

Datos sintéticos y publicidad, desafíos legales en el uso de inteligencia artificial en campañas publicitarias

Malena Cárdenas Tello*

Resumen.- El artículo analiza los principales retos legales vinculados al uso de datos sintéticos desde la protección de datos personales, el principio de veracidad en la publicidad y su uso con el objetivo de simular un testeo publicitario en “consumidores” antes de que la campaña sea lanzada al mercado. Se revisarán sus implicancias legales, así como la regulación sobre esta materia a nivel mundial. Finalmente se abordarán las principales ventajas y desventajas que enfrentaría su uso.

Palabras clave.- Datos sintéticos - Campañas Publicitarias - Principio de Veracidad.

Abstract.- The article analyzes the main legal challenges associated with the use of synthetic data, focusing on personal data protection, the principle of truthfulness in advertising, and the use of such data to simulate advertising testing on “consumers” prior to the public launch of a campaign. The article analyzes its legal implications, as well as the emerging international regulatory approaches to this issue. It also discusses that mail advantages and potential drawbacks of its implementation.

Keywords.- Synthetic Data - Advertising Campaigns - Principle of Truthfulness.

* Magíster en Propiedad Intelectual e Industrial por la Universidad de Alicante, España. Abogada por la Universidad de San Martín de Porres. Actualmente se desempeña como Corporate Legal Counsel en BELCORP.
Las opiniones vertidas en este artículo son de exclusiva responsabilidad de la autora y no representan la posición de la empresa en la que labora.

I. Introducción

En un contexto donde la inmediatez en obtener resultados es cada vez más apreciada por el consumidor, las empresas tienen el desafío de aprovechar al máximo el tiempo entre el desarrollo de productos o servicios de alta calidad y comunicar las bondades que ofrecen con campañas publicitarias atractivas y que logren conectar con el público.

Así, el uso de inteligencia artificial se ha convertido en una herramienta cada vez más potente para lograr estos objetivos, aplicándola en diversas etapas del proceso de desarrollo de productos y servicios.

Especial atención empieza a cobrar en el campo de la publicidad, específicamente en el desarrollo de campañas publicitarias, en donde con el objetivo de optimizar tiempos se empiezan a utilizar distintos recursos, entre ellos podemos encontrar a los datos sintéticos, aquellos datos creados utilizando inteligencia artificial, para mejorar el proceso de validación del mensaje que se quiere comunicar al público, reduciendo significativamente el tiempo y costo que ello implicaría, sobre todo cuando el anunciante tiene más de una opción de ruta a seguir.

En este escenario se propone crear grupos de perfiles sintéticos creados con inteligencia artificial, diseñados específicamente para simular el comportamiento de consumidores reales. Para lo cual, en lugar de testear la mejor ruta publicitaria a seguir para el lanzamiento de una campaña publicitaria frente a un grupo de 400 consumidores, el uso de datos sintéticos generados con inteligencia artificial podría generar 400 “consumidores sintéticos” tomando en consideración distintos patrones como género, edad, intereses de consumo y reacciones emocionales.

Actualmente algunas empresas de investigación de mercado y comportamiento del consumidor ofrecen los servicios de obtención de resultados de análisis del usuario sobre la base de datos sintéticos. Sin embargo, su uso, aun cuando a primera impresión puede parecer una opción bastante favorable, no está exento de desafíos desde el punto de vista jurídico.

Conforme desarrollaremos en el presente artículo, es importante considerar las implicancias legales en especial desde el frente de la privacidad y sesgos, a fin de que su uso sea seguro y el adecuado cumplimiento de regulaciones como el Reglamento General de Protec-

ción de Datos (GDPR) y la implementación de políticas internas con el fin de garantizar la correcta anonimización y trazabilidad de los datos.

II. Concepto de Datos Sintéticos

Los datos sintéticos son datos artificiales que se generan utilizando un modelo especialmente diseñado, por ejemplo, modelos de aprendizaje automático (ML, por sus siglas en inglés *machine learning*), los cuales se pueden obtener entrenando un modelo con un conjunto de datos personales reales de origen, de modo que puedan imitar las características, patrones y estructura de los datos personales originales. Como resultado, el análisis de datos sintéticos puede producir resultados similares a los obtenidos con los datos de origen (Personal Data Protection Commission Singapore, 2024)¹.

Si bien el concepto de datos sintéticos puede ser muy técnico, para efectos del presente artículo nos referiremos únicamente a los datos sintéticos utilizados para entrenar un modelo con perfiles de preferencias de consumidores.

Así, este uso de datos sintéticos se les conoce como datos sintéticos tabulares debido al formato que se usa, los cuales deben ser consistentes con las estadísticas recogidas de usuarios reales, a fin de entrenar modelos de análisis de comportamientos.

Para ahondar en el tema de los datos sintéticos tabulares, se refieren a los datos generados artificialmente para replicar la distribución de un conjunto de datos reales y producir muestras sintéticas de alta fidelidad, a la vez que logren preservar la privacidad².

Además, en la medida que la data usada no es real, estos datos encajan dentro de tipo de dato totalmente sintético.

Actualmente, los datos sintéticos han despertado mayor interés en distintos sectores industriales debido a que son considerados como un tipo de tecnología de mejora de la privacidad (PET, por sus siglas

¹ Personal Data Protection Commission Singapore (2024). Proposed guide synthetic data generation <https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/other-guides/proposed-guide-on-synthetic-data-generation.pdf>

² SHI, R., WANG, Y., DU, M., SHEN, X., & WANG, X. (2025). A comprehensive survey of synthetic tabular data generation. arXiv.org. <https://arxiv.org/abs/2504.16506>.

en inglés *Privacy Technology*), en específico dentro de la categoría de ofuscación de datos, según la Organización para la Cooperación y el Desarrollo Económico (OCDE, 2023)³, de esta forma es posible tratar y analizar información para entrenar a la inteligencia artificial sin revelar los datos personales utilizados para el entrenamiento y con ello lograr una ventaja competitiva frente a otros competidores.

A diferencia de los métodos más clásicos de PET, como por ejemplo la técnica de anonimización de datos existe el riesgo de que los datos personales sean revelados aprovechando alguna brecha de seguridad, con los datos sintéticos se eliminaría este problema de raíz⁴ ya que los datos personales para entrenar al ML no serán compartidos.

III. Regulación comparada de los datos sintéticos

Hasta el momento el Perú no cuenta con ninguna norma que regule expresamente el uso de datos sintéticos, a diferencia de otras jurisdicciones como se desarrollará en el presente apartado; no obstante, se debe tener presente que cuando los datos sintéticos son creados a partir del uso de una base de datos personales, su tratamiento sí deberá cumplir con la Ley de Protección de Datos Personales⁵, así como su Reglamento⁶, puesto que en el referido escenario, los datos personales pasarán por un proceso de anonimización.

En contraste con el escenario que se presenta en Perú, la tendencia, en los últimos años, a nivel global ha sido que las autoridades competentes consideran necesario pronunciarse sobre el uso de los datos sintéticos, mediante la emisión de lineamientos para su uso de forma segura. A continuación, citaremos algunos casos a modo de ejemplo:

³ Organización para la Cooperación y el Desarrollo Económicos (OCDE). (2023). *Emerging privacy enhancing technologies: Current regulatory and policy approaches*.

⁴ Se indica “eliminaría” en condicional, debido a que más adelante profundizaremos sobre los posibles riesgos que implica el mal diseño de datos sintéticos.

⁵ Ley de Protección de Datos Personales, Ley N. 29733. Publicada el 3 de julio de 2011 en el Diario Oficial El Peruano.

⁶ Decreto Supremo 016-2024-JUS, que aprueba el Reglamento de la Ley N. 29733, Ley de Protección de Datos Personales.

Guía sobre la Generación de Datos Sintéticos – PDPC Singapur

- Esta guía desarrolla ampliamente el concepto de datos sintéticos, entendiéndolos como aquellos generados mediante modelos matemáticos que incluya el uso de inteligencia artificial, además precisa que existen tipos de datos sintéticos, enfocándose principalmente en aquellos datos completamente sintéticos, que se general sin usar data real, lo cual difiere de los datos parcialmente sintéticos, que combinar datos reales con generados.

También profundiza acerca de sus beneficios y principales riesgos, como sería el riesgo de reidentificar los datos de origen utilizados para generar datos sintéticos, para lo cual se deberán emplear técnicas como la adición de ruido o reducción de granularidad a fin de mitigar riesgos de reidentificación.

La guía emite además una serie de buenas prácticas para que las empresas que los empleen reduzcan los riesgos de seguridad, mediante la implementación de medidas técnicas y control de gobernanza del riesgo. Finalmente, brinda ejemplos de métodos de generación de datos sintéticos⁷.

Guía sobre IA y Protección de Datos- ICO UK

- En este documento se desarrolla el uso, en general, de inteligencia artificial en relación con los principios que deben cumplirse al procesar datos personales.
- La guía reconoce la existencia de los datos sintéticos como un mecanismo para mejorar la privacidad de los datos utilizados como fuente de entrenamiento de sistemas que emplean inteligencia artificial, los cuales a su vez generan datos sintéticos.
- Así, en caso las empresas utilicen datos sintéticos tiene el deber de documentar y auditar todo el proceso de generación y uso de datos con el fin de que puedan demostrar que garantizan la protección de los datos originales empleados como entrenamiento.

⁷ Personal Data Protection Commission. (2024). Proposed Guide on Synthetic Data Generation. <https://www.pdpc.gov.sg/help-and-resources/2024/07/proposed-guide-on-synthetic-data-generation>

Un punto no menor, es que las empresas que emplean los datos sintéticos también tienen el deber los modelos empleados, a fin de que eviten sesgos discriminatorios o resultados inexactos.

También refuerza la idea que, si los datos sintéticos no permiten identificar a personas específicas, no se considerarán datos personales y, por ende, las obligaciones de protección de datos no le aplicarán.

No obstante, en caso sí se lograra inferir información sobre aquellos datos personales originales a partir de los datos sintéticos, entonces sí deberán cumplir con las leyes de protección de datos⁸.

IV. Uso de los datos sintéticos en el lanzamiento de campañas publicitarias

No es novedoso señalar que una de las áreas en donde más se emplean herramientas entrenadas con inteligencia artificial es el marketing, ello ya que estas herramientas permiten analizar grandes volúmenes de data y personalizar mensajes publicitarios de manera mucho más exacta y rápida, frente a los métodos tradicionales.

De hecho, es común encontrar tanto empresas transnacionales como pequeños creadores de contenido en distintas redes sociales que emplean herramientas accesibles al público en general, logrando mejorar el alcance de su publicidad y comunicaciones o ajustando su estrategia en tiempo real según la respuesta que obtengas de la audiencia.

Sin embargo, poco se habla de que algunas empresas están empezando a incorporar dentro de sus procesos de testeo de rutas para el lanzamiento de campañas publicitarias herramientas más sofisticadas de evaluación, no tan difundidas, como por ejemplo aquellas que utilizan datos sintéticos.

Este tipo de herramienta permite que la empresa que la emplea pueda prever distintos escenarios de comportamiento de sus consumidores antes de que se implemente en la realidad, lo cual le generará un evidente

⁸ Information Commissioner's Office. (2023). Guidance on AI and Data Protection. <https://ico.org.uk/media2/ga4lfb5d/guidance-on-ai-and-data-protection-all-2-0-38.pdf>

beneficio en cuanto a la reducción de costos y en especial a riesgos de no obtener los resultados esperados si es que acaso se prescindiera de esta evaluación.

A efectos del presente artículo, es importante precisar que cuando mencionamos testeo de rutas, nos referimos únicamente al uso de datos sintéticos en la etapa de validación del mensaje publicitario y no respecto al testeo de los beneficios o de seguridad por el que debe pasar un producto antes de ser lanzado al mercado para no incurrir en infracciones de las normativas regulatorias, publicitarias y de protección al consumidor.

De esta forma, los datos sintéticos se generan sobre la base de datos personales como pueden ser ubicaciones demográficas y tendencias de comportamiento, los cuales son recopilados de una muestra representativa del público consumidor al que destinará su producto o servicio. Con ello, mediante el entrenamiento de la inteligencia artificial, es posible obtener una cantidad representativa y significativa de datos sintéticos, los cuales permitirán probar cada una de las rutas publicitarias propuestas y predecir el comportamiento del público objetivo.

El uso de datos sintéticos ha generado un creciente entusiasmo, ya que al ser generados de forma artificial permiten ampliar significativamente la cantidad de datos personales que se necesitan tratar y procesar con el fin de agilizar el proceso de testeo, sin tener que enfrentar las limitaciones o riesgos que, algunas veces, puede presentar el tratamiento de datos personales, en especial cuando se trata de vulneración a la seguridad de la información.

Los datos sintéticos también permiten explorar escenarios que, de usarse con datos personales, podrían resultar costosos y difíciles de evaluar. Por ejemplo, se podría usar datos sintéticos para probar distintas comunicaciones de un mismo mensaje o respecto a 3 rutas distintas en el diseño del empaque del producto, en un tiempo bastante acotado.

En definitiva, esta forma de utilizar los datos sintéticos permite recrear una suerte de audiencia sintética, la cual, mediante el procesamiento de distintas variables como lo son el historial de compras, el consumo de medios y análisis de sentimiento u otros comportamientos digitales, permite crear perfiles sintéticos que se comportan como consumidores

reales y permite obtener pronósticos precisos sobre la preferencia del público objetivo⁹.

Dentro de las principales ventajas que ha traído a la industria del marketing este tipo de uso de datos sintéticos, encontramos que se puede ver reducido significativamente el tiempo que toma experimentar distintos escenarios de comportamiento de los consumidores, al realizar una investigación de mercado sustituyendo las respuestas humanas con modelos predictivos o quizá utilizar panelistas humanos únicamente para las preguntas claves y complementar dicha información con muestras sintéticas obteniendo con ello resultados representativos. (Kalidas & Ostler, 2023¹⁰).

Sin perjuicio de ello, se debe considerar que existen ciertas limitaciones inherentes a su uso, ya que los datos sintéticos podrían reflejar sesgos presentes en los datos originales que se utilizan de entrenamiento. Por ello, se recomienda que la data obtenida de forma sintética sea utilizada de forma complementaria, incluso de requerirlo, haciendo una validación de los resultados con seres humanos para comparar el nivel de exactitud en las respuestas y así asegurar que la data obtenida sea confiable. (Kalidas & Ostler, 2023¹¹).

De hecho, de un estudio realizado por Qualtrics en donde se encuestaron a 3 000 investigadores de mercado de 14 países, se obtuvieron los siguientes resultados: 87% de aquellos encuestados que han utilizado respuestas sintéticas obtuvieron una alta satisfacción con los resultados arrojados. Además, un 71% de los encuestados coincide en que, dentro de tres años, la mayoría de la investigación de mercado se realizará utilizando respuestas sintéticas.

⁹ BRITO, M. Synthetic data: A new frontier in data-driven innovation. <https://www.britopian.com/research/synthetic-data/>

¹⁰ KALIDAS, A., & OSTLER, J. (2023). What is synthetic sample- and is it all it's cracked up to be? Kantar. <https://www.kantar.com/inspiration/analytics/what-is-synthetic-sample-and-is-it-all-its-cracked-up-to-be>

¹¹ KALIDAS, A., & OSTLER, J. (2023). What is synthetic sample- and is it all it's cracked up to be? Kantar. <https://www.kantar.com/inspiration/analytics/what-is-synthetic-sample-and-is-it-all-its-cracked-up-to-be>

Además, se reconoce que el correcto uso de datos sintéticos podría vencer ciertos desafíos a los que las empresas se enfrentan, como las limitaciones presupuestarias, los riesgos de privacidad y escasez de datos.

Finalmente, el 89% de los encuestados reconocieron estar usando herramientas impulsadas por inteligencia artificial en sus tareas diarias y el 83% de las empresas planean aumentar significativamente su inversión en herramientas de inteligencia artificial¹².

Pero, el uso de datos sintéticos no se encuentra libre de riesgos ni consideraciones claves que se deben seguir, lo cual implica que un uso inadecuado podría gatillar diversos riesgos. Entre los principales riesgos identificados están aquellos que vulnerarían la privacidad de la información utilizada para entrenar a la inteligencia artificial, los cuales eventualmente podrían ser reidentificados, o también la reproducción de sesgos existentes en los datos originales, que puedan existir al momento de crearlos, lo cual podría llevar a obtener resultados discriminatorios.

Si bien estos riesgos serán desarrollados más adelante, es importante tener en el radar estos puntos.

En principio, el desarrollar datos sintéticos que sirvan como muestras representativas implica que sean diseñados por profesionales expertos tanto en el uso de herramientas de inteligencia artificial, como en estadística y diseño de datos; de lo contrario la herramienta empleada podría no procesar correctamente los datos de muestra y terminar cayendo en las ya conocidas alucinaciones de los modelos de lenguaje de gran tamaño (LLM por sus siglas en inglés *Large Language Model*).

De igual forma, cuando se determine utilizar modelos de aprendizaje automatizados se debe tomar en cuenta que estos empleen medidas de seguridad suficientemente eficientes para resguardar la confidencialidad de los datos personales ingresados al momento de su entrenamiento; caso contrario, se podrían ver expuestos a brechas de seguridad que derive en las divulgaciones de la información consecuencia de una posible reidentificación, es decir comunicar los datos personales utilizados.

¹² Qualtrics. (2024). AI to drive massive changes to market research in 2025, Qualtrics Report Says. Qualtrics.

Otra consideración al momento de emplear datos sintéticos es que estos sean entrenados con datos personales exactos y representativos, de lo contrario, se corre el riesgo de obtener data inexacta o incluso que conlleven sesgos desde su generación.

Por ello, la calidad del insumo que se utilice, en este caso los datos de entrenamiento deben ser de alta calidad y representativos a fin de garantizar que los modelos de inteligencia artificial empleados logren reproducir datos confiables y útiles, en especial en el caso de testeo de piezas publicitarias, en donde las decisión se basan en patrones de comportamiento de consumidores.

Sobre este punto, consideramos relevante compartir una publicación realizada por Kantar, en la cual se realizó una prueba comparativa entre una muestra de datos sintética versus respuestas humanas reales, para lo cual emplearon un LLM estándar desarrollado por OpenAI como es GPT-4, lo que permitió en este estudio analizar las diferencias en exactitud y sesgos entre datos reales y datos sintéticos.

Los resultados obtenidos en esta prueba concluyeron que GPT-4 tiene un fuerte sesgo positivo en comparación con las obtenidas de los humanos encuestados, en especial en aquellas relacionadas a afirmaciones subjetivas como lo es la emoción que le puede generar un producto; mientras que cuando se trata de afirmaciones objetivas, como lo sería el precio de un producto, el resultado es más preciso.

Así, aún cuando los datos sintéticos pueden reaplicar patrones, existen ciertas limitaciones como los sesgos inherentes a los datos originales (Kalidas & Ostler, 2023)¹³.

Finalmente, un hecho no menor, es que los resultados de preferencia obtenidos con el empleo de datos sintéticos no deberán ser utilizados para comunicar supuestos resultados de percepción de consumidores sobre el uso de los productos.

En efecto, lamentablemente suelen darse casos en los que se observan piezas publicitarias engañosas que comunican porcentajes de resultado

¹³ KALIDAS, A., & OSTLER, J. (2023). What is synthetic sample- and is it all it's cracked up to be? Kantar. <https://www.kantar.com/inspiration/analytics/what-is-synthetic-sample-and-is-it-all-its-cracked-up-to-be>

sobre un supuesto beneficio, cuando en realidad se tratan de estadísticas de preferencia de los consumidores.

Sin embargo, si se emplean datos sintéticos para obtener estos resultados estadísticos, se estaría cometiendo un acto de engaño, toda vez que, mediante el empleo de data no obtenida en su integridad de un estudio de preferencia de los consumidores, se buscaría inducir a error sobre los beneficios de un producto.

Este tipo de acciones en el Perú podrían ser calificados como un acto de engaño, conforme a lo regulado a nivel peruano en el Decreto Legislativo N.1044, Ley de Represión de la Competencia Desleal.

V. Reflexión final: ¿es necesario regular el uso de datos sintéticos en Perú?

Hasta el momento hemos analizado el uso de los datos sintéticos como herramienta de evaluación en el lanzamiento de campañas publicitarias, así como los beneficios y riesgos que su uso inadecuado podría implicar. De igual forma hemos visto como algunas jurisdicciones han considerado necesario pronunciarse de forma directa, como en el caso de la Autoridad de Singapur, o de forma indirecta, como se muestra en la guía emitida por la Autoridad de datos del Reino Unido.

Frente a ello, resulta pertinente cuestionar si es necesario que en el Perú se regule de forma expresa el uso de datos sintéticos, tomando en cuenta que ya se vienen empleando numerosas herramientas de inteligencia artificial no solo en la elaboración de piezas publicitarias, sino en la mayor parte del desarrollo de productos.

Desde la evaluación desarrollada en este artículo, consideramos que, si bien, no será necesario emitir una norma específica que regule como tal el uso de datos sintéticos, es recomendable que se emita una guía que permita empezar a conocer con mayor profundidad los datos sintéticos, incluyendo definiciones claras sobre el concepto, criterios para su creación, técnicas para reducir el riesgo de brechas de seguridad y recomendaciones sobre gobernanza de procesos.

Esto permitirá que tanto grandes como pequeñas empresas puedan incorporar los datos sintéticos de forma segura en sus procesos, sin dejar de lado aquellos mecanismos de mejora de la privacidad de los datos que se utilicen como fuente de entrenamiento.

En ese sentido, el contar con esta guía fomentaría la innovación y competitividad de la industria mediante el uso de datos sintéticos, logrando optimizar tiempo y reduciendo las posibilidades de que los datos personales de entrenamiento sean revelados como consecuencia de una brecha de seguridad e infrinjan la Ley de Datos Personales.

Consideramos que esta propuesta seguiría la ruta ya trazada por otras jurisdicciones, pero adaptándolo al contexto regulatorio peruano, sin que ello implique frenar el uso de herramientas innovadoras.

VI. Conclusiones

El uso de datos sintéticos en la industria del marketing viene cobrando mayor relevancia, en especial como una herramienta para agilizar la etapa de validación del mensaje publicitario, lo cual permitiría a las empresas evaluar diferentes estrategias sin que implique un uso de datos personales como tal.

Así, los resultados evaluados a lo largo del presente artículo arrojan un panorama alentador para su uso, desplegando múltiples de beneficios como podría ser la mejora en la eficiencia de procesos de análisis y reducción de costos versus la disminución de riesgos de privacidad.

Sin embargo, su uso no está exento de posibles riesgos que podría enfrentar, los cuales podrían ir desde sesgos en los modelos que se entrenen hasta reidentificación de los datos personales usados como entrenamiento; lo cual, a su vez, impactaría en el objetivo de testeo que se pretende obtener.

Por ello, diversas jurisdicciones vienen orientando a los usuarios a través de la emisión de guías o lineamientos de buenas prácticas, como la guía publicada por la PDPC de Singapur o el ICO del Reino Unido, ofrecen una orientación para la generación de datos sintéticos de forma responsable con el objetivo de reducir los riesgos de seguridad que su uso conllevaría.

Sin perjuicio de ello, en tanto los datos sintéticos configura una herramienta que se encuentra en constante evolución y que los avances en el uso de inteligencia artificial son cada vez más rápidos, es evidente que de implementarse una regulación rígida podría limitar la evolución, lo cual no sería la forma más adecuada de enfrentar los riesgos de privacidad u otros riesgos que aun no hayan sido detectados.

En especial en aquellos países como el Perú, donde no se observa un vasto desarrollo ni se han identificado plenamente cuáles serían los riesgos. Por ello, sería más beneficioso optar por tener un enfoque flexible y que sirva para educar al mercado.

De esta forma, podría ser de mayor utilidad promover la implementación de guías explicativas o recomendaciones y buenas prácticas para su uso, incluyendo incluso pautas para su aplicación considerando la privacidad, seguridad, mitigación de sesgos y auditoría de riesgos, con el objetivo de que las empresas puedan adentrarse al mundo de los datos sintéticos de forma responsable.

Consideramos que esta propuesta lograría fomentar la innovación sin perder de vista la obligación de respetar los datos personales.