UAX Universidad Alfonso X el Sabio

Informe de conclusiones de la Jornada de trabajo sobre IA y ciberseguridad.

Proyecto PECIEE.

Dr. Jorge Calvo Martín

Ingeniero en Informática. Mención en Criptografía.
Coordinador Grado en Computación e Inteligencia Artificial.
Facultad Business & Tech.
Universidad Alfonso X el Sabio.

Financiado por la Unión Europea NextGenerationEU

GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES

Plan de Recuperación, Transformación y Resiliencia

incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD

Índice

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD

1. Introducción y contexto

El presente informe tiene por objeto recoger las principales conclusiones derivadas de la reunión de investigadores interuniversitarios en la Jornada de Trabajo sobre Inteligencia Artificial y Ciberseguridad, en el marco de la Actividad 11 del proyecto de investigación PECIEE (Proyecto Estratégico de Ciberseguridad: Investigación, Educación y Empresas), que tiene que objetivo impulsar la cultura de la ciberseguridad a través de la investigación, la educación y la colaboración con empresas, incluyendo acciones de divulgación dirigidas a la ciudadanía.

La Jornada de Trabajo sobre IA y Ciberseguridad se llevó a cabo los días 26 y 27 de enero de 2026 en la sede de la UAX en Madrid (Campus de Arapiles). El evento logró reunir y conectar a grupos de investigación expertos de la Universidad de Oviedo, Universidad Alfonso X el Sabio (UAX), Universidad de León, Universidad Complutense de Madrid (UCM), Universidad de Islas Baleares (UIB) y Universidad de A Coruña (UdC). En total se reunieron 25 participantes especializados.

La Jornada de Trabajo sobre IA y Ciberseguridad tuvo como objetivo principal presentar líneas de investigación activas y contrastadas en la intersección entre inteligencia artificial, análisis de datos y ciberseguridad, con especial atención a escenarios dinámicos y de alta complejidad, donde los enfoques tradicionales resultan insuficientes.

La jornada se diseñó en dos bloques diferenciados para fomentar tanto la divulgación como la profundización técnica: una sesión inaugural el primer día que incluyó una masterclass del Aula de Ciberseguridad y Criptografía Postcuántica, y propició la primera interacción entre los asistentes, y una segunda jornada dedicada íntegramente a sesiones técnicas de investigación sobre la intersección entre Inteligencia Artificial y Ciberseguridad. A diferencia de otros encuentros con foco empresarial, esta jornada tuvo un **carácter marcadamente investigador y técnico**, orientado a presentar líneas de investigación activas y contrastadas en la intersección entre inteligencia artificial, análisis de datos y ciberseguridad, estado del arte y retos futuros.

El programa abordó de forma transversal dispositivos móviles, detección avanzada de amenazas, protección de menores, análisis forense, tráfico de red y gobernanza del riesgo asociado a la IA, conectando investigación básica, aplicada y consideraciones éticas y sociales.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD

2. Ciencia de Datos y Servicios de IA para modelos de IA en ciberseguridad

Esta ponencia fue impartida por **Noemí de Castro García**, de la Universidad de León, y constituyó una de las intervenciones de mayor profundidad científica de la Jornada de Trabajo sobre Inteligencia Artificial y Ciberseguridad, apoyándose directamente en líneas de investigación consolidadas y publicaciones recientes de alto impacto centradas en la aplicación de técnicas avanzadas de *machine learning* y *data science* a problemas reales de ciberseguridad.

Su intervención se estructuró en torno a un problema ampliamente reconocido en la literatura: la **detección de malware en entornos dinámicos**, donde las amenazas evolucionan de forma constante y los modelos de aprendizaje automático tradicionales pierden eficacia con el paso del tiempo. En este contexto, Noemí disertó sobre el fenómeno conocido como *concept drift,* definido como el cambio en la distribución de los datos a lo largo del tiempo, debido, entre otros factores, a la aparición de nuevas familias de malware, modificaciones en el comportamiento de las ya existentes o cambios en el entorno de ejecución.

El núcleo técnico de la ponencia se apoyaba en varias publicaciones relacionadas, como en el artículo *Application of Transfer Learning to Online Models in Malware Detection*, donde se analiza críticamente la utilización de *transfer learning* como estrategia para mejorar el rendimiento de los modelos de detección de malware en escenarios afectados por *concept drift*.

Tradicionalmente, los enfoques basados en *machine learning* han demostrado ser más adaptativos que los sistemas basados en firmas, pero presentan una dependencia crítica de datos etiquetados. En la práctica, obtener etiquetas fiables y actualizadas en ciberseguridad resulta costoso, lento y, en muchos casos, inviable, especialmente cuando se trata de muestras nuevas o de malware poco conocido. Para paliar este problema, se han propuesto modelos online, capaces de actualizarse dinámicamente conforme llegan nuevos datos. Sin embargo, estos modelos siguen requiriendo un flujo continuo de etiquetas para mantener su fiabilidad.

Es en este punto donde el **transfer learning** se presenta como una alternativa prometedora. La idea fundamental consiste en aprovechar el conocimiento aprendido en conjuntos de datos históricos o relacionados (origen) para mejorar el rendimiento del modelo en un nuevo conjunto de datos (destino), sin necesidad de disponer de etiquetas en este último. En teoría, este enfoque permitiría mejorar la detección de **nuevas familias de malware** reutilizando patrones aprendidos en el pasado.

No obstante, uno de los aportes más relevantes del trabajo es su análisis crítico de los resultados obtenidos. Los experimentos muestran que los algoritmos de *transfer learning*

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD

producen resultados inconsistentes: aunque en determinados escenarios se observan mejoras significativas en el rendimiento, estas no se mantienen de forma generalizada entre distintos modelos, algoritmos o configuraciones experimentales. Esta conclusión resulta especialmente relevante, ya que introduce un contrapunto necesario frente a una visión excesivamente optimista del *transfer learning* en ciberseguridad.

La ponencia destacó, por tanto, que no existe una solución universal, y que la eficacia del *transfer learning* depende en gran medida de la relación real entre los dominios de origen y destino, así como de la naturaleza concreta del *concept drift* presente en los datos. Este enfoque riguroso y basado en evidencia experimental aporta valor al debate científico y evita la adopción acrítica de técnicas de moda.

Una extensión natural de esta línea de investigación, también abordada en la intervención, es la necesidad de **gestionar la incertidumbre asociada a las predicciones de los modelos de IA**. En trabajos posteriores, como *Conformal prediction for labelling and updating online models in the presence of concept drift in cybersecurity*, la autora ha explorado el uso de **predicción conforme** como mecanismo para dotar a los modelos de una estimación cuantitativa de la confianza de sus decisiones.

Este aspecto es especialmente crítico en ciberseguridad, donde una decisión errónea puede derivar en **falsos positivos** costosos o, peor aún, en la **no detección de amenazas reales**. Incorporar medidas de incertidumbre permite diseñar sistemas más robustos, que combinen la automatización con la supervisión humana, alineándose con enfoques de IA responsable y confiable.

La ponencia también conectó estas investigaciones con trabajos más recientes sobre **IA generativa aplicada a la gestión de alertas tempranas** de ciberseguridad, como el artículo *Generative Artificial Intelligence and Machine Translators in Spanish Translation of Early Vulnerability Cybersecurity Alerts*, donde se analiza cómo los modelos de traducción automática y sistemas generativos pueden mejorar la difusión temprana de alertas de vulnerabilidades, especialmente en contextos multilingües. Sin embargo, es imprescindible ser cautelosos, puesto que, aunque la IA generativa puede acelerar la comunicación y reducir barreras idiomáticas, también introduce riesgos asociados a errores semánticos, ambigüedad o pérdida de precisión técnica, que pueden tener consecuencias relevantes en la gestión de incidentes.

La ponencia deja de manifiesto la importancia de modelos adaptativos y conscientes del cambio, la necesidad de evaluar la fiabilidad y la incertidumbre de las predicciones y el valor de combinar técnicas avanzadas de IA con criterios de gobernanza, supervisión y responsabilidad.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD

3. Dimensión social y ética de la IA en ciberseguridad

La protección de los menores en el entorno digital fue el eje central de la tercera ponencia, que aportó una dimensión social y ética esencial a la jornada. **Ana Lucila Sandoval** Orozco, investigadora postdoctoral de la UCM presentó soluciones basadas en IA orientadas a detectar y mitigar riesgos específicos que afectan a los menores en el ciberespacio.

Se abordaron problemáticas como el ciberacoso, la exposición a contenidos inapropiados, la manipulación emocional y otras formas de violencia digital. La ponencia mostró cómo técnicas de inteligencia artificial, en particular el procesamiento del lenguaje natural y el análisis de patrones de comportamiento, pueden utilizarse para identificar situaciones de riesgo de forma temprana.

En la ponencia se habló del proyecto HEROES (Holistic Empowerment to counter Online abuse of children and adolescents) y la continuación en su línea de investigación ALUNA, una iniciativa internacional de gran envergadura coordinada por la Universidad Complutense de Madrid (UCM), cuyo objetivo principal es proteger a menores en el ciberespacio frente al abuso sexual infantil y la trata de personas mediante el uso de inteligencia artificial (IA) y tecnologías avanzadas de ciberseguridad.

ALUNA parte de una constatación clave: el incremento sostenido de los delitos de abuso y explotación sexual infantil, intensificado por la digitalización acelerada y el uso masivo de plataformas online, exige nuevas estrategias que combinen tecnología avanzada, cooperación internacional y atención prioritaria a las víctimas. En este contexto, el proyecto adopta un enfoque *child-centred*, en el que las decisiones tecnológicas, legales y operativas se subordinan a la protección, el bienestar y la dignidad de niños, niñas y adolescentes, alineándose con los principios europeos de derechos fundamentales y protección de la infancia

El diseño conceptual de ALUNA se estructura en torno a tres pilares estrechamente interrelacionados. En el ámbito de la prevención, la investigación se orienta a identificar factores de riesgo digitales y sociales que facilitan la captación de menores por parte de redes delictivas. Para ello, el proyecto desarrolla metodologías de análisis de comportamiento online, herramientas de concienciación digital y estrategias de detección temprana que permiten anticipar situaciones de riesgo en plataformas de comunicación, redes sociales y entornos digitales de alta exposición.

En el pilar de la investigación criminal, ALUNA profundiza en el uso de inteligencia artificial, análisis forense digital y explotación de datos para fortalecer la actuación de las fuerzas y cuerpos de seguridad (LEAs). Las líneas tecnológicas incluyen la detección automática de patrones delictivos, la clasificación inteligente de evidencias digitales y el apoyo a investigaciones transnacionales, un aspecto crítico dado el carácter global de estos delitos. Estas soluciones se diseñan en colaboración directa con policías nacionales

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD

e internacionales y organizaciones especializadas, lo que garantiza su aplicabilidad operativa y su adecuación a los marcos legales de distintos países.

El tercer pilar, la asistencia a las víctimas, representa uno de los elementos más distintivos de ALUNA. La línea de investigación integra conocimientos de psicología, trabajo social, derecho y criminología para desarrollar protocolos y herramientas que reduzcan la revictimización, faciliten la denuncia y mejoren la atención posterior al rescate. El objetivo es que la tecnología no solo sirva para perseguir delitos, sino también para humanizar los procesos de apoyo, mejorar la coordinación entre instituciones y asegurar que las necesidades emocionales, sociales y legales de los menores sean atendidas de forma prioritaria.

ALUNA presta especial atención a la armonización de prácticas y marcos legales, promoviendo la cooperación entre países y la transferencia de conocimiento entre investigación académica y práctica policial. Esta dimensión resulta clave para mejorar la eficacia de las investigaciones y para trasladar los resultados del proyecto a políticas públicas y protocolos institucionales, tanto a nivel nacional como europeo.

En conjunto, la línea ALUNA demuestra que la lucha contra el abuso y la explotación sexual infantil requiere **soluciones tecnológicas avanzadas integradas en un marco ético, legal y social sólido**, donde la inteligencia artificial actúe como herramienta de apoyo y no como fin en sí misma. Su enfoque holístico y centrado en la infancia ofrece un modelo sostenible para afrontar uno de los mayores desafíos de la ciberseguridad contemporánea.

Además, el proyecto ha desarrollado **aplicaciones de código abierto y guías prácticas** orientadas a aumentar la seguridad digital de los niños y adolescentes, así como a concienciar a familias, educadores y profesionales sobre los riesgos existentes en el ciberespacio.

Como conclusión, la ponencia de Ana Lucila Sandoval Orozco puso de manifiesto el potencial de la inteligencia artificial para mejorar la protección de menores en el ciberespacio, siempre que su uso se realice con criterios de rigor técnico, responsabilidad ética y coherencia legal. Las publicaciones asociadas a la intervención ilustran cómo es posible combinar investigación académica, desarrollo tecnológico y análisis social para abordar uno de los retos más delicados y relevantes de la ciberseguridad actual.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD

4. Aplicaciones de la IA en Ciberseguridad: gestión de tráfico de red e identificación de artefactos forenses

La ponencia impartida por **José Manuel Vázquez Naya**, investigador del CITIC (Centro de Investigación en Tecnologías de la Información y las Comunicaciones) de la Universidad de A Coruña, aportó a la Jornada de Trabajo sobre Inteligencia Artificial y Ciberseguridad una sólida visión investigadora, metodológica y de transferencia tecnológica, centrada en la aplicación avanzada de la IA al análisis de tráfico de red y a la informática forense digital.

Durante su intervención, José Manuel presentó los avances más recientes desarrollados por su equipo en el CITIC, articulados en torno a una idea central: la necesidad de **modelos capaces de aprender representaciones generales del tráfico de red**, de forma análoga a lo que los *foundation models* han supuesto en ámbitos como el procesamiento del lenguaje natural o la visión artificial. Esta visión se encuentra plenamente respaldada por el artículo recientemente publicado en la revista Computer Networks, titulado *"Network Traffic Foundation Models: A systematic review"* (2026), en el que Vázquez Naya participa como coautor. Este trabajo constituye una revisión sistemática de referencia sobre el emergente paradigma de los *Network Traffic Foundation Models* (NT-FMs), y proporciona el marco conceptual que da coherencia a muchas de las líneas expuestas durante la ponencia.

En dicho artículo se plantea que el modelo tradicional de aprendizaje automático aplicado al tráfico de red (basado en entrenar modelos específicos para cada tarea, entorno o conjunto de datos) presenta importantes limitaciones en términos de generalización, escalabilidad y coste de mantenimiento. Frente a ello, los NT-FMs proponen un enfoque de *train-once, adapt-anywhere*, en el que un modelo se preentrena sobre grandes volúmenes de tráfico sin etiquetar, capturando patrones fundamentales del comportamiento de red, para posteriormente adaptarse a múltiples tareas de seguridad y gestión con cantidades mínimas de datos adicionales.

Esta idea se reflejó en la ponencia al abordar la evolución de las técnicas de descubrimiento de activos de sistemas operativos. El investigador explicó cómo su grupo ha progresado desde algoritmos clásicos de *machine learning*, concretados en la herramienta *open-source* fingerAI, hacia modelos más avanzados que aprenden directamente de representaciones ricas del tráfico. Este tránsito conceptual se alinea con la transición descrita en el artículo, desde enfoques basados en características estadísticas hacia modelos end-to-end entrenados sobre datos crudos de tráfico.

Un avance especialmente relevante presentado en la ponencia fue la adopción de **arquitecturas tabulares basadas en Transformers** para el análisis de tráfico de red, una línea que representa la traslación directa de técnicas exitosas en NLP al dominio de las comunicaciones. Este enfoque demuestra cómo los Transformers pueden aprender

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD

"semántica de tráfico" a partir de secuencias de paquetes, flujos y campos protocolarios, sin necesidad de ingeniería manual de características.

La ponencia alcanzó su punto más innovador con la presentación del proyecto NetHermes, concebido como una evolución natural del paradigma de los NT-FMs descrito en *Computer Networks*. NetHermes se orienta a transformar los Centros de Operaciones de Seguridad (SOC) mediante el desarrollo de un modelo multimodal de tráfico de red y lenguaje natural, capaz de procesar de manera conjunta paquetes de red y consultas expresadas en lenguaje natural. Este enfoque conecta directamente con la aparición de *copilotos de red* basados en modelos fundacionales, que permitan a los analistas interactuar con datos de tráfico complejos mediante lenguaje natural, reduciendo la dependencia de interfaces rígidas y consultas manuales complejas. La ponencia dejó claro que NetHermes no persigue únicamente mejorar la detección automática, sino facilitar la comprensión, exploración y toma de decisiones en entornos SOC altamente exigentes.

En conjunto, la ponencia de José Manuel Vázquez Naya mostró cómo la investigación en inteligencia artificial aplicada a ciberseguridad está transitando desde soluciones puntuales hacia **modelos fundacionales de propósito general**, capaces de aprender representaciones profundas del tráfico de red y reutilizarlas en múltiples contextos.

5. Impacto del desarrollo de la IA en la ciberseguridad

Esta ponencia fue realizada por **José Manuel Matalobos Veiga**, investigador de la Universidad Alfonso X el Sabio y ofreció una visión e integradora del impacto de la inteligencia artificial en la ciberseguridad.

En una primera parte de la ponencia, Juan Manuel realizó un recorrido sintético pero clarificador por la evolución histórica del concepto de inteligencia artificial, desde sus orígenes teóricos y simbólicos hasta los enfoques actuales basados en datos y aprendizaje automático. Este recorrido permitió contextualizar el actual auge de la **inteligencia artificial generativa**, que representa un punto de inflexión tanto tecnológico como social. El ponente destacó que modelos generativos de gran escala han ampliado de forma radical las capacidades de la IA, permitiendo la generación de texto, código, imágenes o contenido multimedia con un nivel de calidad sin precedentes. No obstante, se subrayó que este salto cualitativo no debe interpretarse únicamente en términos de oportunidades, sino también de nuevos riesgos y desafíos.

En este sentido, la ponencia dedicó una atención especial a los retos actuales y futuros de la inteligencia artificial, analizando de forma estructurada cuestiones clave como la

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD

aplicabilidad real de los sistemas de IA, la precisión y fiabilidad de los resultados, y las limitaciones inherentes a modelos entrenados sobre grandes volúmenes de datos heterogéneos. Se señaló que, si bien la IA generativa ofrece grandes ventajas en términos de automatización y eficiencia, sus resultados no están exentos de errores, alucinaciones o inconsistencias que pueden ser especialmente problemáticas en contextos críticos como la ciberseguridad.

Otro tema importante a resolver en el futuro son los problemas de propiedad intelectual asociados a los resultados generados por sistemas de IA. El ponente destacó la complejidad actual de determinar la autoría y los derechos sobre contenidos producidos de forma automática, así como los conflictos derivados del uso de datos de entrenamiento protegidos por derechos de autor. Estas cuestiones adquieren una especial relevancia en entornos profesionales y académicos, donde la trazabilidad y la legitimidad de los resultados son fundamentales.

Asimismo, se analizaron los sesgos presentes en los sistemas de inteligencia artificial, derivados tanto de los datos utilizados para su entrenamiento como de las decisiones de diseño de los modelos. En estrecha relación con ello, se abordó el problema de los contenidos nocivos, incluyendo desinformación, discursos de odio o generación de material potencialmente ilegal, que puede verse facilitado por el uso no controlado de sistemas generativos.

La reflexión se completó con un análisis del marco regulatorio emergente, destacando la necesidad de avanzar hacia modelos de regulación y supervisión que equilibren innovación y protección de derechos. En este contexto, se enfatizó la importancia de la supervisión humana, la auditoría de sistemas de IA y la transparencia en su funcionamiento. Finalmente, se abordó el consumo energético asociado a los grandes modelos de IA, señalando que la sostenibilidad se perfila como uno de los retos estratégicos clave en el desarrollo futuro de estas tecnologías.

Tras este marco general, la ponencia se centró específicamente en el **impacto de la inteligencia artificial en la ciberseguridad**, articulado en torno a tres grandes ejes claramente diferenciados.

El primer eje abordado fue el de los **ataques basados en inteligencia artificial**. El ponente explicó cómo los actores maliciosos están comenzando a incorporar técnicas de IA para automatizar, escalar y sofisticar sus ataques. Ejemplos de esta tendencia incluyen el uso de IA generativa para campañas de *phishing* altamente personalizadas, la creación automática de código malicioso o la optimización dinámica de ataques en función del comportamiento de las víctimas. Este uso ofensivo de la IA reduce las barreras de entrada a ataques complejos y plantea nuevos desafíos para los mecanismos de detección tradicionales.

El segundo eje analizado fue el del **incremento de la superficie de ataque asociada a la propia inteligencia artificial**. Los sistemas de IA introducen nuevos vectores de ataque

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD

específicos, como la manipulación de datasets de entrenamiento o la explotación de vulnerabilidades en modelos de datos. En este sentido, se enfatizó que proteger sistemas modernos implica no solo defender infraestructuras y aplicaciones, sino también asegurar los modelos de IA que las sustentan, así como los datos y procesos asociados a su ciclo de vida.

El tercer y último eje se centró en la **ciberdefensa basada en inteligencia artificial**. La ponencia destacó el enorme potencial de la IA para mejorar la detección de amenazas, la correlación de eventos, la respuesta ante incidentes y la gestión de la seguridad en entornos complejos. Se mencionó el avance hacia Centros de Operaciones de Seguridad más inteligentes y asistidos por IA, capaces de priorizar alertas, reducir falsos positivos y apoyar la toma de decisiones. No obstante, se subrayó que esta ciberdefensa basada en IA debe diseñarse con criterios de robustez, explicabilidad y control humano, evitando una automatización acrítica que pueda introducir nuevos riesgos.

Como conclusión, la ponencia ofreció una **visión estructurada y equilibrada del papel de la inteligencia artificial en la ciberseguridad**, integrando una reflexión histórica, tecnológica y estratégica. Su intervención permitió contextualizar las aportaciones técnicas de la jornada dentro de un marco más amplio, destacando que el futuro de la ciberseguridad estará inevitablemente ligado al desarrollo responsable, regulado y supervisado de la inteligencia artificial.

6. Conclusiones

La Jornada de Trabajo sobre Inteligencia Artificial y Ciberseguridad ha permitido compartir un conjunto de líneas de investigación prioritarias que emergen de manera transversal a las distintas ponencias.

Una primera conclusión relevante es la constatación de que la IA aplicada a ciberseguridad debe abordarse como un problema científico de sistemas complejos, donde confluyen aspectos algorítmicos, estadísticos, operativos, legales y sociales. Las intervenciones han mostrado que los avances más sólidos no provienen de la aplicación aislada de técnicas de moda, sino de enfoques rigurosos que analizan de forma crítica sus supuestos, limitaciones y condiciones de aplicabilidad. Esto abre una línea de investigación clara en torno a la evaluación científica de modelos de IA en escenarios reales y no estacionarios, superando benchmarks estáticos y proponiendo métricas que incorporen adaptación, robustez e incertidumbre.

En este sentido, la problemática del cambio en la distribución de los datos (*concept drift*) emerge como uno de los ejes centrales de investigación. Las contribuciones han puesto de manifiesto que ni el aprendizaje online ni el *transfer learning* constituyen soluciones universales, y que su eficacia depende de la relación estructural entre dominios, del tipo de deriva y del contexto operativo. De ello se deriva una línea de trabajo prioritaria en la

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD

caracterización formal del *concept drift* en ciberseguridad, así como en el diseño de mecanismos híbridos que combinen adaptación automática, estimación de incertidumbre y supervisión humana. La integración de técnicas como la predicción conforme apunta hacia modelos más confiables y científicamente interpretables.

Otra conclusión de fuerte calado investigador es la necesidad de avanzar hacia modelos fundacionales específicos del dominio de la ciberseguridad, en particular en el análisis de tráfico de red. La transición desde soluciones puntuales hacia arquitecturas capaces de aprender representaciones generales del comportamiento de red plantea retos abiertos en términos de escalabilidad, disponibilidad de datos, transferencia entre tareas y evaluación de la generalización real. Esto abre una agenda de investigación orientada a los Network Traffic Foundation Models, que incluye la definición de esquemas de representación del tráfico, protocolos de preentrenamiento, y marcos experimentales que eviten la reutilización circular de los mismos conjuntos de datos.

Desde una perspectiva más amplia, la jornada ha subrayado que la inteligencia artificial introduce nuevas clases de vulnerabilidades y superficies de ataque, lo que sitúa la seguridad de los propios sistemas de IA como un objeto de estudio prioritario. El análisis de ataques basados en IA, ataques contra modelos (envenenamiento de datos, evasión, manipulación adversaria) y la protección del ciclo de vida completo de los sistemas inteligentes se configura como una línea de investigación emergente, aún poco sistematizada, que requiere modelos teóricos, experimentación controlada y colaboración interdisciplinar.

La dimensión social y ética abordada en la jornada, especialmente en relación con la protección de menores, permite extraer una conclusión clave desde el punto de vista investigador: la ciberseguridad centrada en la persona no es un complemento, sino un eje estructural de la investigación en IA aplicada. Proyectos como ALUNA muestran que la integración de inteligencia artificial en contextos sensibles exige marcos metodológicos que incorporen principios *child-centred*, evaluación de impacto ético, y articulación entre tecnología, derecho y ciencias sociales. Esto define una línea de investigación clara en IA responsable aplicada a ciberseguridad social, donde los criterios de éxito no se limitan a métricas técnicas, sino que incluyen reducción del daño, protección de derechos y mejora de la intervención institucional.

Finalmente, la reflexión sobre la evolución histórica de la IA y el auge de la IA generativa ha permitido situar la investigación en ciberseguridad dentro de un escenario de transformación tecnológica acelerada, marcado por retos aún abiertos: fiabilidad de resultados, sesgos, propiedad intelectual, regulación y sostenibilidad energética. Desde esta perspectiva, se identifica como línea estratégica la investigación en gobernanza de sistemas de IA en ciberseguridad, que combine modelos técnicos, regulación emergente, auditoría algorítmica y mecanismos de supervisión humana.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD

En conjunto, la jornada no solo ha presentado resultados relevantes, sino que ha contribuido a compartir líneas de investigación futura en inteligencia artificial y ciberseguridad. La colaboración interuniversitaria evidenciada en este encuentro se perfila como un elemento clave para abordar estos retos desde una perspectiva verdaderamente interdisciplinar y de largo alcance.

Bibliografía

De Castro García, N., & Escudero García, D. (2025).
Application of transfer learning to online models in malware detection. En *Lecture Notes in Computer Science* (Springer)

Escudero García, D., & De Castro García, N. (2025).
Conformal prediction for labelling and updating online models in the presence of concept drift in cybersecurity. *Journal of Information Security and Applications, 93*.

Pérez-Jove, R., Munteanu, C. R., Dorado, J., Pazos, A., & Vázquez-Naya, J. (2026). *Network traffic foundation models: A systematic review*. *Computer Networks, 276*, 111998. Elsevier.

Román Martínez, J., Triana Robles, D., El Oualidi Charchmi, M., Salamanca Estévez, I., & De Castro García, N. (2025). Generative artificial intelligence and machine translators in Spanish translation of early vulnerability cybersecurity alerts. *Applied Sciences, 15*(8), 1–18.

Sandoval Orozco, A. L., et al. (2024). *HEROES project: Integrated technological and social approaches to combat child sexual abuse and trafficking in human beings*. Universidad Complutense de Madrid.

Vázquez-Naya, J., Pérez-Jove, R., Dorado, J., Pazos, A., & Munteanu, C. R. (en prensa). *Tabular transformer-based fingerprinting for network traffic analysis*. *Cybersecurity*.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD