ARI 50/2019
14 de mayo de 2019

Implicaciones sobre el uso de la inteligencia artificial en el campo de la ciberseguridad

Javier Alonso Lecuit | Miembro del Grupo de Trabajo sobre Ciberpolítica del Real Instituto Elcano

Tema

La aplicación de la inteligencia artificial (IA) al ciberespacio plantea un complejo debate ético y normativo sobre el aumento cualitativo y cuantitativo de las amenazas y de las contramedidas basadas en IA.

Resumen

En el campo de la ciberseguridad, la IA aporta importantes mejoras mediante el análisis algorítmico aplicado a gran cantidad de información, infiriendo resultados basados en el contexto y en el aprendizaje adquirido a partir de situaciones anteriores. Las capacidades de la IA, sus algoritmos, se pueden aplicar de forma similar por quienes crean inseguridad en las sociedades avanzadas y por quienes las protegen. La confrontación directa entre algoritmos de IA y su escalada pueden llevar a un punto en el que la intervención humana podría quedar relegada a un segundo plano. La respuesta a esta situación promueve un debate de ámbito internacional sobre la necesidad de regular las características y el uso de la IA, principalmente desde un plano ético, pero también desde el punto de vista normativo y de control de su empleo, sin por ello limitar los beneficios aportados por la innovación en IA a la sociedad.

A pesar de existir unanimidad sobre la necesidad de una normativa internacional, resulta particularmente complicado concretar una hoja de ruta que establezca los pasos a seguir. La inacción, es decir, dejar que las fuerzas del mercado establezcan las reglas de juego conduciría a la desprotección de los derechos fundamentales de la seguridad de los individuos y naciones similar a la experimentada actualmente a escala mundial en materia de la privacidad. Este ARI se aproxima al impacto potencial de la IA, a los efectos de su empleo malicioso y la necesidad de controles y contramedidas que carecen de un marco regulatorio mundial

Análisis

La IA abarca un amplio espectro de usos, desde la conducción autónoma, la navegación no asistida de drones, el reconocimiento y clasificación de imágenes, la comprensión del lenguaje o el análisis de datos, entre otros[1]. Los resultados que puede conseguir en tareas muy concretas y especializadas alcanzan resultados netamente superiores a los obtenidos por las personas en tareas repetitivas y continuadas en el tiempo, evaluadas en velocidad de respuesta, tasa de errores, nivel de complejidad, etc. Los beneficios

[1] En este ARI se utiliza el concepto de IA en un sentido amplio, sin realizar distinción entre las distintas subcategorías posibles, tales como *machine learning*, *deep learning* o redes neuronales.

que puede aportar a la sociedad y la economía son incuestionables en sectores tales como la asistencia sanitaria, el consumo de energía, la automoción, la agricultura, el cambio climático o la gestión del riesgo financiero, por señalar los más relevantes, por ahora.

En el ciberespacio, es por todos conocida la relevancia que ha alcanzado la IA en la totalidad de plataformas de Internet: buscadores, redes sociales, comercio *online*, contenidos multimedia, etc. Igualmente conocida es la capacidad de influencia lograda gracias a su alcance mundial, el número de usuarios y la ingente información (en su mayoría datos de carácter personal) que nutren los algoritmos de IA. El resultado es habitualmente deseado y positivo para los usuarios; por ejemplo, al recibir recomendaciones de compra o al mantener en contacto a grupos de personas. Sin embargo, en ocasiones la IA conduce a situaciones perjudiciales, como al inducir pautas de consumo no deseadas o difundir información falsa a través de redes sociales con propósitos desestabilizadores.

En el campo de la ciberseguridad, las aplicaciones de la IA contribuyen a anticipar y neutralizar amenazas o gestionar incidentes de ciberseguridad con mayor rapidez y efectividad mediante el análisis de gran cantidad de información de contexto sin la necesidad de intervención humana altamente especializada. Ejemplos habituales son la detección de nuevo *malware* y virus mediante el reconocimiento de patrones anómalos en el funcionamiento de las aplicaciones, el aislamiento de los sistemas antes de ser infectados, el reconocimiento y neutralización de un ciberataque en sus fases iniciales (por ejemplo, aislando el sistema de un ataque de *ransomware* antes de que los archivos sean cifrados) o la detección y bloqueo automático de *phishing*, *spam*, intrusiones en la red o actuaciones fraudulentas.

Otro caso de uso es la mejora de la seguridad en el proceso de autenticación y acceso de usuarios a los sistemas mediante la aplicación de IA ponderando dinámicamente su perfil, la criticidad de la aplicación, el emplazamiento de los puntos de acceso a la red o el nivel de amenaza del momento. Los algoritmos deciden caso por caso solicitar distintos factores de autenticación (entre otros, biométricos), analizar el comportamiento del usuario o las características del dispositivo o modificar, si fuera necesario, sus privilegios en el sistema.

Asimismo, la IA aplicada al procesado de lenguaje natural facilita la automatización de actividades de ciberinteligencia; por ejemplo, la evaluación de la naturaleza, velocidad y gravedad de las amenazas mediante el análisis sistemático y continuado de fuentes abiertas en Internet, en particular las amenazas dirigidas contra un determinado objetivo o empresa, mediante la detección en la Internet visible y la invisible (*darknet*) de trazas de información que señalen la preparación de ataques o fugas de información sobre empleados y sistemas. En todos estos casos, es imprescindible entrenar los algoritmos de IA con gran cantidad de información actualizada, tanto en la fase de aprendizaje como posteriormente en producción, para lo que necesita importantes recursos en procesamiento, memoria y energía. La intervención humana es indispensable para asegurar el adecuado calibrado de los algoritmos y para detectar pronósticos erróneos o falsos positivos, que podrían acarrear graves consecuencias. Por consiguiente, la IA facilita la automatización de las funciones habituales en ciberseguridad, a pesar de que

la intervención humana siga siendo necesaria. Es previsible que su empleo se generalice y se convierta en una mercancía (*commodity*), lo que reducirá los costes operativos y la necesidad de personal altamente especializado, aspectos importantes en un contexto de amenazas crecientes en número y complejidad —en numerosos casos, promovidas por organizaciones muy profesionalizadas o por Estados hostiles— donde la escalabilidad de los medios resulta un factor económico y operativo determinante.

El empleo malicioso de la IA

La IA es una tecnología que amplía el rango y alcance de las ciberamenazas; mejora notablemente la eficiencia, escalabilidad y efectividad de los ciberataques, que superan en determinadas funciones la capacidad humana; potencia el anonimato y el distanciamiento psicológico del atacante, y facilita una rápida distribución e implantación de los algoritmos entre ciberdelincuentes. También permite mejorar las capacidades para proteger el ciberespacio; por ejemplo, los *hackers* éticos utilizan la IA para detectar vulnerabilidades y, sin embargo, estos mismos algoritmos pueden emplearse para poner a prueba y reforzar nuevo *malware*. Otro ejemplo del uso dual es la alteración de los algoritmos de IA con el fin de incorporar puertas traseras o vulnerabilidades o tomar control sobre el comportamiento de las aplicaciones. Por consiguiente, la alteración con fines maliciosos de algoritmos de IA utilizando sistemas de IA o de la información de referencia que los entrena y el aprovechamiento de sus vulnerabilidades son aspectos que considerar por su notable impacto en la seguridad.

Informes pioneros como "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation"[2] señalan que la tecnología de IA con fines maliciosos puede utilizarse en los ámbitos de la ciberseguridad, de la seguridad física y de la seguridad de los Estados o bien coordinar en la misma operación ataques a más de un ámbito, lo que agrava y crea nuevas amenazas que se pueden concretar en ataques más efectivos, precisos y selectivos cuya autoría es difícil de atribuir. En efecto, la IA facilita la coordinación de varias acciones contra un mismo objetivo, llevadas a cabo simultáneamente en el mundo físico (por ejemplo, el ataque mediante drones autónomos a una infraestructura crítica), en el ciberespacio (por ejemplo, el control de los sistemas informáticos y la red de comunicaciones de la compañía atacada) y hacia el Estado (por ejemplo, una campaña de desinformación sobre el incidente en Internet que difunda imágenes y audio sintetizado que emula la declaración de las autoridades), lo cual multiplica el impacto del ataque.

Ejemplos de uso malicioso de la IA en el ámbito de la ciberseguridad son la automatización de ataques de *phishing* personalizados basados en ingeniería social, la automatización del descubrimiento de vulnerabilidades, la automatización y sofisticación de la intrusión en redes, ataques de denegación de servicio que imitan la navegación de personas reales, la automatización de tareas en el cibercrimen (por ejemplo, el diálogo y el procesamiento de transacciones a víctimas de *ransomware*), la selección y

[2] Future of Humanity Institute, University of Oxford, Centre for the Study of Existential Risk, University of Cambridge, Center for a New American Security, Electronic Frontier Foundation y OpenAI, "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation", febrero de 2018, https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf

priorización de objetivos a partir de su comportamiento en la red o la alteración de algoritmos de IA.

Asimismo, el uso dual de la IA plantea nuevas amenazas a ciudadanos de determinados Estados mediante sistemas de vigilancia y neutralización de disidentes, la vigilancia y evaluación individualizada del comportamiento social de los ciudadanos en Estados totalitarios, la elaboración de noticias falsas respaldadas por vídeos y audios sintetizados que imitan a personas reales y dirigidos a microaudiencias, campañas de desinformación individualizadas dirigidas a colectivos o individuos, campañas de denegación o de manipulación al acceso de la información en medios e internet… Por último, cabe la posibilidad de utilizar esta dualidad en el terreno de la defensa activa, en respuesta a ciberataques y a operaciones militares realizadas con sistemas autónomos o coordinadas con operaciones en el ciberespacio. Este nuevo escenario plantea interrogantes éticos, jurídicos y diplomáticos; por ejemplo, sobre la legitimidad de la acción o sobre la atribución a un equipo de personas o a un sistema automatizado de IA.

La necesidad de contramedidas y controles

En ausencia de contramedidas, el uso malicioso de la IA puede agravar significativamente las ciberamenazas modificando sustancialmente su impacto con ataques más efectivos, dirigidos o difíciles de atribuir, incrementando las vulnerabilidades de los sistemas y reduciendo barreras de entrada a los agresores en aspectos tales como el coste, disponibilidad de sistemas de IA o los menores requisitos de especialización técnica de los atacantes.

En el campo de la ciberseguridad, cabría citar dos tipos de contramedidas: la centralización de los sistemas de información en grandes redes corporativas que cuenten con una estricta monitorización y protección frente a amenazas en todos sus elementos (incluyendo los terminales de usuario) y programas, respaldados por sistemas de IA, y, en segundo lugar, un uso intensivo de servicios virtualizados en la nube. En ambos casos, los usuarios se benefician de las economías de escala y la escalabilidad tanto en el uso de recursos (computación, licencias, etc.) como en el empleo de los medios de ciberseguridad más avanzados respaldados por IA y personal altamente cualificado, recursos que no están al alcance de las compañías particulares.

Asimismo, es previsible que las funciones y algoritmos de uso más habitual en el campo de la seguridad (reconocimiento biométrico o análisis dinámico de redes) se ofrezcan comercialmente a través de modelos de provisión virtualizados en la nube. Esto permitirá, por una parte, contar con los recursos de computación necesarios en cada caso y asegurar la actualización y entrenamiento de los algoritmos de IA, así como establecer puntos de control al uso malicioso de estas capacidades por la delincuencia u organizaciones terroristas sin por ello limitar las ventajas que aportan los desarrollos de código abierto en materia de IA a la comunidad de desarrolladores. Otro plano de actuación complementario que facilita establecer puntos de control muy relevante en la seguridad de la IA es la definición de estándares y esquemas de certificación de sistemas y aplicaciones de IA. En el campo de la seguridad de los Estados, las contramedidas técnicas y puntos de control en curso se enfocan al análisis sistemático

mediante IA de la información distribuida en redes sociales y medios para la detección de bots y noticias falsas, la participación de los proveedores de las plataformas en el filtrado de noticias falsas o el cifrado de las comunicaciones en Internet u otras medidas de ofuscación que dificultan la interceptación masiva en Estados totalitarios.

En relación con el control de esta dualidad, preocupa la asimetría existente entre la creciente facilidad con que cuenta la ciberdelincuencia para la ejecución de un ciberataque respaldado por tecnología de IA frente a la limitada capacidad de respuesta para contrarrestarlo de la que disponen ciudadanos y pequeñas empresas. Esta situación puede acelerar la adopción de medidas regulatorias, pero la tarea a la que se enfrentan legisladores, reguladores y desarrolladores para controlar el desarrollo malicioso de la IA no es fácil por muchas razones. En primer lugar, porque la frontera entre un uso legítimo e ilegítimo de la IA resulta difícil de delimitar en muchas ocasiones; por ejemplo, el mismo sistema de IA diseñado para la identificación de *phishing* puede prevenir los ataques o automatizarlos dirigiéndolos a objetivos específicos. En estos casos, los investigadores y expertos en el uso de la IA en ciberseguridad pueden explorar buenas prácticas o modelos abiertos de innovación y de licencias que limiten el uso de algoritmos de IA con fines maliciosos, así como promover practicas responsables (por ejemplo, la publicación controlada de sus vulnerabilidades o de sus procedimientos de verificación formal o la inclusión en los sistemas de IA de medidas de seguridad que limiten la copia o cesión de sus algoritmos).

Estas contramedidas y controles orientados a limitar el uso malicioso de la IA en el ámbito de la ciberseguridad son parte del debate general sobre el cómo y cuándo abordar la regulación sobre el uso de la IA que asegure unos principios éticos y jurídicos a ciudadanos, empresas y Estados. En este sentido, la Comisión Europea inició un estudio a mediados de 2018 con el propósito de establecer unas directrices éticas como base para el uso y el desarrollo legal de la IA y ha publicado en abril de 2019 el informe "Ethics Guidelines for Trustworthy Artificial Intelligence", elaborado por un grupo de expertos convocado por la CE en 2018[3]. La Comisión recomienda unas directrices sobre el uso de la IA fiable y centrada en el ser humano que cuenten con apoyo extracomunitario. El informe ha obtenido el respaldo del Parlamento Europeo y el Consejo en la comunicación publicada el 8 de abril de 2019[4].

Las directrices establecen como requisitos fundamentales y de carácter general a los sistemas de IA: (1) propiciar sociedades equitativas apoyando la intervención humana y los derechos fundamentales sin disminuir, limitar o desorientar la autonomía humana; (2) ser capaces de resolver errores o incoherencias durante todas las fases del ciclo de vida útil de los sistemas de IA; (3) asegurar la privacidad y el control por parte de los ciudadanos en la gestión de sus datos; (4) garantizar la transparencia mediante la trazabilidad de los sistemas; (5) fomentar la diversidad, la no discriminación y la equidad teniendo en cuenta el conjunto de capacidades, competencias y necesidades humanas;

[3] Grupo de Alto Nivel sobre Inteligencia Artificial, "Ethics Guidelines for Trustworthy Artificial Intelligence", abril de 2019, https://ec.europa.eu/futurium/en/ai-alliance-consultation

[4] Comisión Europea, "Generar confianza en la inteligencia artificial centrada en el ser humano", 8 de abril de 2019, https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52019DC0168&from=ES

(6) mejorar el cambio social y la sostenibilidad, y (7) contar con mecanismos que garanticen la responsabilidad y auditabilidad de los sistemas de IA y de sus resultados.

Con el fin de asegurar un desarrollo ético en línea con los principios establecidos, la Comisión establece tres líneas de acción a muy corto plazo (mediados de 2019): (1) el lanzamiento mediante el programa Horizonte 2020 de cuatro redes o centros de excelencia enfocados a retos científicos o tecnológicos tales como la *explicabilidad* y la interacción avanzada entre seres humanos y máquinas; (2) la creación de polos de innovación digital centrada en el uso de IA en la fabricación y utilización de macrodatos, y (3) el desarrollo de un modelo de intercambio y uso de datos públicos con hincapié en los sectores de transporte, atención sanitaria y fabricación industrial (Industria 4.0).

Varias organizaciones internacionales (por ejemplo, el IEEE[5]) llevan a cabo iniciativas similares, previas a la adopción de normativas y la definición de estándares internacionales o esquemas de certificación.

En el ámbito de la ciberseguridad, Enisa, la Agencia Europea para la Seguridad de la Redes y Sistemas de Información, publicó en marzo de 2019 el informe "Towards a framework for policy development in cybersecurity. Security and privacy considerations in autonomous agents"[6], donde desarrolla distintas consideraciones sobre la seguridad y privacidad de sistemas de IA. En relación con el uso malintencionado de sistemas de IA, Enisa indica la posibilidad de que estos puedan enmascararse como agentes legítimos para evitar su detección y que los desarrolladores habrán de ofrecer mecanismos que garanticen que el sistema lleva a cabo sus tareas asegurando la confidencialidad, integridad y disponibilidad de los datos procesados. Asimismo, debería verificarse su integridad a lo largo del ciclo de vida para que no sea posible contaminar el sistema de IA durante su funcionamiento[7].

El informe identifica el reparto de tareas. El desarrollador ofrecerá pruebas de que se ha adoptado un enfoque de seguridad desde el diseño documentando el desarrollo del *software*, la gestión de la calidad y los procesos para la gestión de la seguridad de la información. El proveedor diseñará y entregará el producto configurado de modo que las funcionalidades básicas sean las mínimas necesarias para llevar a cabo las operaciones del sistema, respaldadas por prácticas de seguridad claras. El fabricante documentará el diseño general del sistema de IA, incluyendo su arquitectura, funcionalidades y protocolos y toda aquella información que permita implementar y

[5] IEEE Standards Association, 2017. "The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems", https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52019DC0168&from=ES

[6] Enisa, "Towards a framework for policy development in cybersecurity - Security and privacy considerations in autonomous agents", 14 de marzo de 2019, https://www.enisa.europa.eu/publications/considerations-in-autonomous-agents

[7] Hay que tener en cuenta que el fabricante no puede controlar totalmente el comportamiento de los sistemas de IA, porque depende de la programación, el aprendizaje y la percepción del entorno. Asimismo, en la mayoría de los agentes autónomos avanzados no existe una distinción clara entre la fase de aprendizaje y la de operación al llevar a cabo un entrenamiento continuo y autónomo a lo largo de su ciclo de vida.

(cont.)

desplegar el sistema de IA de la manera más segura posible. Los componentes y servicios utilizados de terceros, así como los servicios que ofrezca el sistema de IA, deberán atenerse a los principios de seguridad por diseño y por defecto, de aplicación y verificables a lo largo del ciclo de vida.

Sobre privacidad, Enisa señala la necesidad de aplicar principios garantes de la minimización en la captura y uso de datos[8], la protección, retención, agregación y reutilización de la información[9] y en relación con la opacidad[10] de los algoritmos y muestra la necesidad de asignar responsabilidades y competencias a las diferentes agencias reguladoras para el desarrollo de estos aspectos.

Conclusión

A pesar de que la IA se encuentre todavía en una fase preliminar de desarrollo, ha propiciado avances en el campo de las tecnologías de la información e Internet, como el despliegue de las grandes plataformas *online*, con evidente éxito, aunque también conlleva importantes riesgos en materia de privacidad. Esta dualidad, la posibilidad de que el desarrollo de la IA afecte no sólo a la seguridad del ciberespacio, sino también a su inseguridad, es un motivo de preocupación entre la cibercomunidad.

Es, por lo tanto, necesario incorporar contramedidas y controles que limiten esos riesgos y potencien el uso de la IA en beneficio del ser humano. El desarrollo de los anteriores presenta notables complicaciones técnicas, regulatorias y éticas, lo que retrasa su incorporación a las políticas públicas y posteriores desarrollos normativos. El desarrollo de la IA libre de mecanismos de control agudizaría la asimetría actual entre quienes protegen su privacidad, negocio o reputación frente a los ciberataques y los que amenazan los derechos fundamentales y la seguridad de los individuos, empresas y naciones. La aplicación de la IA podría decantar la victoria del lado de los segundos de forma irreversible, especialmente para Estados y empresas poco resilientes, si no se encuentran, ensayan y codifican contramedidas y controles adecuados.

[8] Los algoritmos exigen grandes cantidades de datos para llevar a cabo el aprendizaje de los agentes autónomos; los sistemas de IA tienden a recopilar tantos datos como sea posible en lugar de una muestra estadísticamente relevante, ya que en un entorno complejo y cambiante al sistema le resulta difícil predecir la utilidad de una información determinada.

[9] Los agentes autónomos procesan datos en tiempo real que podrían borrarse después de su uso sin afectar su funcionalidad; sin embargo, se conservan durante largos períodos de tiempo por diferentes motivos (registro de funcionamiento, auditabilidad, explotación para otras finalidades, etc.).

[10] Los procesos de aprendizaje y operación de los sistemas de IA operan como una *caja negra*, es decir, no proporcionan explicaciones directas sobre el proceso de obtención del resultado. Por consiguiente, resulta complejo demostrar la legalidad, equidad y transparencia del procesamiento de datos de carácter personal.

Las autoridades u organizaciones internacionales, al igual que las europeas que se mencionan en el ARI, están ya valorando la contención de los riesgos de la IA para prevenirlos o mitigarlos en la medida de lo posible, porque no se puede garantizar la inmunidad frente a ellos. Antes de poder adoptar normativas, es necesario desarrollar mejor la definición de los estándares internacionales o esquemas de certificación; resulta complicado equilibrar la necesidad de implantar principios, normas y jurisprudencia de ámbito internacional que cuenten con el respaldo de la industria y los Estados sin por ello frenar la innovación o la libre competencia comercial entre distintas regiones en el campo de la IA.