El Comité de Ética informa

Los datos sintéticos y sus potenciales problemas éticos

Si hay un tema del que en nuestra industria se esté hablando mucho últimamente es el de los datos sintéticos. Pero *¿qué son los datos sintéticos?* Esta es una pregunta fácil y difícil de responder al mismo tiempo. La definición sencilla sugiere que los datos sintéticos se refieren al uso de IA para crear simulaciones de participantes de investigación, de manera que se pueda realizar la investigación con ellos en lugar de con personas "reales". La definición más compleja abarca datos generados por cualquier tipo de algoritmo y que no son datos primarios. Es decir, que no es algo tan nuevo, ya que hemos estado utilizando algunas formas de datos sintéticos durante muchos años. He aquí algunos ejemplos:

- Añadir "ruido" para proteger el anonimato: Durante décadas, las autoridades censales han añadido ruido a los registros del censo para que la información pueda compartirse con investigadores sin riesgo de que estos puedan identificar a los individuos.
- Imputación de datos: A veces tenemos datos faltantes, por ejemplo, cuando las personas no responden a una pregunta o si no se les preguntó. En su forma más simple, utilizando el algoritmo básico, el valor faltante se reemplaza por la media de la muestra. Estos valores imputados son datos sintéticos.
- Fusión de datos: La fusión de datos se refiere a tomar dos conjuntos de datos, por ejemplo, un conjunto de datos de consumo de medios y un conjunto de información sobre compras, y combinarlos para crear un solo conjunto de datos que parece contener datos de personas que proporcionaron datos de medios y de compras.

Pero ¿qué nuevos usos de datos sintéticos están surgiendo?

En general, todos los LLM ("large language models" por sus siglas en inglés, modelos de lenguaje de gran tamaño, en español) son datos sintéticos: buscan predecir lo que una persona real diría si se le hace una pregunta específica. Se le puede pedir a un LLM que adopte diferentes personalidades para que responda la misma pregunta con diferentes respuestas tanto en un estudio cualitativo como en preguntas abiertas de un cuantitativo, sustituyendo a participantes reales por virtuales (a veces llamados datos aumentados).

Estos datos sintéticos ofrecen ventajas como la generación rápida y barata de grandes volúmenes de datos, y pueden tener gran interés en estudios exploratorios o para testar o evaluar propuestas de cuestionarios o metodologías.

Sin embargo, los datos sintéticos plantean varias preocupaciones éticas:

- Validez: Existen dudas razonables sobre si los datos sintéticos pueden proporcionar resultados válidos y confiables. La capacidad de los LLMs para generar respuestas plausibles pero no necesariamente precisas es una preocupación muy real.
- Dependencia de datos reales: Los modelos de datos sintéticos se entrenan con datos reales. Si los datos reales no son de calidad o están sesgados, la precisión y la relevancia de los datos sintéticos será baja.
- Privacidad y anonimato: Aunque los datos sintéticos pueden proteger la identidad de los individuos, existe el riesgo de que los datos puedan ser revertidos para identificar a personas específicas.
- Impacto económico: Las empresas que generan datos sintéticos podrían competir deslealmente con aquellas que recopilan y venden datos reales, lo que plantea cuestiones sobre el uso justo de la información y la remuneración.

Para aprovechar los beneficios de los datos sintéticos y mitigar sus riesgos, desde el Comité de Ética recomendamos:

- Evaluación rigurosa: Implementar métodos estrictos para evaluar la validez y confiabilidad de los datos sintéticos en diferentes contextos y aplicaciones.
- Ética y transparencia: Establecer directrices éticas claras sobre el uso de datos sintéticos y asegurar que los clientes comprendan la procedencia y las limitaciones de estos datos.
- Protección de la privacidad: Desarrollar técnicas robustas para asegurar que los datos sintéticos no puedan revertirse para identificar a individuos.
- Educación y formación: Capacitar a los investigadores y profesionales del mercado en el uso adecuado y ético de los datos sintéticos.

En resumen, mientras que los datos sintéticos ofrecen un futuro prometedor para la investigación de mercado, es crucial abordar de manera proactiva las preocupaciones éticas y metodológicas para garantizar que su uso sea beneficioso y justo para todos los actores involucrados.

Si tiene dudas sobre esta u otras cuestiones éticas, no dude en contactar con el Comité de Ética en el correo electrónico: etica@ia-espana.org. Se garantiza la confidencialidad.

Para más información sobre datos sintéticos:

https://researchworld.com/articles/synthetic-data-a-game-changer-in-data-driven-decision-making

https://newmr.org/blog/synthetic-data-an-overview-a-taxonomy-some-faqs/

https://newmr.org/blog/syntheticdata/

https://www.mrs.org.uk/campaign/id/synthetic?MKTG=SYNTHETIC