Cuando el enemigo está dentro

Cada día, millones de correos electrónicos intentan engañar a sus destinatarios mediante técnicas como el *phishing*. Si bien la tecnología logra detener entre el 90% y el 99% de estos ataques, el pequeño porcentaje que logra pasar las barreras técnicas suele bastar para poner en peligro tanto a las organizaciones como al eslabón más débil de la cadena: el ser humano. Investigadores españoles han creado un nuevo algoritmo que puede prevenirlo.

[ILUSTRACIÓN: NANZEEBA IBNAT / ISTOCK]

¿Por qué algunas personas, incluso advertidas por las herramientas de seguridad y los antivirus, siguen haciendo clic en un enlace fraudulento? ¿Qué nos hace confiar más en un correo que nos promete un estupendo premio que en una alerta que nos invita a ser precavidos?

Cada día, millones de correos electrónicos intentan engañar a sus destinatarios mediante técnicas como el *phishing*, un tipo de ciberataque que busca obtener información confidencial —contraseñas, datos bancarios o credenciales corporativas— mediante la manipulación psicológica. Si bien la tecnología logra detener entre el 90% y el 99% de estos ataques, el pequeño porcentaje que logra pasar las barreras técnicas suele bastar para poner en peligro tanto a las organizaciones como al eslabón más débil de la cadena: el ser humano. Por eso, el siguiente gran avance pasa por comprender ese 1% (y protegerlo).

El análisis del comportamiento de los usuarios ha sido, precisamente, el centro de la investigación del proyecto EVE (*Emotions and Vulnerabilities Exposed and Protected*), que ha dado como resultado un algoritmo capaz de aumentar nuestra ciberprotección. Integrado en una plataforma de neurociberseguridad, este algoritmo predice la vulnerabilidad del factor humano ante un ciberataque, basándose en diversas variables psicológicas. Se trata de un nuevo y pionero enfoque en ciberseguridad que une neurociencia, tecnología y psicología para predecir el riesgo humano ante un ataque digital y protegernos, así, de nosotros mismos.

Un cerebro preparado para sobrevivir… pero no en Internet

Nuestro cerebro está programado para reaccionar con rapidez ante amenazas físicas. Cuando percibimos un peligro —un ruido repentino o una sombra inesperada—, se activa la amígdala, que pone en marcha una respuesta automática de defensa. Este mecanismo, conocido como sesgo de negatividad, nos ha permitido sobrevivir durante millones de años.

Sin embargo, en el entorno digital el sistema no se activa porque no tenemos factores de supervivencia: no hemos generado una respuesta instintiva a las amenazas, cosa que sí sucede cuando oímos el rugido de un león, incluso aunque nunca hayamos estado en la sabana. Como leer un correo electrónico aparentemente no pone en riesgo nuestra supervivencia, el cerebro no enciende la alarma emocional. Y cuando lo hace, suele ser en la dirección equivocada: el miedo se orienta hacia las consecuencias de no actuar, no hacia el ataque.

Mensajes como “Su cuenta será bloqueada si no actualiza sus datos” o “Ha perdido 1.000 euros de su cuenta bancaria, pulse aquí para recuperarlos” provocan miedo al castigo, no al engaño. Ese miedo “secuestra” la

atención y deja todo el peso de la decisión al pensamiento racional, más lento y exigente. Si además estamos cansados, distraídos, estresados o bajo presión, nuestra capacidad para analizar el mensaje disminuye y el clic se vuelve casi inevitable.

Personalidad y contexto

El modelo científico que sustenta el algoritmo EVE se apoya en tres variables psicológicas y de personalidad clave, a las que se añaden variables contextuales, como la carga de trabajo, la multitarea, la presión del tiempo o el nivel de implicación con el asunto del correo. Todo ello puede aumentar o reducir nuestra capacidad para procesar la información de forma crítica.

La primera de esas variables es el Sistema de Inhibición del Comportamiento (BIS). Mediado por la ansiedad, quienes puntúan alto reaccionan con miedo ante posibles castigos y son más vulnerables a mensajes del tipo "si no actúas ahora, pierdes algo".

Otro elemento a tener en cuenta es el Sistema de Activación del Comportamiento (BAS). Está asociado a la impulsividad y la búsqueda de recompensa, y los usuarios responden más fácilmente a mensajes que prometen beneficios inmediatos ("gana un premio", "aprovecha la oferta").

Finalmente, interviene la Necesidad de Cognición (NC), que mide la tendencia a disfrutar del pensamiento complejo. Las personas con NC suelen analizar más y caer menos, aunque pueden ser víctimas de correos que apelan a la curiosidad intelectual ("descubre más", "lee este informe exclusivo").

Integrando estos componentes, EVE ha generado un perfil dinámico de vulnerabilidad, que no pretende etiquetar a las personas sino entender en qué condiciones concretas cada individuo es más propenso a caer. Lejos de ser estático, el algoritmo entrena según la toma de decisiones y el comportamiento humano. Si cambia el nivel de vulnerabilidad, se alerta al usuario y a la organización de la transformación, porque lo que se pretende es un círculo virtuoso.

Este empieza con la validación del autodiagnóstico, de cómo somos, las variables psicológicas y rasgos de personalidad. A partir de ahí, se despliega un mecanismo que se basa en simulaciones de *phishing* para validar esa hipótesis continuamente. Así se genera un sistema de alerta, denominado semáforo, que se contrarresta con una serie de microhistorias que explican los procesos cognitivos por los que se ha caído o no en esa simulación. Y vuelta a empezar. En este círculo virtuoso, el algoritmo está siempre entrenando y aprendiendo del comportamiento de los usuarios.

Ciencia aplicada a una ciberseguridad más humana

Nuestro equipo probó el modelo en dos fases: primero con estudiantes universitarios y, después, con empleados de empresas, para aproximarse a contextos laborales reales. Cada participante completó pruebas de personalidad y se enfrentó a simulaciones de phishing mientras los investigadores medían tiempos de reacción, emociones evocadas y decisiones tomadas. Con esos datos, el algoritmo aprendió a predecir patrones de conducta y puntos débiles, y el usuario recibía píldoras de aprendizaje personalizadas sobre ciberseguridad.

Pero no todo el mundo tiene que recibir la misma formación para enfrentar una amenaza que llega por correo electrónico. Una persona con alta ansiedad no debería recibir el mismo entrenamiento que otra más impulsiva. Así, la formación adaptada al perfil psicológico puede reducir significativamente el riesgo de caer en la trampa y, sobre todo, evitar la falsa sensación de seguridad que generan los cursos genéricos.

De esta forma, el proyecto EVE marca un cambio de paradigma: entender la ciberseguridad no solo como un

desafío técnico, sino como un fenómeno profundamente humano. Los ciberdelincuentes no atacan máquinas: atacan emociones. Por eso, los sistemas del futuro deberán aprender a protegernos también de nuestras propias vulnerabilidades.

El proyecto EVE ha sido desarrollado por la Universidad Pontificia Comillas junto a TechHeroX, Ticsmart, Softcom, la Universidad Autónoma de Madrid e INCIBE, y ha sido cofinanciado por el Instituto Nacional de Ciberseguridad (INCIBE) a través de la Compra Pública de Innovación, con fondos del Plan de Recuperación, Transformación y Resiliencia de la Unión Europea – NextGenerationEU, dentro del programa de Compra Pública Innovadora del INCIBE.

Alhaddad, M., Mohd, M., Qamar, F. y Imam, M. "*Study of student personality trait on spear-phishing susceptibility behavior*" en *International Journal of Advanced Computer Science and Applications* (2023, vol. 14, n.º 5).

Frauenstein, E. D. y Flowerday, S. "*Susceptibility to phishing on social network sites: A personality–information-processing model*" en *Computers & Security* (2020, vol. 94, 101862).

Kadena, E. y Gupi, M. "*Human factors in cybersecurity: Risks and impacts*" en *Security Science Journal* (2021, vol. 2, n.os 2–3, pp. 19–30).

Kavvadias, A. y Kotsilieris, T. "*Understanding the role of demographic and psychological factors in users' susceptibility to phishing emails: A review*" en *Applied Sciences* (2025, vol. 15, n.º 4, 2236).

Khadka, K. y Ullah, A. B. M. S. "*Human factors in cybersecurity: An interdisciplinary review and framework proposal*" en *International Journal of Information Security* (2025, vol. 24, pp. 119–135).

López-Aguilar, P., Urruela, C., Batista, E. y Solanas, À. "Phishing vulnerability and personality traits: Insights from a systematic review" en *Computers in Human Behavior Reports* (2025, 100784).

Pollini, A., Macchi, L., Tosi, F. y Arcuri, G. "*Leveraging human factors in cybersecurity: An integrated human factors (HF) approach*" en *Frontiers in Psychology* (2022, vol. 12, 706422).

Renaud, K. y Goucher, W. "*Exploring cybersecurity-related emotions and finding that they are linked to lived experiences*" en *Humanities and Social Sciences Communications* (2021, vol. 8, n.º 1, 81).