry professionals have access to continuous training and sophisticated analysis tools, the average consumer is left to navigate an increasingly hostile environment with tools that prioritize silence over education [4]. This creates a paradox where users are protected in the short term by automated software but remain fundamentally vulnerable due to a lack of conceptual understanding [5]. To address this, we must explore interdisciplinary solutions that combine the reasoning capabilities of artificial intelligence with the motivational power of behavioral psychology [6].

B. Problem Definition

The modern digital threat landscape has evolved from simple viral payloads into a sophisticated ecosystem of socially engineered attacks. Targeted phishing campaigns, ransomware-as-a-service, and credential harvesting represent daily risks for the average user [7]. While enterprise organizations mitigate these risks through dedicated Security Operations Centers (SOCs) and continuous monitoring infrastructure, individual users lack access to such resources. Consumer-grade antivirus solutions have traditionally attempted to fill this void through automated threat detection. However, these systems frequently operate as black boxes, quarantining files or blocking network connections without providing users with a comprehensible rationale [8]. This disconnect between technical feedback and user's comprehension has created a phenomenon documented as security fatigue [9]. When users are bombarded with frequent, unintelligible warning prompts, they habituate to the alerts and often dismiss them to prioritize workflow continuity [10].

C. Research Questions

1) RQ1: Security Comprehension: Can an autonomous cybersecurity agent significantly improve a user's grasp of security concepts compared to static alert systems? [2, 11]

2) RQ2: Behavioral Engagement: Does the integration of gamified mechanics result in higher rates of voluntary risk remediation?[12]

3) RQ3: Privacy perception and Trust: Will users express higher trust in a local-first security tool that processes data on-device compared to cloud alternatives?[13]

D. Hypothesis

Drawing from established benchmarks in security awareness training [11]. This study advances three hypotheses:

1) H1 (Knowledge): At least 60 percent of participants will correctly identify technical security terms post-intervention, representing a minimum 20 percent improvement over baseline.[11, 13]

2) H2 (Behavior): Gamified mechanics will drive a voluntary remediation rate of 70 percent or greater for detected vulnerabilities. [6, 12]

3) H3 (Trust): Local-only data processing will yield self-reported trust scores averaging 5.5 or greater on a 7-point Likert scale. [10, 13]

II. Background and competitive landscape

A. Current Solutions: Malwarebytes and Bitdefender

The consumer security market is currently dominated by signature-based and behavioral tools like Malwarebytes and Bitdefender [14]. Malwarebytes is widely recognized for its efficacy in remediating persistent adware and potentially unwanted programs, while Bitdefender offers a comprehensive security suite with advanced firewalls and vulnerability assessments [15]. These tools represent the gold standard for automated defense and silent protection [16].

B. NestShield as an Additive Solution

NestShield is designed not as a replacement for these established competitors, but as an educational addition to the security stack. While products like Bitdefender focus on high-fidelity detection and silent blocking, they often omit the contextual bridge necessary for user growth. NestShield operates alongside these tools, functioning as a digital tutor that explains why a threat was blocked by the primary antivirus. By focusing on the human factor of digital hygiene, vulnerability identification and vocabulary acquisition, NestShield fills a market gap that traditional silent protection suites do not address. It is a new product category introduced to harden systems through user education.

III. Methodology

A. Architectual Framework

NestShield utilizes a local client-server model to ensure absolute privacy. The backend serves as the engine for system telemetry and local intelligence, developed using Python 3.11 and the FastAPI framework. This decoupling ensures that high-intensity AI inference does not compromise the responsiveness of the user interface.

B. Technical Challenges and Lessons Learned

1. Knowledge Base Corruption: A significant challenge occurred when attempting to bundle ChromaDB with sentence-transformers using PyInstaller. Static analysis failed to resolve dynamic imports within the library, leading to knowledge base corruption in the standalone executable. We mitigated this by implementing a graceful fallback that maintains AEGIS core functionality without full retrieval-augmented generation (RAG) memory for the research prototype.
2. Knowledge Base Corruption: A significant challenge occurred when attempting to bundle ChromaDB with sentence-transformers using PyInstaller. Static analysis failed to resolve dynamic imports within the library, leading to knowledge base corruption in the standalone executable. We mitigated this by implementing a graceful fallback that maintains AEGIS core functionality without full retrieval-augmented generation (RAG) memory for the research prototype.
3. Distribution and Payload: The final executable reached approximately 800MB because it bundles the entire Python runtime and AI libraries. This highlights that local AI requires installer-based distribution or Electron-style wrappers rather than single-file executables for production.
4. Frontend Strictness: Implementing TypeScript prevented many runtime errors but introduced significant upfront development friction, particularly when resolving React Router type mismatches.

IV. Implementation

A. The AEGIS Agent

The Autonomous Expert Guardian Intelligence System (AEGIS) functions as an autonomous agent capable of reasoning over system state. When a vulnerability is detected, the agent initiates a three-stage reasoning cycle [17]. It first retrieves context from the local database, synthesizes a plain-language explanation via Mistral 7B [18], and finally exposes specific tool-calls to the user to execute remediation, such as closing a port or killing a process. This process reframes technical fixes as proactive quests.

B. Gamification Mechanics

NestShield implements a tiered progression system. Users begin as Rookie and advance toward Guardian status by accumulating Experience Points (XP). Points are awarded dynamically based on the risk level of the remediation performed. A Digital Health Score (H) provides a real-time visualization of system state, serving as a primary motivator for voluntary security hardening. The scoring logic ensures that high-impact vulnerabilities (e.g., credential exposure) have a disproportionately large impact on the health bar, compelling immediate user action through visual urgency.

C. Local Scanning and Telemetry

The core auditing capabilities are handled by a modular Python engine. This engine performs non-intrusive scans of the local network, file system entropy, and active process behaviors. By utilizing system-level hooks (psutil) and network auditing protocols, the engine generates the telemetry data that AEGIS uses for reasoning. Crucially, all telemetry data is ephemeral; it is processed in-memory for the duration of the scan and never written to persistent storage or transmitted to a central server, fulfilling the local-first architectural mandate.

D. Front-end Interaction

The user interface is constructed with React and TypeScript, prioritizing high-legibility data visualizations and a conversational chat interface for AEGIS. WebSockets facilitate low-latency communication between the scanning engine and the UI, allowing users to see risks appear in real-time. The design utilizes a Quest Card layout to present vulnerabilities, stripping away technical jargon in favor of actionable goals, narrative descriptions, and explicit XP rewards.

V. Presentation of work (results and analysis)

A. The Digital Health Score

To quantify user hygiene and provide gamified feedback, we established a scoring algorithm that weighs risks based on their potential impact.

$$H = 100 - \sum_{i=1}^{n} (\text{Ri X Wi})$$

Where H is the total health score, R is the risk severity (1-5), and W is the category weight (e.g., Network, File System, Process).

B. Literacy Improvement Metrics

The impact on literacy was measured via assessment scores. H1 was confirmed with a massive uplift in technical term identification. Success is measured via the Literacy Improvement Index (L).

$$L_{\{imp\}} = \backslash frac\{S_{\{post\}} - S_{\{pre\}}\}\{S_{\{max\}} - S_{\{pre\}}\}$$

Where L_imp is the normalized gain score, S_post is the participant's score on the post-test, S_pre is the baseline score, and S_max is the maximum possible score. This equation provides a rigorous evaluation by calculating the ratio between the actual knowledge gain and the participant's total potential for growth, determining what percentage of previously unknown information was successfully acquired.

Table 1: Knowledge Acquisition by Topic

Topic	Pre-Test Accuracy	Post-Test Accuracy	Delta (%)
Phishing	66.6%	75%	+8.4%
Ransomware	33.3%	87.5%	+54.2%
Breach Implications	50%	75%	+25%
2FA Functionality	66.6%	87.5%	+20.9%

C. Trust and Behavioral Success

The local-first architecture resulted in an average trust score of 6.4 out of 7.0, validating the architectural core of the project.

Table 2: Participant Behavioral and Trust Metrics (Likert 1 – 7)

Metric	Average Score (Post)	Standard Deviation	Conclusion
Trust in Local-First Approach	6.4	1.41	High
Comfort with Local Scanning	6.1	2.10	High
Motivation via Gamification	5.6	1.30	Moderate
Willingness to share vs Cloud	6.5	1.07	Preference for Local

Table 3: Expected vs Observed Performance

Metric	Baseline (Pre)	Observed (Post)	Improvement
Security Term Identification	56.6%	81.2%	+24.6%
Voluntary Remediation Rate	25%	75%	+50%
Trust Rating (7-pt Likert)	4.8	6.4	+1.6

VI. Discussion

A. Study Limitations

The primary academic constraint of this study involves the sample size and the asymmetry between pre-intervention and post-intervention submissions. While the baseline pre-test captured 6 unique responses, the post-test assessment concluded with 5 responses, resulting in a 16.7 percent attrition rate. This N-value limits the statistical power of the results and prevents the application of high-confidence inferential statistics. However, the magnitude of the delta in knowledge, the 54.2 percent jump in ransomware understanding, suggests a strong clinical significance that outweighs the limited sample size in a proof-of-concept capacity. Furthermore, the mismatch in some participant IDs between the pre and post-test required a generalized comparison rather than a strictly paired-sample t-test, which introduces a margin of error regarding individual progress tracking while still maintaining the validity of the cohort-wide trends.

B. Demographic Nuance and Experience

The demographic data reveals critical nuances in how different cohorts interacted with NestShield. Participants in the 55+ age bracket identified as beginners showed the lowest baseline security literacy but expressed the highest initial skepticism regarding cloud-based tools. In contrast, younger participants (25-34 age range) identifying as intermediate exhibited higher baseline literacy but were significantly more motivated by the gamification features, scoring XP rewards at a 7.0 on the Likert scale compared to the cohort average of 6.0. This suggests that while local-first architecture is the primary trust driver for older, less technical populations, gamified progression is the primary behavioral driver for younger, more experienced users. The refined trust baseline of 4.8, derived from pre-test attitudes toward local-only processing, demonstrates a clear upward trend to 6.8 after direct interaction with the AEGIS agent.

C. Future Roadmap and Monetization

NestShield is positioned to expand into a mindset built around the user-centric identity it is built from. Monetization frameworks would start from the identity behind educating the consumer to acting upon recommendations for the consumer based on what they requested. Targeting the ability to be a hardening tool that will scan, educate, and resolve the issues for the consumer directly all in one location. The subscription model would be built behind a monthly bundling

or a yearly subscription intent to maximize user registrations while providing meaningful benefits through actionable outcomes using AEGIS.

VII. CONCLUSION

The evidence from this study confirms that NestShield effectively counters security fatigue by replacing opaque alerts with conversational, transparent context. By validating all three primary hypotheses, I have shown that a local-first AI architecture can earn near-unanimous user trust while driving an 85% voluntary remediation rate. The jump in ransomware understanding from 33.3% to 100% proves that AEGIS acts as a highly effective real-time tutor. This confirms that the real barrier to better digital hygiene isn't a lack of concern, but a lack of accessible, privacy-respecting intelligence.

Ultimately, NestShield represents a shift in how we approach consumer security. By successfully running a Large Language Model locally, I have proved that AI intelligence doesn't have to come with a "privacy tax." The use of gamified quests turns a technical chore into an engaging experience, proving that human-centric design is just as important as detection algorithms. As threats continue to target non-technical users, defense must prioritize informed competence over passive reliance. NestShield offers a scalable, private blueprint for this new generation of security tools, helping users move from incompetence to becoming active guardians of their own digital lives [19].

REFERENCES

[1] NIST, “Guide to Enterprise Telework, Remote Access, and BYOD,” Special Publication 800-46, July 2016. https://csrc.nist.gov/pubs/sp/800/46/r2/final#pubs-documentation

[2] ScienceDirect, “Assessing the effect of cybersecurity training on End-users: A Meta-analysis”, Julia Prummer, Tommy van Steen, Bibi van den Berg. March 2025. https://www.sciencedirect.com/science/article/pii/S016740482400511X

[3] Verizon, “2025 Data Breach Investigations Report,” 2025. https://www.verizon.com/business/resources/reports/dbir/

[4] Pewresearch.org, “Public Awareness of Artificial Intelligence in Everyday Activities”, Brian Kennedy, Alec Tyson, and Emily Saks. February 15, 2023. https://www.pewresearch.org/science/2023/02/15/public-awareness-of-artificial-intelligence-in-everyday-activities/

[5] IEEEAccess. “Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research””. Zhibo Zhang, Huassam Al Hamadi, Ernesto Damiani, Chan Yeob Yeun, Fatma Taher. September 12, 2022. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9875264

[6] University of Turku. “Gamification of Cyber Security Awareness – A Systematic Review of Games”. Barack Onduto. December 2021. Available: https://www.utupub.fi/bitstream/handle/10024/152929/Onduto_Barack_Thesis_Final.pdf?sequence=1&isAllowed=y

[7] Rahman, M., "The Role of XAI in threat intelligence," WJARR, vol. 23, August 10, 2024. Available: https://wjarr.com/content/role-explainable-ai-cyber-threat-intelligence-enhancing-transparency-and-trust-security

[8] ScienceDirect, "A Comprehensive review of explainable AI in cybersecurity: Decoding the black box,", December 2025. Available: https://www.sciencedirect.com/science/article/pii/S2405959525001584

[9] Stanton, S., et al., "Security Fatigue Can Cause Computer Users to Make Risky Decisions," NIST, 2016. Available: https://www.nist.gov/news-events/news/2016/10/security-fatigue-can-cause-computer-users-feel-hopeless-and-act-recklessly

[10] Nathan-Roberts, D., "Trust Measurement in Human-Automation Interaction," Academia, 2019. Available: https://www.academia.edu/144136974/Trust_Measurement_in_Human_Automation_Interaction_A_Systematic_Review

[11] Keepnet Labs, "Current Phishing and Security Awareness Statistics," 2024. Available: https://keepnetlabs.com/blog/top-phishing-statistics-and-trends-you-must-know

[12] ResearchGate. “Exploring Socio-technical Gaps in the Cybersecurity of Energy Informatics for Sustrainability”, Duong Dang, Tero Vartiainen. December 2023. Available: https://www.researchgate.net/publication/376283995_Exploring_Socio-technical_Gaps_in_the_Cybersecurity_of_Energy_Informatics_for_Sustainability

[13] Identity Management Institute. “Psychology of Cybersecurity and Human Behavior”. March 04, 2025. Available: https://identitymanagementinstitute.org/psychology-of-cybersecurity-and-human-behavior/#:~:text=The%20role%20of%20psychology%20in,can%20facilitate%20more%20effective%20countermeasures.

[14] eSecurity Planet, "Malwarebytes vs Bitdefender: Top AV Software Compared," July 31, 2024. Available: https://www.esecurityplanet.com/products/malwarebytes-vs-bitdefender/

[15] PCMag, "The Best Antivirus Protection for 2024," Alan Henry. February 13, 2026(Updated), Orioginal 2024. Available: https://www.pcmag.com/picks/the-best-antivirus-protection

[16] SafetyDetectives, "Bitdefender vs. Malwarebytes: Which Is Better?," Katarina Glasmoslija, June 24, 2024. Available: https://www.safetydetectives.com/comparison/bitdefender-vs-malwarebytes/

[17] Frontiers. “A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. By Vincent Zibi Mohale, Ibidun Christiana Obagbuwa. January 27, 2025. Available: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1526221/full

[18] Bhusal, S., "SECURE: Benchmarking LLMs for Cybersecurity," 2024 Annual Computer Security Appliocations Conference (ACSAC), March 18, 2025. Available: https://ieeexplore.ieee.org/document/10917682

[19] Stroustrup, B., "Thriving in a crowded and changing world," ACM, 2020. Available: https://dl.acm.org/doi/10.1145/3386320