Un viaje hacia la comprensión del cerebro digital

El aumento de la automatización y la complejidad en la inteligencia artificial plantea preocupaciones sobre su confiabilidad, exacerbadas por la opacidad de los algoritmos de aprendizaje profundo. La transparencia es crucial para generar confianza.

La confianza en el contexto de entornos digitales o entre humanos y agentes artificiales ha ganado considerable atención en las últimas décadas (Taddeo & Floridi, 2011). A medida que la inteligencia artificial (IA) se va automatizando, tareas que incluyen la toma de decisiones rutinarias también aumentan las preocupaciones sobre la confiabilidad. Se suma a ello el uso de una clase de inteligencia artificial que utiliza el aprendizaje profundo (DL), un sistema algorítmico de redes neuronales profundas que en general permanecen opacas y ocultas a la comprensión humana. Este problema se conoce como el problema de la caja negra en la IA.

Los observadores pueden presenciar las entradas (A) y salidas (C) de estos procesos complejos y no lineales, pero no el funcionamiento interno (B). La forma en la que la IA llega a su conclusión es opaca y oculta a la vista. Desconocer cómo la IA llega a sus conclusiones, abre la cuestión de hasta qué punto podemos confiar en estos sistemas. A medida que delegamos cada vez más la toma decisiones y confiamos cada vez más en la IA para salvaguardar bienes humanos importantes, como la seguridad, la atención médica y la protección, la confianza se vuelve un factor determinante. La capacidad que existe de “abrir la caja negra” haciendo que el proceso de decisión complejo y no lineal sea comprensible para los observadores humanos, es limitada, al menos en su estado actual es menos opaco para la mayoría de los observadores. Aunque por el momento, nuestra relación con la tecnología a menudo es de dependencia más que de confianza. A medida que los resultados en el uso del aprendizaje profundo nos resulten satisfactorios, nuestra confianza en estos sistemas aumentará. Como ocurre en el diagnóstico y detección de algunas enfermedades. Por lo tanto, la confiabilidad se basa en los juicios sobre las condiciones bajo las cuales uno deposita su confianza en otro o el hecho de que uno responde de cierta manera a que se le haya confiado algo.

¿Podemos confiar en la ‘Caja Negra’ de la IA?

Cuando hablamos acerca de la confianza aparece la transparencia, esta palabra no solo puede asignarse a este tipo de sistemas. Ciertamente, se pueden hacer comparaciones entre las redes neuronales de DL y los cerebros humanos y, en muchos sentidos, la cognición humana también puede verse como una caja negra (Burrell, 2016; Castelvecchi, 2016). Sin embargo, inspeccionar el funcionamiento interno de estos dispositivos y sistemas podría no ser posible por razones prácticas y no alcanza el tipo de transparencia requerido para la confianza (Dahl, 2018). Así como el conocimiento pleno del funcionamiento interno de las funciones cerebrales de aquellos a quienes hemos confiado algo de valor en sí mismo no aumentaría nuestra confianza ni justificaría juicios sobre su confiabilidad, también el conocimiento completo de los procesos de toma de decisiones no basta para alcanzar el nivel adecuado de transparencia necesaria para la confianza.

La incapacidad de explicar por qué se

cometieron errores o como se llegó a determinadas conclusiones presenta numerosos desafíos y socava la confianza en estos sistemas

Los sistemas DL e IA son cada vez más utilizados debido a su velocidad y sofisticación, por su gran potencial para realizar predicciones complicadas sobre una serie de dominios. Casos como Deep Patient, un sistema de DL utilizado en medicina para predecir enfermedades futuras, ilustran esta complejidad. Aunque DL puede superar a los humanos en ciertos aspectos de la toma de decisiones, su funcionamiento interno sigue siendo en gran medida incomprensible. Sin embargo, también son mucho más difíciles de entender sin una transparencia sobre su funcionamiento. Actualmente, debido a su creciente confiabilidad y precisión, estos sistemas se están empleando para tomar decisiones que involucran bienes e interés humanos importantes, como la seguridad nacional, la atención médica, el transporte, las finanzas y los sistemas de información. Todo son ventajas, hasta que incluimos en esta ecuación la falta de transparencia, que presenta dilemas éticos, especialmente cuando la IA hace diagnósticos engañosos (Bleicher, 2017). La incapacidad de explicar por qué se cometieron errores o como se llegó a determinadas conclusiones presenta numerosos desafíos y socava la confianza en estos sistemas.

La comprensión de cómo funcionan estos sistemas es fundamental para generar confianza en su uso. No se trata solo de revelar su funcionamiento interno, sino de comprender las razones y motivaciones detrás de sus decisiones. Este nivel de interpretabilidad es esencial para evaluar si estas decisiones son coherentes con nuestros intereses y valores.

¿Podemos descifrar la 'Caja Negra' de la IA?

La explicabilidad de la IA o XAI, se aborda desde modelos desarrollados para abordar el problema de la caja negra en la IA haciendo que procesos como el DL sean más transparentes, interpretables y explicables. XAI proporciona modelos simplificados para mejorar la comprensión de los procesos de toma de decisiones de la IA, sus fortalezas y debilidades, y cómo podría comportarse en el futuro. Proporcionando explicaciones tanto a nivel global como local. Esto se puede lograr con modelos interpretativos con aproximaciones lineales o utilizando métodos interpretativos *post hoc*, como explicaciones en lenguaje natural (Páez, 2019). Tras estos modelos, existen dos enfoques principales en XAI: explicar "qué" hace la IA y explicar "por qué" lo hace. Ambos son relevantes para diferentes partes interesadas, ya que algunos buscan entender el proceso en sí mismo, mientras que otros quieren comprender las razones detrás de una decisión específica.

La confianza en la IA no solo se basa en su competencia para producir resultados, sino también en la comprensión de cómo y por qué toma ciertas decisiones. Algunos modelos de XAI hacen transparente el funcionamiento interno de la caja negra, mientras que otros explican las decisiones sin revelar necesariamente el proceso completo. Las explicaciones de "por qué" son particularmente importantes para generar confianza en la IA, ya que muestran cómo las decisiones están alineadas con nuestros intereses y valores.

La confianza en la IA también puede desarrollarse a través de la experiencia y el testimonio de expertos (Dahl, 2018). La verificación social, donde un grupo de expertos evalúa y respalda la tecnología, puede ayudar a aumentar la confianza pública en la IA. Sin embargo, es importante reconocer que la confianza en la IA no solo depende de su funcionamiento, sino también de nuestra confianza en aquellos que la desarrollan y utilizan. XAI ofrece un camino hacia una comprensión más profunda y una mayor confianza en la IA, pero sigue siendo

un campo en desarrollo que enfrenta desafíos en cuanto a la interpretación y la aplicación práctica.

¿Quién tiene la llave de la 'Caja Negra' de la IA?

La confianza en la tecnología, especialmente en la inteligencia artificial (IA), es un tema complejo que involucra consideraciones éticas y sociales. Algunos argumentan (Pitt, 2010) que la confianza se reserva exclusivamente para las personas y no se aplica a la tecnología, que simplemente son artefactos. Sin embargo, esta visión simplista no tiene en cuenta el creciente poder e influencia de la IA en nuestras vidas, lo que hace necesario considerar las interacciones dentro de un contexto sociotécnico más amplio (Floridi y Sanders, 2004).

Al entender la tecnología como parte de un sistema sociotécnico, que combina lo técnico y lo social, podemos abordar mejor el problema de la confianza en la IA. Esto implica considerar a todos los actores involucrados en el desarrollo, implementación y uso de la tecnología, así como sus roles, intereses y experiencias. Por ejemplo, en un sistema de diagnóstico médico que utiliza IA, los médicos, pacientes, técnicos y diseñadores son todos parte de un sistema sociotécnico interconectado (Zednik, 2019).

La confianza en la IA también puede desarrollarse a través de la experiencia y el testimonio de expertos

La confianza en la IA no se reduce a una relación entre dos agentes o entre un agente y un artefacto, sino que se basa en una red de confianza que se desarrolla dentro del sistema sociotécnico (Durante, 2010). Esto implica confiar en expertos y organizaciones que utilizan la IA, así como en el propósito y objetivo compartido del sistema en su conjunto.

Abordar el problema de la confianza en la IA también implica responder a preocupaciones escépticas sobre la transparencia y la explicabilidad de la tecnología. Si bien la explicabilidad de la IA puede ser limitada, al integrar la IA en un sistema sociotécnico más amplio, podemos justificar su uso incluso en decisiones de alto riesgo

Conclusión

Para que la IA actúe en nuestro nombre requiere que tengamos para creer que la IA es confiable, por lo que se necesita transparencia sobre cómo opera la IA y alcanza sus resultados y predicciones para que podamos juzgar que la IA es confiable. Por otro lado, las dificultades que vienen asociadas al uso del aprendizaje profundo amenaza con la capacidad de emitir tales juicios a pesar de los esfuerzos en sistemas como XAI. Sin embargo, aunque la IA sigue siendo opaca, se ha observado como podemos tener buenas razones para confiar en su uso al aportar resultados muy satisfactorios. Del mismo modo, tampoco se debe olvidar otras vías de investigación, como la auditoria de la IA basada en la ética, pues son fundamentales para este proyecto y deben explorarse a fondo (Mökander & Floridi, 2021).

Bleicher, A. (2017): «Demystifying the black box that is AI» en *Scientific American*. Disponible en:

https://www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai/

Burrell, J. (2016): «How the machine 'thinks': Understanding opacity in machine learning algorithms» en *Big Data & Society*, 3(1). Disponible en: https://doi.org/10.1177/2053951715622512

Castelvecchi, D. (2016): «Can we open the black box of AI?» en *Nature*, 538, 21–23. Disponible en: https://doi.org/10.1038/538020a

Dahl, E. S. (2018): «Appraising black-boxed technology: The positive prospects» en *Philosophy and Technology*, 31(4), 571–591. Disponible en: https://doi.org/10.1007/s13347-017-0275-1

Durante, M. (2010): «What is the model of trust for multi-agent systems? Whether or not e-trust applies to autonomous agents» en *Knowledge Technology & Policy*, 23, 347–366.

Floridi, L. y Sanders, J. W. (2004): «On the morality of artificial agents» en *Minds and Machines*, 14, 349–379.

Mökander, J. y Floridi, L. (2021): «Auditoría basada en la ética para desarrollar una IA confiable» en *Mentes y Máquinas. Springer Science and Business Media BV*. Disponible en: https://doi.org/10.1007/s11023-021-09557-8

Páez, A. (2019): «The pragmatic turn in explainable artificial intelligence (XAI)» en *Minds and Machines*, 29(3), 441–459. https://doi.org/10.1007/s11023-019-09502-w

Taddeo, M. y Floridi, L. (2011): «The case for e-trust» en *Ethics and Information Technology*, 13, 1–3. Disponible en: https://doi.org/10.1007/s10676-010-9263-1

Zednik, C. (2019): «Solving the black box problem: A normative framework for explainable artificial intelligence» en *Philosophy & Technology*, 34, 265–288. Disponible en: https://doi.org/10.1007/s13347-019-00382-7