ARI 128/2020
10 de noviembre de 2020

La ciberseguridad y su relación con la inteligencia artificial

Ana Ayerbe | Directora del Área de Negocio TRUSTECH de Tecnalia, *Basque Research and Technology Allliance* (BRTA)| @AnaAyerbe

Tema

La inteligencia artificial y la ciberseguridad pueden utilizarse tanto para reforzarse como para deteriorarse mutuamente.

Resumen

La ciberseguridad y la inteligencia artificial guardan una estrecha relación. Primero, las técnicas de inteligencia artificial (IA) pueden utilizarse para mejorar la ciberseguridad y resiliencia de productos, servicios, sistemas y, por ende, de las empresas y la sociedad (enfoque de defensa). Segundo, la IA está empezando a ser utilizada por cibercriminales y otro tipo de ciberatacantes para poner en riesgo la ciberseguridad y perpetrar diferentes tipos de ataques y generar noticias falsas (enfoque de ataque). Finalmente, los sistemas de IA son, a su vez, susceptibles de sufrir ciberataques, por lo que se deben desarrollar sistemas de IA seguros, que preserven la privacidad, en los que podamos confiar y que sean aceptados por sus usuarios (enfoque de confianza). Dada la interacción entre ambas, es necesario que las Estrategias de Ciberseguridad, de Inteligencia Artificial y de I+D+i se coordinen para crear técnicas, métodos y herramientas que faciliten el diseño, desarrollo, validación y despliegue de sistemas basados en la IA con un enfoque multicriterio que considere la ciberseguridad del dato, del modelo y del resultado.

Análisis

En 1950 Alan Turing definía las condiciones que debía cumplir una máquina para poder considerarla inteligente, pero fue realmente en 1956 cuando John McCarthy acuñó el término *inteligencia artificial* (IA) para referirse a máquinas que ejecutasen tareas características de la inteligencia humana y resolviesen problemas y lograsen objetivos de una forma similar a como lo hacían las personas.

Aunque las investigaciones en IA continuaron durante la década de 1970 y parte de la de 1980, pocos querían invertir su dinero en una tecnología que no estaba ofreciendo resultados palpables. Fue en 1996, el día en que el ordenador Deep Blue de IBM se impuso en una partida de ajedrez al entonces campeón del mundo Kaspárov, cuando se comenzó a ver que la IA ofrecía posibilidades de aplicación práctica. Ya en 2012 se empezó a hablar de *deep learning* al crear Google, un sistema capaz de identificar gatos en imágenes, y en 2015 AlphaGo se convirtió en la primera máquina en ganar a un jugador profesional del juego chino go.

Precisamente el mayor conocimiento sobre el funcionamiento del cerebro adquirido en los últimos años, junto con los avances en microelectrónica, el aumento de la potencia de computación, así como la posibilidad de acceder a grandes cantidades de datos y la conexión ubicua entre sistemas, han posibilitado los grandes avances en IA que se están dando actualmente. Esto ha motivado que la IA sea uno de los términos más utilizados en la actualidad, hasta generar la impresión de que un sistema que no haga uso de la IA en alguna de sus variantes (*machine learning*, *deep learning*…) no pueda considerarse un sistema relevante[1].

La IA ofrece múltiples posibilidades de aplicación[2] y, sin embargo, no dejan de surgir noticias referentes a decisiones erróneas tomadas por sistemas de IA o sobre conclusiones a las que han llegado sistemas de IA cuyo proceso de obtención es ininteligible para los humanos[3]. Esto demuestra que dichos sistemas de IA no han sido diseñados asegurando la imparcialidad y la transparencia en la toma de decisiones. Al igual que muchas otras tecnologías, la IA puede usarse tanto para el bien como para el mal[4]. Por este motivo, vamos a desgranar el posible buen y mal uso de la IA en el ámbito de la ciberseguridad y los peligros que puede representar su utilización si la IA no ha sido diseñada de una forma segura.

Inteligencia artificial en apoyo de la ciberseguridad

La IA puede utilizarse para ayudar a los profesionales de la seguridad a tratar la cada vez mayor complejidad de los sistemas modernos de IT, industria 4.0, infraestructura del Internet de las Cosas (IoT por sus siglas en inglés)…, así como la ingente cantidad de datos creados por ellos, e intentar estar por delante de los ciberatacantes. La ciberseguridad se enfrenta a múltiples retos, como son la detección de intrusiones, la protección de la privacidad, la defensa proactiva, la identificación de comportamientos anómalos o la detección de amenazas sofisticadas, pero, sobre todo, a las cambiantes amenazas que aparecen continuamente. Debido a ello, se están explorando métodos basados en la IA que faciliten el análisis y la toma de decisiones en tiempo real para una rápida detección y reacción ante ciberataques. También se está empleando la IA para desarrollar sistemas autoadaptables y que permitan automatizar las respuestas ante ciberamenazas.

Realmente, la IA puede utilizarse en todas las etapas de una seguridad integral inteligente: identificación, protección, detección, respuesta y recuperación ante incidentes; en este sentido, la ciberseguridad puede considerarse un dominio más de

[1] Miguel González San Emeterio (2020), "La inteligencia de las sumas y las restas", *Inspiring Blog Tecnalia*, 23 de julio de 2020, http://blogs.tecnalia.com/inspiring-blog/2020/07/23/la-inteligencia-las-sumas-restas.

[2] Jesús López Cobo (2020), "Los datos nunca duermen: analítica de datos en tiempo real", *Inspiring Blog Tecnalia*, 12 de marzo de 2020, http://blogs.tecnalia.com/inspiring-blog/2020/03/12/los-datos-nunca-duermen-analitica-datos-tiempo-real.

[3] Donna Lu (2019), "I can predict if you'll die son but we've no idea how it works", *NewScientist*, 11/XI/2019, https://www.newscientist.com/article/2222907-ai-can-predict-if-youll-die-soon-but-weve-no-idea-how-it-works.

[4] Eneko Osaba Icedo (2020), "Inteligencia artificial: un mundo bajo constantes críticas", *Inspiring Blog Tecnalia*, 5 de marzo de 2020, http://blogs.tecnalia.com/inspiring-blog/2020/03/05/inteligencia-artificial-mundo-constantes-criticas.

aplicación de la IA, como pueden serlo el de la energía, transporte, industria o salud. De hecho, este no es un ámbito nuevo de aplicación de la IA, sino que ya lleva tiempo utilizándose para desarrollar soluciones que puedan detectar y atajar ciberamenazas complejas y sofisticadas a la vez que evitar fugas de datos. Tal y como indica ENISA, se debe seguir investigando la utilización de IA en la inteligencia de ciberamenazas (*cyber threat intelligence*, CTI) para reducir el número de pasos manuales en los análisis realizados y validar dichos análisis, esto es, apoyando la CTI a lo largo de todo el ciclo de vida de la gestión y mitigación de los riesgos de seguridad[5].

En la actual situación de pandemia debida al COVID-19, lo que se ha observado es la gran capacidad que han mostrado los cibercriminales para adaptarse rápidamente al nuevo contexto vulnerable de teletrabajo aprovechando las conexiones a internet de los hogares para acceder a datos y sistemas de la empresa. Los cibercriminales han personalizado los vectores de ataque con métodos avanzados de robo de credenciales, ataques de *phishing* muy orientados, sofisticados ataques de ingeniería social y técnicas avanzadas de ocultación de *malware*, entre otros. En la medida en que estas técnicas se vayan combinando cada vez más con la IA, los ataques serán más difíciles de detectar y tendrán un mayor éxito, según el citado informe de ENISA.

La utilización de la IA por parte de los ciberatacantes

La IA ya se está utilizando en aplicaciones del mercado para conocer los patrones de comportamiento de usuarios y diseñar campañas comerciales utilizando *software* de IA a disposición de todo el mundo, por lo que sería muy ingenuo pensar que los cibercriminales no lo estén utilizando también, en el caso más sencillo, para conocer mejor a sus víctimas e identificar el mejor momento en el que realizar una acción delictiva con las mayores garantías de éxito.

La utilización de la IA depende del perfil de los ciberatacantes, que van desde los más inofensivos, asociados a la cibermalicia, a los más peligrosos, como pueden ser los relacionados con el ciberterrorismo, el ciberespionaje o la ciberguerra. La misma variedad puede encontrarse en el nivel de sofisticación y complejidad de los ciberataques, que varía mucho de unos a otros. Detrás de los ciberataques más peligrosos y sofisticados susceptibles de utilizar la IA pueden estar grupos muy especializados, financiados por determinados Estados y cuyos ataques pueden estar dirigidos hacia infraestructuras críticas de otro país o a generar campañas de desinformación. Si analizamos a los potenciales atacantes desde otra perspectiva, la de su forma de operar, pueden utilizar la IA tanto para explotar vulnerabilidades conocidas como para encontrar otras desconocidas o crearlas.

Además de la utilización del uso ofensivo de la IA por parte de los ciberatacantes para conocer patrones de comportamiento de las futuras víctimas, también puede utilizarse para romper más rápidamente contraseñas y *captchas*, construir *malware* que evite la detección, esconderse donde no puedan ser encontrados, adaptarse lo antes posible a

[5] ENISA (2020), “Research Topics. From January 2019 to April 2020”, https://www.enisa.europa.eu/topics/threat-risk-management/threats-and-trends/etl-review-folder/etl-2020-cybersecurity-research.

las contramedidas que puedan tomarse, así como para la obtención automática de información utilizando métodos de procesamiento del lenguaje natural (*natural language processing*, NLP) y la suplantación y generación de audios, vídeos y textos falsos. Los atacantes también están utilizando redes generativas conflictivas (*generative adversarial networks*, GAN) [6] para imitar patrones de tráfico de comunicaciones normales con el objetivo de distraer la atención de un ataque y encontrar y extraer datos sensibles rápidamente[7].

¿Son vulnerables los sistemas de IA a los ciberataques?

Los atacantes pueden utilizar sistemas de IA no sólo para tomar sus decisiones, sino también para manipular las decisiones que toman otros. De forma simplificada, un sistema de IA no deja de ser un sistema de *software* en el que se utilizan datos, modelos y algoritmos de procesamiento. En este sentido, un sistema de IA que no haya sido desarrollado con ciberseguridad desde el diseño puede ser vulnerable a ciberataques que tengan como objetivo los datos, el modelo o el algoritmo de la IA y provocar resultados no deseados o decisiones equivocadas en el sistema. Diferentes empresas han sufrido ya ataques a sistemas comerciales de IA, como es el caso de Microsoft, que ha observado en los últimos cuatro años un importante incremento en este tipo de ataques[8], o de Tesla[9], Google[10] y Amazon[11], por citar algunos.

La red de innovación del proyecto SPARTA, dentro de su programa de investigación SAFAIR, ha identificado diferentes tácticas de ataque que pueden llevarse a cabo sobre los sistemas de IA, tanto durante la fase de entrenamiento del sistema como durante la fase de operación, que se muestran en el siguiente cuadro.

[6] Thomas Klimek (2018), "Generative Adversarial Networks: What Are They And Why We Should be Afraid", https://www.cs.tufts.edu/comp/116/archive/fall2018/tklimek.pdf.

[7] Veronica Combs (2020), "3 ways criminals use artificial intelligence in cybersecurity attacks", *TechRepublic*, 7/X/2020, https://www.techrepublic.com/article/3-ways-criminals-use-artificial-intelligence-in-cybersecurity-attacks.

[8] Ram Shankar Siva Kumar y Ann Johnson (2020), "Cyberattacks against machine learning systems are more common than you think", 22 de octubre de 2020, https://www.microsoft.com/security/blog/2020/10/22/cyberattacks-against-machine-learning-systems-are-more-common-than-you-think.

[9] Evan Ackerman (2019), "Three Small Stickers in Intersection Can Cause Tesla Autopilot to Swerve Into Wrong Lane", *IEEE Spectrum*, 1/IV/2019. https://spectrum.ieee.org/cars-that-think/transportation/self-driving/three-small-stickers-on-road-can-steer-tesla-autopilot-into-oncoming-lane.

[10] Tiernan Ray (2018), "Google's image recognition AI fooled by new tricks", *ZDNet*, 30/XI/2018, https://www.zdnet.com/article/googles-best-image-recognition-system-flummoxed-by-fakes.

[11] Jesús Díaz (2018), "Alexa can be hacked-by chirping birds", *Fast Company*, 28/IX/2018, https://www.fastcompany.com/90240975/alexa-can-be-hacked-by-chirping-birds.

Cuadro 1. Tácticas de ataque

Acceso a los datos		
Envenenamiento	Indirecto	
	Directo	Inyección de datos
		Manipulación de los datos: manipulación de etiquetas
		Manipulación de los datos: manipulación de la entrada
		Corrupción de la lógica
Evasión	Gradiente	Un solo paso
		Iterativo
		Libre de gradiente
Oracle	Extracción	
	Inversión	
	Inferencia de pertenencia	

Fuente: Erkuden Rios (Tecnalia)[12]

El ataque denominado “acceso a los datos” consiste en que el atacante puede acceder a todo o parte de los datos de entrenamiento y utilizar esos datos para crear un modelo sustituto.

En el caso de los ataques de tipo “envenenamiento”, se alteran los datos o el modelo de forma indirecta o directa. En el primer supuesto, los atacantes sin acceso a los datos preprocesados utilizados por el modelo envenenan los datos antes del preprocesamiento.

En el “envenenamiento directo” pueden encontrarse las siguientes tácticas de ataque:

- Inyección de datos: inserción de muestras conflictivas (maliciosas) en los datos de entrenamiento originales para cambiar la distribución subyacente de datos sin cambiar las características o etiquetas de los datos de entrenamiento iniciales.
- Manipulación de datos (etiquetas): modificación por el atacante de las etiquetas de salida de los datos de entrenamiento originales.
- Manipulación de datos (entrada): modificación por el atacante de los datos de entrenamiento originales.
- Corrupción de la lógica: manipulación del algoritmo de *machine learning* para alterar el proceso de aprendizaje y el propio modelo.

Por otro lado, en los ataques de tipo “evasión”, el atacante resuelve un problema de optimización restringido para encontrar una perturbación en la entrada que provoque la

[12] Erkuden Rios (2020), “AI Systems Threat Analysis Mechanisms and Tools”, D 7.1, 31 de julio de 2020, Tecnalia para el proyecto SPARTA (**H2020-SU-ICT-2018-2020**).

clasificación errónea que desea. Bajo este tipo de ataque pueden encontrarse las siguientes tácticas:

- Gradiente, un solo paso: utiliza algoritmos de búsqueda basados en gradiente.
- Gradiente, iterativo: utiliza de forma iterativa algoritmos de búsqueda basados en gradiente.
- Libre de gradiente: necesita acceder a los valores de confianza dados por el modelo para ser efectivo.

Finalmente, en los ataques de tipo "Oracle", un atacante utiliza una interfaz o API del modelo para presentarle entradas, poder observar sus salidas y obtener de esta forma la información deseada. Pueden distinguirse las siguientes tácticas:

- Extracción: el atacante extrae los parámetros o estructura del modelo a partir de observaciones realizadas sobre las predicciones del modelo, incluyendo normalmente las probabilidades devueltas para cada clase.
- Inversión: las características inferidas pueden permitirle al atacante reconstruir los datos utilizados para entrenar al modelo, incluida información personal que viole la privacidad de un individuo.
- Inferencia de pertenencia: los atacantes utilizan los resultados devueltos por consultas realizadas al modelo para determinar si puntos de datos concretos pertenecen a la misma distribución que el conjunto de datos de entrenamiento y explotar las diferencias en la confianza dada por el modelo sobre puntos que fueron o no vistos durante el entrenamiento.

Según un informe de Gartner para el 2022, el 30% de todos los ciberataques de IA utilizarán envenenamiento de los datos de entrada, robo del modelo de IA o muestras conflictivas para atacar a los sistemas de IA[13]. Todas las tácticas de ataque identificadas en SPARTA han sido recogidas en una base de conocimiento de amenazas a la inteligencia artificial con el fin de apoyar la recogida, estructuración y reutilización de conocimiento acerca de las amenazas a los sistemas de IA y cómo actuar para minimizar el impacto de estas ciberamenazas. En esta línea, organizaciones como Microsoft, Mitre, Bosch, IBM, Nvidia, Airbus, Pricewaterhouse y el SEI de Carnegi Mellon, junto con otras cuatro entidades, han lanzado recientemente la Adversarial ML Threat Matrix, un *framework* abierto diseñado para ayudar a los analistas de seguridad a detectar, responder y solucionar amenazas que puedan producirse contra los sistemas de *machine learning*[14].

Abordando la ciberseguridad de la inteligencia artificial

Si queremos minimizar la deuda técnica[15], es decir, el esfuerzo adicional que habrá que realizar en el futuro para resolver los problemas generados, en este caso al desarrollar sistemas de IA sin tener en cuenta la ciberseguridad, debemos actuar durante todo el

[13] David Cearly y otros (2019), "Top 10 Strategic Technology Trends for 2020", Gartner, octubre de 2019, https://www.gartner.com/en/doc/432920-top-10-strategic-technology-trends-for-2020.

[14] Citado en nota 8.

[15] Jean-Louis Letouzey y Declan Whelan (2016), "Introduction to the Technical Debt Concept", https://www.agilealliance.org/wp-content/uploads/2016/05/IntroductiontotheTechnicalDebtConcept-V-02.pdf.

ciclo de desarrollo de la IA y a lo largo de la cadena de suministro para crear sistemas de IA seguros y justos. Para este fin, necesitamos continuar investigando las amenazas a la IA y compartir la información que se vaya generando. Es imprescindible diseñar los sistemas de IA teniendo en cuenta la seguridad, privacidad, imparcialidad y explicabilidad desde el diseño y por defecto, realizar pruebas que demuestren que el sistema cumple estas propiedades, velar por que las cumpla durante su operación y, en caso de producirse incidentes, realizar un análisis forense pormenorizado.

No debemos olvidar que los incidentes relacionados con la IA disminuyen la confianza en estos sistemas y su aceptación por parte de los usuarios. Si tenemos en cuenta que, según una encuesta realizada a 28 organizaciones, 25 de ellas no sabían cómo proteger sus sistemas de IA, se necesita trabajar para facilitar esta tarea a los implementadores de la IA y que lleguen sistemas seguros a los usuarios[16]. En esta línea, la Comisión Europea ha creado una lista de verificación para IA de confianza (ALTAI), que traduce los principios de una IA de confianza en una lista de comprobación que pueden utilizar los desarrolladores y proveedores para realizar una autoverificación de sus sistemas[17]. Aun así, debe dotarse a los desarrolladores de métodos, técnicas y herramientas que les permitan desarrollar estos sistemas con ciberseguridad y privacidad desde el diseño. Para diseñar, desarrollar, validar y desplegar sistemas de IA, debe velarse por la calidad del dato, del modelo y del resultado (las “3 C”). De este modo, los aspectos que deben tenerse en cuenta a la hora de desarrollar un sistema de IA son:

- Privacidad: debe asegurarse que el modelo de IA puede utilizar cualquier fuente de datos y que mantiene la privacidad del dato.
- Equidad: debe velarse por que el modelo de IA no se vea afectado por sesgos implícitos en la muestra de datos y no discrimine o favorezca determinadas salidas sin razón aparente.
- Trazabilidad: en el caso de que se produzca un fallo en el sistema, debe poderse analizar cuál ha sido el motivo, depurar responsabilidades y poner los medios necesarios para que no vuelva a suceder.
- Robustez: en caso de ciberataque, se debe saber hasta qué punto podemos fiarnos o no de la salida del sistema.
- Fiabilidad: debemos saber si la salida del modelo es fiable y qué sucede si su entrada cambia mínimamente.
- Causalidad: debemos saber si podemos influir en la salida del modelo actuando sobre los datos de entrada.
- Explicabilidad y transparencia: debemos intentar que el usuario pueda entender cómo funciona el modelo y qué ha visto el sistema en los datos para llegar a determinadas conclusiones.
- Gobernanza del dato: se debe garantizar el uso lícito, eficiente y eficaz de la información.

[16] Ram Shankar Siva Kumar y otros (2020), “Adversarial Machine Learning Industry Perspectives”, Microsoft, 21 de mayo de 2020, https://arxiv.org/pdf/2002.05646.pdf.

[17] Comisión Europea (2020), “Assesment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment”, 17 de julio de 2020, https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

Conclusiones

Existe una estrecha relación entre la inteligencia artificial y la ciberseguridad, con múltiples facetas que deben ser consideradas, por lo que una Estrategia Nacional de Inteligencia Artificial debe estar alineada y coordinada en ciertos ámbitos con la Estrategia Nacional de Ciberseguridad. No debemos olvidar que la ciberseguridad es un dominio más de aplicación de la IA, pero que, al mismo tiempo, los sistemas de IA también pueden ser sistemas vulnerables, por lo que se tienen que desarrollar con privacidad y seguridad desde el diseño y hay que velar por su ciberseguridad a lo largo de todo el ciclo de vida del sistema y considerando toda la cadena de suministro.

La Estrategia Nacional de Inteligencia Artificial debe apoyar la I+D+i en inteligencia artificial para crear técnicas, métodos y herramientas que faciliten el diseño, desarrollo, validación y despliegue de sistemas basados en IA con un enfoque multicriterio que considere la ciberseguridad del dato, del modelo y del resultado.

Por otro lado, debe impulsarse la I+D+i orientada a la aplicación de la IA en los nuevos sistemas de inteligencia de ciberamenazas, de forma que las organizaciones estén más preparadas y puedan reaccionar mejor frente a los ciberataques de vanguardia o de día cero. Integrar la ciberseguridad y privacidad desde el diseño implica fomentar una cultura de la ciberseguridad en las organizaciones para que se pueda capacitar a las personas, adoptar ciclos de vida ciberseguros para el desarrollo de los sistemas de IA y pensar en las respuestas ante posibles incidentes.

Finalmente, debemos asegurar la aplicación responsable y ética de la IA en todos los dominios, pero muy especialmente en ámbitos como la robótica, nuevas aplicaciones propiciadas por el 5G, sistemas automatizados de toma de decisiones y sistemas de vigilancia.