Hackear con un *prompt*
Amenazas, defensas y lo que el Derecho aún no nombra

Ricardo J. Rodríguez

rjrodriguez@unizar.es ※ @RicardoJRdez ※ www.ricardojrodriguez.es

Universidad Zaragoza

Dpto. de Informática e Ingeniería de Sistemas
Universidad de Zaragoza

14 de septiembre, 2026
Jornada Inauguración Máster Europeo EMILDAI
Facultad de Derecho, Universidad de León

$whoami

- **Profesor Titular @ UNIZAR**
- **Director Cátedra Inetum de Ciberseguridad y Cloud**
- **Líneas de investigación**:
  - Análisis de software y binarios
  - Forensia digital
  - Seguridad de sistemas (IoT, ICS, cloud)
  - ML aplicado a ciberseguridad

Universidad Zaragoza

$whoami

- **Profesor Titular @ UNIZAR**
- **Director Cátedra Inetum de Ciberseguridad y Cloud**
- **Líneas de investigación**:
  - Análisis de software y binarios
  - Forensia digital
  - Seguridad de sistemas (IoT, ICS, cloud)
  - ML aplicado a ciberseguridad
- **Equipo de investigación** – *¡hacemos cosas chulas!* 🙂
  - https://reversea.me
  - https://twitter.com/reverseame/
  - https://t.me/reverseame

Universidad Zaragoza 1474

“Hace pocos años, automatizar un ataque sofisticado requería años de experiencia. Hoy basta con un prompt bien construido.”

Datos recientes: OpenAI, Anthropic y Google han documentado en sus informes de 2024-2025 abusos reales de sus modelos por parte de actores estatales y criminales

¿Estamos preparados (técnica *y* jurídicamente) para defender lo que ya está siendo atacado con IA?

Un caso real (no una hipótesis)

Espionaje orquestado por IA, septiembre de 2025

Informe de Anthropic (nov. 2025):

- Un actor estatal usó un agente de IA (*Claude Code*) contra una treintena de organizaciones: tecnológicas, financieras, químicas, gubernamentales
- Reconocimiento, explotación, movimiento lateral, exfiltración: **el 80–90 % lo ejecutó la IA**
- El humano intervino en 4–6 puntos de decisión por campaña
- Los filtros del modelo se sortearon **troceando** la tarea y fingiendo ser una empresa de *pentesting*

Y unos meses antes. . .

Un único individuo, sin conocimientos avanzados, extorsionó a 17 organizaciones con el mismo tipo de agente: el modelo elegía qué robar, calculaba el rescate y redactaba las notas

¿Quién es autor?
¿Quién, cómplice?
¿Y el proveedor del modelo?

Fuentes:
Anthropic. *Disrupting the first reported AI-orchestrated cyber espionage campaign*, nov. 2025; *Threat Intelligence Report*, ago. 2025. Informes análogos: OpenAI (oct. 2024, jun. 2025), Google GTIG (ene. y nov. 2025).

Agenda

1. Una doble inflexión: IA y ciberseguridad
2. IA como amenaza
3. IA como defensa
4. Donde la técnica va por delante de la norma
5. Conclusiones

Universidad Zaragoza
1474

Agenda

Universidad Zaragoza

Una doble inflexión

Por qué este tema, y por qué ahora

La ciberseguridad ya era difícil:

- Superficie de ataque que no para de crecer (cloud, IoT, OT)
- Cadena de suministro opaca
- Escasez crónica de talento

La IA generativa cambia el equilibrio:

- Reduce barrera técnica del atacante
- Acelera la detección y la respuesta
- Introduce nuevos activos a proteger (modelos, datos)

La pregunta

Si la técnica cambia a esta velocidad, ¿puede el Derecho regularla, juzgarla y probarla al mismo ritmo?

Spoiler: hacen falta perfiles que hablen **ambos idiomas**

Universidad Zaragoza

Una doble inflexión

El mapa del territorio (visto desde la trinchera técnica)

Análisis de binarios y malware	Forensia digital (memoria/disco)	ML adversarial y defensivo
Seguridad ICS e infraestructuras	Cloud y APIs (GDPR, BOLA)	IoT, NFC, dispositivos

- Cada una de estas casillas tiene hoy una **pregunta jurídica abierta**: autoría, prueba, responsabilidad, jurisdicción
- Lo que sigue: tres frentes donde la IA está cambiando las reglas, contados desde la investigación

Universidad Zaragoza

Agenda

Universidad Zaragoza

IA como amenaza

Caso 1: dominios maliciosos generados por IA

Contexto

El malware moderno no se conecta a una IP fija: usa miles de dominios generados algorítmicamente (DGAs) cada día para hablar con su servidor

- Históricamente: las DGAs eran detectables por su patrón "poco humano" (`xkjqwphq.com`)
- Defensa típica: clasificadores ML entrenados sobre estos patrones
- **Hoy:** los LLMs generan dominios indistinguibles de los reales

Lo que estamos investigando

Cómo **atacar** y cómo **defender** clasificadores de DGA en presencia de LLMs adversariales

Lecturas adicionales:

Pelayo-Benedet, Rodríguez, Gañán. *The Machines are Watching: LLMs for DGA detection*. JISA 2025 (Q2).

Pelayo-Benedet, Rodríguez, Gañán. *RAMPAGE: framework para reproducibilidad en detección de DGAs*. ESWA 2025 (Q1).

IA como amenaza

Caso 1: ¿cuáles de estos dominios los ha generado un algoritmo?

Dominio	¿Generado?	Cómo
xkjqwphqzmv.com		
secure-invoice-review.com		
heavengoodness.net		
nordic-cloudbackup.net		
qwvmrhngxlbmqf.biz		
medialoginhub-support.com		

Ejemplos ilustrativos; los dominios no apuntan a infraestructura real conocida

Universidad Zaragoza

IA como amenaza

Caso 1: ¿cuáles de estos dominios los ha generado un algoritmo?

Dominio	¿Generado?	Cómo
`xkjqwphqzmv.com`	✔	DGA clásica (aleatoria)
`secure-invoice-review.com`	✔	**LLM**
`heavengoodness.net`	✔	DGA de diccionario
`nordic-cloudbackup.net`	✔	**LLM**
`qwvmrhngxlbmqf.biz`	✔	DGA clásica (aleatoria)
`medialoginhub-support.com`	✔	**LLM**

Todos.

Los tres “normales” los ha generado un LLM en menos de un segundo. Un clasificador entrenado con dominios “raros” como el primero o el quinto **no los ve** – y un ser humano, tampoco

Ejemplos ilustrativos; los dominios no apuntan a infraestructura real conocida

Universidad Zaragoza

IA como amenaza

No solo dominios... también código y comportamiento

- El malware de hoy se ofusca, se polimorfiza, y crecientemente se **genera**
- Caracterizar el panorama real (familias, técnicas, plataformas) sigue siendo trabajo de campo
- Nuestro grupo lo ha mapeado en **macOS**, en **IoT** y en entornos **Windows**

La pregunta que esto le deja al Derecho

Si el código lo escribió un modelo a partir de un *prompt*, ¿quién es el autor del delito? ¿Quién responde: quien pide, quien despliega, quien entrena?

Lecturas adicionales:
Lastanao Miró, Carrillo, Rodríguez. *Characterizing TTPs in the macOS Threat Landscape*. C&S 2026 (Q1).
Carrillo-Mondéjar, Suárez-Tangil, Costin, Rodríguez. *Exploring Shifting Patterns in Recent IoT Malware*. ECCWS 2024.

Agenda

Universidad Zaragoza

IA como defensa

Caso 2: LLMs para leer lo que nadie puede leer

El problema

Cuando llega un binario sospechoso o un firmware sin código fuente, alguien debe leer ensamblador para entender qué hace. Es lento, caro y escasea quien sabe hacerlo.

Lo que estamos investigando:

- ¿Pueden los LLMs descompilar binarios a código legible *y correcto*?
- ¿Pueden hacer triaje de firmware completo y señalar qué módulos requieren atención humana?

La pregunta que esto le deja al Derecho

Si un modelo “explica” qué hace un programa, ¿qué valor probatorio tiene esa explicación? ¿Quién la verifica y cómo se impugna?

Lecturas adicionales:

Pelayo-Benedet, Bolgorte, Rodríguez. *Does Representation Matter? IRs for LLM-based Binary Decompilation*. BAR 2026.
Huici, Yousefnezhad, Rodríguez, Costin. *Automating Firmware Vulnerability Triage via High-Level Representations*. BAR 2026.

IA como defensa

Caso 2: lo que el modelo dice. . . y lo que el código hace

Fragmento del binario

```
loop:
  mov  al, [esi]
  xor  al, 0x5A
  mov  [edi], al
  inc  esi
  inc  edi
  dec  ecx
  jnz  loop
  ret
```

Lo que hace: descifra un *buffer* con XOR de un solo byte.
Nada más

Resumen generado por el LLM

> *"La función descifra una cadena de configuración mediante XOR con la clave* `0x5A`*,* **la envía al servidor de control mediante una petición HTTPS** *y* **borra el buffer para dificultar el análisis forense**.*"*

- La primera mitad es **correcta**
- La segunda es **inventada**: no hay red ni borrado en el código
- **Suena igual de segura en ambas mitades**

Imaginad este párrafo en un informe pericial sin que nadie haya comprobado el código ensamblador...

Nota: ejemplo construido para la charla, representativo del tipo de error que observamos en nuestros experimentos

IA como defensa

Lo que ya funciona... y lo que aún no

Lo que sí

- Resúmenes de funciones desconocidas
- Triaje inicial: “mira aquí primero”
- Asistencia a analistas senior
- Reducción del tiempo del primer hallazgo

Lo que aún no

- Reemplazar el juicio humano
- Garantizar correctud sin verificación
- Explicar de forma trazable *por qué* concluye algo
- Sostener esa conclusión ante un tribunal

Conclusión honesta: el LLM es un asistente, no un perito.
*Pero **multiplica** la capacidad de un equipo pequeño*

Universidad Zaragoza
1474

Agenda

Universidad Zaragoza

Donde la técnica va por delante de la norma

Caso 3: forensia de memoria, o la evidencia que se evapora

Por qué importa

Cuando una intrusión deja poco rastro en disco, la memoria RAM es la última frontera de evidencia. Y se esfuma al apagar el equipo

Nuestra contribución reciente:

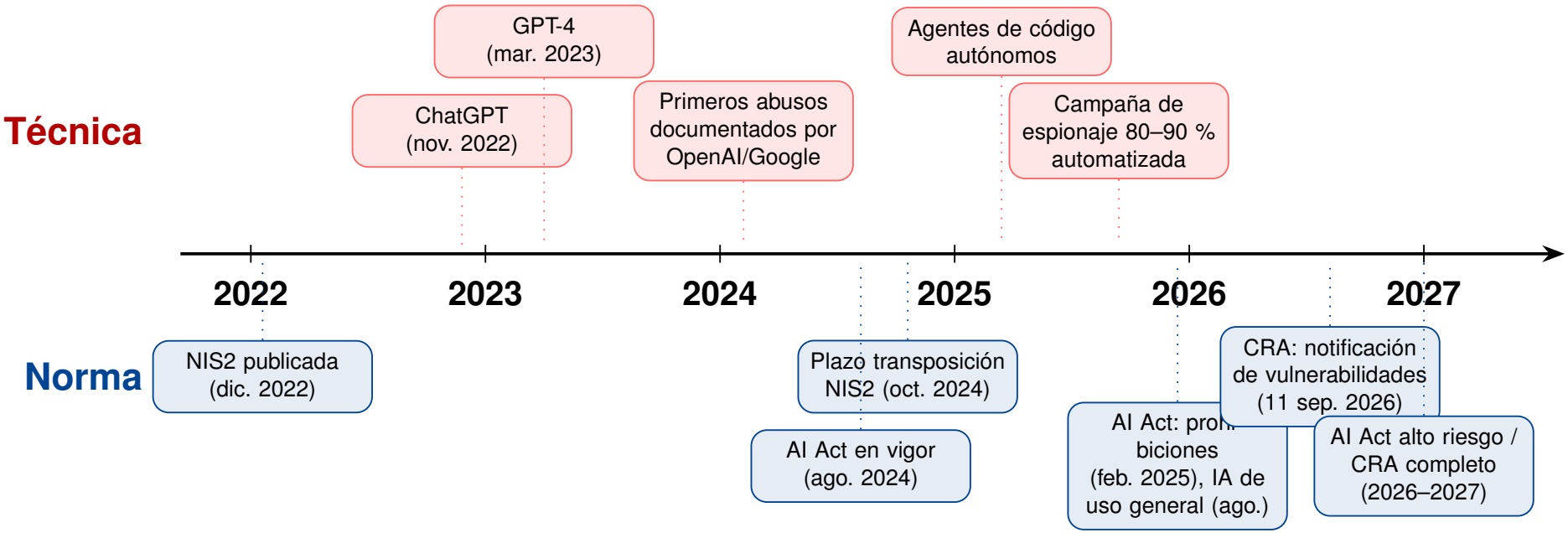
- Análisis estructural del *heap* de Windows NT
- Detección de *DLL hijacking* desde volcados de memoria
- Workflows de investigación alineados con MITRE ATT&CK
- Extracción de claves criptográficas de procesos en vivo

*El reto jurídico: cadena de custodia y garantías sobre una evidencia **volátil**, extraída con herramientas que cambian más rápido que la jurisprudencia*

Universidad Zaragoza
1474

Donde la técnica va por delante de la norma

Dos relojes que no van a la misma velocidad



- La obligación del CRA de notificar vulnerabilidades explotadas entró en vigor **hace tres días**. La campaña automatizada de arriba es de hace un año
- Entre publicar una norma y aplicarla pasan 2–3 años. En ese tiempo el atacante ha cambiado de herramientas dos veces

Universidad Zaragoza

Donde la técnica va por delante de la norma

Caso 4: infraestructuras críticas

Por qué importa

Los protocolos industriales (Modbus, SunSpec, DNP3) se diseñaron sin pensar en ciberseguridad. Y siguen funcionando en plantas eléctricas, agua, transporte

Nuestra contribución reciente:

- **MOSTO**: auditoría de seguridad de dispositivos ICS sobre Modbus/TCP
- *Datasets* para entrenar IDS en subestaciones eléctricas
- Detección de ataques BOLA en APIs REST a partir de su especificación OpenAPI

Aquí el Derecho ya ha llegado... sobre el papel

NIS2, Cyber Resilience Act, AI Act: obligaciones ambiciosas sobre sistemas que técnicamente no siempre pueden cumplirlas. Traducir entre ambos mundos es un oficio por cubrir

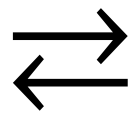
Agenda

Universidad Zaragoza

Conclusiones

El doble filo, en una imagen

IA ofensiva
Más rápido, más barato, más accesible



IA defensiva
Mejores detectores, mejor asistencia humana

- La carrera atacante–defensor se acelera; la norma llega después
- Cada avance técnico abre una pregunta jurídica: autoría, prueba, responsabilidad, cumplimiento
- Nadie puede responderlas desde una sola disciplina

Universidad Zaragoza
1474

Conclusiones

El recurso más escaso: quien habla los dos idiomas

Lo que el mundo necesita:

- Juristas que entiendan qué es (y qué no es) capaz de hacer la técnica
- Técnicos que entiendan qué exige (y por qué) la norma
- Mentalidad crítica frente a la IA: ni pánico ni fe ciega

Lo que tenéis por delante:

- Un máster diseñado justo en esa intersección
- Un campo profesional en plena expansión (y con déficit de perfiles)
- Problemas abiertos de sobra para investigar

Los casos de hoy no eran ciencia ficción: son artículos publicados

y cada uno de ellos sigue teniendo su mitad jurídica sin escribir

Universidad Zaragoza 1474

Conclusiones

Mensaje final

La IA no resuelve la ciberseguridad,
pero sí redefine quién la hace y cómo

El Derecho y la técnica, juntos, son la única respuesta sensata

Universidad Zaragoza

Hackear con un *prompt*
Amenazas, defensas y lo que el Derecho aún no nombra

Ricardo J. Rodríguez

rjrodriguez@unizar.es ※ @RicardoJRdez ※ www.ricardojrodriguez.es

Universidad Zaragoza

Dpto. de Informática e Ingeniería de Sistemas
Universidad de Zaragoza

14 de septiembre, 2026
Jornada Inauguración Máster Europeo EMILDAI
Facultad de Derecho, Universidad de León