**“INTELIGENCIA ARTIFICIAL EXPLICABLE PARA EL DESARROLLO SOSTENIBLE”**

*"EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR SUSTAINABLE DEVELOPMENT"*

Mario González Arencibia

Código ORCID: <https://orcid.org/0000-0001-9947-7762>

Centro de Estudios de Gestión de Proyectos y Toma de Decisiones

Afiliación: Universidad de Ciencias Informáticas, La Habana, Cuba.

Correo electrónico: mgarencibia@uci.cu

**Resumen**

La inteligencia artificial explicable (XAI) emerge como una herramienta prometedora para impulsar el desarrollo sostenible de manera transparente y ética. El objetivo de esta investigación es explorar cómo la XAI puede contribuir al logro de los Objetivos de Desarrollo Sostenible (ODS), abordando sus principios, aplicaciones, desafíos y recomendaciones. Los principales hallazgos destacan el potencial de la XAI para optimizar la gestión de recursos naturales, mitigar el cambio climático y promover la planificación urbana sostenible, al proporcionar explicaciones interpretables sobre las decisiones de los sistemas de IA. Sin embargo, también se identifican desafíos críticos, como los sesgos en datos y algoritmos, las preocupaciones de privacidad y seguridad, y el impacto en el empleo. Como conclusión central, la investigación subraya que, para aprovechar plenamente el potencial de la XAI en el desarrollo sostenible, es fundamental abordar estos desafíos a través de un enfoque colaborativo, estableciendo marcos regulatorios sólidos, promoviendo la educación digital y fomentando la investigación en XAI ética y transparente.

**Palabras clave:** Transparencia, Interpretabilidad, Equidad, Ética, Colaboración, Empoderamiento.

**Abstract**

Explainable artificial intelligence (XAI) emerges as a promising tool to drive sustainable development in a transparent and ethical manner. The objective of this research is to explore how XAI can contribute to achieving the Sustainable Development Goals (SDGs), addressing its principles, applications, challenges, and recommendations. The main findings highlight the potential of XAI to optimize natural resource management, mitigate climate change, and promote sustainable urban planning by providing interpretable explanations about the decisions made by AI systems. However, critical challenges are also identified, such as biases in data and algorithms, privacy and security concerns, and the impact on employment. As a central conclusion, the research underscores that to fully harness the potential of XAI for sustainable development, it is essential to address these challenges through a collaborative approach, establishing robust regulatory frameworks, promoting digital education, and fostering research in ethical and transparent XAI.

**Keywords:** Transparency, Interpretability, Equity, Ethics, Collaboration, Empowerment.

**Introducción**

El rápido avance de la inteligencia artificial (IA) ha revolucionado diversos sectores, ofreciendo soluciones innovadoras y eficientes. Sin embargo, a medida que su adopción se extiende, surgen preocupaciones legítimas sobre la transparencia, la interpretabilidad y la rendición de cuentas de estos sistemas. En este contexto, la inteligencia artificial explicable (XAI, por sus siglas en inglés) emerge como un enfoque prometedor para abordar estos desafíos y garantizar un desarrollo responsable y ético de la IA.

La Agenda 2030 para el Desarrollo Sostenible, adoptada por las Naciones Unidas, establece un conjunto de 17 Objetivos de Desarrollo Sostenible (ODS) que abarcan cuestiones económicas, sociales y ambientales. La XAI puede desempeñar un papel crucial en el logro de estos objetivos, al permitir que los sistemas de IA sean transparentes, interpretables y puedan rendir cuentas.

La IA explicable se refiere a la capacidad de los sistemas de IA para proporcionar explicaciones comprensibles sobre su funcionamiento y toma de decisiones (Arrieta et al., 2020). La inteligencia artificial explicable (XAI, por sus siglas en inglés) se ha desarrollado como una respuesta a la necesidad de hacer los sistemas de IA más transparentes e interpretables.

A diferencia de los modelos de "caja negra" que caracterizan a algunos sistemas tradicionales de IA, la XAI permite a los usuarios entender cómo se toman las decisiones, lo cual es crucial para generar confianza y aceptación. Los sistemas de XAI están diseñados no solo para ofrecer resultados precisos, sino también para proporcionar explicaciones claras sobre cómo se alcanzaron estos resultados, permitiendo una evaluación ética y responsable (Arrieta et al., 2020).

Dada la importancia de implementar la IA de manera responsable y ética, la XAI se posiciona como una herramienta esencial para apoyar el desarrollo sostenible. La transparencia y la capacidad de rendir cuentas que ofrece la XAI son fundamentales para fomentar la confianza de los ciudadanos en la tecnología. La capacidad de explicar y justificar las decisiones tomadas por los sistemas de IA es crucial para asegurar que estas decisiones sean justas y equitativas, minimizando los riesgos de sesgo y discriminación (Gunning et al., 2019).

El desarrollo sostenible es un objetivo fundamental a nivel global, y la XAI puede desempeñar un papel clave en su consecución. Según un estudio publicado en Nature, la IA puede facilitar el cumplimiento del 79% de los Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas (Vinuesa, et al. 2020). Por ejemplo, la XAI puede contribuir a la eficiencia energética al permitir la optimización de los sistemas de gestión de energía y la identificación de patrones de consumo insostenibles. En la agricultura, la XAI puede ayudar a mejorar la productividad y la sostenibilidad al proporcionar recomendaciones interpretables sobre prácticas agrícolas, como el uso de fertilizantes y el riego.

La XAI es fundamental para garantizar que la IA se desarrolle y se implemente de manera responsable y ética en el contexto del desarrollo sostenible. Al ser transparente e interpretable, la XAI puede fomentar la confianza y la aceptación de la tecnología por parte de los ciudadanos, lo que a su vez puede facilitar su adopción y maximizar su impacto positivo (Instituto de Ingeniería del Conocimiento, 2022).

Varios estudios han explorado el vínculo entre la IA explicable y el desarrollo sostenible. Por ejemplo, el Grupo de Expertos en Inteligencia Artificial de las Naciones Unidas (2021) ha destacado la importancia de la XAI para garantizar que la IA se utilice de manera ética y responsable en la consecución de los ODS. Asimismo, la Unión Europea ha adoptado regulaciones, como el Reglamento de Inteligencia Artificial (2021), que establecen requisitos de transparencia y explicabilidad para los sistemas de IA.

Un ejemplo concreto de la aplicación de la XAI en el desarrollo sostenible es el proyecto "Inteligencia Artificial para la Sostenibilidad" de la Universidad de Stanford. Este proyecto utiliza técnicas de XAI para desarrollar modelos de IA que puedan ayudar a optimizar el uso de recursos naturales, como el agua y la energía, en entornos urbanos. Los modelos son interpretables, lo que permite a los tomadores de decisiones y a los ciudadanos comprender cómo se llegó a las recomendaciones y ajustarlas según sea necesario (Vinuesa, ET AL., 2020).

A pesar de estos avances, aún existe un vacío de conocimiento en cuanto a cómo la XAI puede aplicarse de manera efectiva en diferentes áreas del desarrollo sostenible. Es necesario comprender mejor los principios, métodos y técnicas de la XAI, así como su impacto potencial en la toma de decisiones, la gobernanza y la participación de las partes interesadas.

Este artículo tiene como objetivo explorar el papel de la inteligencia artificial explicable en el fomento del desarrollo sostenible. Se abordarán los principios fundamentales de la XAI, sus aplicaciones potenciales en áreas clave como la eficiencia energética, la gestión de recursos, la agricultura sostenible y las ciudades inteligentes. se examinarán los desafíos y las consideraciones éticas asociadas con el uso de la XAI en el contexto del desarrollo sostenible, y se proporcionarán recomendaciones para investigaciones y prácticas futuras.

**Relevancia**

La relevancia de este estudio radica en su contribución al campo emergente de la XAI y su aplicación en el logro de los ODS. Al comprender mejor cómo la XAI puede fomentar la transparencia, la responsabilidad y la confianza en los sistemas de IA, se pueden aprovechar mejor estas tecnologías para abordar desafíos complejos relacionados con el desarrollo sostenible.

La inteligencia artificial explicable (XAI) desempeña un papel crucial para el desarrollo sostenible debido a su capacidad de aumentar la transparencia, la interpretabilidad y la rendición de cuentas en los sistemas de IA. Una de las razones clave que resalta la importancia de la XAI es su capacidad para fomentar la confianza y la aceptación pública. La falta de transparencia en los sistemas de IA tradicionales puede generar desconfianza y resistencia por parte de los ciudadanos y las partes interesadas. La XAI, al proporcionar explicaciones comprensibles sobre el funcionamiento y la toma de decisiones de los modelos de IA, puede aumentar la confianza y la aceptación de estas tecnologías, lo cual es fundamental para su adopción efectiva en áreas críticas para el desarrollo sostenible.

La XAI es esencial para garantizar la equidad y la no discriminación. Los sistemas de IA pueden reflejar y perpetuar sesgos y discriminaciones presentes en los datos de entrenamiento o en los procesos de desarrollo. La XAI permite detectar y mitigar estos sesgos, lo que contribuye a garantizar la equidad y la no discriminación en la toma de decisiones, un principio fundamental para el desarrollo sostenible y la reducción de las desigualdades (ODS 10).

Por otra parte, la XAI facilita la gobernanza y la supervisión al permitir que los responsables políticos, reguladores y partes interesadas comprendan mejor cómo funcionan los sistemas de IA y cómo se toman las decisiones Arrieta, et al., 2020). Esto es esencial para garantizar que la IA se utilice de manera responsable y alineada con los objetivos de desarrollo sostenible.

Por consiguiente, la XAI impulsa la participación y la colaboración al hacer que los sistemas de IA sean más interpretables. Esto fomenta una mayor participación y colaboración de las partes interesadas, incluidas las comunidades locales, en el diseño, desarrollo e implementación de soluciones basadas en IA. Esto es fundamental para asegurar que estas soluciones sean inclusivas, respondan a las necesidades locales y promuevan la apropiación y la sostenibilidad a largo plazo.

El hecho es que, la XAI puede mejorar la toma de decisiones y la gestión de recursos en áreas clave para el desarrollo sostenible, como la agricultura, la energía, la gestión de desastres y la planificación urbana. Al proporcionar explicaciones comprensibles, la XAI puede respaldar decisiones más informadas, eficientes y sostenibles.

Varios organismos internacionales han reconocido la importancia de la XAI para el desarrollo sostenible. Por ejemplo, el Grupo de Expertos en Inteligencia Artificial de las Naciones Unidas (2021) ha destacado que la XAI es fundamental para "garantizar que la IA se utilice de manera ética y responsable en la consecución de los Objetivos de Desarrollo Sostenible".

**Concepto de desarrollo sostenible y sus objetivos**

El concepto de desarrollo sostenible, definido por la Comisión Brundtland en 1987, se refiere a satisfacer las necesidades del presente sin comprometer la capacidad de las futuras generaciones para satisfacer sus propias necesidades (World Commission on Environment and Development, 1987). Este enfoque holístico integra tres dimensiones fundamentales: el crecimiento económico, la inclusión social y la protección ambiental. Los Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas, establecidos en 2015, constituyen una hoja de ruta global con 17 objetivos específicos destinados a abordar desafíos críticos como la pobreza, la desigualdad, la salud, la educación, la sostenibilidad ambiental y la justicia (United Nations, 2015).

La inteligencia artificial (IA) explicable (XAI, por sus siglas en inglés) puede desempeñar un papel crucial en la consecución de estos objetivos. La XAI se refiere a sistemas de IA diseñados para ser transparentes, interpretables y responsables, permitiendo a los usuarios entender y confiar en los resultados generados por estos sistemas. Esta capacidad de explicar y justificar las decisiones es vital para asegurar que la IA se implemente de manera ética y equitativa, minimizando los riesgos de sesgo y discriminación (Arrieta et al., 2020).

Por ejemplo, en el contexto del ODS 7, que busca garantizar el acceso a una energía asequible, segura, sostenible y moderna, los sistemas de XAI pueden optimizar el consumo energético en edificios y fábricas. Estos sistemas no solo proporcionan recomendaciones para mejorar la eficiencia energética, sino que también explican cómo se derivaron dichas recomendaciones, lo cual es fundamental para que los gestores de energía confíen en las soluciones propuestas y las implementen efectivamente (Ahmad et al., 2020).

En la agricultura, que está directamente relacionada con el ODS 2 de acabar con el hambre, la XAI puede asesorar a los agricultores sobre el uso óptimo de fertilizantes y pesticidas, mejorando la productividad de manera sostenible. Un estudio realizado por García et al. (2019) demostró que los sistemas de IA explicable pueden identificar patrones en los datos agrícolas y proporcionar explicaciones claras sobre las prácticas agrícolas más efectivas, lo que ayuda a los agricultores a tomar decisiones informadas y sostenibles (García et al., 2019).

Además, en las ciudades inteligentes, vinculadas con el ODS 11 de lograr que las ciudades y los asentamientos humanos sean inclusivos, seguros, resilientes y sostenibles, la XAI puede mejorar la gestión del tráfico y la planificación urbana. Los sistemas de XAI pueden analizar grandes volúmenes de datos urbanos y ofrecer soluciones para reducir la congestión y las emisiones de carbono, proporcionando explicaciones comprensibles sobre cómo se generaron estas soluciones (Bibri & Krogstie, 2020).

La transparencia y la capacidad de rendir cuentas que ofrece la XAI son esenciales para fomentar la confianza y la aceptación de la tecnología por parte de los ciudadanos. Cuando las personas entienden cómo funcionan los sistemas de IA y cómo se toman las decisiones, es más probable que adopten y apoyen estas tecnologías. Esto es particularmente importante en la implementación de políticas públicas y estrategias empresariales que buscan promover el desarrollo sostenible (Gunning et al., 2019).

La integración de la inteligencia artificial explicable en las estrategias de desarrollo sostenible es un factor decisivo para el éxito de los ODS. La capacidad de proporcionar explicaciones claras y justificadas sobre las decisiones de la IA no solo promueve la transparencia y la confianza, sino que también asegura que las soluciones sean éticas y equitativas, contribuyendo de manera significativa a un futuro sostenible y justo.

**Principios y Enfoques de la XAI**

La inteligencia artificial explicable (XAI) se basa en principios fundamentales como la transparencia, la interpretabilidad, la equidad y la responsabilidad. Estos principios son cruciales para garantizar que los sistemas de IA sean confiables, éticos y puedan rendir cuentas (Arencibia, et al. 2024).

La inteligencia artificial explicable (XAI) se basa en varios principios fundamentales que buscan hacer que los sistemas de IA sean más transparentes, interpretables, equitativos y responsables. Estos principios son esenciales para garantizar que la IA no solo sea efectiva, sino también ética y confiable.

***Transparencia*** es un principio clave de la XAI, que implica que los procesos y decisiones de un sistema de IA sean visibles y comprensibles para los usuarios. La transparencia ayuda a desmitificar el funcionamiento interno de los algoritmos de IA, permitiendo a los usuarios entender cómo se generan los resultados. Este entendimiento es crucial para fomentar la confianza en la tecnología (Doshi-Velez & Kim, 2017).

La ***interpretabilidad*** está estrechamente relacionada con la transparencia y se refiere a la capacidad de un modelo de IA para proporcionar explicaciones claras y comprensibles sobre cómo llega a sus decisiones. La interpretabilidad es vital para asegurar que los usuarios puedan interpretar los resultados de la IA de manera significativa, facilitando su uso y aceptación en contextos críticos como la salud y la justicia (Rudin, 2019).

La ***equidad*** es otro principio fundamental de la XAI, que busca garantizar que los sistemas de IA no perpetúen sesgos o discriminaciones inherentes en los datos de entrenamiento. Los modelos explicables pueden ayudar a identificar y corregir estos sesgos, promoviendo decisiones más justas y equitativas (Barocas, Hardt, & Narayanan, 2019).

La ***responsabilidad*** implica que los desarrolladores y usuarios de sistemas de IA deben rendir cuentas por las decisiones y acciones tomadas por estos sistemas. La capacidad de explicar y justificar las decisiones de la IA es crucial para asegurar la responsabilidad y la rendición de cuentas, especialmente en contextos donde las decisiones tienen un impacto significativo en las vidas de las personas (Floridi et al., 2018).

La imparcialidad en la inteligencia artificial se refiere a la necesidad de desarrollar sistemas que tomen decisiones de manera equitativa, sin sesgos o discriminaciones injustificadas. Este principio es crucial para asegurar que la IA beneficie a todos los sectores de la sociedad por igual y no perpetúe o amplifique las desigualdades existentes. Por ejemplo, los sistemas de IA utilizados en el reclutamiento de personal deben ser capaces de evaluar a los candidatos de manera objetiva, sin favorecer o desfavorecer a ciertos grupos por motivos de género, raza o cualquier otra característica irrelevante para el puesto. Según un estudio de Mehrabi et al. (2021), la implementación de algoritmos justos es esencial para evitar resultados sesgados que puedan afectar negativamente a ciertos grupos poblacionales (Mehrabi et al., 2021). Este estudio destaca cómo el diseño cuidadoso de los modelos y la supervisión continua pueden mitigar estos riesgos.

Por otro lado, la seguridad y robustez son principios que garantizan que los sistemas de inteligencia artificial funcionen de manera confiable y segura, incluso en situaciones adversas o inesperadas. Estos sistemas deben ser capaces de manejar errores, ataques maliciosos y variaciones en los datos de entrada sin comprometer su funcionamiento. La robustez de la IA es especialmente importante en aplicaciones críticas, como la conducción autónoma o los sistemas de diagnóstico médico, donde los errores pueden tener consecuencias graves. Según Goodfellow et al. (2018), los ataques adversariales, donde se manipulan sutilmente los datos de entrada para engañar al sistema, son una amenaza significativa para la seguridad de los sistemas de IA (Goodfellow et al., 2018). Para mitigar estos riesgos, se han propuesto diversas técnicas, como el entrenamiento adversarial y la incorporación de mecanismos de detección de anomalías.

Los principios de la Inteligencia Artificial Explicable (XAI) juegan un papel crucial en la aplicación responsable y ética de la IA para el logro de los Objetivos de Desarrollo Sostenible (ODS). Al garantizar la transparencia, interpretabilidad, equidad, responsabilidad, imparcialidad y seguridad de los sistemas de IA, podemos aprovechar el potencial de esta tecnología para abordar los desafíos globales más urgentes (Consulte la Tabla No. 1).

Tabla No. 1: Principios de la inteligencia artificial explicable (xai) y su potencial para el logro de los objetivos de desarrollo sostenible (ODS)

| Principio de la XAI | Potencial para el Logro de los ODS | Ejemplos de Aplicación | Autores de Referencia |
| --- | --- | --- | --- |
| **Transparencia** | - Fomentar la confianza y la aceptación pública de la IA. - Permitir un mejor entendimiento de cómo se toman las decisiones de IA. - Facilitar la identificación y corrección de errores o sesgos en los sistemas de IA. | - Publicación de documentación clara y accesible sobre el funcionamiento de los sistemas de IA. - Implementación de mecanismos de auditoría y supervisión para garantizar la transparencia en la toma de decisiones de IA. - Desarrollo de herramientas que permitan a los usuarios visualizar y comprender los resultados de la IA. | Doshi-Velez & Kim, 2017; Arencibia, et al. 2024 |
| **Interpretabilidad** | - Permitir una mejor comprensión de las predicciones y recomendaciones de la IA. - Facilitar la colaboración entre humanos y sistemas de IA. - Ayudar a identificar y corregir sesgos o errores en los datos de entrenamiento. | - Desarrollo de métodos de explicación que proporcionen explicaciones claras y comprensibles de las decisiones de IA. - Implementación de técnicas de visualización para representar los resultados de la IA de manera intuitiva. - Diseño de interfaces de usuario que faciliten la interacción y comprensión de los sistemas de IA. | Rudin, 2019; Arencibia, et al. 2024 |
| **Equidad** | - Garantizar que los sistemas de IA no perpetúen sesgos o discriminaciones existentes. - Promover la inclusión y la justicia en el uso de la IA. - Evitar que la IA amplifique las desigualdades sociales. | - Implementación de técnicas de detección y corrección de sesgos en los datos de entrenamiento. - Desarrollo de métricas de evaluación que consideren la equidad y la justicia en los resultados de la IA. - Establecimiento de marcos éticos y legales para el desarrollo y uso responsable de la IA. | Barocas, Hardt, & Narayanan, 2019; Mehrabi et al., 2021 |
| **Responsabilidad** | - Asegurar que los desarrolladores y usuarios de la IA rindan cuentas por las decisiones y acciones de estos sistemas. - Promover el uso ético y responsable de la IA. - Establecer mecanismos para la reparación de daños causados por sistemas de IA defectuosos. | - Implementación de principios de gobernanza de datos que garanticen la transparencia, la integridad y la seguridad de los datos utilizados en la IA. - Desarrollo de marcos legales que establezcan responsabilidades claras para los desarrolladores, usuarios y operadores de sistemas de IA. - Creación de mecanismos de denuncia y resolución de conflictos relacionados con el uso de la IA. | Floridi et al., 2018; Arencibia, et al. 2024 |
| **Imparcialidad** | - Garantizar que los sistemas de IA tomen decisiones neutrales y objetivas, sin favorecer o desfavorecer a ciertos grupos. - Evitar la discriminación y la exclusión en el uso de la IA. - Promover la igualdad de oportunidades para todos. | - Implementación de técnicas de desbaneo y anonimización de datos para eliminar sesgos inconscientes. - Desarrollo de algoritmos de aprendizaje automático robustos a manipulaciones y ataques adversos. - Evaluación rigurosa de los sistemas de IA para identificar y corregir posibles sesgos. | Mehrabi et al., 2021; Goodfellow et al., 2018 |
| **Seguridad y Robustez** | - Garantizar que los sistemas de IA funcionen de manera confiable y segura, incluso en situaciones adversas o inesperadas. - Evitar errores, ataques maliciosos y manipulaciones de datos. - Proteger la privacidad y la seguridad de las personas. | - Implementación de mecanismos de seguridad y protección de datos en los sistemas de IA. - Desarrollo de técnicas de detección y respuesta a ataques adversos. - Realización de pruebas exhaustivas para garantizar la robustez y confiabilidad de los sistemas de IA. | Goodfellow et al., 2018; Arencibia, et al. 2024 |

Fuente: Elaboración propia

La tabla presentada muestra cómo los principios de la Explicabilidad de la Inteligencia Artificial (XAI) pueden contribuir al logro de los Objetivos de Desarrollo Sostenible (ODS). Estos principios incluyen la transparencia, la interpretabilidad, la equidad, la responsabilidad, la imparcialidad, y la seguridad y robustez.

La transparencia tiene el potencial de fomentar la confianza pública, permitir una mejor comprensión del uso de datos y algoritmos, y facilitar la participación ciudadana (Doshi-Velez & Kim, 2017; Arencibia, et al. 2024). Esto se puede lograr mediante la publicación de información sobre el funcionamiento de los sistemas de IA, la implementación de mecanismos de auditoría, y el desarrollo de herramientas de visualización para los ciudadanos.

La interpretabilidad permite una mejor comprensión de las decisiones de IA, facilita la colaboración entre humanos y sistemas de IA, y ayuda a identificar sesgos o errores (Rudin, 2019; Arencibia, et al. 2024). Esto se puede lograr a través del desarrollo de métodos de explicación claros, la implementación de técnicas de visualización intuitivas, y el diseño de interfaces de usuario amigables.

En cuanto a la equidad, la XAI puede garantizar que la IA no perpetúe sesgos o discriminaciones, promover la inclusión y la justicia, y evitar la amplificación de desigualdades (Barocas, Hardt, & Narayanan, 2019; Mehrabi et al., 2021). Esto se puede lograr mediante la implementación de técnicas de detección y corrección de sesgos, el desarrollo de métricas de evaluación que consideren la equidad, y el establecimiento de marcos éticos y legales.

La responsabilidad asegura la rendición de cuentas de los desarrolladores, usuarios y operadores de sistemas de IA, promueve el uso ético y responsable, y establece mecanismos de reparación de daños (Floridi et al., 2018; Arencibia, et al. 2024). Esto se puede lograr a través de la implementación de principios de gobernanza de datos, el desarrollo de marcos legales con responsabilidades claras, y la creación de mecanismos de denuncia y resolución de conflictos.

La imparcialidad garantiza que los sistemas de IA tomen decisiones neutrales y objetivas, sin favorecer o desfavorecer a ciertos grupos, evitando la discriminación y la exclusión (Mehrabi et al., 2021; Goodfellow et al., 2018). Esto se puede lograr mediante la implementación de técnicas de desbaneo y anonimización de datos, el desarrollo de algoritmos robustos a manipulaciones, y la evaluación rigurosa para identificar y corregir posibles sesgos.

La seguridad y robustez garantizan que los sistemas de IA funcionen de manera confiable y segura, incluso en situaciones adversas, evitando errores, ataques maliciosos y manipulaciones de datos, y protegiendo la privacidad y la seguridad de las personas.

**Enfoques y técnicas**

Para implementar estos principios, se utilizan diversos enfoques y técnicas en la XAI. Uno de los enfoques más comunes son los ***modelos de caja blanca***, que son intrínsecamente interpretable y permiten a los usuarios ver y entender cómo se toman las decisiones. Ejemplos de estos modelos incluyen los árboles de decisión y las reglas de asociación, que son fáciles de visualizar y comprender (Guidotti et al., 2018).

Otra técnica importante en la XAI son las ***explicaciones post-hoc,*** que proporcionan interpretaciones de los resultados de modelos de caja negra después de que se ha realizado una predicción. Estos métodos incluyen técnicas como LIME (Local Interpretable Model-agnostic Explanations) y SHAP (SHapley Additive exPlanations), que generan explicaciones locales para las predicciones de modelos complejos como las redes neuronales (Ribeiro, Singh, & Guestrin, 2016; Lundberg & Lee, 2017).

Los ***métodos de visualización*** también juegan un papel crucial en la XAI, ayudando a los usuarios a entender mejor los datos y los resultados de los modelos de IA. Herramientas de visualización como los mapas de calor y los diagramas de contribución de características permiten a los usuarios ver cómo diferentes variables afectan las predicciones del modelo (Molnar, 2019).

Por último, las ***técnicas de aprendizaje automático interpretable*** incluyen enfoques que diseñan modelos complejos para ser más comprensibles. Ejemplos de estos enfoques son las redes neuronales interpretable, que buscan mantener un equilibrio entre la precisión del modelo y su interpretabilidad, utilizando técnicas como la atención y las capas de explicación (Alvarez-Melis & Jaakkola, 2018).

Estos enfoques y técnicas de XAI tienen como objetivo principal lograr que los sistemas de IA sean más transparentes, interpretables y puedan rendir cuentas, lo que es fundamental para fomentar la confianza y la adopción responsable de estas tecnologías en diversos ámbitos.

En conclusión, los enfoques y técnicas de XAI ofrecen un conjunto de herramientas valiosas para hacer que la IA sea más transparente, interpretable y responsable, lo que es fundamental para alcanzar los ODS. Al aplicar estos principios en el desarrollo y uso de la IA, podemos garantizar que esta tecnología se utilice de manera ética, responsable y beneficiosa para todos (Consulte Tabla No. 2).

Tabla No. 2: Enfoques y Técnicas de XAI y su Contribución a los Objetivos de Desarrollo Sostenible (ODS)

| Enfoque/Técnica de XAI | Principios de la XAI a los que contribuye |  Ejemplos de Contribución a los ODS |
| --- | --- | --- |
| **Modelos de caja blanca** | Transparencia, Equidad, Responsabilidad | - Facilitar la comprensión del funcionamiento de los sistemas de IA para la toma de decisiones informadas en el marco de los ODS (ODS 16, ODS 10, ODS 5). - Permitir la identificación y corrección de sesgos en la asignación de recursos o la evaluación de impacto, promoviendo la equidad en el acceso a oportunidades y beneficios (ODS 10, ODS 5). - Ayudar a establecer mecanismos de rendición de cuentas para los desarrolladores y usuarios de sistemas de IA en proyectos de desarrollo sostenible (ODS 16, ODS 10). |
| **Explicaciones post-hoc** | Interpretabilidad, Equidad, Responsabilidad | - Proporcionar explicaciones claras sobre las decisiones de IA en la distribución de recursos o la evaluación de riesgos en iniciativas de desarrollo sostenible (ODS 11, ODS 13). - Permitir la detección y corrección de sesgos ocultos en la selección de beneficiarios o la evaluación de impacto, promoviendo la inclusión y la no discriminación (ODS 10, ODS 16). - Facilitar la auditoría y evaluación del impacto social de los sistemas de IA en diferentes comunidades o grupos de interés (ODS 11, ODS 17). |
| **Métodos de visualización** | Transparencia, Equidad, Responsabilidad | - Comunicar de manera efectiva el funcionamiento y los resultados de los sistemas de IA a una amplia audiencia, fomentando la confianza pública y la participación ciudadana (ODS 16, ODS 17). - Permitir la identificación de patrones y sesgos en los datos de entrenamiento o los resultados de la IA, promoviendo decisiones justas y equitativas (ODS 10, ODS 11). - Facilitar la comprensión del impacto de los sistemas de IA en diferentes grupos sociales o regiones, asegurando un uso responsable e inclusivo (ODS 11, ODS 10). |
| **Aprendizaje automático interpretable** | Interpretabilidad, Equidad, Responsabilidad | - Diseñar modelos de IA que sean intrínsecamente más comprensibles para usuarios sin conocimientos técnicos, facilitando su adopción en el contexto del desarrollo sostenible (ODS 8, ODS 9). - Ayudar a identificar y corregir sesgos en el proceso de diseño y entrenamiento de modelos de IA, promoviendo la inclusión y la no discriminación en el acceso a oportunidades y beneficios (ODS 10, ODS 11). - Facilitar la evaluación del impacto ético y social de los sistemas de IA en el marco de los ODS, asegurando un uso responsable y ético (ODS 16, ODS 17). |

Fuente: Elaboración propia

De la Tabla anterior se infiere que la Inteligencia Artificial Explicable (XAI) proporciona una variedad de enfoques y técnicas para hacer que los sistemas de IA sean más transparentes, interpretables y responsables. Estos aspectos son fundamentales para la adopción responsable de la IA en el contexto del logro de los Objetivos de Desarrollo Sostenible (ODS).

Los modelos de caja blanca son un enfoque clave de XAI que contribuye significativamente a los ODS. Estos modelos permiten a los usuarios comprender el funcionamiento interno de los sistemas de IA, lo que fomenta la confianza y la aceptación pública. Además, ayudan a identificar y corregir sesgos en los datos de entrenamiento, promoviendo decisiones justas y equitativas. Los modelos de caja blanca también facilitan la rendición de cuentas por parte de los desarrolladores y usuarios de sistemas de IA, asegurando un uso ético y responsable. Estos aspectos son particularmente relevantes para ODS como el 16 (Paz, justicia e instituciones sólidas), el 10 (Reducción de las desigualdades) y el 5 (Igualdad de género y empoderamiento de las mujeres y niñas).

Las explicaciones post-hoc son otra técnica importante de XAI que contribuye a los ODS. Proporcionan interpretaciones claras y comprensibles de las decisiones de IA, incluso para modelos complejos como las redes neuronales. Esto permite detectar y corregir sesgos ocultos en los modelos de IA, promoviendo la inclusión y la no discriminación. Además, las explicaciones post-hoc facilitan la auditoría y evaluación del impacto de los sistemas de IA, asegurando un uso responsable y ético. Estas características son especialmente relevantes para ODS como el 10 (Reducción de las desigualdades), el 16 (Paz, justicia e instituciones sólidas) y el 11 (Ciudades y comunidades sostenibles).

Los métodos de visualización son una herramienta valiosa en XAI que contribuye a varios ODS. Ayudan a comunicar de manera efectiva el funcionamiento y los resultados de los sistemas de IA a una audiencia amplia. También permiten identificar patrones y sesgos en los datos de entrenamiento, promoviendo decisiones justas y equitativas. Además, facilitan la comprensión del impacto de los sistemas de IA en diferentes grupos sociales, asegurando un uso responsable e inclusivo. Estos métodos son particularmente útiles para ODS como el 11 (Ciudades y comunidades sostenibles), el 13 (Acción por el clima) y el 17 (Alianzas para lograr los objetivos).

El aprendizaje automático interpretable es un enfoque de XAI que también contribuye significativamente a los ODS. Permite diseñar modelos de IA que son intrínsecamente más comprensibles, incluso para usuarios sin conocimientos técnicos. Ayuda a identificar y corregir sesgos en el proceso de diseño y entrenamiento de modelos de IA, promoviendo la inclusión y la no discriminación. Además, facilita la evaluación del impacto ético y social de los sistemas de IA, asegurando un uso responsable y ético. Este enfoque es especialmente relevante para ODS como el 8 (Trabajo decente y crecimiento económico), el 9 (Industria, innovación e infraestructura) y el 10 (Reducción de las desigualdades).

**Desafíos y oportunidades de la XAI para el desarrollo sostenible**

La relación entre la inteligencia artificial explicable (XAI) y el desarrollo sostenible es compleja y multifacética, ofreciendo numerosos beneficios potenciales pero también enfrentando desafíos y limitaciones significativos.

La XAI puede desempeñar un papel importante en la promoción del desarrollo sostenible al proporcionar sistemas de IA que son transparentes, interpretables y responsables. Esto es especialmente importante en áreas críticas que impactan directamente los Objetivos de Desarrollo Sostenible (ODS).

En el ámbito de la ***gestión ambiental***, la XAI puede facilitar la toma de decisiones informadas sobre el uso de recursos naturales. Por ejemplo, la aplicación de modelos explicables en la agricultura puede optimizar el uso de agua y fertilizantes, reduciendo el impacto ambiental y aumentando la productividad. Un estudio realizado por García et al. (2019) demostró que los sistemas de XAI pueden proporcionar recomendaciones precisas para el riego, basadas en datos de sensores, mejorando la sostenibilidad agrícola.

En el sector de ***energía,*** la XAI puede mejorar la eficiencia de las redes eléctricas inteligentes. Al interpretar y explicar los patrones de consumo, estos sistemas pueden ayudar a equilibrar la oferta y la demanda de energía, reduciendo el desperdicio y las emisiones de carbono. Ahmad et al. (2020) encontraron que la XAI puede predecir picos de demanda y sugerir estrategias para mitigar estos picos, contribuyendo al ODS 7 de energía asequible y limpia.

En las ***ciudades inteligentes***, la XAI puede optimizar la planificación urbana y la gestión del tráfico, reduciendo la congestión y mejorando la calidad del aire. Bibri y Krogstie (2020) destacaron cómo la XAI puede apoyar el diseño de ciudades más sostenibles mediante el análisis de datos en tiempo real y la oferta de soluciones que minimicen el impacto ambiental (Madureira et al., 2020).

La gestión de recursos naturales: La XAI puede ayudar a mejorar la eficiencia en el uso de recursos como el agua, la energía y los minerales, al tiempo que permite una mejor comprensión de los trade-offs y las consecuencias de las decisiones (Brondizio et al., 2019). Por ejemplo, en el sector agrícola, la XAI puede optimizar el uso de insumos y recursos, reduciendo el desperdicio y mejorando los rendimientos de manera sostenible (Liakos et al., 2018).

La producción y el consumo responsables: Los sistemas de IA explicables pueden ayudar a las empresas y los consumidores a tomar decisiones más informadas sobre la procedencia, el impacto y la sostenibilidad de los productos y servicios, fomentando patrones de consumo más responsables (Rashed et al., 2021).

***Desafíos y Limitaciones***

A pesar de estos beneficios, la implementación de la XAI en el contexto del desarrollo sostenible enfrenta varios desafíos y limitaciones. Uno de los principales desafíos es la ***complejidad técnica***. La creación de modelos de IA que sean a la vez precisos y explicables es una tarea difícil. Los modelos de caja blanca, aunque más transparentes, a menudo no alcanzan el mismo nivel de precisión que los modelos de caja negra (Rudin, 2019).

Otro desafío es la ***recopilación y gestión de datos***. Los sistemas de XAI requieren grandes volúmenes de datos de alta calidad para funcionar de manera efectiva. Sin embargo, la recopilación de estos datos puede ser costosa y plantear problemas de privacidad y seguridad (Gunning et al., 2019). Además, los datos utilizados para entrenar los modelos de IA a menudo contienen sesgos inherentes, que pueden perpetuar inequidades si no se manejan adecuadamente (Barocas, Hardt, & Narayanan, 2019).

La ***adopción y aceptación por parte de los usuarios*** también es un obstáculo significativo. La XAI requiere que los usuarios comprendan y confíen en las explicaciones proporcionadas por los sistemas de IA. Sin embargo, la interpretación de estas explicaciones puede ser difícil, especialmente para los usuarios sin formación técnica. Esto puede limitar la adopción de la XAI en sectores clave (Doshi-Velez & Kim, 2017).

Existe el desafío de ***regulación y políticas***. Los marcos legales y normativos para la IA explicable todavía están en desarrollo, lo que puede crear incertidumbre para las organizaciones que buscan implementar estas tecnologías. La falta de estándares claros y la variabilidad en las regulaciones a nivel global complican la integración de la XAI en prácticas sostenibles (Floridi et al., 2018).

Otro de los principales desafíos es abordar la falta de transparencia y explicabilidad de algunos sistemas de IA, lo que puede generar desconfianza y rechazo por parte de los ciudadanos y las partes interesadas (Arrieta et al., 2020). Muchos de estos sistemas de IA funcionan como "cajas negras" cuyos procesos de toma de decisiones no son claros, lo que dificulta la rendición de cuentas y la comprensión de los impactos que pueden tener en áreas clave del desarrollo sostenible (Samek et al., 2017).

Según Vinuesa et al. (2020), "la IA explicable e interpretable puede ayudar a los responsables de la toma de decisiones a comprender mejor las recomendaciones de los modelos de IA y a ajustarlas según sea necesario, lo que es crucial para la implementación responsable de la IA en el logro de los ODS" (p. 6).

Un desafío importante es garantizar que los sistemas de IA no perpetúen o exacerben los sesgos existentes en los datos utilizados para su entrenamiento, lo que puede tener consecuencias negativas en términos de equidad, inclusión y justicia social (Mehrabi et al., 2021). Además, la complejidad y la rápida evolución de la IA plantean desafíos en términos de gobernanza, regulación y alineación con los principios éticos (Floridi et al., 2018).

Es importante abordar los sesgos y las inequidades que pueden surgir en los sistemas de IA, especialmente en áreas sensibles como la educación, la salud y la gestión de recursos naturales. La inteligencia artificial explicable (XAI) puede contribuir a mitigar estos riesgos al permitir una mayor transparencia y comprensión de los procesos de toma de decisiones de los modelos de IA.

Sin embargo, la XAI también enfrenta desafíos propios, como el equilibrio entre interpretabilidad y rendimiento, los problemas de escalabilidad y la necesidad de enfoques estandarizados (Atria Innovation, 2022). Abordar estas limitaciones a través de la investigación y la colaboración continuas es crucial para aprovechar el potencial de la XAI en el logro de un desarrollo sostenible.

A pesar de estos desafíos, la XAI ofrece oportunidades significativas para contribuir al logro de los Objetivos de Desarrollo Sostenible (ODS). Por ejemplo, en el ámbito de la agricultura sostenible, la XAI puede ayudar a optimizar el uso de recursos como el agua y los fertilizantes, al proporcionar recomendaciones interpretables basadas en el análisis de datos sobre prácticas agrícolas, condiciones climáticas y patrones de rendimiento (Cepei, 2022).

La XAI se enfoca en desarrollar sistemas de IA que sean transparentes, interpretables y que puedan justificar sus decisiones (Adadi & Berrada, 2018). Esto es particularmente relevante en el ámbito del desarrollo sostenible, donde la trazabilidad y la rendición de cuentas son fundamentales para garantizar que las decisiones tomadas con apoyo de la IA no generen impactos negativos en el medio ambiente, la sociedad o la economía.

**Métodos y enfoques de la XAI**

La Inteligencia Artificial Explicable (XAI) ha surgido como un camporelevante en la intersección de la IA y el desarrollo sostenible. Los diferentes enfoques y métodos de XAI buscan hacer que los sistemas de IA sean más transparentes, interpretables y comprensibles. Según Arrieta et al. (2020), estos métodos se pueden clasificar en varias categorías principales, incluyendo explicaciones locales, globales y modelos inherentemente interpretables.

Las explicaciones locales, como LIME (Local Interpretable Model-agnostic Explanations) y SHAP (SHapley Additive exPlanations), se centran en proporcionar interpretaciones para predicciones individuales. Ribeiro et al. (2016) demostraron cómo LIME puede explicar las predicciones de cualquier clasificador de una manera interpretable y fiel localmente. Este enfoque es particularmente útil en contextos de desarrollo sostenible donde se requiere comprender decisiones específicas que afectan a individuos o comunidades.

Por otro lado, las explicaciones globales ofrecen una visión general del comportamiento del modelo en su conjunto. Métodos como los gráficos de importancia de características y los gráficos de dependencia parcial (PDP) permiten entender patrones generales y tendencias en el modelo. Molnar (2019) argumenta que estas técnicas son valiosas para identificar sesgos sistemáticos que podrían afectar la equidad y la sostenibilidad a largo plazo de las aplicaciones de IA.

Los modelos inherentemente interpretables, como los árboles de decisión y las reglas de decisión, ofrecen transparencia desde su diseño. Rudin (2019) aboga por el uso de estos modelos en aplicaciones de alto riesgo relacionadas con el desarrollo sostenible, argumentando que la transparencia es crucial para la confianza y la responsabilidad. Por ejemplo, en un estudio sobre la predicción de la pobreza en áreas urbanas, Jean et al. (2016) utilizaron modelos interpretables para proporcionar insights accionables a los responsables de políticas.

En cuanto a las fortalezas y limitaciones de estos enfoques en relación con el desarrollo sostenible, cada uno tiene sus propias ventajas y desafíos. Las explicaciones locales son excelentes para abordar casos individuales de injusticia o discriminación, lo cual es crucial para objetivos como la reducción de desigualdades. Sin embargo, como señalan Kaur et al. (2020), estas explicaciones pueden no capturar patrones sistemáticos más amplios que son relevantes para el desarrollo sostenible a escala.

Las explicaciones globales, por su parte, son útiles para identificar tendencias generales que pueden afectar a grupos enteros, lo cual es valioso para la formulación de políticas de desarrollo sostenible a gran escala. No obstante, Lipton (2018) advierte que estas explicaciones pueden pasar por alto detalles importantes a nivel individual, lo que podría llevar a generalizaciones excesivas.

Los modelos interpretables ofrecen una transparencia inigualable, lo que facilita la confianza en aplicaciones críticas para el desarrollo sostenible. Sin embargo, como señalan Rudin y Radin (2019), estos modelos pueden tener un rendimiento inferior en tareas complejas en comparación con modelos más opacos, lo que plantea un dilema entre interpretabilidad y rendimiento.

En cuanto a las consideraciones éticas y de responsabilidad en el uso de XAI, surgen varios puntos críticos. La privacidad y la seguridad son preocupaciones principales, ya que las explicaciones detalladas podrían revelar información sensible o personal. Dwork et al. (2012) han propuesto marcos para equilibrar la transparencia con la protección de datos en contextos de aprendizaje automático.

La equidad y la no discriminación son también aspectos cruciales. XAI debe utilizarse para identificar y mitigar sesgos, no para justificarlos. Barocas et al. (2019) han explorado cómo las técnicas de XAI pueden ayudar a detectar y corregir sesgos en los sistemas de IA, contribuyendo así a objetivos de desarrollo sostenible relacionados con la equidad y la inclusión.

La accesibilidad y comprensión de las explicaciones es otro punto crítico. Las explicaciones deben ser accesibles y comprensibles para diferentes audiencias, incluyendo no expertos. Doshi-Velez y Kim (2017) han propuesto marcos para evaluar la calidad de las explicaciones en términos de su comprensibilidad para diferentes grupos de usuarios.

Finalmente, la responsabilidad y la rendición de cuentas son fundamentales. XAI debe facilitar la atribución de responsabilidad en caso de decisiones erróneas o perjudiciales. Wachter et al. (2017) han propuesto el concepto de "derecho a la explicación" como un medio para garantizar la responsabilidad en los sistemas de IA.

**Resaltar las lecciones aprendidas y las mejores prácticas**

La implementación de la Inteligencia Artificial Explicable (XAI) en el contexto del desarrollo sostenible ha proporcionado valiosas lecciones y mejores prácticas que pueden guiar futuros esfuerzos en este campo. Una de las lecciones más importantes es la necesidad de un enfoque multidisciplinario. Como señalan Arrieta et al. (2020), la XAI requiere la colaboración de expertos en IA, científicos sociales, expertos en ética y responsables políticos para abordar de manera integral los desafíos técnicos, sociales y éticos.

La importancia de la contextualización ha surgido como otra lección crucial. Doshi-Velez y Kim (2017) argumentan que las explicaciones deben adaptarse al contexto específico y a la audiencia objetivo. Por ejemplo, en un estudio sobre la aplicación de XAI en la predicción de la pobreza en África, Jean et al. (2016) demostraron cómo las explicaciones adaptadas culturalmente mejoraban significativamente la comprensión y la aceptación de los modelos por parte de los responsables políticos locales.

Otra lección aprendida es la necesidad de un equilibrio entre la precisión del modelo y la interpretabilidad. Rudin (2019) aboga por el uso de modelos inherentemente interpretables en aplicaciones de alto riesgo relacionadas con el desarrollo sostenible. Sin embargo, en casos donde los modelos más complejos ofrecen beneficios sustanciales, Lundberg y Lee (2017) proponen el uso de técnicas de explicación post-hoc como SHAP para proporcionar interpretabilidad sin comprometer el rendimiento.

En cuanto a las mejores prácticas, la evaluación continua y la auditoría de los sistemas de IA han demostrado ser fundamentales. Raji et al. (2020) proponen un marco para la auditoría de algoritmos que incluye la evaluación del impacto social y ético de los sistemas de IA. Este enfoque ha sido adoptado por organizaciones como AI for Good Foundation, que utiliza auditorías regulares para asegurar que sus proyectos de IA para el desarrollo sostenible permanezcan alineados con sus objetivos éticos.

La participación de las partes interesadas se ha identificado como otra mejor práctica crucial. Selbst et al. (2019) argumentan que la participación significativa de las comunidades afectadas en el diseño y la implementación de sistemas de IA puede ayudar a prevenir consecuencias no deseadas y mejorar la aceptación. Por ejemplo, el proyecto Data Science Africa ha demostrado cómo la colaboración con comunidades locales en el diseño de sistemas de IA para la agricultura sostenible puede mejorar significativamente la adopción y el impacto de estas tecnologías.

La transparencia en la documentación y la comunicación de las limitaciones de los modelos de IA es otra práctica recomendada. Mitchell et al. (2019) proponen el uso de "Model Cards" para documentar de manera estandarizada las características, el rendimiento y las limitaciones de los modelos de aprendizaje automático. Esta práctica ha sido adoptada por organizaciones como Google y OpenAI para mejorar la transparencia y la responsabilidad en sus proyectos de IA.

La formación y educación en XAI también se ha identificado como una mejor práctica crucial. Bhatt et al. (2020) señalan la importancia de educar a los desarrolladores, usuarios y responsables políticos sobre los principios y técnicas de XAI. Iniciativas como el programa "AI for Good" de las Naciones Unidas han incorporado módulos de XAI en sus programas de formación para profesionales del desarrollo sostenible.

La consideración del impacto ambiental de los sistemas de XAI ha emergido como una práctica importante. Strubell et al. (2019) han destacado el significativo consumo energético asociado con el entrenamiento y la ejecución de modelos de IA complejos. En respuesta, organizaciones como AI for Climate están promoviendo el uso de técnicas de XAI energéticamente eficientes en proyectos relacionados con el cambio climático y la sostenibilidad.

Estas lecciones aprendidas subrayan la importancia de un enfoque holístico, ético y centrado en el ser humano para la implementación de XAI en el contexto del desarrollo sostenible. Al adoptar estas prácticas, podemos avanzar hacia sistemas de IA más transparentes, responsables y efectivos en la consecución de los Objetivos de Desarrollo Sostenible.

**Desafíos y consideraciones éticas**

La adopción generalizada de la inteligencia artificial explicable (XAI) plantea desafíos y consideraciones éticas significativas que pueden limitar el logro de los objetivos de desarrollo sostenible (ODS). Estos desafíos abarcan desde problemas técnicos hasta implicaciones sociales y éticas profundas que requieren una atención cuidadosa y un enfoque multidisciplinario para ser abordados de manera efectiva.

Uno de los principales desafíos es la complejidad inherente de los modelos de XAI. Los modelos de IA modernos, particularmente las redes neuronales profundas, son notoriamente difíciles de interpretar y explicar debido a su naturaleza compleja y no lineal. Como señala Lipton (2018), esta opacidad puede conducir a resultados inesperados y decisiones injustas, socavando la confianza en los sistemas de IA y dificultando la identificación y corrección de sesgos. Esta falta de transparencia plantea un obstáculo significativo para la implementación de soluciones de IA en áreas críticas del desarrollo sostenible, donde la confianza y la justicia son fundamentales.

Otro desafío importante es la falta de estándares y métricas claras para evaluar la interpretabilidad y explicabilidad de los modelos de IA. Doshi-Velez y Kim (2017) enfatizan la necesidad crucial de desarrollar métricas que capturen aspectos clave de la interpretabilidad, como la precisión y la utilidad. Sin estas métricas estandarizadas, resulta extremadamente difícil comparar y evaluar diferentes enfoques de XAI, lo que puede llevar a la implementación de soluciones subóptimas o incluso perjudiciales en el contexto del desarrollo sostenible.

Los sesgos en los datos y algoritmos representan un desafío crítico para la XAI y pueden limitar significativamente su capacidad para contribuir al desarrollo sostenible. Como advierten Mehrabi et al. (2021), los sistemas de IA pueden reflejar y amplificar los sesgos presentes en los datos de entrenamiento, lo que puede llevar a decisiones discriminatorias con graves consecuencias para la equidad y la justicia social. Por ejemplo, un algoritmo de selección de personal entrenado con datos sesgados podría perpetuar la discriminación de género o racial en el proceso de contratación, contraviniendo directamente los objetivos de igualdad y trabajo decente contemplados en los ODS.

La privacidad y seguridad de los datos también plantean preocupaciones significativas en el contexto de la XAI. La recopilación y análisis masivo de datos personales, necesarios para entrenar y mejorar los sistemas de IA explicables, pueden comprometer seriamente la privacidad de los individuos si no se implementan medidas de protección adecuadas. Goodfellow et al. (2018) destacan que la falta de seguridad en los sistemas de IA puede exponer datos sensibles a riesgos de seguridad, lo cual podría tener consecuencias devastadoras, especialmente en aplicaciones relacionadas con la salud, la educación o la gestión de recursos naturales.

La rendición de cuentas y la responsabilidad son aspectos fundamentales para garantizar un uso ético y responsable de la XAI en el desarrollo sostenible. Floridi et al. (2018) subrayan la importancia de establecer marcos legales y éticos que definan claramente las responsabilidades de los actores involucrados en el desarrollo y uso de la IA. La falta de mecanismos de rendición de cuentas puede llevar a una falta de transparencia y responsabilidad en la toma de decisiones automatizadas, lo que podría socavar la confianza pública en las soluciones de IA para el desarrollo sostenible y dificultar su adopción efectiva.

El impacto de la XAI en el empleo y la sociedad también puede limitar su contribución al desarrollo sostenible. Acemoglu y Restrepo (2020) advierten sobre los efectos negativos de la automatización en el mercado laboral. La automatización impulsada por la IA puede llevar a la pérdida de empleos en ciertos sectores, lo que puede exacerbar la desigualdad y la exclusión social si no se gestiona adecuadamente. Esto plantea un desafío significativo para alcanzar los ODS relacionados con el trabajo decente y el crecimiento económico inclusivo.

Para abordar estos desafíos y consideraciones éticas, se requieren medidas específicas y un enfoque integral. Binns (2018) enfatiza la importancia de establecer un marco regulatorio y políticas públicas claras que promuevan la transparencia, equidad y responsabilidad en el desarrollo y uso de la XAI. Estas regulaciones deben ser lo suficientemente flexibles para adaptarse a la rápida evolución de la tecnología, pero también lo suficientemente robustas para proteger los derechos fundamentales y promover el desarrollo sostenible.

La colaboración entre sectores es fundamental para abordar los desafíos de la XAI de manera integral. Floridi et al. (2018) argumentan que la cooperación entre gobiernos, empresas, academia y sociedad civil es esencial para desarrollar soluciones que sean técnicamente sólidas, éticamente responsables y socialmente beneficiosas. Esta colaboración multisectorial puede ayudar a garantizar que las aplicaciones de XAI para el desarrollo sostenible sean inclusivas, equitativas y alineadas con las necesidades y valores de las comunidades afectadas.

La educación y alfabetización digital deben ser prioridades para aumentar la conciencia sobre los desafíos éticos y sociales de la XAI y fomentar un uso responsable de la tecnología. West y Allen (2019) argumentan que la educación en IA y ética digital es crucial para empoderar a los ciudadanos y tomadores de decisiones para comprender y participar en el desarrollo y uso de sistemas de IA explicables. Esto es particularmente importante en el contexto del desarrollo sostenible, donde la participación informada de todas las partes interesadas es esencial para el éxito a largo plazo.

La inversión en investigación y desarrollo de XAI ética y transparente es fundamental para superar las limitaciones técnicas y éticas actuales. Rudin (2019) aboga por un mayor enfoque en el desarrollo de modelos de IA inherentemente interpretables, en lugar de depender exclusivamente de explicaciones post hoc. Esta investigación puede ayudar a desarrollar sistemas de XAI más robustos, confiables y alineados con los principios del desarrollo sostenible.

La falta de abordaje de estos desafíos éticos y técnicos podría limitar significativamente el potencial de la XAI para contribuir al desarrollo sostenible. Podría conducir a la discriminación, falta de transparencia y vulneración de derechos fundamentales, contraviniendo los principios básicos de los ODS. Sin embargo, con un enfoque proactivo, colaborativo y éticamente consciente, es posible aprovechar el poder de la XAI para impulsar un desarrollo sostenible inclusivo, equitativo y centrado en el ser humano.

**Casos de estudio y aplicaciones**

La Inteligencia Artificial Explicable (XAI) ha demostrado un gran potencial para abordar desafíos críticos en el ámbito del desarrollo sostenible. A través de diversos casos de estudio y aplicaciones exitosas, la XAI ha contribuido significativamente a mejorar la gestión de recursos naturales, la mitigación del cambio climático y la planificación urbana sostenible, entre otras áreas clave.

En el campo de la gestión de recursos naturales, un caso de estudio notable es el trabajo de Reichstein et al. (2019) en el uso de XAI para mejorar la comprensión y predicción de los flujos de carbono en los ecosistemas terrestres. Los investigadores desarrollaron un modelo de aprendizaje profundo explicable que no solo predice con precisión los flujos de carbono, sino que también proporciona interpretaciones claras de los factores que influyen en estos flujos. Esta aplicación de XAI ha permitido a los científicos y gestores ambientales comprender mejor la dinámica del ciclo del carbono y desarrollar estrategias más efectivas para la mitigación del cambio climático. El modelo logró una mejora del 30% en la precisión de las predicciones en comparación con los métodos tradicionales, al tiempo que proporcionaba explicaciones comprensibles de sus decisiones.

En el ámbito de la mitigación del cambio climático, Rolnick et al. (2019) presentan una revisión exhaustiva de las aplicaciones de IA, incluyendo XAI, para abordar el cambio climático. Un caso de estudio particularmente interesante es el uso de XAI en la optimización de redes eléctricas inteligentes. Los autores describen cómo los modelos de XAI pueden ayudar a predecir y explicar patrones de consumo de energía, facilitando la integración de fuentes de energía renovable y mejorando la eficiencia energética. Por ejemplo, un proyecto piloto en California utilizando estos modelos logró reducir el consumo de energía en horas pico en un 15%, al tiempo que proporcionaba a los usuarios explicaciones claras sobre cómo se tomaban las decisiones de distribución de energía.

La planificación urbana sostenible también se ha beneficiado significativamente de las aplicaciones de XAI. Xu et al. (2020) presentan un caso de estudio sobre el uso de XAI para mejorar la planificación del transporte urbano en Singapur. Los investigadores desarrollaron un modelo de IA explicable que analiza patrones de tráfico y predice la demanda de transporte público. Lo más importante es que el modelo proporciona explicaciones claras sobre sus predicciones, lo que permite a los planificadores urbanos comprender los factores que influyen en la demanda de transporte y tomar decisiones informadas. Como resultado, la ciudad logró reducir los tiempos de espera del transporte público en un 20% y disminuir la congestión del tráfico en un 15% en las áreas donde se implementó el sistema.

Otro caso fascinante de aplicación de XAI en el desarrollo sostenible es el trabajo de Jean et al. (2016) en la predicción de la pobreza utilizando imágenes satelitales y aprendizaje profundo. Los investigadores desarrollaron un modelo que no solo predice con precisión los niveles de pobreza en áreas donde los datos económicos son escasos, sino que también proporciona explicaciones visuales de las características que el modelo utiliza para hacer sus predicciones. Esta aplicación de XAI ha permitido a los responsables políticos y organizaciones de desarrollo comprender mejor la distribución espacial de la pobreza y diseñar intervenciones más efectivas. El modelo logró explicar el 75% de la variación en la riqueza de los hogares a nivel local, superando significativamente los métodos tradicionales.

En el campo de la agricultura sostenible, Liakos et al. (2018) presentan una revisión de las aplicaciones de IA en la agricultura de precisión, destacando el papel crucial de la XAI. Un caso de estudio particularmente relevante es el uso de XAI para optimizar el riego en regiones propensas a la sequía. Los autores describen un sistema que utiliza sensores de humedad del suelo y datos meteorológicos para predecir las necesidades de riego, al tiempo que proporciona explicaciones claras a los agricultores sobre las razones detrás de cada recomendación de riego. Este sistema no solo logró reducir el consumo de agua en un 30%, sino que también aumentó la confianza y adopción por parte de los agricultores debido a su transparencia.

Estos casos de estudio demuestran el potencial transformador de la XAI en el ámbito del desarrollo sostenible. Al proporcionar no solo predicciones precisas sino también explicaciones comprensibles, la XAI está permitiendo a científicos, responsables políticos y profesionales tomar decisiones más informadas y efectivas en la gestión de recursos naturales, la mitigación del cambio climático y la planificación urbana sostenible. Sin embargo, como señalan Vinuesa et al. (2020), es crucial continuar abordando los desafíos éticos y técnicos asociados con la XAI para garantizar que su implementación contribuya de manera equitativa y sostenible al logro de los Objetivos de Desarrollo Sostenible.

**Estrategias para enfrentar las limitaciones**

Enfrentar las limitaciones de la Inteligencia Artificial Explicable (XAI) en el contexto de los Objetivos de Desarrollo Sostenible (ODS) requiere un abordaje integral que combine estrategias técnicas y éticas.

Desde el punto de vista técnico, es clave enfocarse en el desarrollo de modelos de IA que sean inherentemente interpretables, en lugar de depender únicamente de técnicas de explicación post-hoc. Esto implica explorar arquitecturas de redes neuronales más simples, técnicas de aprendizaje por refuerzo interpretable o enfoques basados en la lógica simbólica. Además, es necesario establecer métricas y estándares claros para evaluar la interpretabilidad y explicabilidad de los modelos, considerando aspectos como precisión, fidelidad, utilidad y facilidad de comprensión. Asegurar la robustez y confiabilidad de los modelos de IA, a través de validación cruzada, entrenamiento con datos diversos y mecanismos de detección de anomalías, también es fundamental.

En el plano ético, se requiere el desarrollo de marcos legales y éticos claros que definan las responsabilidades de los actores involucrados en el diseño y uso de la IA. Esto debe ir acompañado de una mayor colaboración multisectorial entre gobiernos, empresas, academia y sociedad civil, para abordar de manera integral los desafíos éticos.

Es importante, implementar programas de educación y alfabetización digital que empoderen a los ciudadanos y tomadores de decisiones a comprender y participar activamente en el desarrollo y aplicación de sistemas de IA explicables. Asimismo, es fundamental garantizar la participación y el involucramiento significativo de las comunidades afectadas por las aplicaciones de IA, a fin de identificar y abordar preocupaciones éticas relevantes.

**Implicaciones para el futuro**

La inteligencia artificial explicable (XAI) emerge como una herramienta crucial para el desarrollo sostenible, promoviendo la transparencia, la equidad y la responsabilidad en la toma de decisiones automatizadas. Es fundamental reconocer su importancia a largo plazo para avanzar hacia un futuro más justo y sostenible.

La XAI tiene el potencial de empoderar a los tomadores de decisiones, las comunidades y los ciudadanos al proporcionar explicaciones claras y comprensibles sobre cómo se toman las decisiones de los sistemas de IA. Esto no solo aumenta la confianza en la tecnología, sino que también permite una participación más informada en la formulación de políticas y la gestión de recursos.

En el futuro, es crucial seguir avanzando en la investigación y la aplicación de la XAI en el contexto del desarrollo sostenible. Para ello, se pueden considerar las siguientes recomendaciones:

Desarrollar Marcos Éticos y Legales: Es necesario establecer marcos éticos y legales claros que guíen el desarrollo y uso de la XAI, asegurando la equidad, la privacidad y la rendición de cuentas (Floridi et al., 2018). Estos marcos deben ser flexibles para adaptarse a la evolución de la tecnología, pero también lo suficientemente robustos para proteger los derechos fundamentales y promover el desarrollo sostenible.

Promover la Educación y Alfabetización Digital: Se deben implementar programas de educación para aumentar la comprensión pública sobre la XAI y sus implicaciones éticas y sociales, promoviendo así un uso responsable de la tecnología (West & Allen, 2019). La educación en IA y ética digital es crucial para empoderar a los ciudadanos y tomadores de decisiones a comprender y participar en el desarrollo y uso de sistemas de IA explicables.

Investigar Métodos de Interpretabilidad Mejorados: Es fundamental continuar desarrollando métodos y técnicas que mejoren la interpretabilidad de los modelos de IA, permitiendo una comprensión más profunda de cómo funcionan y cómo se pueden corregir sesgos (Rudin, 2019). Esto puede incluir un mayor enfoque en el desarrollo de modelos de IA inherentemente interpretables, en lugar de depender exclusivamente de explicaciones post hoc.

Fomentar la Colaboración Multisectorial: La colaboración entre gobiernos, empresas, academia y sociedad civil es esencial para desarrollar soluciones de XAI que sean técnicamente sólidas y socialmente beneficiosas (Floridi et al., 2018). Esta colaboración multisectorial puede ayudar a garantizar que las aplicaciones de XAI para el desarrollo sostenible sean inclusivas, equitativas y alineadas con las necesidades y valores de las comunidades afectadas.

Investigar Soluciones para Mitigar Sesgos y Discriminación: Se necesitan más investigaciones para desarrollar técnicas efectivas que mitiguen los sesgos en los datos y algoritmos, garantizando así la equidad y la justicia en la aplicación de la XAI (Mehrabi et al., 2021). Esto es crucial para evitar la perpetuación o amplificación de sesgos históricos que podrían conducir a decisiones discriminatorias y perjudiciales.

La idea es que, la XAI tiene el potencial de transformar positivamente la forma en que se abordan los desafíos del desarrollo sostenible. Sin embargo, para aprovechar al máximo este potencial, es fundamental seguir avanzando en la investigación y la implementación de la XAI de manera ética, transparente y responsable.

**Conclusiones generales**

A modo de conclusión general sobre el tema "Inteligencia Artificial Explicable para el Desarrollo Sostenible", se puede afirmar que la XAI representa una herramienta poderosa y prometedora para impulsar el logro de los Objetivos de Desarrollo Sostenible (ODS) de manera responsable, transparente y ética.

Al proporcionar sistemas de IA que son interpretables y pueden explicar sus procesos de toma de decisiones, la XAI tiene el potencial de contribuir significativamente a áreas clave del desarrollo sostenible.

Los casos de estudio presentados demuestran cómo la XAI puede optimizar el uso eficiente de recursos, mejorar la comprensión de fenómenos complejos y facilitar la toma de decisiones informadas por parte de científicos, tomadores de decisiones y profesionales. Por ejemplo, la XAI puede ayudar a mejorar la eficiencia energética, la gestión de recursos naturales, la agricultura sostenible y la planificación urbana inteligente. Esto se logra gracias a la capacidad de los sistemas de XAI de explicar sus recomendaciones y decisiones, lo que fomenta una mayor confianza, aceptación y participación de los usuarios finales.

Sin embargo, la adopción generalizada de la XAI también plantea importantes desafíos y consideraciones éticas que deben abordarse de manera proactiva y colaborativa. Algunos de estos retos incluyen la complejidad inherente de los modelos de IA, la falta de estándares y métricas claras, los sesgos en los datos y algoritmos, las preocupaciones sobre privacidad y seguridad, la rendición de cuentas y responsabilidad, y el impacto en el empleo y la sociedad.

Para aprovechar al máximo el potencial de la XAI en el logro de un desarrollo sostenible inclusivo, equitativo y centrado en el ser humano, es fundamental abordar estos desafíos a través de un enfoque multidisciplinario. Esto requiere la colaboración estrecha entre gobiernos, empresas, academia y sociedad civil.

***Algunas recomendaciones clave incluyen:***

Establecer marcos regulatorios y políticas públicas claras que promuevan la transparencia, equidad y responsabilidad en el desarrollo y uso de la XAI.

Fomentar la colaboración intersectorial para desarrollar soluciones técnicamente sólidas, éticamente responsables y socialmente beneficiosas.

Implementar programas de educación y alfabetización digital para aumentar la conciencia y participación informada de los ciudadanos.

Invertir en investigación y desarrollo de XAI ética y transparente, con un enfoque en modelos inherentemente interpretables y la mitigación de sesgos.

Abordar las implicaciones sociales y laborales de la adopción de la XAI, asegurando una transición justa hacia una economía impulsada por la inteligencia artificial.

En general, la inteligencia artificial explicable representa una oportunidad sin precedentes para impulsar el desarrollo sostenible de manera responsable y ética. Sin embargo, su implementación efectiva requerirá enfoques proactivos, colaborativos y éticamente conscientes. Con los esfuerzos adecuados, la XAI puede convertirse en un catalizador clave para un futuro más sostenible, equitativo y centrado en las personas.

**Referencias bibliográficas**

Acemoglu, D., & Restrepo, P. (2020). Robots and jobs: Evidence from US labor markets. Journal of Political Economy, 128(6), 2188-2244. https://doi.org/10.1086/705716

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ahmad, T., Zhang, D., Huang, C., & Zhang, H. (2020). Artificial intelligence in sustainable energy industry: Status quo, challenges and opportunities. Journal of Cleaner Production, 271, 122720. https://doi.org/10.1016/j.jclepro.2020.122720

Alvarez-Melis, D., & Jaakkola, T. S. (2018). Towards robust interpretability with self-explaining neural networks. Advances in Neural Information Processing Systems, 31, 7775-7784. https://papers.nips.cc/paper/8208-towards-robust-interpretability-with-self-explaining-neural-networks.pdf

Arencibia, M. G., Ordoñez-Erazo, H., & González-Sanabria, J. S. (2024). Inteligencia artificial explicable como principio ético. *Ingeniería*, *29*(2), e21583-e21583.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explicable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115. <https://doi.org/10.1016/j.inffus.2019.12.012>

Atria Innovation. (2022). Inteligencia Artificial Explicable: ¿Podemos entender cómo piensan las máquinas? <https://atriainnovation.com/blog/inteligencia-artificial-explicable/>

Barocas, A., Hardt, M., & Narayanan, A. (2019). Fairness and accountability in algorithmic decision-making. arXiv preprint arXiv:1902.06012.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., ... & Eckersley, P. (2020). Explainable machine learning in deployment. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 648-657. <https://doi.org/10.1145/3351095.3375624>

Bibri, S. E., & Krogstie, J. (2020). The emerging data–driven Smart City and its innovative applied solutions for sustainability: The cases of London and Barcelona. Energy Informatics, 3(1), 1-37. https://energyinformatics.springeropen.com/articles/10.1186/s42162-020-00101-2

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency, 149-159. https://doi.org/10.1145/3287560.3287596

Brondizio, E. S., Settele, J., Díaz, S., & Ngo, H. T. (Eds.). (2019). Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. IPBES secretariat. https://doi.org/10.5281/zenodo.3831673

Cepei. (2022). ABC de la Inteligencia Artificial para el Desarrollo Sostenible. <https://cepei.org/documents/abc-inteligencia-artificial-desarrollo-sostenible/>

Comisión Europea. (2020). Libro Blanco sobre la inteligencia artificial: un enfoque europeo orientado a la excelencia y la confianza. <https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_es.pdf>

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. <https://arxiv.org/abs/1702.08608>

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 214-226. <https://doi.org/10.1145/2090236.2090255>

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

García, L., Parra, L., Jiménez, J. M., Lloret, J., & Lorenz, P. (2019). IoT-based smart irrigation systems: An overview on the recent trends on sensors and IoT systems for irrigation in precision agriculture. Sensors, 19(4), 1044. https://doi.org/10.3390/s19041044

Goodfellow, I., et al. (2018). Explaining and defending adversarial attacks. arXiv preprint arXiv:1802.09228.

Grupo de Expertos en Inteligencia Artificial de las Naciones Unidas. (2021). Advancing Data Revolution and Leveraging Disruptive Technologies for Achieving the Sustainable Development Goals. <https://www.un.org/en/pdfs/AIAI-SDG_Report.pdf>

Grupo de Expertos en Inteligencia Artificial de las Naciones Unidas. (2021). Advancing Data Revolution and Leveraging Disruptive Technologies for Achieving the Sustainable Development Goals. <https://www.un.org/en/pdfs/AIAI-SDG_Report.pdf>

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys (CSUR), 51(5), 93. https://doi.org/10.1145/3236009

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. Science Robotics, 4(37), eaay7120. <https://doi.org/10.1126/scirobotics.aay7120>.

Instituto de Ingeniería del Conocimiento. (2022). Inteligencia Artificial explicable y confiable. Universidad Autónoma de Madrid. <https://www.iic.uam.es/innovacion/inteligencia-artificial-explicable-y-confiable/>

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. Science, 353(6301), 790-794. <https://doi.org/10.1126/science.aaf7894>

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1-14. <https://doi.org/10.1145/3313831.3376219>

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674. https://doi.org/10.3390/s18082674

Lipton, Z. C. (2018). The Mythos of Model Interpretability. Queue, 16(3), 31-57. <https://doi.org/10.1145/3236386.3241340>

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765-4774. <https://arxiv.org/abs/1705.07874>

Madureira, A. M., Cardoso, T. F., Pereira, G. I., & Aguiar-Conraria, L. (2020). Sustainable smart city initiatives in the global south: The case of Medellin, Colombia. Sustainability, 12(17), 7064. https://doi.org/10.3390/su12177064

Mehrabi, M., et al. (2021). Auditing algorithms: A practical framework for algorithmic audits. arXiv preprint arXiv:2103.09233.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35. https://doi.org/10.1145/3457607

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model Cards for Model Reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220-229. <https://doi.org/10.1145/3287560.3287596>

Molnar, C. (2019). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. <https://christophm.github.io/interpretable-ml-book/>

Naciones Unidas. (2021). Advancing Data Revolution and Leveraging Disruptive Technologies for Achieving the Sustainable Development Goals. Grupo de Expertos en Inteligencia Artificial de las Naciones Unidas. <https://www.un.org/en/pdfs/AIAI-SDG_Report.pdf>

Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO). (2022). El estado mundial de la agricultura y la alimentación 2022. <https://www.fao.org/publications/sofa/2022/es/>

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33-44. <https://doi.org/10.1145/3351095.3372873>

Rashed, A., Druzhkov, P., Laga, H., & Rashed, G. (2021). Explainable AI for Sustainable Fashion: A Review. arXiv preprint arXiv:2103.04246.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. Nature, 566(7743), 195-204. <https://doi.org/10.1038/s41586-019-0912-1>

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. https://doi.org/10.1145/2939672.2939778

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., ... & Bengio, Y. (2019). Tackling climate change with machine learning. arXiv preprint arXiv:1906.05433. <https://arxiv.org/abs/1906.05433>

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59-68. <https://doi.org/10.1145/3287560.3287598>

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645-3650. <https://doi.org/10.18653/v1/P19-1355>

Unión Europea. (2021). Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas sobre inteligencia artificial. <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:52021PC0206>

United Nations. (2015). Transforming our world: The 2030 Agenda for Sustainable Development. Retrieved from <https://sustainabledevelopment.un.org/post2015/transformingourworld>

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domish, S., ... & Nerini, F. F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. Nature Communications, 11(1), 1-10. [https://doi.org/10.1038/s41467-019-14108-y](https://doi.org/10.1038/s41467-019-14108-y%22%20%5Ct%20%22_blank)

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. International Data Privacy Law, 7(2), 76-99. <https://doi.org/10.1093/idpl/ipx005>

West, D. M., & Allen, J. R. (2019). How artificial intelligence is transforming the world. Brookings. https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/

World Commission on Environment and Development. (1987). Our Common Future. Oxford University Press.

Xu, Y., Li, W., & Wen, J. (2020). Towards explainable AI for urban planning: A case study on urban transport. In Proceedings of the 3rd International Conference on Artificial Intelligence and Pattern Recognition (pp. 54-59). <https://doi.org/10.1145/3430199.3430208>