EUROPEAN INTERNATIONAL UNIVERSITY

EIU Journal of Action Research

Measuring What Training Cannot See:

VISHWAR as a Longitudinal Framework for Human Cyber Risk Assessment

Mayukh Paul

Security Researcher & Architect, PR3CORIA
MS in Cybersecurity, Katz School of Science and Health
Yeshiva University

Abstract

Organizations routinely measure security awareness. They rarely measure what happens after it. Completion rates and post-training quiz scores are administratively useful, but they say almost nothing about whether a person will recognize a spear-phishing attempt three months later, or hold firm against an MFA push-flooding attack on a stressful afternoon. The gap between what gets measured and what actually needs to be understood is wider than the field tends to acknowledge.

This paper documents the early development and pilot implementation of VISHWAR, a simulation-driven behavioral cybersecurity framework designed to generate longitudinal telemetry on human susceptibility to social engineering attacks. The framework is organized around the Human Vulnerability Management Lifecycle (HVML) — a model that positions human risk not as a fixed knowledge deficit but as a dynamic operational state that shifts with context, cognitive load, and time. Pilot data from a cohort of 20 participants surfaced uneven susceptibility across attack vectors, persistent vulnerability concentrated in MFA fatigue and vishing scenarios, and a gradual but uneven improvement in reporting behavior across simulation cycles. The work is explicitly early-stage; the findings are treated as directional observations rather than validated conclusions. What the pilot suggests is that behavioral measurement — longitudinal, vector-specific, and individual-level, may be a more productive frontier for human-centric cybersecurity than the awareness-training

paradigm currently dominant in organizational practice. This paper is less a report of findings than a reflection on what it takes to begin building that kind of measurement infrastructure.

Keywords: behavioral cybersecurity, human vulnerability measurement, simulation-based training, longitudinal risk assessment, social engineering

1. Introduction and Problem Context

There is a persistent mismatch in how organizations manage human cyber risk. The dominant model — deliver awareness training, record completion, move on — treats susceptibility as something that can be corrected once and set aside. Run the phishing simulation. Send the reminder. Check the box. The underlying assumption is that once trained, a person stays trained.

It is a convenient assumption. It is not a well-supported one.

What the behavioral literature actually shows is that susceptibility to social engineering varies not just across people but within the same person over time. Someone who correctly flagged a phishing simulation in January may fall for a more contextually tailored attempt in spring, particularly if they are under pressure or simply haven't thought about it since. MFA fatigue works differently: it floods a user with push authentication prompts until exhaustion produces approval — and exhaustion is not a knowledge gap that training can address. Vishing succeeds most consistently through urgency and trust-signal construction, both of which are contextual rather than informational.

This is less a failure of training than a failure of measurement. The field has built extensive infrastructure to track whether training has occurred. Very few organizations have built anything comparable to track whether behavioral risk is actually changing and fewer still measure it over time.

That gap motivates this research. The central question is practical: how can organizations move beyond one-time awareness metrics toward continuous, behaviorally grounded measurement of human cyber risk? VISHWAR was built as a direct attempt to work through that question. This paper documents where the effort currently stands — what the pilot revealed, what it didn't, and what the experience of building and running it taught me about the limits and possibilities of this kind of approach.

2. Background and Rationale

The human element appears as a primary attack vector in almost every major cybersecurity incident report in recent years. The Verizon Data Breach Investigations Report has consistently attributed the majority of confirmed breaches to human action — phishing, credential misuse, social engineering — across multiple annual editions. The organizational response, by and large, has been to invest more heavily in training programs. More modules, more simulations, better content. The measurement infrastructure surrounding all of this has largely not kept pace.

The field has optimized for dashboard stability rather than behavioral fidelity. Annual phishing simulations produce numbers like click rates, report rates, completion percentages that are legible and defensible in compliance contexts. They are also, in important ways, temporally shallow. They capture a snapshot. They do not capture drift.

Several research traditions are relevant here, and they tend to be cited in cybersecurity contexts without being fully absorbed. Protection motivation theory, originally developed in health communication, suggests that how individuals respond to perceived threat is shaped by perceived efficacy, self-efficacy, and threat appraisal in complex, non-uniform ways, not simply by whether they received correct information. Cognitive load research complicates the picture further: susceptibility appears to be partly state-dependent, varying with workload, stress, and attentional context rather than remaining fixed. Vishwanath and colleagues' (2016) work on automaticity is particularly instructive — experienced users are

not reliably more resistant; habituated pattern-recognition can, in some conditions, be more exploitable than deliberate evaluation.

None of this invalidates awareness training as a practice. It does challenge the assumption that training is sufficient as a risk management strategy and it raises a harder question. If susceptibility drifts, shifts with context, and varies within the same individual over time, what would it actually mean to measure it?

3. The VISHWAR Framework and HVML

VISHWAR is a simulation-driven behavioral assessment framework. The acronym is secondary to what the system does: generate structured, longitudinal data on how individuals respond to simulated attack scenarios — across time, across attack types, and across repeated cycles of observation.

It was developed during graduate cybersecurity research as a direct response to the measurement gap described above. The distinction between a training platform and a measurement platform matters here. VISHWAR is not primarily designed to teach. It is designed to observe repeatedly over time and to surface behavioral patterns in human susceptibility that a single annual simulation cannot reveal. That orientation shapes every design decision in the framework.

The platform currently operates across four attack vectors: phishing, vishing, MFA fatigue, and pretexting. Each scenario is designed for contextual plausibility within ethical constraints. Participants are enrolled with informed consent in an ongoing security research program; they are not notified of specific scenario timing or content. This is not deception in a harmful sense, but it does preserve enough behavioral realism to make the telemetry meaningful.

VISHWAR is organized around the Human Vulnerability Management Lifecycle (HVML), a five-phase model that structures the assessment process across repeated cycles:

Baseline Assessment — an initial simulation cycle to establish individual and cohort-level susceptibility profiles across vectors, before any targeted intervention.

Behavioral Telemetry Collection — structured logging of interaction behaviors across scenarios: click events, credential submissions, MFA response patterns, reporting latency, and debrief observations.

Risk Stratification — cross-vector analysis to identify individuals and clusters where susceptibility is concentrated, persistent, or particularly vector-specific.

Targeted Micro-Intervention — scenario-specific, behaviorally grounded feedback rather than generic remediation training. The goal is to respond to what the telemetry actually shows, not what a training calendar prescribes.

Longitudinal Reassessment — repeat simulation cycles to evaluate whether behavioral patterns shift, stabilize, or degrade over time — and to distinguish genuine behavioral change from temporary performance effects.

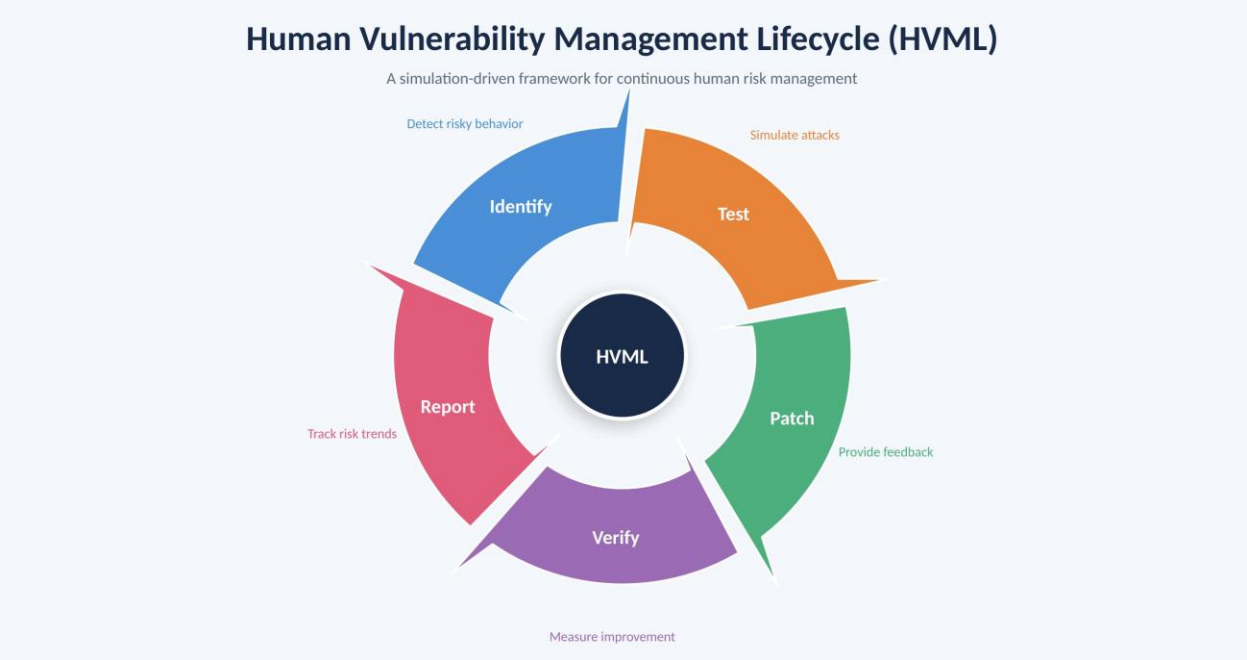

Figure 1 — HVML Lifecycle Diagram *—The HVML framework structures longitudinal human cyber risk assessment across five recurring phases: Baseline Assessment, Behavioral Telemetry Collection, Risk Stratification, Targeted Micro-Intervention, and Longitudinal Reassessment. The*

cyclical structure emphasizes continuous behavioral evaluation rather than one-time awareness remediation, with each phase generating operational outputs that inform subsequent assessment cycles.

The conceptual logic underlying HVML borrows from how technical vulnerability management operates. Security teams do not patch a CVE once and assume the system is permanently safe. They track, reassess, and prioritize based on ongoing risk posture. The argument here is that human susceptibility warrants analogous treatment, not because people are equivalent to software vulnerabilities, but because the one-time-remediation logic is equally insufficient in both domains.

What distinguishes this approach from existing simulation tools is the emphasis on the longitudinal record. A single simulation event produces one data point. A sequence of events across time, vectors, and shifting contextual frames starts to build something more useful: a behavioral pattern. The aim is to accumulate enough of that record to say something meaningful about how susceptibility evolves and whether anything done during the process actually changes it.

4. Pilot Methodology

The pilot was conducted with a cohort of 20 participants in a graduate academic environment. All participants were enrolled with informed consent in a security research program involving periodic simulated attack scenarios. They knew the program existed; they did not know when specific scenarios would arrive or what form they would take.

Three simulation cycles were run across the study period. Each cycle drew from at least three of the four vectors, though not every vector appeared in every cycle. Scenarios were re-framed across cycles, varying contextual presentation while preserving the core behavioral demands of each attack type to reduce purely recognition-based avoidance without eliminating the realism the telemetry depends on.

Telemetry collection varied by vector. For phishing, the system logged click events, credential entry attempts, and reporting submissions with timestamps. MFA fatigue scenarios captured push-response behavior and response latency. Vishing was assessed through structured debriefs after each interaction, since automated telemetry is limited there. Pretexting relied on behavioral observation and post-scenario notes.

The analysis was descriptive throughout. Given the cohort size, inferential statistics would have been inappropriate, and pursuing them would have misrepresented what the pilot was actually designed to do. The goal at this stage was to test the telemetry infrastructure, identify what behavioral patterns emerged, and surface the methodological problems that would need to be addressed in the next iteration. Findings should be read accordingly as directional observations from an early-stage exploratory effort, not as validated conclusions.

4.1 Telemetry Standardization: Current State and Forward Plan

The most consequential methodological limitation of the current implementation is the inconsistency of telemetry granularity across vectors. Phishing scenarios generate automated, machine-readable event logs with consistent structure. MFA fatigue produces push-response timing data — structured but narrower in scope. Vishing relies almost entirely on structured post-interaction debriefs, which are richer in qualitative texture but harder to normalize into a cross-vector schema. Pretexting sits closest to ethnographic observation: interpretively valuable, but the least tractable for automated analysis.

This matters because the longitudinal comparison the framework promises depends on comparing behavioral responses meaningfully across vectors and across cycles. As long as telemetry schemas differ substantially by attack type, cross-vector analysis rests on imperfect translations between different kinds of data — some automated, some manual, some behavioral and some self-reported. That is not a fatal limitation at the pilot stage, but it becomes one at scale.

The primary technical priority for the next iteration is a unified behavioral event schema — a common data structure capable of representing interaction events across all four vectors with consistent fields for event type, response latency, decision outcome, and confidence proxy. For vectors where automated logging is inherently limited, structured elicitation protocols will be developed to capture the same behavioral dimensions through different collection mechanisms. The goal is not to pretend that a vishing debrief and a phishing click log are the same kind of data. It is to ensure that the analytical layer above them operates on a consistent representation, enabling comparison without misrepresenting what was actually measured. This schema does not yet exist in full. It is named here as a design requirement rather than a solved problem.

5. Preliminary Observations and Findings

A number of patterns became visible across the three cycles that a single-point simulation would not have surfaced. None of them are conclusive at this sample size. Several of them are worth documenting.

5.1 Susceptibility Did Not Transfer Cleanly Across Vectors

One of the more consistent early observations was how poorly susceptibility in one vector predicted performance in another. Participants who showed low click rates in phishing scenarios showed no reliable advantage when MFA fatigue scenarios were deployed. The behavioral demands of the two attack types are genuinely different — phishing exploits attention and judgment; MFA fatigue works on habituation and cognitive load under pressure and the data reflected that distinction. What became clearer over repeated cycles was that susceptibility appears to be partly vector-specific, not a general trait that transfers uniformly across attack types. A profile that looks strong in phishing may mask real exposure elsewhere.

Figure 2 — Cross-Vector Susceptibility Comparison *— Participant susceptibility levels across phishing, vishing, MFA fatigue, and pretexting scenarios during cycle one. The figure illustrates that low susceptibility in one vector did not reliably predict low susceptibility in another, suggesting substantial divergence in individual behavioral profiles across attack types.*

5.2 MFA Fatigue and Vishing Showed Persistent Vulnerability

Across all three cycles, MFA fatigue and vishing scenarios produced the highest and most consistent susceptibility rates in the cohort. MFA fatigue in particular showed limited response to the targeted micro-interventions delivered after cycle one. In retrospect, this was not surprising: the mechanism it exploits — cognitive load accumulation under push-notification pressure is not something awareness feedback can reliably address. Knowing that MFA flooding is a technique does not change how someone responds to their fifteenth push notification at the end of a long day. The persistence of this pattern across multiple cycles was one of the operationally important observations from the pilot.

Vishing showed similarly strong persistence, with somewhat higher individual-level variability. Voice-based social engineering appears to remain compelling even when participants are enrolled in a program where they know simulations are occurring which itself is worth noting as a limitation of what awareness-oriented preparation can achieve for this vector.

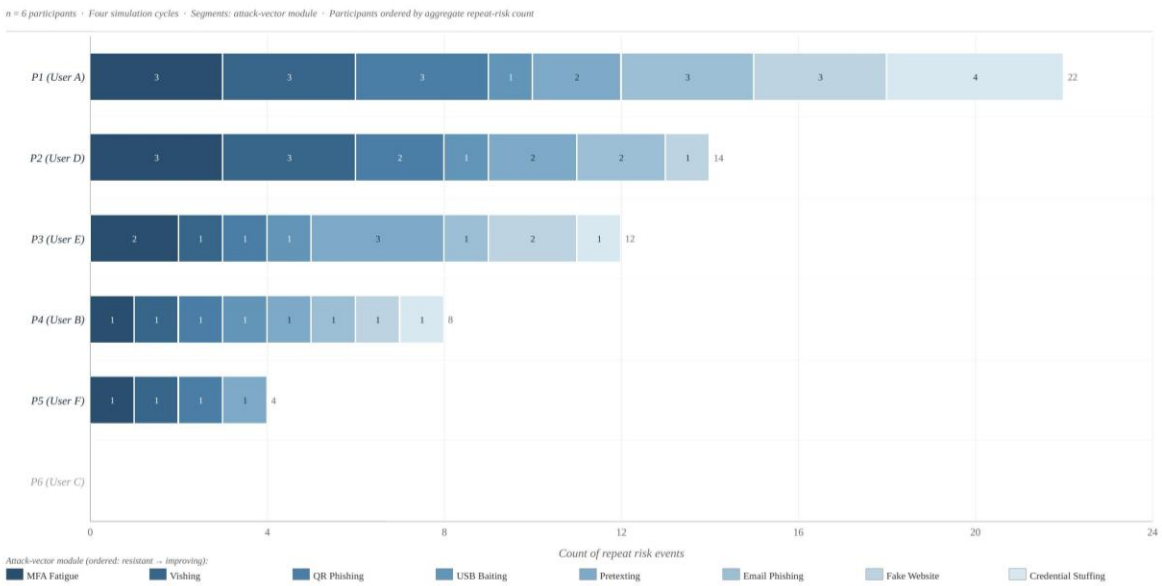

Figure 3 — Longitudinal Susceptibility Drift Across Cycles *— Susceptibility rates across phishing, MFA fatigue, and vishing scenarios over three simulation cycles. The figure illustrates a gradual decline in phishing susceptibility over time, while MFA fatigue and vishing susceptibility remained comparatively elevated and stable, suggesting uneven behavioral adaptation across attack vectors. Lighter participant-level trajectories demonstrate substantial heterogeneity within the cohort.*

5.3 Phishing Click-Through Rates Declined — But Unevenly

Phishing click-through rates showed a downward trajectory across the three cycles, moving from a cohort average of roughly 35% in cycle one to around 18% by cycle three. That is a real shift, and it is worth acknowledging. But it is a misleading summary if read alone. The decline was unevenly distributed across the cohort — a persistent subset showed high click rates across all three cycles, while others improved sharply after the first and then plateaued. The aggregate metric improved. The high-risk tail did not. Any framework that reports only the mean is obscuring the part of the distribution that most needs attention.

5.4 Reporting Behavior Improved Over Time

Proactive reporting — participants flagging suspected simulation scenarios without being prompted roughly doubled between cycle one and cycle three. This was one of the more encouraging observations, because it suggests something more than avoidance: an active

cognitive posture toward suspected threats. Whether that posture would generalize outside a simulation context, or whether it reflects program-specific priming rather than durable behavioral change, is genuinely unknown from the current data. But the direction seems right.

5.5 Behavioral Heterogeneity Was the Core Finding

Step back from any individual metric and what stands out most clearly is the degree of heterogeneity within a cohort of 20. Individual-level variation was substantial enough that cohort-level statistics consistently obscured more than they revealed. There was no such thing as a typical susceptibility profile. That is not a methodological limitation — it is exactly what longitudinal behavioral measurement is supposed to surface. The field's reliance on aggregate metrics is, at least partly, a product of not having the infrastructure to track individual patterns over time. This pilot suggested that infrastructure is buildable, and that what it reveals looks quite different from what aggregate reporting currently shows.

6. Reflection on Challenges and Learning

Running this pilot clarified several things that theoretical design work had left underdeveloped.

6.1 Ethical Dimensions of Simulation-Based Behavioral Research

The ethical structure of this research required more deliberate attention than the initial design anticipated — both because ethics-related decisions materially shaped what data could be collected, and because longitudinal simulation research surfaces tensions that single-point studies largely avoid.

All participants enrolled through an informed-consent process that disclosed the nature of the research program, the types of simulated scenarios that would be used, and their right to withdraw at any point without academic or professional consequence. Consent was

framed at the program level rather than the scenario level — participants knew they were part of an ongoing behavioral security research effort, but were not notified of specific scenario timing or content. This distinction is ethically significant. Scenario-level pre-notification would eliminate the behavioral realism the research depends on; program-level disclosure preserves enough contextual awareness to satisfy the core conditions of informed participation. Whether this balance is philosophically satisfying depends on how one weighs scientific fidelity against full transparency — and reasonable disagreement exists on that question.

The scope of deception here is deliberately narrow. Participants are not told when a simulation will occur, or what form it will take. They are not deceived about the existence of the research, its purposes, or their right to exit. No scenario involves contact with individuals outside the study, and none involves financial, reputational, or professional harm as a design mechanism. These constraints are ethically meaningful; they also mean the scenarios are probably less aggressive than real-world attacks, which has implications for ecological validity discussed elsewhere.

Participant distress was a live consideration throughout, particularly in MFA fatigue scenarios, which involve repeated interruptions over a compressed period. Debriefs were conducted after every cycle, and participants were encouraged to raise concerns at any point. No participant formally withdrew, though informal feedback indicated that MFA scenarios were experienced as more aversive than phishing or pretexting scenarios. A scenario realistic enough to be aversive is, in a sense, doing exactly what it is supposed to do — but that creates a research relationship cost that needs active management in any study of significant duration.

What the pilot reinforced is that the research relationship in longitudinal simulation work is itself a methodological variable. Participants who feel repeatedly deceived — even within a disclosed research frame — may disengage, behave strategically, or develop orientations toward the simulation program that contaminate the telemetry. Maintaining the relationship

over time requires ongoing transparency about the program's progress, honest communication about what the data is showing, and genuine responsiveness to participant feedback. These are not ethical formalities. They are preconditions for methodologically sound longitudinal behavioral research.

Institutional safeguards in the pilot included IRB-equivalent ethics review within the academic context, data anonymization protocols, and storage of identifiable behavioral records in a restricted-access system. No raw telemetry was linked to identifiable personal information outside the immediate research team. These protocols will require formal scaling before the research moves into organizational settings.

6.2 The Learning–Fatigue Ambiguity

The most persistent conceptual difficulty in interpreting the pilot data was distinguishing behavioral learning from behavioral fatigue. When phishing click-through rates decline across simulation cycles, two interpretations remain defensible: participants are genuinely developing better recognition capacity, or they have grown sufficiently fatigued with the simulation program that they click less on anything that feels like a test. The current design cannot cleanly separate these — and the difference matters enormously for how the findings should be read.

This ambiguity is not merely inconvenient. It is a validity threat, and one that future study designs need to address directly. Blind simulation cycles — delivered without any prior indication that an assessment period is active — would test whether behavioral improvement holds when participants are not in a heightened-awareness state. Randomized scenario timing would reduce participants' ability to model when a simulation is likely to arrive and adjust behavior accordingly. Decoy non-security interactions, inserted into the experimental environment, could help distinguish selective vigilance from generalized behavioral change. Control cohorts — comparable groups not enrolled in the simulation program — would allow behavioral trajectories to be compared against an unexposed baseline. The most defensible approach would combine several of these, but

doing so requires a study scale and duration substantially beyond what this pilot could support.

Interpreting small-cohort behavioral data without overclaiming also required more active discipline than I anticipated. Multiple times during analysis, patterns emerged that felt significant until the sample size reasserted itself. The habit of distinguishing observation from evidence is easier to describe than to consistently apply when you have spent months building a system and watching it run. I mention this not as a caveat but as a methodological observation: early-stage action research is particularly exposed to the pull of premature interpretation, and that pull is worth naming explicitly.

What the pilot confirmed, despite all of this, is that the longitudinal approach surfaces things that point-in-time assessments cannot. The persistent MFA fatigue risk, the disconnect between aggregate CTR improvement and individual-level stagnation, the vector-specificity of susceptibility profiles — none of these are visible from a single annual simulation. That seems worth the methodological difficulty.

7. Implications for Practice

Even at this preliminary stage, the pilot data pushes back against a few assumptions built into standard organizational security programs.

The first is that aggregate improvement is not the same as general improvement. Organizations that rely on cohort-level CTR reduction as a success indicator may be declaring victory at precisely the moment when their highest-risk individuals are being left unchanged. A risk management orientation — as distinct from a compliance orientation — attends to the tail of the distribution, not the mean. The individuals who showed no improvement across three simulation cycles are exactly the ones that matter most. Standard aggregate reporting provides no mechanism to identify them.

The second concerns coverage. Most simulation programs are phishing-centric, partly because phishing scenarios are technically easier to deploy at scale and partly because

click-through rates are administratively clean to report. But if MFA fatigue and vishing represent persistently high-yield vectors for attackers — which current threat intelligence increasingly suggests — then programs weighted toward phishing produce a systematically distorted picture of where human risk actually lives. What gets measured gets managed. What doesn't get measured tends not to.

There is also a third implication, and arguably the most practically significant: behavioral telemetry enables early identification of individuals whose susceptibility remains consistently high across time and attack types. These are not necessarily people with knowledge gaps. They are people whose behavioral response, under realistic conditions, stays vulnerable — regardless of what training they have completed. Identifying them early, and offering targeted vector-specific support rather than more generic awareness content, is what a genuine risk management approach to human cyber vulnerability looks like. Most organizations do not have that infrastructure today.

8. Future Research Directions

The current work's limitations are a reasonably clear guide to what needs to happen next.

Cohort expansion is the most immediate priority. Twenty participants is sufficient to pilot methodology and observe directional patterns; it is not sufficient to support comparative analysis across organizational roles, technical backgrounds, or work-context differences. The next phase will expand to a larger and more occupationally diverse cohort, with attention to whether vector-specific susceptibility patterns hold differently for technical versus non-technical staff, or for individuals working in high-deadline environments compared to lower-pressure ones. These are operationally relevant differences, not just demographic ones.

The longitudinal window also needs to extend substantially. Three simulation cycles in a single study period captures behavioral patterns at a moment in time — it does not capture how those patterns evolve across months or years, or how they respond to real-world security incidents that occur between cycles. A twelve-month minimum longitudinal design

is probably necessary to say anything meaningful about behavioral drift. Longer would be better, and would begin to address the learning-versus-fatigue ambiguity that the current design cannot resolve.

The telemetry infrastructure requires standardization. A unified behavioral data schema that supports consistent cross-vector and cross-cycle comparison is a prerequisite for the comparative analysis the framework is designed to enable. The current implementation works; it is not yet consistent enough to support the kind of longitudinal comparison that would make the work compelling at scale.

There is also a conceptual question the current work has not seriously addressed: what would a practically actionable, longitudinally sensitive human risk score actually look like? Technical vulnerability management has reasonably mature frameworks for translating security signals into prioritized operational decisions. The human equivalent is largely underdeveloped. Whether the HVML model, scaled and empirically validated, can serve as a foundation for something analogous is the longer-term question this research is working toward.

8.1 Toward a Longitudinal Human Risk Score: Theoretical Foundations

One of the more practically significant questions this research raises — without yet answering — is what a longitudinal human risk score would actually look like: a quantified representation of an individual's behavioral susceptibility profile that is sensitive to change over time, vector-specific in its composition, and actionable as a risk management input.

Technical vulnerability management has reasonably mature precedents. The Common Vulnerability Scoring System translates complex vulnerability characteristics into a normalized score that supports prioritization and remediation tracking. The appeal of something analogous for human risk is clear: organizations need mechanisms to prioritize attention, and a behavioral risk profile that evolves over time offers something that a point-in-time click rate fundamentally cannot.

The challenge is that human susceptibility does not reduce cleanly to a single number without losing something important. Any serious scoring framework would need to represent at least the following: vector-specific susceptibility profiles, since — as the pilot demonstrated — susceptibility in phishing does not predict susceptibility in MFA fatigue; temporal drift, since a score derived from behavior three months ago may misrepresent current risk posture; and repeat-risk persistence, meaning the degree to which high susceptibility in one cycle predicts high susceptibility in subsequent cycles. This last dimension may be more operationally relevant than any single-cycle score, precisely because it encodes behavioral momentum rather than a snapshot. Reporting latency and the ratio of proactive to reactive security behaviors could contribute additional signal.

Confidence weighting is also necessary. A score derived from three simulation events is less reliable than one derived from eighteen, and treating both with equal confidence produces false precision. Longitudinal behavioral trajectories — the direction and rate of change across cycles — are arguably the most actionable output of the framework, more useful than any point-in-time value because they tell you something about where a person is heading, not just where they currently stand.

There is a difficulty worth naming directly: reducing a person to a score, even a carefully constructed one, carries real institutional risk. Scores become normalized. They get applied in contexts their designers did not anticipate. A human risk score that informs targeted, supportive intervention is useful; one used for punitive action, hiring decisions, or public ranking is harmful. The governance and use-policy questions are not secondary concerns to be addressed after the measurement methodology is built — they are part of the design problem. Any serious development of this concept needs to engage with them from the outset.

8.2 Scalability and Enterprise Operationalization

The research as currently structured depends on direct researcher involvement at every stage: scenario design, telemetry collection, debrief facilitation, cycle coordination. That

level of involvement is appropriate for a 20-person pilot. It does not scale to enterprise environments with hundreds or thousands of employees — and if the framework is to be operationally relevant, it needs to account honestly for what scaling requires.

Telemetry automation is the most tractable scaling requirement. The manual components of the current collection process need to be replaced or supplemented by automated data pipelines. For phishing and MFA fatigue scenarios, automated collection is already largely in place; the challenge is extending that infrastructure to more behaviorally complex scenarios without losing the interpretive richness that makes the data meaningful. At enterprise scale, the telemetry infrastructure becomes a data engineering problem as much as a behavioral science one.

Privacy constraints become substantially more complex at scale, and in ways that matter for the framework's legitimacy as much as its legality. Behavioral telemetry collected at the individual level, stored in organizational systems, and used to inform management decisions sits in a fundamentally different relationship to individuals than research program data does. Employees in many jurisdictions have legal protections governing the collection and use of behavioral data in workplace settings. The informed-consent structure that works in a research program — where participation is genuinely voluntary and the researcher has no power over the participant's professional life — does not map cleanly onto an employment relationship. Any enterprise deployment needs explicit data governance policies and meaningful transparency about what is collected and how it is used.

Alert fatigue is a genuine operational risk at scale. A system that flags high-risk individuals for intervention only creates value if someone acts on those flags. In environments where security teams are already overloaded with technical alerts, a stream of human-risk signals risks becoming another ignored dashboard — another mechanism for organizational defensibility rather than operational change. The intervention pipeline matters as much as the measurement infrastructure; a risk stratification system that does not connect to a

lightweight, context-appropriate response process produces actionable intelligence that never gets acted on.

The ethical stakes of scaling also deserve direct attention. In a small pilot, the research relationship depends on direct communication, responsive debriefs, and genuine participation in a study participants understand and trust. At enterprise scale, that relationship is mediated by HR structures, management hierarchies, and organizational power dynamics that can substantially alter how the program is experienced. The risk of behavioral measurement infrastructure being perceived as surveillance — rather than as support — is real, and the conditions under which that distinction holds are organizational and cultural, not merely technical. Those conditions need to be designed for explicitly, not assumed.

None of this argues against scaling. It argues for scaling thoughtfully, with explicit attention to governance structures, privacy protections, and intervention pipelines that make the measurement infrastructure meaningful rather than merely deployable. The pilot provides a methodological foundation. What comes next depends as much on the institutional choices organizations make about how to use it.

9. Conclusion

Security awareness training has become a standard feature of organizational practice. Its ubiquity is not evidence of its effectiveness. The dominant model — train, certify, move on — is epistemically limited in a specific way: it records whether training was delivered, but says very little about whether behavioral risk has actually changed. And almost nothing about whether any change observed at a single point in time will persist.

VISHWAR and the HVML framework represent an attempt to build different infrastructure — oriented around longitudinal behavioral measurement rather than one-time awareness delivery. The pilot work described here is early, explicitly exploratory, and honest about what it cannot yet establish. It is not a validated theory of human susceptibility. It is a set of

observations from a small study, collected through an imperfect but iteratively improving methodology, that suggest the measurement approach is worth taking seriously and building out further.

What the data does support, even at this stage is the core argument: human susceptibility is not static, it is not uniform across attack vectors, and it is not reliably addressed through training alone. Measuring it well — continuously, behaviorally, at the level of the individual — is a harder problem than developing better training content. It is also, this research increasingly suggests, the more important one. The field has spent considerable effort building better simulations. It has spent considerably less effort building the infrastructure to learn from them over time. That is the gap this work is trying to close.

References

Anderson, C. L., & Agarwal, R. (2010). Practicing safe computing: A multimethod empirical examination of home computer user security behavioral intentions. MIS Quarterly, 34(3), 613–643.

Canham, M., Posey, C., & Strickland, D. (2021). Phishing for long-term behavioral change: A longitudinal field experiment. Computers & Security, 111, 102456.

Hadlington, L. (2017). Human factors in cybersecurity: Examining the link between internet addiction, impulsivity, attitudes towards cybersecurity, and risky cybersecurity behaviours. Heliyon, 3(7), e00346.

Heartfield, R., & Loukas, G. (2015). A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks. ACM Computing Surveys, 48(3), 1–39.

Lain, D., Kostiainen, K., & Capkun, S. (2021). Phishing in organizations: Findings from a large-scale and long-term study. In Proceedings of the IEEE Symposium on Security and Privacy (pp. 1793–1810).

Mitnick, K. D., & Simon, W. L. (2002). The Art of Deception: Controlling the Human Element of Security. Wiley.

Posey, C., Roberts, T. L., Lowry, P. B., Bennett, R. J., & Courtney, J. F. (2011). Insiders' protection of organizational information assets: Development of a systematics-based taxonomy and theory of diversity for protection-motivated behaviors. MIS Quarterly, 37(4), 1189–1210.

Verizon. (2024). Data Breach Investigations Report. Verizon Enterprise Solutions. https://www.verizon.com/business/resources/reports/dbir/

Vishwanath, A., Harrison, B., & Ng, Y. J. (2016). Suspicion, cognition, and automaticity model of phishing susceptibility. Communication Research, 45(8), 1146–1166.

Witte, K. (1996). Predicting risk behaviors: Development and validation of a diagnostic scale. Journal of Health Communication, 1(4), 317–341.

Workman, M. (2008). Wisecrackers: A theory-grounded investigation of phishing and pretext social engineering threats to information security. Journal of the American Society for Information Science and Technology, 59(4), 662–674.