***Unidad 1.4: Éticas, sesgos y responsabilidad en Sistemas de Inteligencia Artificial***

***Éticas:*** La ética de la IA es un campo multidisciplinario que estudia cómo optimizar el impacto beneficioso de la inteligencia artificial (IA) mientras se reducen los riesgos y los resultados adversos. Algunos ejemplos de cuestiones éticas de la IA son la responsabilidad y la privacidad de los datos, la imparcialidad, la explicabilidad, la robustez, la transparencia, la sustentabilidad medioambiental, la inclusión, la agencia moral, la alineación de valores, la responsabilidad, la confianza y el mal uso de la tecnología.

A medida que salieron a la luz casos de resultados injustos, surgieron nuevas directrices, principalmente de las comunidades de investigación y ciencia de datos, para abordar las preocupaciones en torno a la ética de la IA. Las compañías líderes en el campo de la IA también mostraron un interés personal en dar forma a estas directrices, ya que ellas mismas comenzaron a experimentar algunas de las consecuencias de no mantener los estándares éticos en sus productos. La falta de diligencia en esta área puede generar exposición reputacional, regulatoria y legal, lo que se deriva en sanciones costosas. Como ocurre con todos los avances tecnológicos, la innovación tiende a superar la regulación gubernamental en campos nuevos y emergentes. A medida que se desarrolle la experiencia adecuada dentro de la industria gubernamental, podemos esperar que las compañías sigan más protocolos de IA, lo que les permitirá evitar cualquier violación de los derechos humanos y las libertades civiles.

Los marcos éticos en el contexto de la inteligencia artificial (IA) son críticos para guiar el diseño, desarrollo y despliegue de sistemas de IA. Estos marcos se basan en filosofías éticas establecidas y son esenciales para abordar las complejidades y desafíos que plantean las tecnologías de IA. La aplicación de principios éticos fundamentales, como el impacto, la justicia y la autonomía, la beneficencia, la no maleficencia, la privacidad, la transparencia y la responsabilidad, se ha propuesto para guiar las discusiones en torno a las implicaciones éticas de la IA. Esta convergencia destaca la necesidad de incorporar consideraciones éticas desde el principio en el proceso de diseño para asegurar que los sistemas de IA beneficien a la sociedad mientras preservan los derechos individuales.

La comunidad académica ha aprovechado el Informe Belmont como medio para orientar la ética dentro de la investigación experimental y el desarrollo algorítmico. Hay tres principios fundamentales que se desprenden del Informe Belmont y que sirven de guía para el diseño de experimentos y algoritmos, que son: Respeto a las personas, beneficencia y justicia.

***Sesgo:*** Los sistemas de IA pueden heredar, al igual que los humanos en el proceso de su educación, y amplificar involuntariamente los sesgos presentes en sus datos de entrenamiento, lo que lleva a resultados discriminatorios en aplicaciones como la contratación, los préstamos y la aplicación de la ley. Garantizar la equidad en los algoritmos de IA es primordial, ya que los juicios injustos pueden surgir de sesgos cognitivos que a menudo no son reconocidos por sus creadores humanos. Abordar estos sesgos requiere una monitorización y validación continuas de los modelos de IA para mantener su confiabilidad.

Los modelos en los que se basan los esfuerzos de IA absorben los sesgos de la sociedad que pueden estar silenciosamente incrustados en las montañas de datos con los que se entrenan. La recopilación de datos con sesgos históricos que reflejen la desigualdad social puede resultar perjudicial para los grupos históricamente marginados en casos de uso como la contratación, la policía, la calificación crediticia y muchos otros.

El sesgo de IA puede provenir de varias fuentes que pueden afectar a la equidad y confiabilidad de los sistemas de IA:

* Sesgo de datos: los sesgos presentes en los datos usados para entrenar modelos de IA pueden conducir a resultados sesgados. Si los datos de entrenamiento representan predominantemente determinados datos demográficos o contienen sesgos históricos, la IA reflejará estos desequilibrios en sus proyecciones y decisiones.
* Sesgo algorítmico: esto ocurre cuando el diseño y los parámetros de los algoritmos introducen el sesgo inadvertidamente. Incluso si los datos son imparciales, la forma en que los algoritmos procesan y priorizan ciertas características sobre otras puede resultar en resultados discriminatorios.
* Sesgo de decisión humana: el sesgo humano, también conocido como sesgo cognitivo, puede filtrarse en los sistemas de IA a través de decisiones subjetivas en el etiquetado de datos, el desarrollo de modelos y otras etapas del ciclo de vida de la IA. Estos sesgos reflejan los prejuicios y sesgos cognitivos de las personas y los equipos involucrados en el desarrollo de las tecnologías de IA.
* Sesgo de IA generativa: los modelos de IA generativa, como los usados para crear texto, imágenes o videos, pueden producir contenido sesgado o inapropiado en función de los sesgos presentes en sus datos de entrenamiento. Estos modelos pueden reforzar los estereotipos o generar resultados que marginan ciertos grupos o puntos de vista.

***Responsabilidad:*** se refiere a la obligación de utilizar esta tecnología de manera ética y segura, minimizando riesgos y asegurando que sea beneficiosa para la sociedad. A medida que las tecnologías de IA evolucionan, la cuestión de la responsabilidad se vuelve cada vez más compleja. La opacidad de muchos modelos de IA complica los esfuerzos para garantizar la transparencia y la equidad en sus procesos de toma de decisiones. Esto plantea preocupaciones sobre quién es responsable de las acciones de los sistemas autónomos, particularmente en áreas de alto riesgo como la atención médica y los vehículos autónomos, donde la seguridad y la confiabilidad son primordiales. No existe una legislación universal y general que regule las prácticas de IA, pero muchos países y estados trabajan para desarrollarlas y aplicarlas a nivel local. En la actualidad existen algunas normas sobre IA, y muchas otras están a punto de entrar en vigor. Para llenar el vacío, han surgido marcos éticos como parte de una colaboración entre especialistas en ética e investigadores para regir la creación y distribución de modelos de IA dentro de la sociedad.

Algunos principios para la responsabilidad en Inteligencia Artificial son:

* Explicabilidad.
* Imparcialidad.
* Fiabilidad.
* Transparencia.
* Privacidad.