SADIO Electronic Journal of Informatics and Operations Research (EJS) Vol. 25 No. 1 (2026) e-ISSN 1514-6774
https://revistas.unlp.edu.ar/ejs | https://doi.org/10.24215/15146774e095

Reinforcement of cyber deception strategies through simulated user behavior

Federico Pacheco[1] and Diego Staino[2]
[1]Universidad Tecnológica Nacional
[2]Instituto Universitario de la Policía Federal Argentina
fpacheco@frba.utn.edu.ar, diegostaino@hotmail.com

Abstract. Cyber deception has emerged as a pivotal defensive strategy in the effort to detect and counter advanced persistent threats and sophisticated attacks. However, the implementation of conventional methods, such as decoy services, frequently lacks the credible support necessary to optimize their effectiveness. Generally, honeypot and honeytoken systems face limitations, including a lack of realism due to static configurations and an absence of human activity. Additionally, the simulation of actions is costly, and profiling is challenging to scale, as well as automation and adaptability. In this paper, we present a tool designed to automate the generation of realistic activities and behaviors of fictitious users, seeking to integrate personalized and coherent patterns of human interaction in cyber deception scenarios to improve the credibility of decoys. Based on the MITRE Engage framework, the tool contributes to the strengthening of defensive operations, addressing a key challenge of cyber deception.

Keywords: Cyber deception, User behavior, Deception technologies.

Refuerzo de estrategias de ciber engaño mediante comportamiento simulado de usuarios

Resumen. El ciber engaño es una estrategia defensiva clave para detectar y contrarrestar las amenazas persistentes avanzadas y los ataques sofisticados. Sin embargo, la implementación de actividades tradicionales, como servicios señuelos, a menudo carecen de un soporte creíbles que refuerce su efectividad. En general, los honeypots y honeytokens tienen limitaciones como la falta de realismo por configuraciones estáticas y nulos rastros de actividad humana. Simular acciones conlleva costos, la creación y mantenimiento de perfiles es de difícil escalabilidad, así como la automatización e integración. En este trabajo presentamos una herramienta diseñada para automatizar la generación de actividades y comportamientos realistas de usuarios señuelos, buscando integrar patrones personalizados y coherentes de interacción humana en escenarios de ciber engaño para mejorar la credibilidad de la operación. Basándonos en el framework MITRE Engage, la herramienta contribuye al fortalecimiento de las operaciones defensivas, abordando un desafío clave de las estrategias de ciber engaño.

Received August 2025; Accepted November 2025; Published February 2026

Palabras clave: Ciber engaño, Comportamiento de usuario, Tecnologías de engaño.

1 Introducción

El ciber engaño (*cyber deception*) es una estrategia de ciberdefensa activa que busca superar la asimetría entre atacante y defensor mediante el uso de técnicas similares a las empleadas por los propios adversarios. Esta estrategia se fundamenta en actividades dirigidas por humanos, con elementos automatizados, que incrementan la diversidad y complejidad de los sistemas, dificultando que los atacantes recopilen información útil (Pacheco, 2022). Consiste en desplegar trampas y señuelos instrumentados en infraestructuras o sistemas que imitan total o parcialmente activos reales de manera que, si un atacante interactúa con estos, se pueden detectar, monitorear, y supervisar sus acciones a tiempo. Su principal objetivo es confundir, ralentizar y exponer a los actores maliciosos en entornos controlados. Los señuelos y elementos ficticios pueden ir desde información, servidores o estaciones de trabajo, carpetas o archivos, cuentas, tokens, personas, dominios, servicios y redes.

Un actor de amenazas tiende a asumir que sus interacciones son legítimas, y el ciber engaño degrada su capacidad de interpretación, forzándolo a interactuar con elementos que no le aportan información útil, y generan incertidumbre sobre hallazgos previos. El ciber engaño permite reducir el tiempo de permanencia del atacante en el sistema, acelera la detección de amenazas y disminuye la fatiga de alertas. Además, permite generar Indicadores de Compromiso (IoC) y detectar Técnicas, Tácticas y Procedimientos (TTPs) con bajo índice de falsos positivos. Entre sus principales ventajas están la detección temprana de amenazas, la desviación de recursos del atacante hacia activos falsos y la recopilación de información para ajustar las defensas (Shimeall & Spring, 2014).

En 2022 se presentó el marco MITRE Engage, que organiza las operaciones de ciber engaño y proporciona una guía estructurada para su implementación. Cuenta con tres fases: planificación, donde se definen objetivos y se diseñan los elementos de engaño; ejecución, que implica la implementación y monitoreo de los señuelos; y análisis, enfocado en evaluar la efectividad de las tácticas empleadas para mejorar estrategias futuras. Las actividades incluidas pueden ser de preparación, exposición, afectación, obtención y comprensión (Morovitz et al., 2022) donde las dos primeras se enfocan en la estrategia, y las demás en el compromiso del adversario. No define detalles de implementación, dejándolos a cada organización según su estrategia y postura de seguridad.

El principal desafío en las estrategias de ciber engaño radica en la credibilidad de los señuelos (*decoys*) ya que adversarios avanzados pueden identificar y evadir trampas que carezcan de un contexto humano realista. La mayoría de las herramientas se centran en la generación de activos técnicos falsos, sin simular una interacción de usuarios, lo que reduce la efectividad del engaño y facilita la detección de señuelos.

Para abordar este problema, presentamos una herramienta de código abierto diseñada para automatizar la creación de perfiles señuelo y la emulación de comportamiento humano en entornos de ciber engaño. A los perfiles señuelo los llamaremos "*honey profiles*", concepto usado en comunidades virtuales y redes sociales online (Wani et al., 2018) que no ha sido aplicado a entornos internos de

organizaciones con estos objetivos. El enfoque se basa en tres ejes. El primero es la personalización de perfiles señuelo, con identidades digitales que incluyen roles laborales, patrones de uso, y hábitos de comunicación. El segundo es la simulación de actividades humanas, mediante generación automatizada de interacciones en plataformas (interacción con terminales, navegación web y acceso a datos o recursos). El tercero es la alineación con MITRE Engage, para la planificación, ejecución y análisis de operaciones.

Este trabajo presenta el modelo, arquitectura, y aplicación de esta herramienta de características innovadoras, y discute los desafíos de su implementación. Se estructura en tres secciones. La primera describe el contexto y estado del arte del ciber engaño, analizando los usos actuales. La segunda presenta las problemáticas asociadas al ciber engaño, exponiendo los desafíos específicos que se buscan resolver. La tercera sección desarrolla la propuesta y diseño de la herramienta, describiendo su arquitectura y funcionamiento, junto con una estrategia para validar su efectividad. Finalmente, se presentan las limitaciones y trabajo futuro, junto con la discusión y conclusiones.

2 Contexto y estado del arte

Las herramientas de software que implementan técnicas de ciber engaño, conocidas como tecnologías de engaño, permiten simular sistemas, datos y credenciales, creando entornos en los que es difícil moverse sin activar alertas. Sus funciones incluyen despliegue de señuelos e integración con operaciones de seguridad para mejorar la detección y respuesta, para minimizar el impacto de los ataques al desviar al atacante hacia entornos controlados. La incorporación de inteligencia artificial (IA) y aprendizaje automático impulsó avances en la generación de señuelos y adaptación de estrategias según el movimiento de los atacantes (Gurtu & Lim, 2025) lo que facilitó la automatización y escalabilidad de efectos más realistas, integrados en las prácticas de la ciberdefensa.

Las soluciones modernas permiten personalizar y simular redes completas, aumentando su credibilidad y eficacia, y se integran con herramientas de seguridad como los sistemas de detección y prevención de intrusiones (IDS/IPS) para mejorar la detección de amenazas y la respuesta a incidentes. El software comercial de ciber engaño suele ser costoso, lo que la hace exclusivo de entornos corporativos y gubernamentales de países desarrollados. Aunque existen herramientas de software libre para el despliegue de honeypots y otros elementos específicos, representan solo una pequeña parte de una estrategia integral de ciber engaño (Pacheco & Staino, 2024).

Pese a su efectividad, las tecnologías de engaño requieren una gestión continua para mantener su credibilidad, lo que deriva en tiempo de personas especializadas. Los atacantes desarrollan continuamente métodos para detectar y evadir estos sistemas, lo que obliga a innovar constantemente en las estrategias defensivas. Además, la implementación y mantenimiento de las estrategias pueden ser costosos, por lo que es clave equilibrar la inversión con la eficacia de las medidas (Liebowitz et al., 2021).

Las organizaciones pueden optar por desarrollar sus propias implementaciones utilizando las herramientas de código abierto disponibles más scripts propios, pero este enfoque es más viable en entornos grandes, con recursos adecuados para afrontar costos y cuestiones técnicas. Sin embargo, la implementación propia no evita la introducción de complejidad adicional en entornos productivos, aumentando riesgos y generando conflictos con las áreas de tecnología e infraestructura, que pueden ver estos proyectos como carga adicional. La combinación entre los altos costos de licenciamiento, dificultades en la implementación interna, y complejidad inherente de estas estrategias, conforman los desafíos fundamentales para su avance.

3 Problemática del ciber engaño

Si un atacante detecta que un activo es falso, los esfuerzos de diseño pierden sentido, de lo cual se deriva que la eficacia del ciber engaño depende de la capacidad de los señuelos para imitar sistemas reales de forma verosímil. MITRE Engage enfatiza la importancia de desarrollar narrativas plausibles en entornos de engaño, reforzando que no basta con generar activos técnicos falsos, sino que deben reflejar acciones típicas de usuarios legítimos, a fin de aumentar la credibilidad de los señuelos y mejorar la efectividad de las estrategias en la protección de infraestructuras digitales.

Los honeypots y honeytokens pueden emular sistemas tecnológicos, pero no alcanza, ya que los señuelos tradicionales suelen ser estáticos y no simulan actividad alguna. La ausencia de un contexto de interacción humana convincente en los mismos facilita que los atacantes detecten la falta de registros de actividad real, lo que reduce la efectividad de las trampas tradicionales. Un sistema puede parecer legítimo a nivel técnico, pero sin registros de actividad humana coherente (accesos a aplicaciones o interacciones con otros usuarios) se vuelve más identificable como engaño y el adversario cambiará sus planes. Si un atacante detecta la trampa, podrá evitarla o ajustar sus tácticas, limitando la oportunidad de recolectar inteligencia y desviarlos. Esto disminuye la detección de ataques dirigidos y debilita el rendimiento general de estas estrategias, limitando su efecto defensivo.

Otra problemática común es el diseño de narrativas coherentes, fundamental para que los señuelos sean efectivos. Un sistema sin registros históricos consistentes, ni inicios de sesión en horarios definidos, consultas a sistemas específicos o navegación web acorde, generan sospechas fundadas. Para mejorar el engaño los activos falsos deben reflejar comportamiento típico de usuarios legítimos y mantener coherencia temporal y contextual. En efecto, las operaciones avanzadas requieren adoptar medidas de desinformación estratégica, en tanto difusión controlada de información falsa, para confundir y retrasar las acciones de los ciber atacantes (Aradi & Bánáti, 2025).

Por otro lado, de existir perfiles de usuario sin ser convincentes, se reduce la credibilidad de los señuelos y se facilita su detección (Zhang & Thing, 2021). La creación de perfiles junto con “interacciones humanizadas” es en sí mismo un proceso manual y costoso, lo que limita su escalabilidad. Definir patrones de actividad para cada perfil requiere tiempo y recursos, lo que lo hace inviable ante grandes volúmenes de datos. La simulación efectiva debe ser variada, coherente y personalizada según el perfil de cada usuario para ajustarse a distintos contextos de amenaza. Sin embargo,

muchas herramientas actuales carecen de la capacidad de generar actividades humanas de manera automática y adaptativa. La mayoría se enfoca en la infraestructura técnica, descuidando la simulación del factor humano, lo que deja una brecha que los atacantes pueden explotar. Muchas soluciones dependen de configuraciones predefinidas, limitando su capacidad para evolucionar ante tácticas adversarias en constante cambio. Además, la simulación manual de perfiles ficticios es costosa y poco escalable, dificultando su aplicación en entornos grandes y dinámicos. Asimismo, no existe estandarización para incorporar acciones humanizadas.

Como propuesta de solución a estos problemas, se diseñó una herramienta, que propone facilitar la creación de perfiles ficticios que generen de forma asistida comportamiento humanizado sobre los activos en los que se desarrolle el ciber engaño. Para estas definiciones, se tomó como referencia el marco de MITRE Engage, que proporciona un enfoque integrado y fundado para la implementación de ciber engaño en distintos sectores y contextos de amenaza.

4 Herramienta propuesta

La herramienta Behavioral User-driven Deceptive Activities Framework (BUDA) se presenta como solución experimental orientada a mejorar la credibilidad y efectividad de las estrategias de ciber engaño mediante la gestión y personalización de *honey profiles*, entendidos como perfiles de personas ficticias similares a los usuarios reales. Esto implica incluir comportamientos en apariencia humanos en los entornos señuelo (o productivos, según la estrategia) aumentando la credibilidad y dificultando su detección por parte de los atacantes.

El diseño se basa en la posibilidad de generar perfiles ficticios personalizados que generan acciones sobre los sistemas objetivo de manera autónoma bajo un conjunto de reglas base definidas (contexto). Así, se simulan patrones típicos de usuarios que podrían ser legítimos en el entorno. Esta capacidad enriquece las narrativas de ciber engaño, lo que fortalece la resiliencia de las operaciones defensivas frente a adversarios avanzados. Esto además requiere introducirse en el campo de estudio de la ciber psicología, que integra las ciencias del comportamiento cibernético con entornos adaptativos para mejorar el ciber engaño, y además investiga e identifica lagunas que incluyen la falta de evaluación empírica y los efectos poco examinados de la cultura organizacional (Ferguson-Walter, Fugate, et al., 2019).

La herramienta pretende aportar una mejora sobre las limitaciones de las actuales soluciones, a través de cuatro ejes. Primero, la automatización de la simulación de rastros de usuarios, permitiendo crear y gestionar perfiles ficticios que imiten de manera pseudo realista el comportamiento humano. Segundo, la personalización y coherencia de una narrativa, adaptando estos perfiles a contextos específicos para garantizar interacciones plausibles dentro del entorno de engaño. Tercero, la escalabilidad, facilitando la generación simultánea de múltiples perfiles y su integración en diversos entornos y sectores. Por último, apoyar a la generación de actividades de ciber engaño alineadas con MITRE Engage, mapeando capacidades con las fases de planificación, ejecución y análisis propuestas por este marco

estratégico.

4.1 **Funcionalidades básicas**

A continuación, se presentan las funcionalidades básicas de la herramienta.

Gestión de narrativas. Las narrativas como parte de una operación de ciber engaño, definen los lineamientos sobre los cuales las actividades serán desarrolladas, y son determinadas por el equipo de ciberseguridad en base a los objetivos planificados para la operación. Posteriormente, cada actividad de los perfiles ficticios tenderá a reforzar la historia detrás del señuelo, token, o actividad de ciber engaño desplegada. Por ejemplo, si el objetivo es simular el uso de un servicio público conocido sobre el que podría realizarse una fuga de datos, los perfiles pueden mostrar interacción contra estos servicios, y dicha actividad guiará a los atacantes hacia la conclusión de que ese servicio es válido en el entorno, permitiendo la detección de esta TTP en el entorno. La planificación mediante el alineamiento con MITRE Engage ordena las capacidades en cada fase del marco estratégico, facilitando la definición de objetivos para las operaciones y la selección de perfiles simulados adecuados al contexto. Esta alineación busca una implementación más estructurada de las tácticas.

Creación de perfiles de usuarios ficticios. Se viabiliza la definición de usuarios ficticios contando con roles específicos y descripción de patrones de comportamiento. Estos perfiles se construyen a partir de datos de contexto reales o definiciones específicas del equipo de ciberseguridad, asegurando que cada usuario simulado sea coherente con el entorno en el que se despliegan los señuelos, o cualquier otra actividad de ciber engaño. Por ejemplo, un perfil falso de un analista de TI puede incluir actividades como acceso a sistemas de gestión de servidores, mientras que el de un empleado administrativo debería interactuar con una intranet o aplicaciones de ofimática. Esta flexibilidad en la distinción de perfiles permite adaptarlos a diferentes sectores industriales y escenarios operativos, logrando obtener sus cualidades correspondientes.

Automatización de generación de actividades. La herramienta genera y ejecuta acciones típicas de usuarios, según los criterios definidos, de manera automatizada y coherente con la narrativa, replicando por ejemplo conductas cotidianas, accesos a archivos, navegación web, consultas a bases de datos o uso de aplicaciones. Esta automatización reduce la carga operativa asociada con la interacción manual con los señuelos, y garantiza que las interacciones sean constantes y verosímiles. La simulación de actividades humanas busca replicar acciones cotidianas para aumentar la credibilidad de los perfiles ficticios. Para mejorar el realismo, se pueden implementan patrones de interacción temporal que simulan rutinas diarias con variaciones naturales, evitando patrones predecibles que podrían ser detectados por atacantes avanzados.

Personalización de comportamientos. Cada perfil generado puede configurarse para reflejar rutinas diarias específicas e interacciones con activos señuelo. Esto puede incluir horarios de trabajo, frecuencia de acceso a ciertos recursos, preferencias tecnológicas, y hábitos de comunicación. Al incorporar esta capa de personalización, se pretende tender a replicar dinámicas humanas complejas, lo que aumenta la credibilidad de los perfiles.

Reportes de actividad e interacción. Al monitorear y registrar de manera continua cómo los perfiles ficticios interactúan con el entorno, se pueden generar reportes que proporcionan información, sobre la interacción con el entorno permitiendo a los equipos de seguridad modificar los criterios y mejorar continuamente las operaciones de ciber engaño. Los registros detallados permiten la integración con herramientas de seguridad existentes, como sistemas SIEM o UEBA, logrando potenciar su capacidad operativa dentro del ecosistema de ciberseguridad preestablecido.

Variabilidad y realismo. Se incorpora la capacidad para introducir una variación basada en una definición de "porcentaje de similitud" sobre las acciones y comportamientos de los perfiles ficticios con relación al contexto y las definiciones realizadas para cada elemento. Esta función evita que los patrones de actividad sean predecibles, reduciendo el riesgo de que los atacantes detecten los señuelos como falsos, y permite poner a prueba los sistemas de detección de anomalías basados en comportamiento de usuarios. Este enfoque provee un caso de uso adicional para aportar valor a las operaciones de ciber engaño desplegadas.

Generación asistida de narrativas y perfiles. Esta función creación permite crear y configurar perfiles y narrativas a partir de un conjunto de definiciones de contexto. Luego, cada objeto puede ser configurado con atributos más detallados, incluyendo el perfil del atacante relacionado con la narrativa o los patrones base de los perfiles de usuarios ficticios, lo que pretende aumentar su realismo y credibilidad.

4.2 **Arquitectura Técnica**

La arquitectura de la herramienta está diseñada para ser modular, facilitando así su implementación y adaptación a diferentes entornos. A continuación, se presenta un diagrama de bloques de su estructura conceptual, continuando con los módulos de configuración principales que la componen.

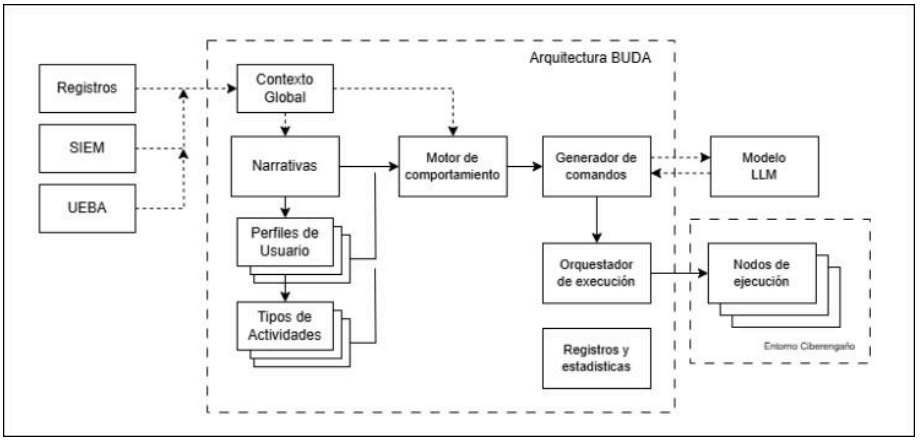

Fig. 1. Diagrama de bloques de la arquitectura conceptual

Módulo de Narrativas. Este módulo tiene como finalidad la creación y gestión de narrativas estratégicas que guíen las actividades de ciber engaño. Estas narrativas permiten estructurar las interacciones de los usuarios ficticios con los activos señuelo, alineando la simulación con objetivos específicos de detección, desviación o recolección de inteligencia sobre los atacantes. Así, se propone una integración con las operaciones de ciber engaño, asegurando que las interacciones simuladas sean coherentes y estratégicamente alineadas con los objetivos de seguridad. La capacidad para definir perfiles de atacantes, gestionar activos señuelo y asignar usuarios ficticios permite diseñar entornos realistas que dificultan la detección de los señuelos por parte de los adversarios. Además, una representación visual de las narrativas facilita el seguimiento y ajuste de las simulaciones a lo largo del tiempo, permitiendo a los operadores de ciberseguridad optimizar sus estrategias en función de la información recopilada. Este módulo está compuesto por tres elementos clave que se detallan a continuación.

Creador de narrativas. Al diseñar una narrativa se asigna un nombre descriptivo y único que facilite su identificación en el sistema. Los objetivos de la narrativa deben estar definidos para asegurar su alineación con las estrategias de ciberseguridad de la organización, y pueden incluir el propósito y alcance del escenario. Se define en este paso el "porcentaje de similaridad" para actividades ejecutadas dentro del contexto de esta narrativa donde 100% implica una ejecución supeditada completamente a los lineamientos y 0% una ejecución que no sigue el contexto ni lineamientos definidos.

Definiciones adicionales. Antes de proceder con la ejecución de una narrativa, es necesario definir los activos señuelo que conformarán el escenario, y establecer el perfil del atacante esperado. Esto se configura con atributos que modelen las acciones esperadas en el entorno, como la motivación (financiera, espionaje, hacktivismo, etc.), el nivel de habilidad (básico, intermedio o avanzado), y las tácticas esperadas (escaneo de redes, movimiento lateral, exfiltración de datos, etc.). Para facilitar la configuración se proporcionan plantillas basadas en tipos de actores de amenazas

conocidos. Las actividades de engaño pueden incluir archivos ficticios confeccionados a tal fin (documentos financieros, reportes estratégicos, etc.), servicios y sistemas simulados, y credenciales falsas creadas para atraer intentos de acceso maliciosos. Por último, la narrativa se relaciona con un nodo de ejecución sobre el que se ejecutarán las actividades diseñadas para dejar rastros sobre ellos.

Definición de límites temporales. Para garantizar el control operativo de la narrativa es necesario establecer una fecha de finalización de su ejecución. Este límite puede definirse de manera fija, con una fecha específica, permitiendo acotar la duración de la simulación, o dispararse ante una condición o serie de condiciones. La asignación de una fecha de cierre asegura que la actividad no se prolongue indefinidamente y facilita la planificación de futuras simulaciones.

Módulo perfiles de usuarios. Este módulo tiene como propósito la configuración de usuarios ficticios realistas que interactúen con los nodos de ejecución basados en las narrativas de ciber engaño. Se compone de cuatro elementos.

Creador de perfiles. Permite diseñar y personalizar los usuarios ficticios, para lo que se definen atributos básicos como el nombre del usuario, su rol en la organización (Gerente de Finanzas, Analista de TI, etc.). Además, cada perfil cuenta con atributos de comportamiento que incluyen los horarios laborales, especificando el inicio y fin de la jornada, almuerzo, pausas, descansos, etc. También se configuran las rutinas diarias, determinando actividades comunes como accesos a archivos, participación en reuniones virtuales y uso de herramientas internas. La frecuencia de ejecución de estas tareas puede ajustarse en niveles (bajo, medio o alto) de acuerdo con el rol y necesidades de la simulación. Para evitar la detección por patrones predecibles, se permite configurar un margen de aleatoriedad en horarios, accesos y actividades. Los usuarios pueden ser creados de forma manual o asistida mediante una consulta al módulo LLM integrado en la herramienta. Cada perfil puede ser asignado a una o más narrativas, para permitir la movilidad de su contexto a otros escenarios, como ocurre en la realidad, lo que facilita la adaptación de las narrativas a diferentes situaciones de amenaza y optimiza el tiempo de configuración.

Biblioteca de perfiles. Un repositorio dinámico que permite gestionar, reutilizar y modificar los perfiles existentes. Proporciona una vista organizada de todos los perfiles disponibles, con opciones de búsqueda por nombre, rol o tipo de actividad, facilitando su administración en escenarios complejos. Cada uno puede ser editado, duplicado o eliminado según las necesidades de la simulación, permitiendo una rápida adaptación a diferentes contextos operativos. Además, pueden exportarse en formatos estándar (JSON) lo que facilita su integración con otras herramientas. Se incluye un sistema de control de versiones, que mantiene un historial de cambios en cada perfil, y permite auditar modificaciones, garantizar la coherencia de las configuraciones y restaurar versiones previas.

Simulación de actividades. Automatiza las acciones de los usuarios ficticios, reforzando la autenticidad de los señuelos. Para esto se generan interacciones

definidas con los nodos de ejecución, incluyendo accesos frecuentes a archivos, bases de datos y sistemas internos. Asimismo, los perfiles pueden interactuar con herramientas corporativas (CRM, ERP, suites de ofimática, etc.) replicando patrones de uso comunes. Otras acciones pueden incluir actividades del sistema operativo como inicio y cierre de sesión, reinicios, y más.

Editor de variabilidad. En base al "porcentaje de similaridad" definido se ajusta la alineación o aleatorización de los comportamientos simulados, proponiendo de esta manera un desfasaje de la línea base. A fin de mejorar la credibilidad, el editor permite modificar los márgenes de las definiciones horarias introduciendo fluctuaciones en los tiempos de actividad, y ajustar el uso de aplicaciones definidas para asegurar la variabilidad en las interacciones.

Módulo de automatización de actividades. Este módulo tiene como finalidad ejecutar las acciones en los sistemas, basadas en las narrativas de ciber engaño y los perfiles creados. Está compuesto por cuatro elementos que se detallan a continuación.

Gestor de actividades. Permite administrar los tipos de actividades a ejecutar. La asignación para cada usuario definirá el marco sobre el que se generen las actividades, como accesos a archivos, consultas a motores de bases de datos o servicios, y uso de aplicaciones como navegadores web o herramientas de oficina. Estos tipos de actividades son consideradas durante la ejecución de las narrativas limitando las interacciones de los perfiles ficticios y simulando a un usuario real en su entorno laboral. Los accesos, modificaciones y consultas de información solo se realizan con sus credenciales y sobre los nodos de ejecución, sin impactar los sistemas fuera de la narrativa.

Generador de comandos. Traduce los tipos de actividades en comandos ejecutables para los nodos. Para esto, se utiliza una integración con un modelo de lenguaje mediante API (OpenAI) o mediante un despliegue local de plataformas para despliegue de modelos de lenguaje (Ollama o LM Studio), lo que permite la generación dinámica de instrucciones adaptadas. A fin de garantizar la usabilidad y confianza en las respuestas generadas, se emplean estrategias de formulación de prompts que minimizan ambigüedades y divergencia en la interpretación de las solicitudes, asegurando que los comandos sean consistentes con la narrativa. Esto se logra mediante una estructuración clara de las solicitudes, la inclusión de contexto relevante y la validación de las respuestas antes de su ejecución, evitando así acciones imprevistas o inconsistencias.

Panel de ejecución. Interfaz centralizada para la supervisión y el control de las actividades automatizadas en tiempo real. Ofrece una vista de las narrativas activas, junto a un desglose detallado de las acciones en curso, y los usuarios ficticios involucrados. Esto permite a los operadores monitorear la evolución de la simulación y realizar ajustes según sea necesario. También cuenta con un sistema de control manual, que permite detener o reiniciar actividades en cualquier momento, junto con la posibilidad de inyectar comandos adicionales a la operación en curso. Finalmente,

mantiene el historial de actividades mediante un registro de las acciones de los usuarios ficticios y los resultados de cada una. Esto permite la auditoría y posteriores análisis, como evaluar la efectividad del engaño, comprender patrones de interacción de los atacantes y mejorar la estrategia de simulación en futuras implementaciones.

Módulo de monitoreo y reportes. Este módulo permite supervisar y analizar las interacciones generadas por los usuarios ficticios, así como la respuesta de los atacantes. Su propósito es proporcionar visibilidad en tiempo real sobre las actividades en el entorno, y facilitar la recopilación de información para evaluar la efectividad de la simulación. Se compone de un panel de estadísticas que proporciona una vista centralizada y en tiempo real de las interacciones en el entorno de ejecución. Permite a los operadores supervisar las acciones de los usuarios ficticios, incluyendo accesos a archivos, uso de aplicaciones, envíos de correos electrónicos y cualquier otra actividad automatizada programada por diseño. Junto al panel de ejecución permiten la visualización completa de las actividades.

4.2.1 Consideraciones de Seguridad

Dado que la infraestructura de engaño puede convertirse en un objetivo, el diseño de debe incorporar medidas de seguridad específicas para evitar que la herramienta sea utilizada como vector de ataque. La comunicación entre el Generador de comandos y los nodos de ejecución se debe proteger mediante comunicación cifrada, asegurando la integridad de las instrucciones automatizadas. Asimismo, la gestión de identidades de los *honey profiles* se debe establecer bajo el principio de menor privilegio (PoLP). Si bien las credenciales de los usuarios ficticios son funcionales y permiten generar actividad verosímil (accesos, *logins*, consultas), sus permisos están limitados granularmente al alcance de la narrativa de engaño. Esto garantiza que, en el caso de que un adversario logre comprometer un perfil simulado, no dispondrá de privilegios administrativos ni capacidad de movimiento lateral efectivo hacia activos productivos críticos fuera del entorno de contención.

4.2.2 Comandos y Control de consistencia

Para garantizar la viabilidad técnica de las acciones simuladas y mitigar la generación de comandos erróneos o "alucinaciones", la herramienta implementa una estrategia de formulación de *prompts* estructurada y contextual. El modelo de lenguaje no recibe instrucciones aisladas, sino que opera bajo una definición de "System Persona" que inyecta dinámicamente las variables del *honey profile* activo (sistema operativo, rol del usuario, software instalado y privilegios). Para asegurar la continuidad narrativa, el *prompt* incluye una ventana de contexto de sobre la narrativa, amenaza y objetivos a cumplir, lo que reduce redundancias ilógicas (como abrir una aplicación que ya no corresponde con el usuario) y asegura una secuencia causal de acciones. A futuro la salida del modelo puede ser sometida a una etapa de saneamiento, validación sintáctica o esquemas de validación que utilicen LLM como Juez antes de su ejecución, de esta manera se descartaría cualquier comando destructivo o fuera de los

límites operativos definidos por la narrativa

4.3 **Implementación**

A continuación, se presentan los detalles para la implementación de la herramienta, incluyendo el proceso de instalación y despliegue, casos de uso básicos, una metodología para evaluación de su efectividad, y sus beneficios esperados. Si bien puede ser empleada de manera independiente y sin marco contextual estratégico, se sugiere la adopción de un enfoque metodológico para su uso (Pacheco & Staino, 2024) a fin de alcanzar resultados más útiles y funcionales. Esto permite abordar de forma estructurada la complejidad inherente a la aplicación de medidas de ciber engaño, manteniendo una perspectiva agnóstica respecto a las tecnologías utilizadas. Este aspecto resulta particularmente relevante, dado que, en principio, es factible implementar estrategias de ciber engaño sin recurrir a herramientas tecnológicamente avanzadas. El enfoque metodológico demanda mayor esfuerzo, pero ofrece mayor profundidad en términos de experiencia generada y conocimiento adquirido, al tiempo que optimiza la implementación de las estrategias, y contribuye a la generación de un marco de trabajo reproducible y escalable, adaptado a las necesidades de cada organización.

Despliegue. Dado que la herramienta está desarrollada en lenguaje Python, se creó para su implementación un paquete *pip*, lo que solo requiere ejecutar el comando: `$ pip install BUDA`. Previo a esto, es conveniente crear un entorno virtual para evitar problemas de dependencias, utilizando el siguiente comando: `$ python3 -m venv venv`.

Para la documentación del proyecto se utilizó *Read the Docs*, una plataforma de alojamiento de documentación de software gratuito de código abierto, que produce documentación escrita con el generador de documentación *Sphinx*. Dicha documentación puede encontrarse en la página: *https://budaframework.readthedocs.io*. Del mismo modo, el código fuente está publicado para su uso libre bajo licencia *GPLv3* en un repositorio de código abierto de la plataforma *GitHub*, que se puede encontrar en la dirección web indicada en las referencias (*BUDA (Behavioral User-driven Deceptive Activities)*, 2025).

Casos de uso. A continuación, se proponen tres casos de uso para la herramienta.

Desvío de ataques y detección temprana. Al generar actividades realistas sobre perfiles falsos en estaciones de trabajo productivas, es posible orientar a los atacantes hacia señuelos con los que realizar detección temprana de dichas actividades, desviado así sus esfuerzos de comprometer activos críticos.

Validación de sistemas de monitoreo de comportamiento. Al simular de forma automatizada interacciones que imitan el comportamiento de usuarios reales, según el porcentaje de similaridad definido, la herramienta puede permitir evaluar y calibrar soluciones de monitoreo (como SIEM o UEBA). Esto ayuda a afinar los parámetros

de detección y ajustar la sensibilidad frente a actividades anormales, mejorando así la capacidad de identificar acciones maliciosas y reducir falsos positivos.

Refinamiento de tácticas de ciber engaño. La herramienta puede utilizarse para experimentar con distintos patrones de interacción, variando la temporalidad y la coherencia de las acciones simuladas. Esto permite afinar las tácticas de engaño y ajustar la "personalidad" de los perfiles falsos, de modo que se adapten a las nuevas técnicas de los atacantes a través del tiempo, sin perder realismo.

Metodología de Evaluación. En tanto herramienta en fase experimental, su evaluación y validación requiere medir su impacto en las estrategias de ciber engaño, lo que se propone analizar desde tres perspectivas: la credibilidad de los honey profiles, es decir, si los perfiles simulados son percibidos como legítimos por los atacantes para minimizar que noten el engaño; la detección y desviación de ataques, midiendo su capacidad para identificar actividades maliciosas y redirigirlas hacia señuelos; y la automatización y escalabilidad, examinando la eficiencia en la generación y mantenimiento de múltiples perfiles sin intervención manual significativa. Para estos se sugiere implementar un entorno simulado con pruebas controladas antes de su despliegue en entornos reales, que incluya una infraestructura que imite una red corporativa con activos digitales, honey profiles que interactúen, y atacantes simulados (Red Team) que intenten comprometer la red. Los experimentos consistirán en analizar cómo la herramienta influye en las decisiones de los atacantes respecto a los señuelos.

Para medir el rendimiento se pueden definir métricas como la tasa de detección, como porcentaje de intentos de ataque identificados; el tiempo de permanencia del atacante, midiendo cuánto interactúan con los señuelos antes de detectar la trampa; y el nivel de interacción, analizando la frecuencia de interacción con perfiles ficticios versus con sistemas reales. Además, se puede evaluar la evasión de la detección a partir del número de atacantes que identifican los señuelos como falsos y la carga operativa para mantener la herramienta funcionando. La efectividad se puede evaluar mediante ataques controlados, donde el Red Team interactúa con activos señuelo y *honey profiles*, y comparando resultados con escenarios sin la herramienta para determinar su impacto en la detección de amenazas. Además, se puede evaluar la credibilidad de los perfiles generados y la calidad de las narrativas, y analizar la diferencia de efectividad entre señuelos estándar y los que integran honey profiles generados.

Beneficios esperados. La herramienta busca ofrecer una mejor cobertura defensiva, y aumentar la capacidad de recopilación de inteligencia de amenazas, sin imponer una carga operativa significativa, superando las limitaciones de los enfoques actuales. Así, se pueden esperar algunos beneficios, a saber. Primeramente, el aumento de la credibilidad de los señuelos, que dificultan su distinción de sistemas legítimos, y reforzando la autenticidad percibida. Segundo, la mejora en la detección de ataques, por la identificación de actividades sospechosas en fases tempranas al interactuar los atacantes con los perfiles y datos simulados, lo que motiva que sus TTPs sean detectadas antes de comprometer activos reales. Tercero, la escalabilidad y reducción

de costos operativos, ya que la automatización elimina la necesidad de crear manualmente señuelos realistas, sumado a la capacidad de integrarse con herramientas existentes, como SIEM y plataformas de inteligencia de amenazas.

4.3.1 Validación de Modelos de Lenguaje

Para la validación experimental de la herramienta y la generación de las pruebas de concepto presentadas, se utilizó el modelo GPT-4, seleccionado por su capacidad de razonamiento en el seguimiento de instrucciones y costo. Sin embargo, el desempeño de la herramienta no debe asumirse uniforme entre distintas arquitecturas de IA. Trabajos futuros deben extender esta evaluación a otros modelos de lenguaje, particularmente alternativas de código abierto y de ejecución local (como Llama 3 o Mistral), para establecer una comparativa rigurosa. Dicho análisis es fundamental para medir variaciones en la consistencia narrativa de los perfiles a largo plazo y la complejidad técnica de los comandos generados, determinando así la viabilidad operativa de modelos con menor cantidad de parámetros en entornos desconectados o con recursos limitados.

5 Limitaciones y trabajo futuro

Existen diversas limitaciones prácticas con este tipo de herramientas. En primer lugar, los atacantes pueden identificar patrones de actividad en los perfiles, y aplicar mecanismos de evasión (Ferguson-Walter, Major, et al., 2019) lo que requiere ser contrarrestado mejorando el motor de generación de actividades para evitar predictibilidad. Por otro lado, la autenticidad de los perfiles en el tiempo requiere consistencia en las interacciones y coherencia con el entorno, lo que puede mejorarse usando IA en otras etapas, con el desafío agregado del entrenamiento y optimización de los modelos. Finalmente, la integración con entornos productivos es en sí mismo un obstáculo, lo que exige pruebas para garantizar compatibilidad y seguridad ante la posible explotación de la propia herramienta. Además, las características de cada organización requieren otros ajustes para satisfacer las necesidades de cada caso, aumentando la complejidad operativa y de configuración. Estas limitaciones subrayan la importancia de un enfoque iterativo y flexible en el desarrollo y despliegue de la herramienta.

Ante estos desafíos se proponen líneas de investigación y desarrollo. Primero, la realización de pruebas en entornos productivos para evaluar efectividad contra ataques reales y recopilar datos sobre el desempeño de honey profiles en operaciones de ciber engaño. Segundo, la mejora en la simulación de comportamiento usando modelos de IA para generar perfiles adaptativos y menos predecibles. Tercero, optimizar el rendimiento y escalabilidad con una arquitectura distribuida para operar en grandes redes, con diferentes niveles de tráfico y actividad. Finalmente, explorar la integración con estrategias de respuesta a incidentes, para automatizar la identificación y mitigación de amenazas en tiempo real, vinculada a sistemas de alerta temprana y análisis forense.

6 Discusión y Conclusiones

Dado que la efectividad del ciber engaño depende en gran medida del despliegue de señuelos verosímiles, se propuso una herramienta que permite aumentar su credibilidad. La principal contribución de este trabajo es la propuesta de un modelo cuya arquitectura modular, más allá de la herramienta desarrollada, sienta las bases para nuevas herramientas que integren comportamiento humano automatizado en estrategias de defensa. Considerando que la resistencia de las organizaciones a incluir tecnologías de ciber engaño entre las medidas defensivas puede reducirse con implementaciones minimalistas y a baja escala, nuestra herramienta permite experimentar sin los costos de licenciamiento de herramientas del mercado, que además no cuentan con las funcionalidades presentadas. Así, este trabajo busca alentar el desarrollo de estrategias de ciber engaño independientes de productos comerciales, mediante metodologías simples y herramientas de código abierto, facilitando la adopción temprana de conocimientos sobre el tema.

En conclusión, la herramienta permite producir actividad automática con honey profiles para añadir realismo a la operación de ciber engaño, lo que incluye generar perfiles ficticios, simular actividades símil humanas, y analizar interacciones con atacantes en base a las conductas de los usuarios existentes. Esto se enmarca en una tendencia creciente en ciberseguridad, que es el aprovechamiento de la IA y automatización para mejorar la defensa activa.

Bibliografía

Aradi, Z., & Bánáti, A. (2025). The Role of Honeypots in Modern Cybersecurity Strategies. 2025 IEEE 23rd World Symposium on Applied Machine Intelligence and Informatics (SAMI)

BUDA (Behavioral User-driven Deceptive Activities). (2025). [Software]. https://github.com/Base4Security/BUDA

Ferguson-Walter, K., Fugate, S., Wang, C., & Patel, T. (2019). Introduction to the Cyber Deception and Cyberpsychology for Defense Minitrack.

Ferguson-Walter, K., Major, M., Van Bruggen, D., Fugate, S., & Gutzwiller, R. (2019). The World (of CTF) is Not Enough Data: Lessons Learned from a Cyber Deception Experiment. 2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC)

Gurtu, A., & Lim, D. (2025). Use of Artificial Intelligence (AI) in Cybersecurity. En Computer and Information Security Handbook (pp. 1617–1624). Elsevier.

Liebowitz, D., Nepal, S., Moore, K., Christopher, C. J., Kanhere, S. S., Nguyen, D., Timmer, R. C., Longland, M., & Rathakumar, K. (2021). Deception for Cyber Defence: Challenges and Opportunities. 2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), 173–182.

Morovitz, M., Raymond, G., Barr, S., & Anderson, L. (2022). MITRE Engage Framework. https://engage.mitre.org

Pacheco, F. (2022). Active cyber defense: Service model for defensive strategies based on the adversary's error. https://doi.org/10.36227/techrxiv.21268458.v1

Pacheco, F., & Staino, D. (2024). Propuesta para implementación de estrategias minimalistas de ciber engaño.

Shimeall, T. J., & Spring, J. M. (2014). Deception Strategies. En Introduction to Information Security (pp. 61–79). Elsevier. https://doi.org/10.1016/B978-1-59749-969-9.00004-3

Wani, M. A., Jabin, S., Yazdani, G., & Ahmadd, N. (2018). Sneak into Devil's Colony- A study of Fake Profiles in Online Social Networks and the Cyber Law (No. arXiv:1803.08810).

Zhang, L., & Thing, V. L. L. (2021). Three Decades of Deception Techniques in Active Cyber Defense—Retrospect and Outlook. Computers & Security, 106, 102288.