MÁQUINAS CON INTELIGENCIA

Inteligencia artificial explicable: la clave para confiar en las máquinas

¿Cómo podemos aprender a confiar en las máquinas? La clave es avanzar en la inteligencia artificial explicable (XAI). ABB repasa los últimos estudios de empresas leviatanas y analiza las investigaciones en marcha en este campo esencial para el negocio.

Jinendra Gugaliya
ABB corporate Research, Process Automation
Bangalore, India

jinendra.gugaliya@in.abb.com

K. Eric Harper
Antiguo empleado de ABB

Los modelos de inteligencia artificial (IA) se extienden por nuestra vida diaria, a menudo sin que nos demos cuenta. Por ejemplo, Netflix utiliza un motor recomendador para sugerir películas a sus usuarios. Este proceso aparentemente sencillo utiliza el aprendizaje automático para ayudar a los algoritmos a automatizar millones de decisiones en función de las actividades del usuario [1]. Facebook utiliza su proceso patentado de «sistema de detección por contacto de visión por computadora» para reconocer objetos a medida que hace una selección de imágenes en cuentas de usuario en busca de logotipos y marcas para que los anunciantes puedan dirigirse a los usuarios con anuncios mediante publicaciones de historias patrocinadas [2]. Más allá del entretenimiento y la publicidad, la ciencia del big data y el aprendizaje automático están colándose en aplicaciones de misión crítica, como la detección y el diagnóstico de enfermedades, la toma de decisiones sobre préstamos y los coches autónomos. Estas aplicaciones pueden tener un impacto crucial en nuestras vidas. La película que va ver podría ser una frivolidad, pero que le den o no una hipoteca es algo sumamente pertinente. He aquí el quid de la cuestión: estos modelos de IA son difíciles de explicar incluso entre los científicos de datos, lo que hace que su aceptación sea problemática. Al no disponer de la capacidad para explicar los modelos, o «explicabilidad», es probable que estos modelos no se acepten y no se utilicen. Por otra parte, el Reglamento general de protección de datos (RGPD) exige una mayor transparencia en el tratamiento de datos y claridad en los procesos de IA, por lo que resulta esencial explicar los modelos [3].

Estos modelos de IA son difíciles de explicar incluso entre los científicos de datos, lo que hace que su aceptación sea problemática.

Tanto si se trata de seres humanos como de máquinas, una explicación de la decisión fomentaría que se comprendiera la motivación que hay detrás de los modelos y crearía un sentido de urgencia, lo que aumentaría la probabilidad de que se aceptaran y aplicaran las

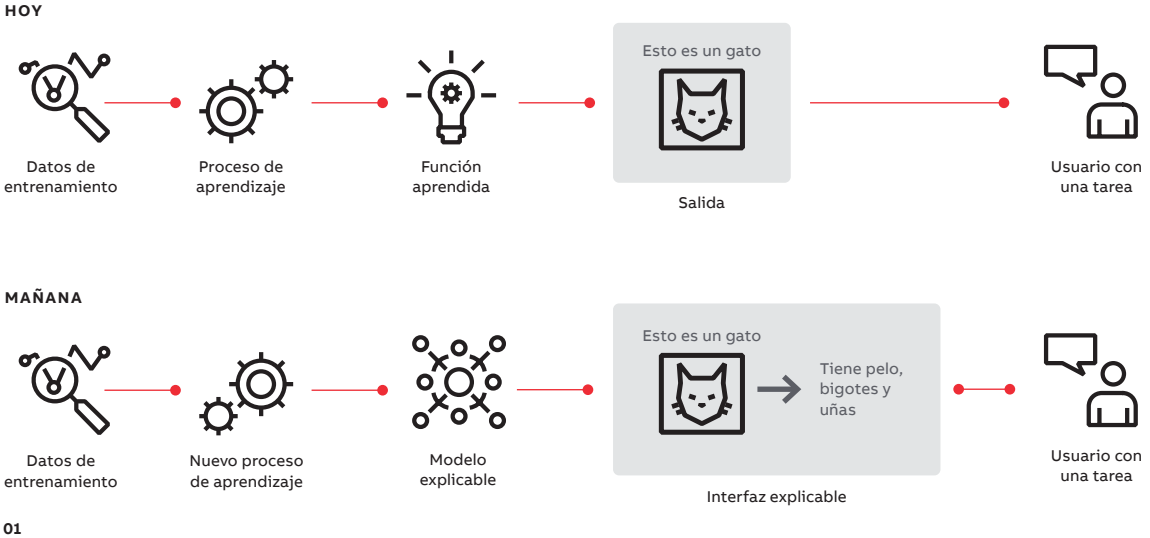
recomendaciones resultantes. Sin esa comprensión, es imposible confiar en las máquinas que construimos. Por lo tanto, los investigadores de ABB han profundizado en la historia de la XAI para permitir la expansión de la IA en el futuro.

Explicación de la primera IA industrial

Esta área de investigación actualmente activa inició con el requisito de que los diagnósticos tenían que ser defendibles. En la década de los 80, Westinghouse inventó y comercializó junto a la Universidad Carnegie Mellon el primer sistema de apoyo a la toma de decisiones basado en el conocimiento y centrado en el funcionamiento [4]. Eric Harper, que contribuyó a los logros citados en este artículo, participó directamente en GenAID, TurbinAID y ChemAID (Artificial Intelligence Diagnostics) y contribuyó a los avances en materia de propiedad intelectual y tecnología de software basándose en una patente clave que se centraba en un sistema operativo [5] con pruebas y mejores prácticas para abordar las condiciones anormales. Las subsiguientes patentes de Westinghouse describen un método y un sistema que permiten ejercer exhaustivamente una base de conocimientos con vistas a confirmar que los resultados previstos proceden de entradas conocidas de datos anormales [6]. Harper creó las herramientas y las técnicas para explorar el conocimiento y rastrear el apoyo para justificar acciones específicas en base a diagnósticos. Esta propiedad intelectual crítica es ahora de dominio público y sus conclusiones siguen siendo relevantes [7].

Innovaciones para explicar los resultados de la IA

Recientemente, una importante consultoría de software ha adoptado ideas para generar un modelo basado en la representación del conocimiento. Su solicitud de patente estadounidense sugiere que una combinación del algoritmo de agrupamiento Kmeans, el análisis de componentes principales, el encadenamiento hacia adelante o hacia atrás y la lógica difusa resolvería el problema general de explicar los resultados de la IA en el mundo real [8]. Utilizando métodos



01

adicionales, una agencia de calificación del crédito que gestiona riesgos de crédito ejecuta modelos de manera exhaustiva con distintas entradas para demostrar con éxito a los clientes cómo entradas distintas dan lugar a resultados distintos, ayudando así a los clientes a comprender las decisiones de crédito adoptadas [9]. En una solicitud de patente de EE.UU., Intel describe una técnica para identificar discrepancias entre los resultados observados en la fase de entrenamiento del aprendizaje automático y los resultados obtenidos durante la operación [10]. Otra solicitud de patente de Intel describe la influencia de las redes neuronales en la IA explicable: este proceso hace un seguimiento exhaustivo de las dependencias y la solidez del apoyo, entre las capas inferiores y superiores de la red neuronal, de vuelta hasta sus características de entrada de una forma sorprendentemente similar a la de un sistema experto de encadenamiento hacia adelante [11]. Google combina estas ideas con sus herramientas y marco de trabajo [12]; IBM creó una plataforma parecida [13].

Tendencias de la IA explicable

Esta creciente atención hacia el estudio de la XAI puede clasificarse en distintas categorías en función del grado de transparencia: desde las de características de caja negra completa (baja transparencia) hasta las de caja blanca (alta transparencia) [14]:

- Sistemas opacos
- Sistemas comprensibles
- Sistemas interpretables

Aunque el diseño y el desarrollo de modelos de XAI aportan muchos beneficios, también pueden suponer importantes costes. El equilibrio se encuentra entre la explicación y la precisión. Este equilibrio y este grado de transparencia estarán impulsados por las necesidades comerciales y por cómo se adopta la aplicación en el mundo real.

La lógica detrás de las decisiones de los modelos de aprendizaje automático es compleja y no evidente. Confiar en las decisiones críticas resultantes sin gobernanza resulta problemático. Esta inquietud ha sido abrumadora desde el principio: las redes neuronales han estado disponibles desde la década de los 80 para el diagnóstico de máquinas, pero no se implantaron. Por ello, es cada vez más importante que la XAI llegue a ámbitos específicos. De este modo, los modelos de XAI pueden someterse a una verificación formal, una capacidad que resulta especialmente importante para las aplicaciones médicas en las que las recomendaciones tienen consecuencias de vida o muerte.

> El desarrollo de la XAI se verá impulsado por las necesidades comerciales y por la forma en que los modelos se adopten en el mundo real.

—
01 El esquema ilustra cómo abrir «cajas negras» con IA explicable hoy y en el futuro. Este modelo ha sido desarrollado por FICO [9].

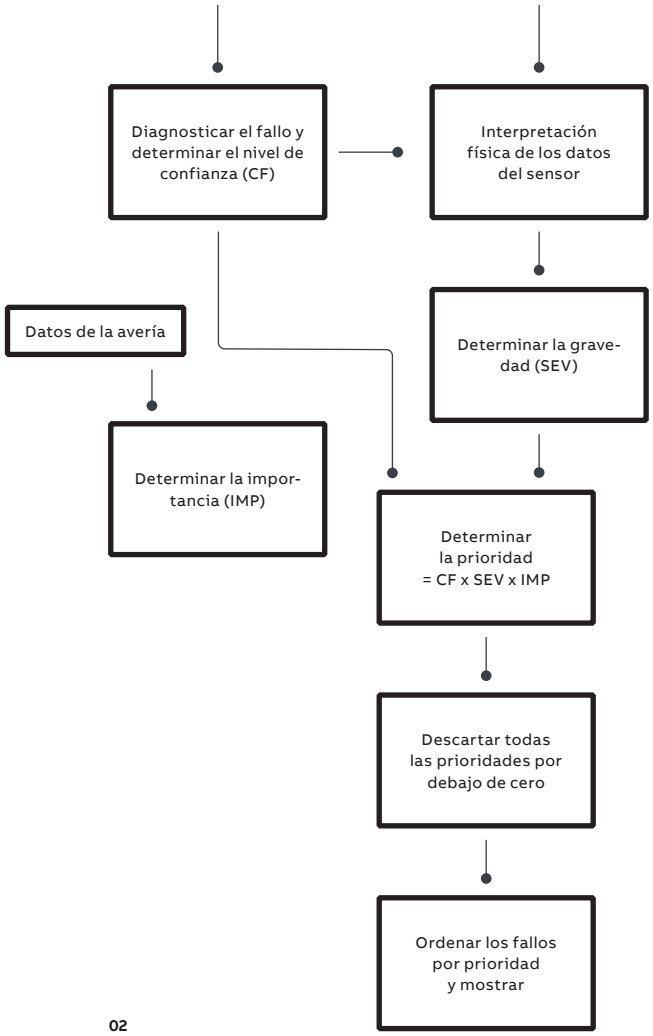
—
02 Se muestra un apoyo práctico para la toma de decisiones extraído de Bellows, et al., [23]. La toma de decisiones se basa en el cálculo de la confianza (CF), la gravedad (SEV) y la importancia (IMP).

—
Nota al pie

1) Go es un juego abstracto de estrategia para dos jugadores.

Es necesario abordar otro reto antes de que los modelos de aprendizaje automático puedan aceptarse sin más: si los datos de entrenamiento no abarcan todo el espacio de la solución se muestra un sesgo [15]. Estos defectos de sesgo podrían detectarse si se realizan pruebas en una amplia gama de condiciones para comparar los puntos fuertes y los puntos débiles de la solución.

Aun así, los modelos de XAI pueden aportar nueva información deslumbrante. Los sistemas de aprendizaje automático actuales se entrenan con millones de ejemplos para poder reconocer patrones de datos que no son obvios para el ser humano. El antiguo sueño de Westinghouse de que los ingenieros pudieran algún día analizar los datos recabados de modelos como GenAID, aplicar métodos de ciencia de datos y descubrir nuevas innovaciones está a punto de cumplirse. Gracias al uso de sistemas de XAI ha surgido nueva información: se puede extraer conocimiento destilado del aprendizaje automático para adquirir nuevas percepciones; por ejemplo, se han desarrollado nuevas estrategias para jugar a Go[1] con aprendizaje automático que ahora utilizan los jugadores humanos.



02

Sin embargo, el apoyo opaco a la toma de decisiones no resulta atractivo para las empresas. Está claro que un banco debe comunicar por qué ha rechazado una solicitud de crédito. Y los modelos de IA deben cumplir la ley aportando pruebas que justifiquen las decisiones adoptadas.

La generalización de estas inquietudes ha llevado a la Unión Europea a adaptar nuevas normativas que implementan un «derecho a la explicación», en virtud del cual el usuario tiene derecho a pedir la explicación detrás de una decisión algorítmica pertinente. Se espera que la XAI proporcione la confianza, la imparcialidad y la seguridad necesarias para que los modelos de aprendizaje automático en modos de entrenamiento y funcionamiento se ganen la confianza de las empresas [16].

Explicar los modelos de IA: situación actual
Actualmente se utilizan dos enfoques para hacer más explicables los modelos de IA. En primer lugar, las estructuras de los modelos se seleccionan con un objetivo intrínseco de interpretación en mente. Alternativamente, los modelos complejos de IA son objeto de ingeniería inversa para hacerlos comprensibles. Sin embargo, diseñar modelos que ofrezcan una explicación que sea fácil de entender puede comprometer la precisión y viceversa; por ejemplo, las redes neuronales profundas (DNN) complejas son precisas, pero no pueden interpretarse. Los algoritmos como la regresión lineal o los modelos basados en árboles de decisiones son mucho más fáciles de explicar, pero menos precisos. Alcanzar el equilibrio entre la precisión y la interpretabilidad de los modelos de IA constituye actualmente el objeto de una intensa investigación [17].

—
Los sistemas de aprendizaje automático se entrenan con millones de ejemplos para que puedan reconocer patrones de datos que no son obvios para el ser humano.

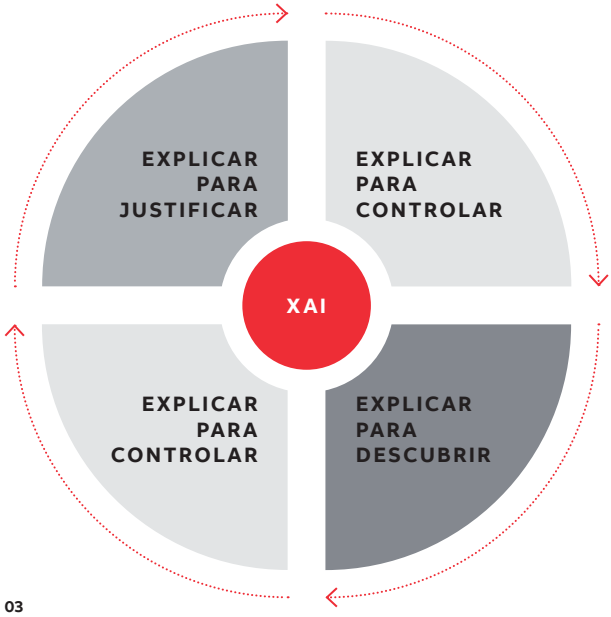
Otro ámbito de investigación de la XAI analiza la diferencia entre la interpretabilidad local y la interpretabilidad global [18] →03. La perspectiva local, a partir de los principios del análisis de sensibilidad (SA), identifica cómo la salida del modelo cambia en caso de perturbaciones en los parámetros de entrada o ajustes. Si bien no ofrece una explicación del valor de la función

en sí misma, el análisis de sensibilidad permite determinar los factores y las configuraciones que explican los resultados del modelo. La perspectiva global utiliza dos técnicas. La propagación de relevancia por capas (LRP) redistribuye la función de predicción hacia atrás, empezando por la capa de salida de la red neuronal y propagándose hacia atrás hasta la capa de entrada. La LRP explica las decisiones del clasificador por descomposición y puede representarse mediante mapas de calor [19]. El sistema de detección de

ABB ofrece valor a los clientes incorporando los avances de la IA para mejorar el mantenimiento predictivo, la optimización y el rendimiento.

intrusiones (IDS) basado en datos es un enfoque contradictorio que se utiliza para identificar las modificaciones mínimas (de las funciones de entrada) necesarias para clasificar correctamente un conjunto dado de muestras mal clasificadas. La magnitud de la modificación visualiza las funciones más relevantes que explican el motivo de la mala clasificación. Los investigadores han combinado la LRP y el IDS para jugar a juegos de Atari impulsados por el aprendizaje profundo por refuerzo (DRL) [20].

La investigación en el ámbito de la XAI se ha ampliado para permitir comparaciones de conjuntos de datos. Se han combinado resúmenes lingüísticos y reglas difusas para diseñar un sistema capaz de comparar numéricamente varios conjuntos de datos [21] y explicar las diferencias



03

04

utilizando el lenguaje natural. La denominada colaboración «hombremáquina» es un factor clave en la investigación de la XAI, donde la usabilidad constituye una consideración importante para el diseño y el uso de modelos. En este caso, los modelos de aprendizaje automático deben permitir al usuario ajustar los modelos de forma interactiva en base al aprendizaje iterativo [22].

Un enfoque práctico hacia el apoyo a la toma de decisiones

La ingeniería nos enseña que cualquier problema complejo puede resolverse dividiendo el trabajo en varios componentes, reconstruyéndolos y comprobándolos de forma independiente e integrándolos de nuevo para obtener una solución completa. En la década de los 80, un sistema Westinghouse para el apoyo a la toma de decisiones y la gestión de activos explicables inventó técnicas [23] para priorizar la reparación de los equipos de una planta mediante la validación de las mediciones y la posterior fusión de las condiciones actuales con el diagnóstico y la recomendación del sistema. Un importante fabricante de equipos eléctricos sigue utilizando este sistema para la supervisión de sus centrales eléctricas y turbogeneradores, y ejecuta el servicio en su centro de diagnósticos eléctricos [24]. Básicamente, los posibles problemas pueden clasificarse conforme a tres dimensiones:

- CF: nivel de confianza en un diagnóstico
- SEV: recíproco del tiempo hasta el fallo
- IMP: costes asociados a los fallos y reparación del máximo daño

Estas dimensiones se calculan para cada uno de los sensores, componentes y sistemas del equipo que contribuyen a un fallo o avería: esta combinación se utiliza para determinar la prioridad de acción para cada fallo posible. La lógica detrás

05

06

—
03 La IA explicable permite a los usuarios comprender y aceptar decisiones. La capacidad de descubrir, controlar y justificar están interconectadas entre sí.

—
04 En ABB, la IA tiene un potencial radical y se está aplicando en muchos ámbitos industriales. Por ejemplo, durante un análisis industrial en una planta, el ingeniero etiqueta patrones de interés que luego pueden utilizarse para entrenar a un clasificador basado en las RNN.

—
05 Al diseñar IA explicable en los diagnósticos de los sensores de ABB, los operadores y los responsables de planta pueden comprender y aceptar lo que hay que hacer.

—
06 Los clientes se benefician de la IA explicable sin necesidad de comprender todos los cálculos que se han utilizado para la toma de decisiones.

de esta clasificación era explicable de cara a los clientes: se podía hacer un seguimiento de las tres dimensiones y de los detalles del diagnóstico a través de los componentes y las lecturas de los sensores. De este modo, los clientes podían aceptar las clasificaciones sin tener que necesariamente ahondar en los detalles.

Relevancia para ABB

Dado que la potencia de computación escaseaba en los 80, se procedió a ajustar las soluciones automatizadas de servicios para obtener el máximo rendimiento con recursos limitados. En la actualidad, se puede calcular cada una de estas características originalmente interpretadas hasta la granulación más fina de los componentes y, a continuación, montarse en función de sus relaciones y dependencias. ABB sobresale en el uso del conocimiento y procesos obtenidos de sus primeros trabajos con soluciones de servicios automatizadas. En la actualidad, ABB confía en numerosas soluciones de supervisión del estado que indican los niveles de confianza cada vez que surgen problemas con los equipos →04–06. Las soluciones ABB Ability™ Advanced Digital Services contienen funciones que abordan el problema crítico del tiempo hasta el fallo.

Gracias a los talentosos ingenieros y científicos de la información de los que dispone, ABB crea soluciones de ciencia de datos para calcular las dimensiones importantes: CF, SEV e IMP, cuando procede. Conocer al detalle el coste global de una reparación tras un fallo permite a ABB diagnosticar problemas a partir de datos, lo que ha dado lugar a nuevos servicios como el ABB Ability™.

Como líder pionero en el dominio de la automatización industrial, ABB ofrece valor a los clientes incorporando los avances de la IA para mejorar el mantenimiento predictivo, la optimización y el rendimiento. Con todo, antes de implementar acciones potencialmente costosas, los operadores y los responsables de planta deben conocer la explicación y el fundamento detrás de las decisiones que generan los modelos de IA y que recomiendan las aplicaciones de operaciones y servicio →05. Al incluir AI explicable en el diseño de sus aplicaciones, ABB destaca en el mercado: esto fomenta la confianza, algo que es ahora más crucial que nunca. Cuando los modelos son explicables, los expertos y los usuarios finales pueden estar seguros de que los resultados no tienen parcialidades, son seguros, legales, éticos y adecuados. •

—

Referencias

[1] Code Academy, “Netflix Recommendation Engine”, Available: https://www.codecademy.com/. [Accessed: May 5, 2020]

[2] A. Razaaq, “Facebook’s New Image Recognition Algorithm Can Scan your Picture for Advertising Opportunities” in *B2C Business to Community*, May 19, 2019, Available: https:// www. business2community. com/. [Accessed May 5, 2020]

[3] A. Woodie, “Opening up Black Boxes with Explainable AI” in *Datanami*, May 30,2018, Available: https:// www.datanami.com/. [Accessed May 5, 2020].

[4] E.D. Thompson, et al., “Process Diagnosis System (PDS) - A 30 Year History” in *Proc. 27th Conf. on Innovative Applications of AI*, Jan. 2015, Available: https://dl.acm.org/doi/10.5555/2888116.2888260. [Accessed: May 5, 2020]

[5] Thompson, et al., “Methods and apparatus for system fault diagnosis and control”, US Patent no. 4,649,515 March 10, 1987.

[6] K.E. Harper, et al., “Expert system tester”, US Patent no. 5,164,912, November 17, 1992.

[7] Y. Lizar, et al., “Implementation of Computer Damage Diagnosis by Expert System Based Using Forward Chaining and Certainty Factor Methods” in *International Journal of Scientific & Technology Research*, vol. 8 issue 6, June 2019, pp. 141 - 144. Available: http://www. ijstr.org/. [Accessed May; 5, 2020].

[8] L. Chung-Sheng, et al., “Explainable Artificial Intelligence”, US Patent Application no. 20190244122, August 8, 2019.

[9] FICO, “How to Make Artificial Intelligence Explainable: A new Analytic Workbench”, in *FICO/blog*, Sept. 18, 2018, Available: https://www.fico.com/blogs/. [Accessed May 5, 2020].

[10] J. Glen, et al., “Misuse Index for Explainable Artificial Intelligence in Computing Environments”, US Patent Application no. 20190197357, June 27, 2019.

[11] K. Doshi, “Mapping and Quantification of Influence of Neural Network Features for Explainable Artificial Intelligence” US Patent Application no. 20190164057, May 30, 2019.

[12] GoogleCloud, “Understand AI Output and Build Trust”, Available: Google Cloud: https://cloud.google.com/. [Accessed May 5, 2020].

[13] A. Mojsilovic, “Introducing AI Explainability 360”, August 8, 2019, Available: IBM Research Blogs: https:// www.ibm.com/blogs/. [Accessed May 5, 2020].

[14] D. Doran, et al., “What does explainable AI really mean? A new Conceptualization of Perspectives”, in *ariXiv*, October 2, 2017, Available: https://arxiv.org/abs/1710.00794. [Accessed May 5, 2020].

[15] N. Mehrabi, et al., “A Survey on Bias and Fairness in Machine Learning” preprint based on work supported by DARPA, in *ariXiv*, September 17, 2019, Available: https://arxiv.org/abs/1908.09635. [Accessed May 5, 2020].

[16] M. Miron, “Interpretability in AI and its relation to fairness, transparency, reliability and trust” in *European Commission HUMANIT*, September 4, 2018, Available: https://ec.europa.eu/jrc/communities/en/community/humaint. [Accessed May 5, 2020].

[17] W.J. Murdoch, et al., “Definitions, methods, and applications in interpretable machine learning”, in *Proc. of the National Academy of Sciences of the United States of America*, vol. 116, issue 44, 22071 October 29, 2019, Available: https://www.pnas.org/.[Accessed May 5, 2020].

[18] A. Adadi and M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)” in *IEEE Access*, September 17, 2018, Available: https://ieeexplore.ieee.org/document/8466590. [Accessed May 5, 2020].

[19] W. Samek, et al., “Evaluating the Visualization of What a Deep Neural Network Has Learned” in *IEEE Transactions. On Neural Networks and Learning Systems*, Nov. 2017. [Abstract]. Available: https://ieeexplore.ieee.org/abstract/document/7552539. [Accessed May 5, 2020].

[20] H. Jo, and K. Kim, “Visualization of Deep Reinforcement Learning using Grad - CAM: How AI Plays Atari Games?” in *IEEE Conf on Games*, August 23, 2019, Available: https://ieeexplore.ieee.org/document/8847950

[21] A. Jain, et al., “Explainable AI for Dataset Comparison” in *IEEE Int. Conf on Fuzzy Systems*, June 26, 2019, Available: https://ieeexplore.ieee.org/document/8858911. [Accessed May 5, 2020].

[22] A. Kirsch, “Explain to whom? Putting the User in the Center of Explainable AI” in *Proc. 1st Int. Workshop on Comprehensibility and Explanation in AI and ML*, October 21, 2018, Available: https:// hal.archivesouvertes. fr/hal-01845135/ [Accessed May 5, 2020].

[23] Bellows, et al., “Automated system to prioritize repair of plant equipment”, US Patent no. 5,132,920, July 21, 1992.

[24] I. Becerra-Fernandez and R. Sabherwa, “Knowledge Application Systems: Systems that Utilize Knowledge” in *Knowledge Management: Systems and Processes*, 2nd ed. New York: Routlege, 2015, pp. 3 - 105