Operationalizing Human Factors in Cybersecurity: The MORPHEUS Approach to Measuring Phishing Susceptibility

Giuseppe Desolda[1], Francesco Greco[1] and Rosa Lanzilotti[1]

[1]Department of Computer Science, University of Bari, Via Orabona 4, 70125 Bari, Italy

Abstract

Human-centric cybersecurity posits that effective defense requires understanding the user not as a liability, but as a dynamic component of the security perimeter. However, the integration of human factors into adaptive defensive systems—a core goal of Symbiotic AI—is hindered by the lack of granular, validated metrics to assess individual vulnerability. Current approaches often rely on monolithic training or generic personality tests that fail to capture the vector-specific nature of susceptibility.

This paper introduces a methodological approach to operationalizing Morpheus [1], a multidimensional framework that maps 50 human factors to specific cyber threats. We present the design and generation process of the *Phishing Susceptibility Questionnaire* (PSQ), a psychometric instrument specifically designed to quantify vulnerability across the primary email attack vectors: Phishing and Spear Phishing. We detail our rigorous "Translation Protocol," which converts abstract psychological constructs (e.g., *Impulsivity*, *Authority Bias*) into observable behavioral policies. By employing mirrored item pairs to ensure data consistency, the PSQ is designed to function as a high-fidelity sensor for user modeling. We further outline a pre-registered validation protocol aimed at establishing the instrument's predictive validity against behavioral ground truth using Signal Detection Theory (d'). We conclude by discussing how this diagnostic framework serves as a foundational layer for next-generation Symbiotic AI systems, capable of moving from generic alerts to personalized, context-aware mitigation strategies.

Keywords

Human-Centric Cybersecurity, Phishing Susceptibility, Psychometrics, Symbiotic AI, Social Engineering, MORPHEUS Framework

CEUR-WS.org/Vol-4329/paper18.pdf

1. Introduction

The prevailing paradigm in human-centric cybersecurity treats the user as a monolithic entity—often described as the "weakest link"—whose behavior must be corrected, for example, through standardized training. This approach, however, has reached a plateau. Despite several defensive strategies, for example, Security Education, Training, and Awareness (SETA) programs [2], success rates for social engineering attacks remain alarmingly high [3]. The failure lies not in the number of defensive strategies (e.g., warning dialogs [4, 5, 6, 7], training [8], automatic detection [9, 10]), but in the lack of *context*: current defenses ignore the complex individual differences—cognitive biases, personality traits, and emotional states—that dictate why a specific user falls for a specific bait at a particular time [11].

This blind spot represents a fundamental barrier to the realization of **Symbiotic AI** in Cyber Social Security. True symbiosis implies a mutual understanding and coupling between the human and the machine agent [12, 13], ensuring the complete safety and security of individuals. While modern AI defenses are adept at analyzing the *technical* vector (e.g., inspecting headers, analyzing payloads), they remain oblivious to the *human* vector. For instance, the defensive AI does not know if the user is clicking a malicious link because they are fundamentally *impulsive*, temporarily *distracted*, or cognitively biased by a specific *authority* figure. Without this "Theory of Mind", the AI cannot intervene effectively; it can only block or alert generically, often causing friction or habituation [14].

1st Workshop on Symbiotic AI for Cyber Social Security: Human–AI Collaboration for Social Resilience and Trustworthy Digital Ecosystems (SAI4CCS 2026), co-located with the 2nd Joint National Conference on Cybersecurity (ITASEC & SERICS 2026), February 09–13, 2026, Cagliari, Italy

giuseppe.desolda@uniba.it (G. Desolda); francesco.greco@uniba.it (F. Greco); rosa.lanzilotti@uniba.it (R. Lanzilotti)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-14

To enable adaptive, symbiotic protection, we must transform the user from a “black box" into a quantifiable variable within the security model. In this paper, we propose a methodological framework to achieve this transformation. We ground our work in Morpheus (Multidimensional Framework for Human Factors in Cybersecurity) [1], a novel theoretical model that maps 50 human factors to specific cyber-threats. Our contribution is the operationalization of this theory into the *Phishing Susceptibility Questionnaire* (PSQ). Unlike broad security intention scales [15], the PSQ is designed to function as a granular “sensor," translating abstract psychological constructs (e.g., *Need for Cognition*, *Trust*, *Cyber-Risk Beliefs*) into measurable micro-behaviors aligned with specific attack vectors (e.g., Phishing, Spear-Phishing, SMishing, Malware Download).

By validating these measures, we aim to provide the structured data layer necessary for Symbiotic AI agents to construct dynamic user risk profiles. This allows for the transition from static “firewalls" to personalized “nudges"—interventions that are timely, context-aware, and psychologically calibrated to the individual’s immediate vulnerability.

The paper is structured as follows: Section 2 outlines the MORPHEUS framework; Section 3 details the translation of theory into psychometric items; Section 4 describes the validation protocol; and Section 5 discusses the architecture of future symbiotic defenses.

2. The MORPHEUS Framework: Modeling the Human Attack Surface

The paradigm of Symbiotic AI in Cyber Social Security rests on a fundamental premise: for a machine agent to effectively protect a human partner, it must possess a high-fidelity model of that partner’s internal state. However, current cybersecurity protocols primarily operate on a “security-by-obscurity" or “security-by-compliance" model, treating the user as a monolithic entity with binary states—either compliant or compromised. This reductionist view fails to capture the complex topology of the “Human Attack Surface," which varies dramatically across individuals and contexts.

To bridge this gap, this research is grounded in Morpheus (Multidimensional Framework for Human Factors in Cybersecurity) [1]. Moving beyond simplistic taxonomies that attribute breaches to generic “human error," MORPHEUS integrates literature from cognitive psychology, behavioral economics, and organizational science to deconstruct vulnerability into a dynamic system. It models the user not as a passive victim, but as an active link in the security chain, where deep-seated internal characteristics (Human Factors) intersect with, and are triggered by, specific external stimuli (Attack Vectors). By adopting this ontological perspective, we can transition from reactive defense to predictive user modeling. An overview of the main components of Morpheus is reported in 1; the next sub-sections briefly describe these components. More details are reported in the original paper [1].

2.1. Mapping Human Factors to Threat Vectors

The first layer of the framework addresses the specificity of vulnerability. A core contribution of MORPHEUS is the assertion that “generic vulnerability" is a fallacy; human risk is strictly vector-dependent. Just as a biological virus requires specific cellular receptors to bind and infect, a social engineering attack requires a specific psychological “hook" to succeed. The framework rigorously maps 50 distinct human determinants—ranging from stable personality traits to transient emotional states—to specific cyber-threats, creating a matrix of “lock-and-key" dependencies.

As illustrated in the mapping excerpt (Table 1), this differentiation reveals distinct, non-overlapping risk profiles. Consider the case of *Authority Bias*, a cognitive heuristic where individuals attribute greater accuracy to the opinions of authority figures. MORPHEUS identifies this as a primary predictor for *Spear-Phishing*, where attackers impersonate executives to trigger an immediate compliance reflex. However, the framework posits that this same bias plays a negligible role in automated *Malware Download* attacks, where other factors, such as *Curiosity* (a facet of Openness to Experience) or *Risk-Taking Behavior*, become the dominant drivers.

Similarly, the framework analyzes the impact of environmental constraints on cognitive performance. The physical limitations of mobile interfaces (e.g., small screens, the inability to hover over links to

inspect URLs) amplify the effects of specific cognitive deficits. Consequently, factors such as *Distraction* and *Impulsivity* are identified as critical, high-weight predictors of *SMishing* (SMS Phishing) susceptibility. In contrast, in a desktop environment, where greater scrutiny is facilitated and multitasking is less physically cumbersome, these factors may have a reduced impact. For a Symbiotic AI agent, this matrix can serve as diagnostic logic, dictating which psychological variables must be monitored in the current technical threat context, thereby allowing for precise resource allocation.

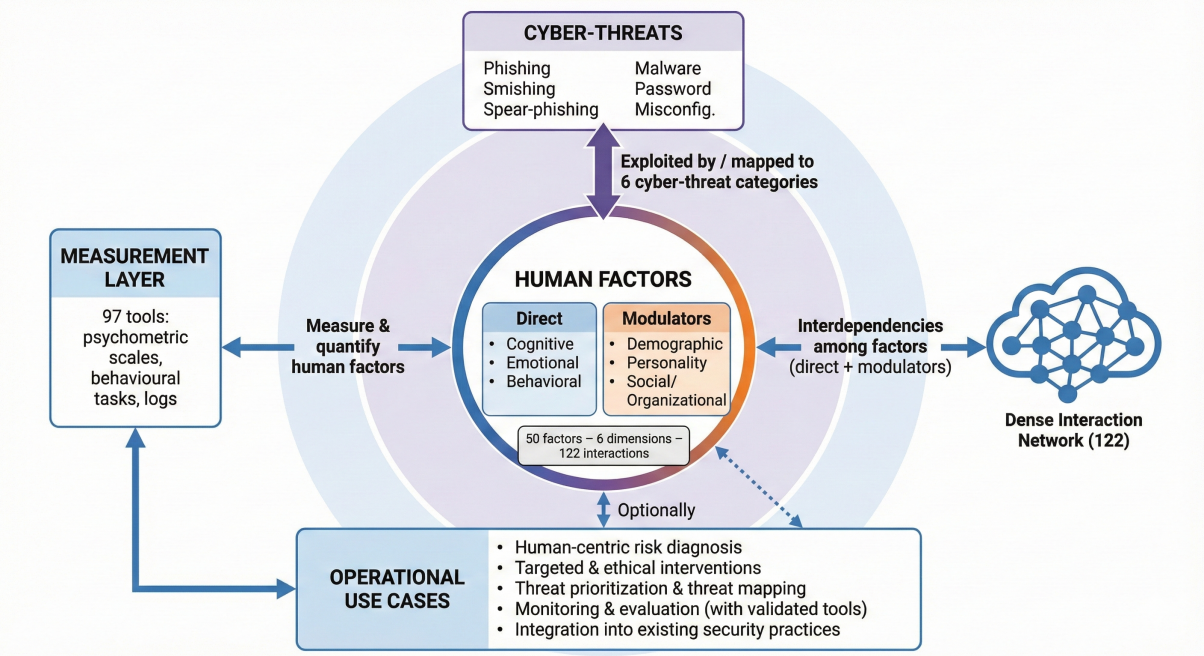


Figure 1: The MORPHEUS Framework structure, illustrating the multidimensional mapping between internal human factors and external attack vectors.

2.2. Discovery of Inter-Factor Interactions and Dynamics

A static map of factors, while useful, is insufficient to predict complex behavior in real-world scenarios. MORPHEUS advances the state of the art by modeling the **synergistic interactions** and non-linear dynamics between these factors. Human traits do not operate in a vacuum; they amplify, dampen, or mask one another depending on the temporal context. The framework explicitly identifies how transient states can act as “multipliers" for stable traits, creating windows of heightened vulnerability that a static profile would miss.

A prime example of this dynamic is the interaction between *Cognitive Fatigue* (a transient state) and *Impulsivity* (a behavioral trait). A user may exhibit average to manageable levels of trait impulsivity under normal conditions. However, MORPHEUS reports that a state of high cognitive fatigue—e.g., induced by a long workday or complex tasks—can make individuals rely on heuristics and act more impulsively [16]. This interaction effectively “simulates" high impulsivity, rendering the user critically vulnerable to urgency-based attacks during specific time windows. An AI agent monitoring only the stable trait would fail to predict this failure; an agent modeling the interaction would anticipate it.

Furthermore, the framework explores the interplay between *Personality* and *Organizational Culture*. For instance, the trait of *Trust* (strongly tied to Agreeableness) is often viewed as a socially positive attribute. However, when situated in an organizational culture that discourages questioning superiors (High Power Distance), a high-trust individual becomes exponentially more vulnerable to social engineering than they would be in a “Zero Trust" culture. These inter-factor dynamics suggest that vulnerability is an emergent phenomenon, arising from the intersection of trait, state, and context. Understanding these non-linearities is crucial for predictive modeling, as it enables the Symbiotic AI to

anticipate risk spikes even in users who appear historically secure based on their static attributes alone.

2.3. The Measurement Challenge: From Fragmented Scales to Unified Assessment

While MORPHEUS provides a robust theoretical ontology, translating this complex, multidimensional map into operational data presents a significant logistical challenge. This is the bottleneck that has historically prevented the adoption of psychological profiling in active cyber-defense. To measure the 50 factors identified by MORPHEUS using traditional psychometric instruments, a security team would be forced to administer a prohibitive battery of disparate tests.

One would need to deploy several solutions to measure different human factors, for example, the *Big Five Inventory* (BFI-10 or NEO-PI-R) to assess personality [17] or the *Barratt Impulsiveness Scale* (BIS-11) to capture behavioral inhibition [18]. Additionally, separate assessments would be required for technical knowledge, digital habits, and organizational commitment. Such an approach is operationally untenable in any real-world organization. Administering hundreds of items across disconnected scales would induce massive survey fatigue, leading to "satisficing" behaviors where users provide low-quality data to finish the task. Furthermore, measuring "general impulsivity" (e.g., via the Abbreviated Impulsiveness Scale [19]) might not be an ecologically valid measure for predicting "clicking impulsivity" in a digital environment.

This measurement gap creates a critical disconnect: we have the theory to understand the user (MORPHEUS), but we lack the sensors to capture the data efficiently. This limitation necessitates the development of a **unified, domain-specific instrument**. We require a single, cohesive tool capable of directly probing this diverse array of factors—cognitive, emotional, and behavioral—within the context of cyber-threats, rather than inferring them from abstract psychological constructs. This requirement is the driving force behind the design of the *Phishing Susceptibility Questionnaire* (PSQ). As detailed in the following section, the PSQ consolidates these multidimensional assessments into a streamlined set of "micro-behavioral" policies, enabling the holistic quantification of the MORPHEUS user model without the burden of a prohibitive testing battery.

Table 1
Excerpt of the MORPHEUS Mapping Matrix. The • symbol indicates a theoretically supported link between the factor and the attack vector, illustrating the granular, vector-dependent nature of human vulnerability.

Human Factor	Category	Phishing	Spear-Phish	SMishing	Malware	Passwords
Authority Bias	Cognitive	•	••			
Impulsivity	Behavioral	•	•	••	•	•
Curiosity	Personality	•	•	•	••	
Social Norms	Social	•	•	•		••
Distraction	Cognitive	•		••		
Trust	Personality	•	••		•	
Fatigue	Cognitive	•		•		••

3. From Theory to Measurement: PSQ Design Methodology

Translating the high-level constructs of MORPHEUS into a data stream usable by an AI agent presents a dual challenge comprising psychometric validity and computational utility. A generic survey question such as "Are you careful online?" generates noise rather than signal, as it relies on the user's subjective and often inflated self-perception. To overcome this limitation, we engineered a structured **Translation Protocol** designed to convert abstract human factors into rigorous, machine-readable metrics. The design of the *Phishing Susceptibility Questionnaire* (PSQ) follows a sophisticated three-stage feature engineering pipeline.

3.1. Scope of the Current Instrument

Although the MORPHEUS framework maps human factors across six distinct vectors (including phishing, SMishing, spear phishing, malware download, password management, and misconfiguration), the operationalization presented in this paper prioritizes the **Email Attack Surface**. Consequently, the current iteration of the PSQ specifically targets *Phishing* (mass-market) and *Spear-Phishing* (targeted) vectors. This scoping decision is driven by the prevalence of email as the primary entry point for advanced persistent threats (APTs) and the need to establish a robust baseline validation before expanding to mobile-specific vectors (SMishing), which introduce complex hardware-dependent variables (e.g., screen size, touch interactions).

3.2. Stage 1: Establishing the Ground Truth (The Behavioral Policy)

The first step in our protocol involves defining a "Behavioral Policy" for each mapped Human Factor. This policy acts as the computable heuristic or "Ground Truth", i.e., the ideal state against which the user's behavior is measured. For instance, when addressing *Impulsivity* in the context of Phishing, we do not simply ask about the trait itself. Instead, we identify the specific behavioral inhibition required to mitigate the risk, such as pausing physically before interacting with a link.

Every policy generated is metadata-tagged with the specific page number and verbatim quote from the scientific literature that justifies it. This evidential lineage allows a future AI agent to provide evidence-based explanations for its interventions. For example, if an agent delays an email delivery, it can justify this action to the user by citing the specific research indicating that impulsivity increases click-rates by a measurable percentage, thereby fostering trust in the human-machine team.

3.3. Stage 2: Contextual Operationalization and Ecological Validity

Abstract policies must be translated into scenarios that resonate with the user's daily experience. We purposely abandoned generic security jargon—such as "check the SSL certificate" or "examine the header"—in favor of "plain English"descriptions of the user experience. This contextualization is vector-specific.

For **Phishing**, the items focus on visual cues present in standard email clients and the cognitive decision-making process at the moment of the click. For **Spear-Phishing**, the focus shifts to the sociological dynamics of the workplace, examining how the user responds to hierarchical pressure from executives or requests that bypass standard internal procedures. This ensures that the instrument measures vulnerability within the ecological constraints of the actual threat landscape.

3.4. Stage 3: Mirrored Item Pairs as Consistency Checks

Self-reported data is notoriously susceptible to biases, particularly acquiescence bias, where users tend to agree with statements regardless of content. To maximize the signal-to-noise ratio for the AI model, we operationalized each policy into a **Mirrored Item Pair**. This consists of a *Positive Item*, which describes the active execution of the defense (e.g., "I verify the sender..."), and a *Negative Item*, which describes the active failure or risky alternative (e.g., "I skip verification and click...").

This dual structure serves a critical function beyond simple scoring: it acts as a mechanism for **Consistency Checking**. If a user agrees with both the positive and negative representations of a behavior—a logical contradiction—the AI agent can flag this data point as unreliable or of low confidence. This prevents the system from making potentially erroneous adaptations based on noisy input, thereby enhancing the robustness of the user model.

The final output consists of 138 questionnaire items (74 phishing, 64 spear phishing). This granular dataset allows the Symbiotic AI to perform **Root Cause Analysis**: distinguishing whether a security failure was caused by a personality trait like *Trust* (which might require skepticism training) or a transient state like *Distraction* (which requires real-time interruption).

Table 2
The Translation Protocol: From abstract Factor to computable Policy, to observable Item Pairs. Ten examples covering different MORPHEUS dimensions across Phishing and Spear-Phishing vectors.

Vector	Factor	Behavioral Policy (AI Ground Truth)	Positive Item (+)	Negative Item (-)
Phishing	*Impulsive Behavior* (Behavioral)	When I have a sudden urge to click an email link, take a breath.	I take a breath and wait a second before clicking a link in an email when I feel a sudden urge.	I skip the wait and click the email link directly when I feel a sudden urge.
	Bias (Halo Effect) (Cognitive)	When I see famous logos in an email, I ignore them and look at the web address instead.	I ignore the logos and look at the actual web address (URL) before trusting an email.	I skip looking at the web address and trust the email directly because of the logos.
	Fear (Emotional)	When threatened with a blocked account in an email, go to the site via browser.	I go to the website via my browser (not the link) to check my account status when an email threatens to block me.	I skip using the browser and click the link in the email directly to check my status.
	Social Proof (Social & Organiz.)	When an email says "everyone is doing it", check first.	I check if it is true before joining, even if the email says everyone else has already signed up.	I skip checking the facts and join directly just because the email says everyone else signed up.
	Extraversion (Social) (Personality)	When getting a social invite via email, check where the link goes before clicking.	I check exactly where the link goes (e.g., by hovering over it) before clicking on social invites in my email.	I skip checking the link destination and click directly on social invites in my email.
Spear-Phishing	*Bias (Authority)* (Cognitive)	When a VIP asks for a favor via email, check their exact email address strictly.	I check the sender's exact email address to ensure it is internal before doing a favor asked by a VIP via email.	I skip checking the address and do the favor directly when a VIP asks via email.
	Agreeableness (Trust) (Personality)	When a colleague asks for sensitive info via email, call them to verify it is really them.	I call the colleague to verify it is really them before sending sensitive information requested via email.	I skip calling the colleague and send sensitive information directly via email.
	Anxiousness (Emotional)	When worried about a client email, check with a supervisor.	I check with a supervisor before replying if I am worried about making a mistake with a client email.	I skip checking with a supervisor and reply directly to the client email when I am worried.
	Organizational Tenure (Social & Organiz.)	When an email uses the wrong internal slang, suspect it is fake.	I suspect the email is fake before replying if it uses the wrong internal slang or jargon.	I skip checking the slang and reply directly to the email even if it sounds odd.
	Internet Usage (Behavioral)	When checking work emails on phone, look for "External" tags.	I look for the "External" tag or label before replying to work emails on my phone.	I skip looking for the tag and reply directly to work emails on my phone.

4. Validation Protocol: Ensuring Diagnostic Precision

The primary purpose of the Phishing Susceptibility Questionnaire (PSQ) is to provide a reliable, granular diagnosis of human cyber-risk. Unlike generic personality tests, the PSQ claims to predict vector-specific security behaviors. Therefore, before operationalization, it requires rigorous psychometric calibration. We have designed a comprehensive two-phase validation protocol ($N_{total} = 1100$) to establish reliability, structural stability, and criterion-related validity.

4.1. Scope: Internal Validity and Vector Control

Validating a new psychometric scale requires strict alignment between the self-reported measure and the behavioral ground truth. Although MORPHEUS maps risks across mobile (SMishing) and desktop contexts, this initial validation is restricted to the **Email Vector** (Phishing and Spear-Phishing). Including mobile vectors would necessitate a multi-device protocol, introducing confounding variables such as screen size and input modality. By isolating the email vector in a controlled desktop environment,

we ensure that measured correlations reflect psychological susceptibility rather than interface usability artifacts.

4.2. Study 1: Scale Calibration and IRT Purification

The pilot study ($N = 500$) is dedicated to "purifying" the instrument. The initial item pool contains 138 items derived from 69 behavioral policies. To reduce this to a parsimonious set (approx. 50–60 items) without compromising diagnostic resolution, we employ a *Randomized Split-Ballot Design.*

Experimental Design To mitigate survey fatigue while testing a large item pool, participants are randomly assigned to one of four conditions based on a 2 (Vector: Phishing vs. Spear-Phishing) $\times$ 2 (Polarity: Positive vs. Negative wording) matrix. This ensures that while the aggregate dataset covers the full pool, each participant responds to a manageable subset (≈ 40 items), preserving data quality.

Analytic Strategy: The "Tournament" Approach Data analysis follows a two-step pipeline to select the optimal item formulation for each policy:

1. **Structure Discovery (EFA):** We use Polychoric correlations with *Principal Axis Factoring* (Promax rotation) to identify the latent dimensionality of the constructs.
2. **Item Selection (IRT):** Using a Graded Response Model (GRM), we generate Item Information Curves for each mirrored pair. For each policy, we retain the single item (Positive or Negative) that exhibits higher discrimination (a-parameter) and information content ($I(\theta)$). This ensures the final PSQ is optimized for differentiating between high- and low-risk users.

4.3. Study 2: Confirmatory Validation and Behavioral Prediction

The second phase ($N = 600$) addresses the fundamental criterion question: *Does the PSQ predict actual security failures?* We employ a **Counterbalanced Predictive Design**, where participants complete the purified PSQ and a high-fidelity behavioral simulation, with order randomized to control for priming effects.

The Validation Battery To establish construct validity, the PSQ is administered alongside:

- *Convergent Measures:* Security Behavior Intentions Scale (SeBIS) [15] and Cyber Vigilance Scale (CVS).
- *Discriminant Measures:* UPPS-P Brief Impulsivity Scale [18], used to demonstrate that PSQ measures vector-specific vulnerability rather than generic personality traits.

Behavioral Benchmark: The Interactive Inbox Participants engage with a simulated webmail client containing 60 emails (30 legitimate, 30 malicious), balanced by difficulty according to the **NIST Phish Scale** taxonomy [20].

Hypotheses and Metrics Performance is quantified using *Signal Detection Theory* (d'). We test three hierarchical hypotheses using Structural Equation Modeling (SEM):

- **H1 (Structural Stability):** Confirmatory Factor Analysis (CFA) will confirm the factor structure extracted in Study 1 (CFI $> .90$, RMSEA $< .08$).
- **H2 (Granular Prediction):** Prediction is vector-specific. High scores on specific PSQ factors (e.g., *Authority Bias*) will predict lower d' specifically on the subset of emails leveraging those cues, exhibiting a stronger correlation than the global risk score.
- **H3 (Incremental Validity):** Hierarchical regression will demonstrate that PSQ scores explain significant additional variance in behavioral performance ($R^2\Delta$) above and beyond demographics and generic scales (SeBIS, Big Five), justifying the need for this specialized instrument.

5. Discussion: From Measurement to Mitigation

The development of the Phishing Susceptibility Questionnaire (PSQ) represents a paradigm shift from “one-size-fits-all" security to “data-driven personalization." By operationalizing the MORPHEUS framework into a validated instrument, we provide organizations with the missing metric: a high-resolution map of the human attack surface. This diagnostic capability transforms how we approach the “Human Factor," enabling three distinct layers of intervention: Targeted Training, Strategic Risk Management, and Symbiotic Defense.

5.1. Layer 1: Precision Training (Next-Gen SETA)

Currently, Security Education, Training, and Awareness (SETA) programs suffer from a phenomenon known as “training fatigue." Employees are often required to read lengthy texts or watch generic videos, regardless of their specific needs. The PSQ enables **Diagnostic-Prescriptive Training**, separating users based on the *nature* of their vulnerability:

- **The Knowledge Deficit:** A user scoring high on *Lack of Technical Knowledge* but low on behavioral risk factors needs education. For them, a module on “How to read a URL domain" is effective.
- **The Behavioral Deficit:** A user scoring high on *Impulsivity* or *Decision Fatigue* often knows the rules but fails to apply them under pressure. For this profile, more technical training is useless. The intervention must be metacognitive: teaching “Stop-Look-Think" protocols or “Cooling-off" strategies.
- **The Social Deficit:** A user scoring high on *Authority Bias* or *Social Norms* requires role-play simulations specifically focused on asserting boundaries with executives.

By tailoring the remediation to the root cause, organizations can maximize the ROI of training with in-person or remote learning [21] while reducing friction for employees who are already secure.

5.2. Layer 2: Organizational Risk Auditing

At the macro level, the PSQ serves as a strategic sensor for Chief Information Security Officers (CISOs). Aggregating PSQ data enables the visualization of the "Psychological Topology" of the entire workforce. For instance, a heatmap might reveal that the Sales Department has high vulnerability in *Social Norms* (indicating a culture of risky compliance), while the IT Department scores high on *Overconfidence*. This data empowers evidence-based policy making. If an entire unit shows high *Cognitive Fatigue* vulnerability, the solution is not a security seminar, but a change in operational workflows or email policies (e.g., restricting after-hours email) to reduce the cognitive load that attackers exploit.

5.3. Layer 3: Enabling Symbiotic AI

Finally, the granular data provided by the PSQ acts as the foundational layer for **Symbiotic AI** systems. While training addresses the human, symbiotic systems use this data to adapt the digital environment *to* the human. Once the PSQ is validated, its scores can feed a **Personalized Risk Engine**.

- **Dynamic Heuristics:** If a user is diagnosed with high susceptibility to *Spear-Phishing* (due to high *Trust*), the AI agent can dynamically tighten the spam filter thresholds specifically for that user’s inbox, blocking “grey-area" emails that would be allowed for a more skeptical user.
- **Psychological Nudging:** The agent can intervene with “Explainable Nudges." Instead of a generic warning, the AI can display: *“I blocked this email because it uses Urgency cues, and your profile indicates you are highly sensitive to time-pressure."*

This represents the ultimate goal of human-centric cybersecurity: a defense system that not only enforces rules but also understands and compensates for the specific psychological blind spots of its partner.

5.4. Ethical Considerations

Deploying psychometric profiling in the workplace raises ethical questions regarding privacy and potential discrimination. Such an approach should not only leverage private and secure software [22, 23], but it is vital that PSQ data is treated as sensitive health data. We advocate for a “Safety Culture" approach: scores should be used solely to configure defensive support and training, never for performance evaluation or hiring decisions. The goal of MORPHEUS is to build a “Digital Shield" tailored to the user, rather than labeling users as liabilities.

6. Conclusion and Future Work

In this work, we have argued that the realization of true Symbiotic AI in the domain of Cyber Social Security is currently hindered by a fundamental data deficit: the lack of granular, scientifically validated metrics regarding the human partner. While defensive algorithms have achieved high sophistication in analyzing technical payloads, their understanding of the user—the intended target of social engineering—has remained static and reductive. By presenting the methodological design of the *Phishing Susceptibility Questionnaire* (PSQ), we propose a blueprint to bridge this gap, transforming the user from an unpredictable "weakest link" into a quantifiable component of the security architecture.

Our primary contribution lies in the rigorous operationalization of the MORPHEUS framework. We have demonstrated that abstract psychological constructs—ranging from personality traits, such as *Conscientiousness*, to transient states, like *Cognitive Fatigue*—can be systematically translated into observable behavioral policies. The resulting instrument, comprising 138 context-aware items across six distinct threat vectors, serves not only as an assessment tool but also as a high-resolution “sensor" for user modeling. By enforcing strict traceability to scientific evidence and employing mirrored item pairs for consistency checking, we have engineered a dataset specifically designed to meet the requirements of Explainable and Robust AI.

The implications of this work extend beyond psychometrics into the design of adaptive interfaces and intelligent agents. A validated PSQ enables the transition from ’one-size-fits-all" security training to dynamic, real-time intervention. It enables a future where a symbiotic agent can distinguish whether a user is at risk due to a lack of knowledge (requiring education) or a temporary lapse in attention (requiring friction), and adapt its defensive posture accordingly.

Our immediate future work is dedicated to the empirical execution of the pre-registered validation protocols described in this study. Following the calibration of the instrument through Exploratory and Confirmatory Factor Analyses (Study 1) and its benchmarking against behavioral signal detection tasks (Study 2), we intend to integrate the PSQ into a live Symbiotic AI testbed. This will enable us to transition from measuring susceptibility to actively mitigating it, thereby validating the hypothesis that personalized, psychologically grounded defenses significantly outperform static protection mechanisms. Ultimately, this research posits that the most resilient firewalls of the future will not be built solely of code, but of a deep, data-driven understanding of the human mind.

Acknowledgments

This work has been supported by the Italian Ministry of University and Research (MUR) and by the European Union-NextGenerationEU, under grant PRIN 2022 PNRR “DAMOCLES: Detection And Mitigation Of Cyber attacks that exploit human vuLnerabilitiES” (Grant P2022FXP5B) CUP: H53D23008140001.

This work is partially supported by the co-funding of the European Union - Next Generation EU: NRRP Initiative, Mission 4, Component 2, Investment 1.3 - Partnerships extended to universities, research centres, companies and research D.D. MUR n. 341 del 5.03.2022 – Next Generation EU (PE0000014 – “Security and Rights In the CyberSpace – SERICS” - CUP: H93C22000620001).

The research of Francesco Greco is funded by the Italian Ministry of University and Research (MUR) and by the European Union - NextGenerationEU, Mission 4, Component 2, Investment 1.1, under

grant PRIN 2022 “DevProDev: Profiling Software Developers for Developer-Centered Recommender Systems”—CUP: H53D23003620006.

Declaration on Generative AI

During the preparation of this work, the author(s) used *Grammarly Pro* in order to fix grammatical errors and improve text quality.

References

[1] G. Desolda, F. Greco, R. Lanzilotti, C. Tucci, Morpheus: A multidimensional framework for modeling, measuring, and mitigating human factors in cybersecurity, arXiv preprint arXiv:2512.18303 (2025).

[2] S. Hu, C. Hsu, Z. Zhou, Security education, training, and awareness programs: Literature review, Journal of Computer Information Systems 62 (2022) 752–764. URL: https://doi.org/10.1080/08874417.2021.1913671. doi:10.1080/08874417.2021.1913671.

[3] Verizon, 2025 Data Breach Investigations Report (DBIR), Technical Report, Verizon, 2025. URL: https://www.verizon.com/business/resources/reports/dbir/, evidenzia l’aumento del phishing guidato da AI e l’abuso di credenziali.

[4] P. Buono, G. Desolda, F. Greco, A. Piccinno, Let warnings interrupt the interaction and explain: designing and evaluating phishing email warnings, in: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, CHI EA ’23, ACM, New York, NY, USA, 2023, pp. 1–6. URL: https://doi.org/10.1145/3544549.3585802. doi:10.1145/3544549.3585802.

[5] G. Desolda, F. Di Nocera, L. Ferro, R. Lanzilotti, P. Maggi, A. Marrella, Alerting users about phishing attacks, in: International Conference on Human-Computer Interaction, Springer, 2019, pp. 134–148.

[6] F. M. Cau, G. Desolda, F. Greco, L. D. Spano, L. Vigano, Can large language models automate phishing warning explanations? a controlled experiment on effectiveness and user perception, 2025. URL: https://arxiv.org/abs/2507.07916. arXiv:2507.07916.

[7] F. Greco, G. Desolda, P. Buono, A. Piccinno, Enhancing Phishing Defenses: The Impact of Timing and Explanations in Warnings for Email Clients, Computer Standards & Interfaces 93 (2025) 103982. URL: https://www.sciencedirect.com/science/article/pii/S092054892500011X. doi:https://doi.org/10.1016/j.csi.2025.103982.

[8] F. Greco, G. Desolda, C. Tucci, A. Esposito, A. Curci, A. Piccinno, Improving phishing resilience with ai-generated training: Evidence on prompting, personalization, and duration, arXiv preprint arXiv:2512.01893 (2025).

[9] G. Desolda, F. Greco, L. Viganò, Apollo: A gpt-based tool to detect phishing emails and generate explanations that warn users, Proceedings of the ACM on Human-Computer Interaction 9 (2025) 1–33.

[10] F. Greco, G. Desolda, A. Esposito, Explaining Phishing Attacks: An XAI Approach to Enhance User Awareness and Trust, 2023. URL: https://ceur-ws.org/Vol-3488/paper22.pdf.

[11] A. Vishwanath, T. Herath, R. Chen, J. Wang, H. R. Rao, Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model, Decision Support Systems 51 (2011) 576–586.

[12] G. Desolda, A. Esposito, R. Lanzilotti, A. Piccinno, M. F. Costabile, From human-centered to symbiotic artificial intelligence: a focus on medical applications, Multimedia Tools and Applications 84 (2025) 32109–32150.

[13] G. Desolda, G. Dimauro, A. Esposito, R. Lanzilotti, M. Matera, M. Zancanaro, A human–ai interaction paradigm and its application to rhinocytology, Artificial Intelligence in Medicine 155 (2024) 102933.

[14] S. Kim, M. S. Wogalter, Habituation, dishabituation, and recovery effects in visual warnings, Human Factors and Ergonomics Society Annual Meeting 53 (2009) 1612–1616. URL: https://journals.sagepub.com/doi/abs/10.1177/154193120905302015. doi:10.1177/154193120905302015.

[15] S. Egelman, E. Peer, Scaling the security wall: Developing a security behavior intentions scale (SeBIS), in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15), ACM, 2015, pp. 2235–2244.

[16] B. Stanton, M. F. Theofanos, S. S. Prettyman, S. Furman, Security fatigue, It Professional 18 (2016) 26–32. URL: https://api.semanticscholar.org/CorpusID:206470587. doi:10.1109/MITP.2016.84.

[17] J. Costa, Paul T., R. R. McCrae, The revised neo personality inventory (neo-pi-r), in: G. J. Boyle, G. Matthews, D. H. Saklofske (Eds.), The SAGE Handbook of Personality Theory and Assessment, volume 2, Sage Publications, Inc., Thousand Oaks, CA, 2008, pp. 179–198. URL: https://doi.org/10.4135/9781849200479.n9. doi:10.4135/9781849200479.n9.

[18] J. H. Patton, M. S. Stanford, E. S. Barratt, Barratt impulsiveness scale–11 (bis-11) [database record], 1995. URL: https://doi.org/10.1037/t05661-000. doi:10.1037/t05661-000.

[19] C. G. Coutlee, C. S. Politzer, R. H. Hoyle, S. A. Huettel, An abbreviated impulsiveness scale constructed through confirmatory factor analysis of the barratt impulsiveness scale version 11, Archives of Scientific Psychology 2 (2014) 1–12. URL: https://doi.org/10.1037/arc0000005. doi:10.1037/arc0000005.

[20] M. P. Steves, K. K. Greene, M. F. Theofanos, The NIST Phish Scale: A method for rating human phishing detection difficulty, Proceedings of the Human Factors and Ergonomics Society Annual Meeting 64 (2020) 661–665.

[21] V. S. Barletta, F. Cassano, A. Marengo, A. Pagano, J. Pange, A. Piccinno, Switching learning methods during the pandemic: A quasi-experimental study on a master course, Applied Sciences 12 (2022). URL: https://www.mdpi.com/2076-3417/12/17/8438. doi:10.3390/app12178438.

[22] M. T. Baldassarre, V. S. Barletta, D. Caivano, A. Piccinno, M. Scalera, Privacy knowledge base for supporting decision-making in software development, in: C. Ardito, R. Lanzilotti, A. Malizia, M. Larusdottir, L. D. Spano, J. Campos, M. Hertzum, T. Mentler, J. Abdelnour Nocera, L. Piccolo, S. Sauer, G. van der Veer (Eds.), Sense, Feel, Design, Springer International Publishing, Cham, 2022, pp. 147–157.

[23] M. Saltarella, G. Desolda, R. Lanzilotti, V. S. Barletta, Translating privacy design principles into human-centered software lifecycle: A literature review, International Journal of Human–Computer Interaction 40 (2024) 4465–4483. URL: https://doi.org/10.1080/10447318.2023.2219964. doi:10.1080/10447318.2023.2219964. arXiv:https://doi.org/10.1080/10447318.2023.2219964.