Implicaciones de la gobernanza de los datos sintéticos en el contexto de la seguridad internacional

Informe de un seminario sobre tecnología y seguridad

Federico Mantellassi

Agradecimientos 2
Sobre el UNIDIR 2
Sobre el Programa de Seguridad y Tecnología 2
Nota 3
Citación 3
Sobre el autor 3
Acrónimos y abreviaturas 4
1. Introducción 5
1.1 Sobre el seminario 6
2. Datos sintéticos y seguridad internacional: contexto de la cuestión 7
2.1 ¿Qué son los datos sintéticos? 7
2.2 Datos sintéticos en el ámbito militar 9
3. Retos e implicaciones para la gobernanza 9
3.1 Datos sintéticos y gobernanza de datos civiles 10
3.2 El papel de las normas 12
3.3 Datos sintéticos y gobernanza internacional de la IA militar 13
a. Sobre la novedad de los retos y la aplicabilidad de los marcos existentes 13
b. Sobre la importancia de un planteamiento multipartito 15
c. Sobre las directrices y la especificidad del contexto 15
d. Sobre las oportunidades de gobernanza en el ámbito militar 16
4. Conclusión 17
Anexo: Programa y participantes 19

Agradecimientos

El apoyo que recibe el UNIDIR de los principales donadores es la base de todas las actividades del Instituto. Esta publicación ha sido financiada por la Unión Europea en el marco del Programa de Seguridad y Tecnología del UNIDIR, que cuenta también con el apoyo de los gobiernos de Alemania, República Checa, Francia, Italia, Noruega, Países Bajos y Suiza, además de Microsoft.

El autor desea expresar su sincero agradecimiento a Wenting He por moderar y organizar el primer panel del evento, así como a Jessica Espinosa Azcárraga por ayudar con la organización. Además, el autor desea agradecer a todos los ponentes su participación, así como al Dr. Giacomo Persi Paoli, Sarah Grand Clément, Wenting He, Calum Inverarity, la Dra. Ana Beduschi y Aldo Lamberti sus comentarios sobre este informe.

Sobre el UNIDIR

El Instituto de las Naciones Unidas de Investigación sobre el Desarme (UNIDIR por sus siglas en inglés) es un instituto autónomo de las Naciones Unidas financiado a través de contribuciones voluntarias. Al ser uno de los pocos institutos de política pública del mundo dedicado al desarme, UNIDIR genera conocimiento y promueve el diálogo y medidas en materia de desarme y seguridad. Con sede en Ginebra, ayuda a la comunidad internacional a desarrollar las ideas prácticas e innovadoras necesarias para hallar soluciones a problemas críticos para la seguridad.

Sobre el Programa de Seguridad y Tecnología

Los avances actuales de la ciencia y la tecnología presentan nuevas oportunidades y desafíos para la seguridad internacional y el desarme. El Programa de Seguridad y Tecnología del UNIDIR tiene por objeto fomentar el conocimiento y concientizar sobre las implicaciones y los riesgos para la seguridad internacional que plantean determinadas innovaciones tecnológicas, así como alentar a las partes interesadas a explorar ideas y desarrollar nuevas formas de pensar para darles solución.

Nota

Las denominaciones empleadas en esta publicación y la presentación de los datos que contiene no implican juicio alguno por parte de la Secretaría de las Naciones Unidas acerca del estatuto jurídico de los países, territorios, ciudades o zonas ni de sus autoridades, ni tampoco respecto de la delimitación de sus fronteras o límites. Las opiniones expresadas en la presente publicación son responsabilidad exclusiva de sus autores. No reflejan necesariamente las opiniones de las Naciones Unidas, de UNIDIR ni de sus funcionarios o donadores.

Citación

Mantellassi, Federico. Implicaciones de la gobernanza de datos sintéticos en el contexto de la seguridad internacional: informe de un seminario sobre tecnología y seguridad. Ginebra, Suiza: UNIDIR, 2024.

Sobre el autor

Federico Mantellassi es investigador del Programa de Seguridad y Tecnología del UNIDIR. Su trabajo se centra en las implicaciones, los riesgos y las oportunidades para la seguridad internacional de la ciencia emergente y los avances e innovaciones tecnológicas. Previamente, Federico era oficial de Investigación y Proyectos en el Centro de Política de Seguridad de Ginebra, donde hizo investigaciones sobre la intersección entre las tecnologías emergentes, la seguridad internacional y la guerra. Tiene una maestría en Inteligencia y Seguridad Internacional por el King's College de Londres y una licenciatura en Estudios Internacionales por la Universidad de Leiden.

Acrónimos y abreviaturas

IA Inteligencia artificial

EU Unión Europea

RGPD Reglamento General de Protección de Datos

IEEE Instituto de Ingenieros Eléctricos y Electrónicos

ISO Organización Internacional de Normalización

ODI Open Data Institute

PET Tecnología de protección de la intimidad

1. Introducción

Los datos son cruciales para la formación y el desarrollo de sistemas de inteligencia artificial (IA). Sin embargo, hay tres cuestiones clave relacionadas con los datos que pueden obstaculizar el desarrollo y la implantación de capacidades y sistemas de IA. En primer lugar, el desarrollo de las tecnologías de IA ha dependido, al menos en parte, de la disponibilidad de grandes conjuntos de datos para entrenar los modelos de IA. En segundo lugar, los datos son un recurso cuya disponibilidad, recopilación, limpieza, uso e intercambio se ven afectados por factores como los costos de recopilación, la falta de datos reales en determinados ámbitos, así como por limitaciones normativas, jurídicas y éticas. En tercer lugar, la calidad, representatividad y diversidad de los datos están directamente relacionadas con el rendimiento, el nivel de sesgo, la precisión y la fiabilidad del modelo de IA. **Se ha propuesto el uso de datos sintéticos —datos generados artificialmente en el mundo digital con propiedades que suelen derivarse de un conjunto original de datos — como solución para resolver algunos de estos problemas con los datos, especialmente para el entrenamiento de modelos de IA**[1]. De hecho, los datos sintéticos pueden ayudar a resolver problemas como los sesgos de los conjuntos de datos, al tiempo que permiten ampliarlos, crearlos, diversificarlos y afinarlos. A los datos sintéticos se les conoce a menudo como tecnología de mejora de la privacidad (PET, por sus siglas en inglés), ya que facilitan el uso y el intercambio de conjuntos de datos sensibles[2]. Los datos sintéticos son especialmente prometedores en sectores como el militar[3]. En este sector, la demanda de capacidades facilitadas por la IA es cada vez mayor, pero escasean los conjuntos de datos diversos y de alta calidad, y las consecuencias de los algoritmos defectuosos pueden ser muy graves. Los datos sintéticos podrían permitir desarrollar capacidades avanzadas de IA sin la necesidad de coleccionar datos del mundo real[4]. Sin embargo, los datos sintéticos no son una panacea y se ha demostrado que pueden agravar muchos de los problemas que pretenden solucionar, lo que ha suscitado discusiones sobre gobernanza y regulación[5].

Los datos sintéticos se encuentran en una relativa "zona gris" en términos de regulación y gobernanza. Los principales marcos regulatorios sobre gobernanza de datos e IA, como la Ley de

IA de la Unión Europea y el Reglamento General de Protección de Datos (RGPD), mencionan los datos sintéticos, si es que acaso lo hacen, de manera superficial. Para algunos, esto implica que los datos sintéticos, como PET, pueden ser tanto una forma de eludir los estrictos marcos regulatorios como una útil herramienta de cumplimiento[6]. Otros señalan que los datos sintéticos conllevan muchos de los mismos riesgos que los datos del mundo real, y que pueden dar lugar a efectos secundarios similares en la precisión, seguridad, imparcialidad y representatividad de los modelos de IA, por lo que insisten en que son necesarios nuevos marcos y planteamientos regulatorios para evitar lagunas de gobernanza y puntos ciegos[7]. En este sentido, es de suma importancia comprender cómo los marcos actuales de gobernanza (civil y militar) y regulación engloban los datos sintéticos, si son adecuados para abordar los riesgos potenciales y si es necesario ajustarlos. Las brechas regulatorias y de gobernanza son especialmente críticas a la hora de integrar rápidamente capacidades basadas en la IA en el ámbito militar. Por lo tanto, es esencial comprender las implicaciones de los datos sintéticos en los debates sobre la gobernanza de la emergente IA militar.

Para explorar los retos que plantea la gobernanza de los datos sintéticos en el contexto de la seguridad internacional, el Programa de Seguridad y Tecnología del UNIDIR organizó un evento titulado “Seminario de Tecnología y Seguridad sobre Datos Sintéticos: Explorar las implicaciones de la gobernanza”.

Este informe ofrece un resumen de los principales temas discutidos y de las conclusiones de los debates. El informe se divide en dos partes, las cuales reflejan la estructura del seminario. En la primera parte ofrece una breve visión general de la tecnología y sus usos en el ámbito militar. La segunda parte presenta los distintos puntos de vista, los problemas y los retos para la gobernanza que pueden plantear los datos sintéticos en el contexto de la seguridad internacional.

1.1 Sobre el seminario

Los Seminarios sobre Tecnología y Seguridad comprenden una serie de eventos organizados por el Programa de Seguridad y Tecnología del UNIDIR centrados en diversas tecnologías facilitadoras. Esta serie de seminarios tiene tres objetivos:

1. Dar a conocer una amplia variedad de tecnologías facilitadoras emergentes y críticas a la comunidad diplomática;
2. Alertar a la comunidad diplomática sobre las posibles implicaciones para la seguridad internacional del desarrollo y uso de dichas tecnologías; y
3. Explorar las posibilidades de gobernanza a través del diálogo y el compromiso entre las distintas partes interesadas.

El 29 de octubre de 2024 se llevó a cabo un Seminario de Tecnología y Seguridad dedicado a la gobernanza de los datos sintéticos. Este evento de que duró medio día consistió en un **Desayuno Tecnológico,** que sirvió de introducción a la tecnología para los representantes políticos, así como un **Diálogo Multilateral sobre Datos Sintéticos** en el que expertos de la industria, de organizaciones internacionales y del mundo académico se reunieron para compartir diversos puntos de vista sobre las dificultades específicas que plantea la gobernanza en el contexto de la seguridad internacional. El seminario tomo lugar virtualmente por la ocasión de la septuagésima novena sesión de la Primera Comisión de la Asamblea General de las Naciones Unidas en 2024.

El programa completo del seminario se encuentra en el anexo del presente informe.

2. Datos sintéticos y seguridad internacional: contextualización del problema

Conclusiones
• Los avances en el campo de la IA generativa y la creciente adopción de la IA en todos los sectores han ampliado la omnipresencia de los datos sintéticos, aumentando no solo la escala y la facilidad con que pueden generarse, sino también su variedad y calidad. • Los datos sintéticos son prometedores y ofrecen posibles soluciones a los problemas relacionados con los datos (sesgo, escasez, calidad, representatividad, privacidad) tanto en el ámbito civil como en el militar. • Las fuerzas armadas recurren cada vez más a los datos sintéticos puesto que cada vez adoptan más capacidades basadas en la IA para entrenar modelos militares para la identificación, los sistemas de puntería, la planificación operativa y táctica, y el desarrollo de escenarios y entornos sintéticos. • A pesar de sus ventajas, los datos sintéticos pueden perpetuar los riesgos que ya plantean los datos, crear nuevos o aumentar la magnitud de sus repercusiones.

2.1 ¿Qué son los datos sintéticos?[8]

Los datos sintéticos pueden definirse como **"aquellos datos creados mediante algoritmos o simulaciones informáticas que reproducen algunas propiedades estructurales y estadísticas**

[8] Esta sección se basa en el trabajo previo realizado por el Programa de Seguridad y Tecnología del UNIDIR sobre Datos Sintéticos y Seguridad Internacional. Para un análisis detallado y en profundidad de lo que son los datos sintéticos, y de los riesgos y oportunidades que suponen para la seguridad internacional, en especial en el contexto de las capacidades militares autónomas que son posibles gracias a la IA, véase

de los datos del mundo real"[9]**.** Existen varios métodos para generar datos sintéticos, donde los conjuntos de datos resultantes pueden ser totalmente sintéticos (con todos los datos generados artificialmente), parcialmente sintéticos (con una pequeña parte de un conjunto de datos reales sustituida por datos sintéticos) o híbridos (en los que se mezclan datos reales y totalmente sintéticos)[10]. En resumen, los datos sintéticos se utilizan sobre todo para **complementar conjuntos de datos** (y tratar de resolver problemas con los datos, como los relacionados con el sesgo o la representatividad), **crear conjuntos de datos cuando no existen** o **eliminar información personal identificable** cuando la sensibilidad de la información así lo requiera. En este sentido, el valor de los datos sintéticos reside en su capacidad para ayudar a resolver **problemas clave con los datos, como el sesgo, la representatividad, la calidad, la escasez y la privacidad.**

Aunque los datos sintéticos no son un concepto novedoso y se utilizan desde hace tiempo, los recientes avances tecnológicos, especialmente en IA generativa, han aumentado de manera drástica la **escala y facilidad** con que pueden producirse, la **diversidad** de tipos de datos que pueden crearse y su **calidad**. Estos avances han disminuido las barreras de acceso a los datos sintéticos y han ampliado de manera significativa el número de personas y organizaciones que ahora pueden utilizarlos sin contar con extenso conocimiento técnico. A su vez, esto ha aumentado su popularidad, y algunas estimaciones indican que el 60 % de todos los datos de entrenamiento de IA serán sintéticos a partir de 2024 [11]. La creciente popularidad de los datos sintéticos es, además, el resultado de una continua necesidad de más datos para entrenar modelos de IA.

Sin embargo, los datos sintéticos no son la panacea, y se ha demostrado que pueden perpetuar, y a veces agravar, los problemas que su uso pretende resolver. De hecho, los datos sintéticos no son intrínsecamente privados, seguros, representativos o imparciales, por lo que requieren mucha consideración y curación para que lo sean. Además, la investigación ha demostrado que el entrenamiento repetitivo de los modelos de IA con datos sintéticos generados a partir de versiones anteriores de ellos mismos puede provocar un "colapso del modelo", de modo que el modelo olvida la distribución de datos subyacente lo que provoca una reducción drástica de la calidad y la precisión de los resultados[12]. Además, la mayor prevalencia de los datos sintéticos podría ampliar la superficie de riesgo en los problemas relacionados con los datos y aumentar la magnitud de las consecuencias negativas[13].

2.2 Datos sintéticos en el ámbito militar

Los datos sintéticos son cada vez más frecuentes en el ámbito militar, donde los problemas relacionados con la escasez, la parcialidad y la sensibilidad de los datos son especialmente graves[14]. Al igual que en el sector civil, el aumento del uso de datos sintéticos en este ámbito está relacionado con el creciente interés de las fuerzas armadas por las soluciones basadas en IA. En este contexto, los datos sintéticos se utilizan principalmente para el **entrenamiento de modelos de IA militar** para la **identificación, la fijación de objetivos** y la **planificación operativa y táctica**, así como para el desarrollo de escenarios y **entornos sintéticos.**

Principalmente, los datos sintéticos pueden ayudar a las fuerzas armadas **a llenar vacíos y a aumentar la calidad de sus conjuntos de datos** —como la creación de imágenes de objetos desde distintos ángulos y en diferentes condiciones— para incrementar el rendimiento de los modelos de IA. Además, los datos sintéticos pueden ayudar en la **gestión de datos,** contribuyendo a reducir los costos asociados al etiquetado y recolecta, y **acelerando el desarrollo** de productos de IA. Los datos sintéticos también pueden utilizarse para crear simulaciones realistas de diversas operaciones militares, como ataques de adversarios. Estas simulaciones permitirían a los Estados probar la eficacia de sus sistemas de IA, desarrollar nuevas estrategias y tácticas, y prepararse para una gama más amplia de amenazas potenciales en un entorno controlado y seguro.

Sin embargo, el uso de datos sintéticos en el ámbito militar padece de los riesgos inherentes a este tipo de datos. De hecho, los datos sintéticos, pese a pretender representar la realidad, pueden perpetuar e incluso reinterpretar los sesgos existentes en los datos originales de los que se derivan. Esa posibilidad supone un riesgo importante, sobre todo en contextos militares delicados en los que las decisiones sesgadas pueden tener graves consecuencias. Además, no elimina el riesgo de reidentificación de personas o de información sensible dentro del conjunto de datos, lo que podría conducir a la revelación de datos militares sensibles. Por otro lado, los ataques de "envenenamiento de datos" por parte de actores maliciosos podrían sesgar el proceso de aprendizaje de los sistemas de IA[15].

3. Retos e implicaciones para la gobernanza

Conclusiones
• El panorama de la gobernanza de los datos sintéticos no está todavía maduro ni en el ámbito civil ni en el militar. Es necesario seguir trabajando para aclarar cómo aplicar los marcos de gobernanza y las normativas existentes a los datos sintéticos, y cómo habría que adaptarlos para cubrir mejor los posibles vacíos.

- No hay un consenso sobre la necesidad de nuevas normativas y marcos específicos para los datos sintéticos.

- Los estándares internacionales son una herramienta importante de la gobernanza tecnológica. Aunque no hay estándares internacionales en materia de datos sintéticos, se está trabajando en su desarrollo, que será decisivo para fomentar la innovación y la adopción responsables de la tecnología.

- Debido a su creciente uso en el ámbito militar para entrenar sistemas de IA, es indispensable abordar los datos sintéticos en los debates sobre la gobernanza de la IA militar. Todavía hay mucho trabajo por hacer para aplicar, adaptar o construir sobre las prácticas establecidas y los conceptos de gobernanza vinculados a los datos en el ámbito militar.

- Los datos sintéticos ofrecen oportunidades para la gobernanza de la IA militar, ya que permiten, en potencia, un mayor intercambio de datos, el desarrollo conjunto de aplicaciones de IA y la elaboración común de directrices para la generación y el uso responsable de datos sintéticos, avanzando así hacia objetivos globales de IA militar responsable.

- La gobernanza de los datos sintéticos en el ámbito militar va a requerir un mayor compromiso de las distintas partes interesadas. Ello implica la cooperación entre los Estados, pero también con el sector privado, que debería de participar activamente en los debates e iniciativas en materia de gobernanza. Fomentar la confianza entre los gobiernos y la industria va a ser fundamental para esta labor.

3.1 Datos sintéticos y gobernanza de datos civiles

Equilibrar los riesgos y las oportunidades de los datos sintéticos va a exigir comprender los retos de la gobernanza. Aunque los datos sintéticos no son necesariamente algo nuevo, **los debates sobre la gobernanza relacionados con su generación y uso están surgiendo ahora tanto en el ámbito civil como en el militar.** Las cuestiones relativas a la situación jurídica de los datos sintéticos, las necesidades regulatorias y los posibles planteamientos de la gobernanza están en estado prematuro y el panorama de la gobernanza se mantiene inmaduro. Actualmente no existe ninguna legislación ni marcos específicos para los datos sintéticos. Algunos marcos regulatorios, como la Ley de Inteligencia Artificial de la UE, mencionan los datos sintéticos brevemente, mientras que solo unos pocos gobiernos han emitido directrices sobre la generación de datos sintéticos[16].

[17] Los conjuntos de datos totalmente sintéticos contienen datos generados en su totalidad por un modelo de IA y no contienen datos del mundo real. El modelo identifica las propiedades y

En el ámbito civil, no existe consenso sobre si los datos sintéticos desafían los marcos regulatorios y de gobernanza existentes de datos, y, si es así, de qué manera. Por ejemplo, se señala que los datos sintéticos podrían cuestionar las categorías de datos personales/no personales, que son la base de las normativas y marcos de gobernanza de datos como el RGPD de la UE. **Se argumenta que estas normativas no están adecuadamente equipadas para abordar las complejidades de los datos sintéticos, que pueden difuminar los límites entre estas categorías.** Dependiendo del tipo de datos sintéticos —completamente sintéticos, parcialmente sintéticos o híbridos— el nivel de información personal presente y el riesgo de reidentificación y, por lo tanto, la aplicabilidad de las leyes de protección de datos, pueden variar considerablemente[17]. Esta ambigüedad crea una inseguridad legal tanto para los creadores como para los usuarios de datos sintéticos. En este sentido, **el creciente uso de datos sintéticos puede ampliar el alcance del paradigma tradicional de datos personales/no personales en la normativa de protección de datos actual.**

Para otros, los datos sintéticos correctamente generados son una PET útil que puede servir de herramienta para cumplir con distintos marcos normativos de datos. Además, los datos sintéticos podrían emplearse para lograr objetivos más amplios de la gobernanza de datos, democratizando el acceso a datos valiosos sin dejar de **proteger la privacidad,** haciendo que los catálogos de datos sean transparentes y facilitando el seguimiento de auditorías de cara a la **rendición de cuentas, mejorando la calidad de los datos**, al contar con una fuente de datos coherente y controlable, y **facilitando el intercambio seguro de datos** a nivel nacional e internacional. Sigue habiendo desacuerdos sobre si los datos sintéticos son una herramienta útil a incentivar o una innovación que podría socavar los mecanismos legales desarrollados para proteger a la sociedad de diversos riesgos relacionados con los datos[18].

Existe una falta de claridad jurídica y normativa con respecto al tratamiento de datos sintéticos. De ahí que algunos argumenten la necesidad de establecer directrices claras que garanticen la

---

patrones estadísticos de un conjunto de datos y genera uno totalmente nuevo. En los datos parcialmente sintéticos se sustituyen ciertas características sensibles seleccionadas de un conjunto de datos y se reemplazan por valores sintéticos, pero conservan algunos datos reales. Los datos sintéticos híbridos combinan unos datos del mundo real y otros totalmente sintéticos, emparejando registros aleatorios de un conjunto de datos real con un registro sintético. Para más información, consulte

[17] Los conjuntos de datos totalmente sintéticos contienen datos generados en su totalidad por un modelo de IA y no contienen datos del mundo real. El modelo identifica las propiedades y patrones estadísticos de un conjunto de datos y genera uno totalmente nuevo. En los datos parcialmente sintéticos se sustituyen ciertas características sensibles seleccionadas de un conjunto de datos y se reemplazan por valores sintéticos, pero conservan algunos datos reales. Los datos sintéticos híbridos combinan unos datos del mundo real y otros totalmente sintéticos, emparejando registros aleatorios de un conjunto de datos real con un registro sintético. Para más información, consulte

transparencia, la **imparcialidad** y la **responsabilidad** en el tratamiento de todo tipo de datos sintéticos, así como una mayor claridad y orientación en lo que respecta a qué datos se han empleado en los modelos fundacionales utilizados para generar datos sintéticos[19]. Entre las propuestas, está:

- la **transparencia**: los datos sintéticos deberían etiquetarse claramente como tales y la información sobre su proceso de generación debería estar disponible;
- la **rendición de cuentas**: deberían desarrollarse medios para establecer unos procedimientos claros que permitan pedir cuentas a los responsables de la generación y el tratamiento de datos sintéticos; y
- la **imparcialidad:** los datos sintéticos deberían incluir algunas garantías de que su generación y uso no conlleve efectos adversos, como la perpetuación de sesgos o la creación de otros nuevos.

Debido a la falta de claridad jurídica, es posible que los datos sintéticos queden al margen de la supervisión regulatoria, además de que pueden acarrear los mismos problemas que los marcos de gobernanza tratan de resolver en relación con los datos del mundo real. **Por lo tanto, un reto clave para la gobernanza en el ámbito civil será garantizar que los datos sintéticos se desarrollen y utilicen de forma que, de quedar fuera del ámbito de aplicación de la normativa actual sobre datos, no perpetúen ni generen nuevos perjuicios.** Es necesario seguir investigando y trabajando para definir con mayor claridad la situación jurídica de los datos sintéticos, así como para identificar las posibles brechas de gobernanza en las normativas y marcos existentes a fin de ofrecer claridad a los desarrolladores y usuarios de datos sintéticos.

3.2 El papel de los estándares

Los estándares (tanto técnicos como no técnicos) son un aspecto importante de la gobernanza de la tecnología civil. Ambos son muy necesarios en el contexto de los datos sintéticos y actualmente constituyen un importante campo de trabajo. **De hecho, no existe ninguna estándar internacional para la generación de datos** sintéticos, ni se han acordado definiciones o puntos de referencia universales para evaluar la calidad y fiabilidad de los datos sintéticos. **Al proporcionar definiciones, metodologías y criterios de evaluación claros, las normas pueden crear interpretaciones y puntos de referencia comunes para evaluar los datos sintéticos**. Los estándares permiten garantizar a las organizaciones que los datos sintéticos que utilizan cumplen con criterios específicos de calidad y privacidad. Además, los estándares de etiquetado y documentación de datos sintéticos, así como los mecanismos de auditoría y seguimiento de su procedencia, podrían ser un elemento clave para garantizar la transparencia, la equidad y la responsabilidad en la generación y el tratamiento de datos sintéticos.

El trabajo para la creación de estándares para datos sintéticos se está desarrollando por distintos actores. Por ejemplo, el Instituto de Ingenieros Eléctricos y Electrónicos (IEEE) lidera una iniciativa para desarrollar un estándar mundial y definir mejores prácticas para la seguridad de la privacidad de los datos sintéticos.[20] La Organización Internacional de Normalización (ISO) está trabajando en iniciativas similares.[21] El Open Data Institute (ODI) ha contribuido al desarrollo de una herramienta llamada "Croissant", un estándar comunitaro que proporciona metadatos legibles por máquina para conjuntos de datos, ayudando a normalizar la documentación de los conjuntos de datos de aprendizaje automático.[22] Los estándares ayudaran a proporcionar un marco y parámetros para la innovación responsable, incentivando las buenas prácticas en el sector privado para la generación, el uso y la innovación de datos sintéticos.

3.3 Datos sintéticos y gobernanza internacional de la IA militar

Los datos tienen cada vez más importancia militar debido a que la IA está cada vez más presente en muchos aspectos de este ámbito. Ante esto y dado que, en el ámbito militar, los datos sintéticos se utilizan principalmente para el entrenamiento y el desarrollo de diversas capacidades de IA, los debates sobre la gobernanza relacionados con el impacto de los datos sintéticos en la seguridad internacional deberían abordarse en el contexto de la gobernanza de la IA militar.

Los datos, y las cuestiones relacionadas, se consideran una de las áreas de trabajo prioritarias para la IA responsable en el ámbito militar.[23] **Sin embargo, estas cuestiones no ocupan un lugar central en las iniciativas en curso sobre gobernanza**, ya que, en gran medida, siguen siendo de alto nivel y carecen de granularidad. Sin embargo, los debates **en torno a los datos sintéticos, su gobernanza y, lo más importante, su potencial impacto en las iniciativas de gobernanza de la IA militar están aún en etapas muy iniciales.** Se han mantenido algunos debates a nivel regional, aunque igualmente incipientes, especialmente en aquellas partes del mundo que están aún en proceso de digitalización, donde las lagunas de datos suponen una importante barrera para acceder a la IA militar y donde alternativas como los datos sintéticos pueden servir de sustitutos.

a. Sobre la novedad de los retos y la aplicabilidad de los marcos existentes

En el ámbito militar, los datos sintéticos plantean problemas similares, con implicaciones parecidas, a las de los datos reales, como los sesgos (tanto su perpetuación, exacerbación o creación), los problemas de fiabilidad y representación, la rendición de cuentas, la rastreabilidad o la falta de explicabilidad entre otros. Por lo tanto, **no se puede seguir excluyendo a los datos sintéticos ni pueden seguir siendo un tema inexplorado en el contexto de la IA militar.** Es importante señalar que esto podría dar lugar a que ciertos tipos de datos queden fuera del ámbito

de las discusiones sobre gobernanza, con la posibilidad de que se perpetúen los riesgos relacionados a los datos y, por lo tanto, que se exacerben en la IA militar.

Por ejemplo, garantizar la *rendición de cuentas* de los datos —un aspecto clave de la responsabilidad— podría complicarse aún más en el contexto de los datos sintéticos. De hecho, el uso de datos sintéticos introduce un eslabón adicional de personas, a veces agentes externos responsables de su generación, lo que dificulta el seguimiento de la responsabilidad directa en caso de errores. En relación con lo anterior, los datos sintéticos podrían agravar los problemas de *explicabilidad*, debido a **la falta de normas consensuadas internacionalmente sobre la generación, el uso y el etiquetado de los conjuntos de datos sintéticos.** La falta de claridad sobre la procedencia de los datos debido a las limitaciones de rastreabilidad podría obstaculizar la capacidad de las auditorías para abordar los sesgos en los conjuntos de datos. Además, la democratización del acceso a datos gracias a los datos sintéticos podría ser una oportunidad para ampliar el acceso para el desarrollo de la IA y a otras capacidades digitales. Por un lado, esto podría ayudar a abordar los problemas relacionados con la brecha digital. Por otra parte, en el contexto de seguridad internacional, también podría actuar como facilitador de **una mayor proliferación de capacidades militares de IA** al bajar el nivel de acceso al desarrollo de modelos avanzados de IA[24]. Sin embargo, más que nuevos retos, se trata de complicaciones adicionales a los retos ya existentes en relación al uso de datos en el ámbito militar y que requieren una atención y aclaración específicas en el contexto de los datos sintéticos.

Por lo tanto, los debates sobre gobernanza en torno al impacto de las nuevas tecnologías en la seguridad internacional deberían centrarse, idealmente, en si los marcos jurídicos y normativos existentes aplican, y de qué manera. En el contexto de los datos sintéticos, es importante analizar si estos retos de gobernanza son nuevos, o si solo complican los retos preexistentes y en qué medida los conceptos vigentes de la gobernanza de datos se aplican a los datos sintéticos. Por lo tanto, en vez de diseñar necesariamente nuevos marcos o planteamientos de gobernanza, la comunidad internacional debería fijarse en las prácticas y conceptos ya establecidos, como la equidad, la responsabilidad, la rastreabilidad y la fiabilidad, y trabajar para aplicarlos o adaptarlos a los datos sintéticos, o usarlos para construir a partir de ellos. En este contexto, la comunidad internacional no parte de cero y puede aprovechar el extenso corpus de trabajo ya existente sobre qué son los "buenos datos" y los conocimientos emergentes sobre cómo son las buenas prácticas de datos en el ámbito militar.

Recuadro 1.
Área de futura investigación: el comercio internacional de conjuntos de datos sintéticos

Un área importante que requiere más investigación es la de las posibles implicaciones del comercio internacional de conjuntos de datos sintéticos, y si debería controlarse, supervisarse o, en algunos casos, restringirse dicho comercio, y cómo hacerlo. De hecho, podría desarrollarse

un mercado para comerciar con conjuntos de datos sintéticos que podrían utilizar agentes maliciosos en el desarrollo de capacidades disruptivas de IA. Por ello, la comunidad internacional debería considerar cómo el comercio de conjuntos de datos sintéticos interactúa con los esfuerzos de no proliferación y control de armas. Habría que estudiar la conveniencia de controlar algunos conjuntos o algún tipo de datos sintéticos mediante herramientas como las listas para el control de las exportaciones.

b. Sobre la importancia de un planteamiento multipartita

Los marcos de gobernanza que permitan un aprovechamiento más efectivo de las ventajas de los datos sintéticos para el ámbito militar serán aquellos que **representen a las múltiples partes interesadas.** Esto implica no solo la cooperación entre los Estados**, sino también una estrecha colaboración con los agentes del sector privado, que deberían de participar en estos debates sobre gobernanza.** El sector privado desempeña un papel fundamental en la IA militar, ya que es el principal responsable de la investigación y el desarrollo de las tecnologías básicas. En este contexto, los datos sintéticos no son una excepción, ya que la mayor parte de las capacidades de generación, así como de prueba y evaluación de conjuntos de datos sintéticos, recae en los agentes de la industria privada. **Esto crea una dependencia adicional a las empresas privadas de tecnología**, en especial para los Estados con menos recursos que pueden no disponer de capacidades independientes de prueba y evaluación de datos sintéticos y que dependen del sector privado para garantizar la calidad de los conjuntos de datos sintéticos. Esta dependencia requiere marcos para las asociaciones público-privadas que den prioridad a **la creación de confianza entre los gobiernos y la industria**. Esta confianza es crucial para garantizar que los actores del sector privado participen en debates sobre gobernanza y adopten prácticas responsables en el desarrollo, implantación y prueba de datos sintéticos para sistemas militares de IA. Además, los planteamientos multipartitas contribuyen a la tan necesaria creación de un lenguaje y un entendimiento común de los datos sintéticos.

c. Sobre las directrices y la especificidad del contexto

Aunque el desarrollo de directrices claras para la generación y el tratamiento de datos sintéticos debería ser un objetivo, se ha observado que, especialmente en un contexto militar, su desarrollo podría ser prematuro. De hecho, las directrices claras suelen basarse en buenas prácticas bien definidas. Sin embargo, **en el caso de los datos sintéticos, es posible que el campo esté aún demasiado incipiente para definir unas mejores prácticas definitivas.** En este contexto, podría resultar contraproducente normalizar los procedimientos de prueba o establecer directrices estrictas antes de conocer a fondo las capacidades, las ventajas, las limitaciones y los riesgos potenciales de la tecnología.

Además, **es posible que unas directrices claras no se adapten a la naturaleza altamente ligada al contexto de evaluar la idoneidad y el nivel de responsabilidad del uso de datos sintéticos en un entorno militar.** De hecho, un determinado conjunto de datos sintéticos podría utilizarse de forma "responsable" en un escenario, mientras que su uso en otro contexto podría considerarse

"irresponsable". Esta naturaleza dependiente del contexto del uso de datos sintéticos dificulta la elaboración de directrices de aplicación universal que aborden con efectividad los matices de los distintos casos.

Además, las métricas de calidad también dependen del contexto. Por ejemplo, la proximidad con la que un conjunto de datos sintéticos representa la realidad se utiliza como indicador clave de su calidad.[25] Sin embargo, en algunos casos, sobre todo en el ámbito militar, lo que se pretende es precisamente desviarse de la realidad. En otras palabras, **el uso de datos sintéticos para representar escenarios sin precedentes que ayuden a la planificación creativa podría ser una de las ventajas que aporta el uso de datos sintéticos en el contexto de las operaciones militares.**

Además, los planteamientos de gobernanza van a requerir tener en cuenta los contextos regionales y nacionales. Debido a las dependencias externas que pueden crear los datos sintéticos —como la necesidad de agentes externos para generar los conjuntos de datos sintéticos y garantizar su calidad— **es especialmente importante garantizar que los conjuntos de datos sintéticos creados fuera de una región determinada reflejen las realidades locales en el contexto del uso previsto.** Esto va a exigir que los parámetros y los supuestos de los conjuntos de datos sintéticos sean transparentes y dejen claro por qué, cómo, para qué y por quién se crean los datos sintéticos.

d. Sobre las oportunidades de gobernanza en el ámbito militar

Los datos sintéticos no solo plantean retos de gobernanza, sino que también presentan **oportunidades, especialmente para la gobernanza de la IA militar.** De hecho, los datos sintéticos pueden facilitar el **intercambio de datos** entre fuerzas armadas y contribuir al **desarrollo común de capacidades militares de IA.** Por ejemplo, el potencial de los datos sintéticos para preservar la privacidad podría hacer posible el compartir conjuntos de datos, algo a menudo deseable, pero que impide el carácter sensible y clasificado de los datos militares. En el ámbito militar, esto presenta un enorme valor tanto a nivel interno como entre organizaciones gubernamentales y naciones.

Los datos sintéticos podrían servir como un "terreno neutral" para proyectos colaborativos de IA militar entre distintas naciones. Usando conjuntos de datos sintéticos que reflejen escenarios del mundo real, pero que no contengan información sensible, **los Estados podrían colaborar para desarrollar y probar sistemas de IA, mejorar la interoperabilidad y compartir las mejores prácticas sin los riesgos asociados al intercambio de datos militares reales.** Además, los Estados pueden colaborar en el desarrollo de conjuntos de datos sintéticos colectivos que puedan utilizarse para entrenar y probar los sistemas de IA con el fin de mejorar la interoperabilidad, una cuestión clave en el desarrollo de las capacidades militares de IA. Esta estrategia de colaboración podría fomentar una mayor cohesión entre las fuerzas aliadas, mejorar

la eficacia de las operaciones conjuntas y contribuir a un entorno internacional más estable y seguro en el contexto de la IA militar.

Además, los datos sintéticos representan una oportunidad para que los Estados desarrollen normas y directrices comunes sobre su generación y uso en el ámbito militar. Debido a su naturaleza incipiente, los debates en torno a los datos sintéticos ofrecen a la comunidad internacional la oportunidad de desarrollar **marcos de responsabilidad compartida.** Varios Estados podrían acordar principios de gobernanza y empezar a compartir buenas prácticas para avanzar juntos de forma sistemática hacia el establecimiento de buenas prácticas de generación y uso de datos sintéticos. Un grupo de trabajo multilateral sobre gobernanza de datos podría, por ejemplo, abordar conjuntamente algunas cuestiones y proporcionar un foro para el desarrollo de procedimientos, procesos y normas de responsabilidad.

4. Conclusión

Los datos sintéticos presentan un potencial significativo para el avance de las capacidades de IA tanto en el ámbito civil como en el militar. Sus ventajas —resolver la escasez de datos, mejorar la privacidad y facilitar la creación de conjuntos de datos más representativos y menos sesgados— la convierten en una herramienta poderosa. Sin embargo, los datos sintéticos no son la panacea y su uso conlleva riesgos inherentes. Para sacar el máximo provecho de esta tecnología, es de suma importancia que los debates sobre gobernanza empiecen a tener en cuenta esta cuestión. Estos esfuerzos se encuentran actualmente en un estado prematuro, tanto en el ámbito civil como en el militar, y siguen existiendo ambigüedades legales y normativas sobre la generación, el tratamiento y el uso de datos sintéticos. Para evitar un vacío legal y normativo que deje sin abordar los riesgos de los datos sintéticos, hay que trabajar en la identificación de las brechas de los marcos existentes y en ofrecer claridad a los usuarios y generadores de datos sintéticos. Para ello, es fundamental la creación de directrices, la elaboración de normas técnicas internacionales y la cooperación con la industria.

En el ámbito militar, las cuestiones relacionadas con los datos siguen sin abordarse completamente en los debates sobre la gobernanza de la IA. En este contexto, hay que centrarse, no solo en asegurarles un lugar dentro de estas iniciativas, sino en considerar específicamente los efectos del uso de datos sintéticos en el ámbito militar. Dado que los datos sintéticos, más que crear un panorama totalmente nuevo, añaden complejidad a los problemas de gobernanza ya existentes, no hay por qué desarrollar nuevos marcos o normativas. Podría bastar con aplicar las mejores prácticas y los conceptos de datos militares a la generación y el uso de datos sintéticos. Por lo que es necesario seguir trabajando en la cuestión para ampliar los marcos emergentes de gobernanza de la IA militar a los datos sintéticos, aclarando cómo pueden aplicarse tales prácticas y conceptos.

A medida que la prevalencia de los datos sintéticos aumenta, no solo se plantean problemas de gobernanza, sino también oportunidades de colaboración internacional que podrían tener efectos

positivos en la gobernanza mundial de la IA militar. Para ello, va a ser fundamental aunar esfuerzos entre las diversas partes interesadas de los Estados y, sobre todo, del sector privado.

De cara al futuro, los datos sintéticos no van a ser la última innovación en la ciencia de datos. Esto subraya la importancia de crear marcos de gobernanza adaptables a futuros desarrollos. Construir estos marcos con la flexibilidad como premisa va a ser esencial para su sostenibilidad y para garantizar su vigencia a medida que surjan nuevas tecnologías y aplicaciones.

Anexo: Programa y participantes

Observaciones preliminares

- **Federico Mantellassi,** investigador del Instituto de las Naciones Unidas de Investigación sobre el Desarme.

Desayuno tecnológico sobre datos sintéticos y seguridad internacional:

- **Dra. Eleonore Fournier-Tombs**, responsable de Acción Anticipatoria e Innovación del Centro de Investigación Política, Universidad de las Naciones Unidas.
- **Calum Inverarity**, investigador sénior del Open Data Institute.

Moderado por Wenting He, investigadora asociada del Instituto de las Naciones Unidas de Investigación sobre el Desarme.

Diálogo multilateral sobre datos sintéticos: Oportunidades y retos para la gobernanza internacional

- **Dra. Jane Pinelis**, ingeniera jefa de IA de la Subdivisión de Ciencias de la Información Aplicada del Laboratorio de Física Aplicada de la Universidad Johns Hopkins.
- **Aldo Lamberti,** fundador y CEO de Syntheticus; experto en la materia de la Comisión Europea; vicepresidente del Grupo de Expertos en Datos Sintéticos del IEEE; experto del grupo de trabajo Norma para los Requisitos de Seguridad y Fiabilidad en Modelos Generativos de Inteligencia Artificial (IA) Preentrenados, del IEEE.
- **Yasmin Afina**, investigadora del Instituto de las Naciones Unidas de Investigación sobre el Desarme; experta de la Comisión Global sobre Inteligencia Artificial Responsable en el Ámbito Militar.
- **Dra. Ana Beduschi,** profesora titular de Derecho con Cátedra Personal en la Universidad de Exeter; directora del Centro de Investigación sobre Ciencia, Cultura y Derecho de la Facultad de Derecho de la Universidad de Exeter.

Moderado por Federico Mantellassi, investigador del Instituto de las Naciones Unidas de Investigación sobre el Desarme