Guía para la gobernanza de los datos de la IA

¿Ha oído alguna vez la frase «la basura entra, la basura sale»? Así es exactamente como funciona la IA en términos de los datos utilizados para entrenarla. Si entran datos malos (es decir, inexactos o incompletos), saldrá una IA maliciosa. Si entran datos buenos (es decir, precisos y completos), saldrá una buena IA. Con IA “buena” nos referimos a la justicia y la precisión.

El problema es la complejidad y la cantidad de datos utilizados para entrenar los modelos de IA y aprendizaje automático. Es mucho lo que hay que gestionar. De ahí la necesidad de gobernanza de los datos de la IA, que es básicamente solo la manera – lo que significa políticas y procesos – que las organizaciones se aseguran de que los datos que alimentan en sus modelos de IA sean seguros, precisos, relevantes y exhaustivos.

Una buena gobernanza de los datos de la IA mejora el rendimiento del modelo, aumenta la fiabilidad, genera confianza y conduce a unos resultados éticos de la IA que no sesgan de ninguna manera. Todas las grandes victorias, ¿verdad?

Siga leyendo para explorar todos los elementos esenciales de la gobernanza de los datos de IA para los datos de entrenamiento, incluidos:

- Los principios y objetivos clave de la gobernanza de los datos de IA para los datos de entrenamiento
- Retos comunes en la gobernanza de los datos de la IA
- Las mejores prácticas para establecer unos marcos de gobernanza de datos de IA efectivos

Componentes clave de la gobernanza de los datos de la IA

Probablemente no sea difícil imaginar lo que comprende la gobernanza de los datos, pero definir y explorar estos términos podría ser útil.

Por supuesto, hay una «calidad de datos» general. Esto puede ser muy subjetivo, pero básicamente se debe a la coherencia, la exhaustividad y la corrección de los datos, lo que significa la eliminación de errores, duplicados e información irrelevante. La calidad también implica datos inadecuados. Por ejemplo, esto se remonta a hace un tiempo, pero en 2016 el chatbot de IA de Microsoft, Tay, se convirtió en un desastre de relaciones públicas cuando empezó a escupir respuestas racistas. ¿Por qué? Porque había aprendido sus valores y lenguaje de Twitter.

Una buena gobernanza de los datos también prioriza el cumplimiento de las normativas de privacidad, como el RGPD o la CCPA. La auditoría está aquí para garantizar la anonimización y minimización de los datos, el consentimiento del usuario y la transparencia sobre el uso de los datos. Las multas del RGPD pueden haber disminuido, pero eso no significa que las organizaciones no deban seguir

estando alerta.

La seguridad es otro aspecto importante de la gobernanza de los datos de la IA. La protección de los datos de entrenamiento confidenciales y privados frente al acceso no autorizado o las vulneraciones implica la implementación de mecanismos de cifrado y control de acceso sólidos y la supervisión de las vulnerabilidades y el uso no autorizado de los datos. Las empresas también tienen que garantizar un almacenamiento y una transmisión seguros de los conjuntos de datos de entrenamiento.

Retos de la gobernanza de los datos de la IA

Ciertas cosas comunes tienden a dificultar la gobernanza de los datos de la IA.

Todos hemos oído hablar de “silos de datos”, por ejemplo. Los silos de datos dificultan la gestión cohesiva de los datos de entrenamiento de la IA, lo que provoca incoherencias e ineficiencias. ¿Qué puede ayudar con los silos de datos? Implementar repositorios de datos centralizados o arquitecturas de lagos de datos para consolidar los conjuntos de datos. También puede usar herramientas y plataformas de integración de datos para optimizar el acceso y garantizar la coherencia.

La falta de estandarización es otro problema. Las diversas fuentes y formatos de datos pueden dificultar la gestión y el control efectivos, reduciendo la interoperabilidad y complicando los flujos de trabajo de preprocesamiento y entrenamiento. A veces es difícil saber de dónde vienen sus datos. La estandarización de los formatos, el etiquetado y los metadatos puede ayudar mucho a facilitarlo. También puede desarrollar prácticas de documentación sólidas y mantener un registro de auditoría claro para todos los procesos de datos. Utilice herramientas que realicen un seguimiento del linaje, las transformaciones y el uso de los datos en todo el ciclo de vida de la IA.

Además, la gran cantidad y la variedad de datos de entrenamiento necesarios para los modelos de IA y ML pueden ser asombrosos. Los datos no estructurados, por ejemplo, (por ejemplo, texto, imágenes, vídeo) presentan su propio conjunto de problemas de almacenamiento y análisis. Invertir en una infraestructura

escalable, como las plataformas basadas en la nube y el almacenamiento de datos por niveles, puede ayudar con ello.

Conclusión

Recuerde: entrada y salida de basura. Garantizar que no está alimentando la basura de sus modelos de IA es un esfuerzo para toda la empresa que requiere una colaboración cruzada masiva. Este tipo de cosas no hace falta decirlo, pero su empresa debería tener políticas completas que cubran la recogida, el almacenamiento, el uso y la conservación de los datos.

Una parte enorme de todo lo anterior es tener la infraestructura de datos adecuada para apoyar sus iniciativas de IA. La plataforma de Pure Storage ayuda a las organizaciones a maximizar el rendimiento y la eficiencia, unificar sus datos, simplificar la gestión del almacenamiento de datos y resolver la imprevisibilidad del crecimiento de la IA. FlashBlade® de Pure Storage® es una solución de almacenamiento certificada para NVIDIA DGX SuperPOD y Pure Storage fue uno de los primeros proveedores de almacenamiento empresarial que trabajó con NVIDIA en soluciones de infraestructura preparadas para la IA certificadas que amplían y aceleran la adopción de la IA.

Obtenga más información sobre cómo puede prepararse para el futuro y acelerar sus resultados de IA con Pure Storage.