La explicabilidad como principio clave para la IA ética

La IA está comenzando a generar cambios sociales importantes y por ello comprender la dimensión ética de su aplicación es clave. La elección de los principios éticos adecuados para los diferentes casos de uso es, por tanto, relevante para conseguirlo. La explicabilidad es, en muchos de estos escenarios, uno de los principios clave de la IA ética por su contribución a la transparencia en la toma de decisiones. Sería deseable incluir la ética en la IA desde el diseño.

La inteligencia artificial está comenzando a transformar la economía, el trabajo, las relaciones personales y la sociedad en todo el planeta. Además, desde hace unos meses, la IA generativa con implementaciones como ChatGPT está suponiendo un cambio social similar al llevado a cabo con el surgimiento de Internet. Los efectos positivos son incontables, pero presenta también algunos desafíos. Uno de esos desafíos es la ética, dimensión ineludible de cualquier actividad profesional y que, en el caso de la IA, presenta retos adicionales frente a otras tecnologías. Seguramente todos recordamos aquellos ejemplos donde un algoritmo de Amazon daba preferencia en la selección de currículum a los hombres, o los reconocimientos visuales erróneos cuando las personas no eran caucásicos o incluso la reciente prohibición de ChatGPT en países como Italia por temas de privacidad y derechos de autor. Todos estos ejemplos, junto a otros muchos, ponen de manifiesto la necesidad de tener en cuenta la ética en todos los procesos donde se aplica la IA.

La ética busca dar respuesta a la pregunta qué debo hacer. Las personas contamos con la libertad para elegir una determinada acción, con la consciencia para reflexionar sobre las posibles consecuencias y con la voluntad para elegir la acción más adecuada. La ética se concibe a veces como un freno que puede ralentizar la innovación, pero se trata precisamente de lo contrario. Si pensamos en la función de los frenos en los automóviles, estos nos proporcionan la capacidad de desplazarnos a una velocidad mayor con la confianza que podemos reaccionar a los imprevistos gracias a ellos. No se trata de no aprovechar los beneficios de esta tecnología sino de mitigar los riesgos y prever posibles problemas en su uso como los mencionados anteriormente. La IA nos proporciona, entre otras capacidades, la de automatizar la toma de decisiones por ello la evaluación de riesgos es fundamental, porque no es lo mismo utilizar la capacidad de toma de decisiones automáticas para recomendar una película que para tomar una decisión médica o conceder un préstamo. Esa posible pérdida de control o dilución aparente de la responsabilidad, por el hecho de que las máquinas tomen decisiones automáticas por nosotros, puede tener importantes implicaciones éticas.

Si revisamos en algo más de detalle que significa la toma de decisiones automática, generalmente se establece una clasificación en tres tipos de sistemas: sistemas de soporte a la toma de decisiones, sistemas semiautónomos y completamente autónomos. En los primeros, el sistema de IA proporciona los resultados, pero la decisión siempre la realiza una persona. En el segundo caso, la toma de decisiones es automática pero la persona puede, en determinadas circunstancias, evaluar y cambiar la decisión. Por último, en los sistemas completamente autónomos, las decisiones son tomadas por el sistema de IA, y la persona solo interviene en la configuración y parametrización del sistema, es decir, en el modo en que se toman el conjunto de todas las decisiones. Son, estos dos últimos tipos de sistemas, los que plantean una reflexión ética más profunda sobre cómo debe establecerse la relación de los seres humanos con los sistemas de IA.

La ética se concibe a veces como un freno que puede ralentizar la innovación, pero se trata precisamente de lo contrario

No se trata de juzgar si la tecnología es aceptable o no, se trata de desarrollar un marco ético que acompañe a los sistemas de IA en su ciclo de vida, desde la concepción, el desarrollo, la programación, la puesta en servicio, el mantenimiento y la explotación, lo cual se conoce como ética por diseño. Para ello es clave analizar los posibles riesgos e impactos de estos sistemas y tratar de anticipar las posibles implicaciones para todas las partes involucradas en cada fase del ciclo de vida. No se trata sólo de contar con expertos en ética o en IA, se trata más bien de un trabajo multidisciplinar, así como una interlocución adecuada entre todas las partes implicadas en el desarrollo de los sistemas de IA: desarrolladores, investigadores, clientes, usuarios, administración y empresas. Esto requiere de cada uno de nosotros un enfoque no de mero observador, sino como agente para guiar el uso ético de la IA en la sociedad.

Los principios éticos

Respondiendo a esa necesidad de actuar como agentes para intentar dar respuesta a los retos éticos de la IA, desde 2017 se han publicado multitud de documentos por parte de gobiernos, empresas privadas, investigadores y diferentes organizaciones que resaltan la importancia de establecer unos principios éticos de la IA. Estos principios tienen como objetivo ayudar a preservar los derechos y libertades de las personas sin frenar la innovación tecnológica. En lo que se refiere a derechos y libertades sin duda están muy influidos por los derechos humanos, pero no es sencillo decidir o establecer cuales son los principios clave o cuales son más importantes que otros, en la mayoría de los casos va a depender del caso de uso o del tipo de empresa.

AIEthicsLab[1], organismo independiente que cuenta con investigadores de distintas disciplinas, ha intentado categorizar los documentos para encontrar principios comunes. Para ello ha desarrollado una herramienta que recoge decenas de documentos que mencionan muchos de estos principios y ha llegado a la conclusión que la mayoría de los principios éticos pueden agruparse en cuatro categorías clave: autonomía humana, no hacer daño, crear beneficios y justicia. En estas categorías se pueden encontrar principios como: la privacidad, el control de riesgos, la explicabilidad, la transparencia y la equidad o la responsabilidad entre otros.

Otros organismos como la UNESCO[2] publicaron a finales de 2021 un documento con las recomendaciones acerca de la ética de la IA. El documento recuerda las repercusiones positivas de la IA, pero alerta también sobre los posibles riesgos y la necesidad de establecer un marco ético para su uso. Asimismo, menciona diez principios éticos entre los que se incluyen responsabilidad y rendición de cuentas, supervisión y decisión humana, protección de datos o transparencia y explicabilidad. Lo que hace la recomendación excepcional es que también establece una serie de políticas de actuación que permite traducir los principios en acciones con respecto el género, la educación, la salud y bienestar social y la gobernanza de los datos. Aunque la UNESCO delega su aplicación en los estados miembros, que a su vez tendrán que incluirlo en las estrategias nacionales de IA, su publicación hace patente la importancia de estos principios a nivel mundial.

La explicabilidad

El principio de explicabilidad no es el único principio clave, pero si es uno de los que contribuye de manera más importante a conformar una IA ética. La explicabilidad tiene que proporcionar las claves para comprender la toma de decisiones del sistema de IA y de esta manera favorecer la transparencia. Por un lado, desde el

punto de vista más técnico, es importante entender qué tipo de algoritmos se han utilizado, ya que no todos los algoritmos pueden ser interpretados igualmente. Por otro lado, es importante también centrarse en el tipo de explicaciones proporcionadas, no ya en un plano técnico, sino adaptadas al contexto de la explicación. Por ejemplo, si se está entrenando un sistema de IA para la concesión de un préstamo, el diseñador del algoritmo, normalmente un científico de datos, estará más interesado en entender las métricas clave del algoritmo y si estas han mejorado desde el último entrenamiento. En cambio, la persona que pidió el crédito, que no sabe nada de algoritmos, va a estar mucho más interesada en una explicación más descriptiva sobre por qué no se le concedió ese préstamo, haciendo referencia por ejemplo al saldo, la antigüedad o el número de cuentas que tiene en ese banco. Es decir, cuando hablamos de explicablidad, no sólo estamos hablando de la parte más tecnológica, referida a la interpretabilidad de los algoritmos, sino también del tipo de explicaciones que se van a proporcionar teniendo siempre en cuenta la persona que va a recibir estas explicaciones. Esta última parte está mucho más relacionada con la psicología de las explicaciones, y durante más de 20 años la psicología cognitiva ha estado investigando cómo las personas generan las explicaciones y cómo evaluar la calidad de éstas. Actualmente se sigue trabajando en líneas de actuación multidisciplinares entre equipos de tecnología y de psicología cognitiva, uno de los estudios más completos al respecto, es la iniciativa de DARPA (Agencia de proyectos de investigación avanzada) al respecto, que resume el trabajo de estos equipos multidisciplinares a lo largo de cinco años.

Sería deseable incluir la ética en la IA desde el diseño, sin la necesidad de una regulación que obligue a ello

Si nos centramos en los algoritmos en términos de interpretabilidad, se pueden dividir básicamente en cajas negras y cajas blancas. Los primeros de ellos tienen en general una interpretabilidad muy baja, es decir, es complicado poder entender como ha sido tomada la decisión, de ahí su nombre. Son algoritmos normalmente más complejos como redes neuronales, *random forest* o *xgboost* entre otros, pero suelen presentar una mejor precisión en sus resultados. Por el contrario, los algoritmos de caja blanca, como su nombre indica, ofrecen una interpretabilidad mayor que permite entender mejor la toma de decisiones llevada a cabo, suelen ser algoritmos de regresión lineal o árboles de decisión. De cara a poder proporcionar una mejor explicabilidad, aun perdiendo precisión, se suelen usar de manera más habitual este tipo de algoritmos por la facilidad a la hora de entender sus resultados y poder explicarlos mejor. No obstante, en los últimos años, las técnicas de interpretabilidad de algoritmos han avanzado mucho y esto está contribuyendo a poder proporcionar una mayor interpretabilidad a los algoritmos de caja negra. SHAP y LIME son técnicas cada vez más usadas que permiten explicar algoritmos complejos mediante la utilización de técnicas más sencillas. Asimismo, en los últimos cinco años se han desarrollado herramientas Open Source y comerciales que facilitan este análisis de interpretabilidad.

En definitiva, queda aún mucho trabajo por hacer respecto a la ética y la IA y solo depende de nosotros, por ello la importancia de actuar como agentes no como meros observadores. La regulación europea en marcha puede ayudar, pero ética y regulación no son lo mismo, y sería deseable incluir la ética en la IA desde el diseño, sin la necesidad de una regulación que obligue a ello.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. C. y Srikumar, M. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center for Internet & Society*, 2020. Disponible en: http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420

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

Morley, J., Floridi, L., Kinsey, L. *et al.* From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science Engineering Ethics* 26, 2141–2168, 2020. Disponible en: https://doi.org/10.1007/s11948-019-00165-5

Olmeda, M. V. e Ibáñez, J. C. (2022): *Manual de ética aplicada en Inteligencia Artificial*. Barcelona, Anaya Multimedia.