REVISTA CUBANA
DE TRANSFORMACIÓN DIGITAL

RNPS 2487 • ISSN 2708-3411
Vol. 2 • Nro. 1 • enero-marzo 2021 • págs. 01-13

EDITORIAL

Breve reseña sobre el estado actual de la Inteligencia Artificial

A Brief Summary about the Current State of Artificial Intelligence

Rafael Bello
rbellop@uclv.edu.cu • https://orcid.org/0000-0001-5567-2638
UNIVERSIDAD CENTRAL "MARTA ABREU" DE LAS VILLAS, CUBA

Alejandro Rosete
rosete@ceis.cujae.edu • https://orcid.org/0000-0002-4579-3556
UNIVERSIDAD TECNOLÓGICA DE LA HABANA "JOSÉ ANTONIO ECHEVERRÍA", CUJAE, CUBA

RESUMEN

El propósito de este artículo es ofrecer al lector una panorámica de la Inteligencia Artificial hoy, sus principales métodos y logros así como su aplicación en la solución de diferentes problemas socio-económicos y científicos. Se presentan algunas de sus tendencias de desarrollo, y los retos que estos pudieran significar para el hombre, al integrarlas plenamente en prácticamente todas las facetas de la vida de la humanidad, y previendo posibles efectos negativos de su desempeño futuro. Esta conceptualización sirve de preámbulo para presentación de los trabajos incluidos en este número de la revista.

Palabras clave: Inteligencia Artificial; transformación digital.

ABSTRACT

The main goal of this paper is to present a global view of the current state of Artificial Intelligence, its most important methods and achievements, as well as its applications in the solution of several socio-economics and scientific problems. Some tendencies of its development are discussed and the consequent challenges for the human society, because of its insertion in all dimensions of the human life, including the prevention of possible negative effects in the future. This conceptual presentation serves as a presentation of the papers included in this number of the journal.

Keywords: *Artificial Intelligence; digital transformation.*

INTRODUCCIÓN

En la actualidad no es necesario buscar revistas científicas muy especializadas para leer sobre Inteligencia artificial (IA), pues es común encontrar en la prensa diaria y en diferentes sitios digitales informaciones positivas que mencionan la importancia del *Big Data,* el papel de los asistentes virtuales como *Siri* y *Alexa,* la predicción de la pandemia de COVID-19 antes que ocurriera, la victoria de inteligencias artificiales sobre humanas en juegos como el ajedrez o el Go. Igualmente, aparecen noticias negativas que mencionan los peligros potenciales de una IA fuera de control que afecten las opciones de empleo, que vislumbran nuevas formas de guerra, que nos expongan a decisiones sin un marco ético y que incluso puedan poner en peligro la misma existencia humana.

La situación era muy diferente hace unos años, cuando la IA a nivel popular solo era asociada a la ciencia ficción. Hoy, ya es muy usual encontrar noticias donde personas muy influyentes en los destinos del mundo hablen de la IA como un factor estratégico en el futuro mundial.

El término "inteligencia artificial" fue creado por el científico estadounidense John McCarthy en la década del 50 (McCarthy, *et al.,* 1955) y ha ido evolucionando en un debate entre la búsqueda de la referencia en lo racional o en lo parecido a lo humano (como indica el Test de Turing) (Rusell y Norvig, 2010). Luego de estudiar varias definiciones, la Comisión Europea hoy lo define como: "sistemas de *software* (y posiblemente también de *hardware)* diseñados por humanos que, ante un objetivo complejo, actúan en la dimensión física o digital: percibiendo su entorno, a través de la adquisición e interpretación de datos estructurados o no estructurados, razonando sobre el conocimiento, procesando la información derivada de estos datos y decidiendo las mejores acciones para lograr el objetivo dado. Los sistemas de IA pueden usar reglas simbólicas o aprender un modelo numérico, y también pueden adaptar su comportamiento al analizar cómo el medio ambiente se ve afectado por sus acciones previas".

Desde una perspectiva más pragmática, la IA se puede ver como una disciplina de la computación que nos ofrece métodos para resolver dos tipos de problemas: (*i)* problemas con algoritmos para resolverlos, pero que por su complejidad computacional, pueden existir instancias del problema con una alta dimensión para los que no es posible usar ese algoritmo, y (*ii)* los que carecen de algoritmo para resolverlos.

Un ejemplo de la primera clase es el problema del viajero vendedor donde un agente de comercio debe recorrer *N* ciudades con el menor costo. Este problema tiene complejidad computacional *N!* por lo que cuando la cantidad de ciudades crece se requiere de mucha potencia de cómputo lo cual es imposible en ciertos casos. Los problemas del segundo tipo se pueden ejemplificar usando dos ejemplos de juegos: el ajedrez y el Go. Ambos juegos son considerados retos para la inteligencia humana (de hecho, al ajedrez se le llama el "juego ciencia", el Go es más complejo aún). Sin embargo, hace ya varios años la computadora fue capaz de ganarles a los grandes maestros del ajedrez y recientemente le ganó a los expertos en Go (Silver, *et al.,* 2016) y (Silver, *et al.,* 2018).

Lo común de ambas situaciones es la inexistencia de un algoritmo específico que pueda usarse para resolverlos. Esto rompe con el esquema tradicional de la computación de partir de unos datos de entrada y producir una salida. El enfoque general de solución con técnicas de IA se basa en una búsqueda inteligente (heurística) en el espacio de soluciones posibles al problema que emplea el conocimiento del dominio de aplicación (Rusell y Norvig, 2010). En la actualidad, una forma de resolver muchos problemas de alta complejidad es usando las metaheurísticas, que ofrecen mecanismos de búsqueda eficientes y eficaces; ejemplos de ellas son los Algoritmos genéticos, la Optimización basada en partículas (*Particle Swarm Optimization)* y la Optimización basada en colonias de hormigas (*Ant Colony Optimization)* (Luke, 2012).

El conocimiento del dominio de aplicación sobre los que trabajan los métodos de la IA puede ser obtenido de diferentes fuentes. La primera de ellas fue los expertos del dominio de aplicación, y posteriormente los datos existentes sobre los problemas resueltos. Esta última vía permite extraer y formalizar ese conocimiento; lo cual ha dado lugar al desarrollo de diferentes técnicas de IA para el descubrimiento de ese conocimiento. La Transformación Digital de la sociedad permite generar de forma continua enormes cantidades de datos, de todo tipo (datos numéricos, textos, voz, imágenes); prácticamente estamos rodeados de equipos que generan datos: los teléfonos, las cámaras, los asociados a Internet de las cosas (IoT), los propios sistemas computacionales que respaldan diferentes tipos de actividades como las transacciones comerciales, bancarias, gestión hospitalaria, gestión de la educación, etc. Esta combinación de muchos, variados y dinámicos datos, y sistemas con una alta independencia, basados en una alta capacidad para la toma de decisiones, marcan las tendencias de desarrollo de la IA en la actualidad.

En la sección siguiente de presenta una breve presentación del estado actual de la IA y a continuación se analizan algunas tendencias que parece que guiarán su desarrollo en los próximos años. Finalmente, se presentan los trabajos incluidos en este número.

SOBRE LA IA EN LA ACTUALIDAD

Los productos y servicios computacionales que utilizan las técnicas de IA ya están en todas partes. Estos pueden aparecer como productos de software independientes como chat bots, buscadores, y sistemas inteligentes de diferentes tipos; o incrustados en equipos (robots avanzados, vehículos autónomos, entornos de IoT, etc.).

Como una disciplina científica, la IA incluye diferentes enfoques y técnicas, entre las que se pueden destacar (Samoili, *et al.,* 2020): el razonamiento automático (búsqueda, optimización, planificación, secuenciación de tareas, representación del conocimiento, manejo de la incertidumbre), el aprendizaje automático (incluido el aprendizaje profundo o "*Deep learning",* la ciencia de datos y el aprendizaje por reforzamiento o refuerzo), y los sistemas ciber-físicos (incluidas la robótica y la internet de las cosas, que comprenden el control, percepción, el procesamiento de los datos recogidos por sensores, y el funcionamiento de los actuadores).

En el primer grupo están los métodos que usan el conocimiento del dominio de aplicación, representado usando algún formalismo, para resolver diferentes tipos de problemas. En el segundo están los métodos para descubrir el conocimiento a partir de la información del dominio de aplicación. Y el tercer grupo se refiere a los artefactos creados usando los sistemas inteligentes desarrollados usando esos métodos de solución de problemas y ese conocimiento.

Según la forma en que se representa el conocimiento (Rusell y Norvig, 2010) las técnicas de IA se clasifican en simbólicas o sub simbólicas. Las primeras se refieren al caso en que el conocimiento del dominio de aplicación se describe utilizando un lenguaje propio de ese dominio que es comprensible para el hombre; ejemplos de estas son las reglas, los árboles de decisión y los casos. Las redes neuronales, y especialmente los modelos computaciones basados en aprendizaje profundo actuales, son ejemplos de la IA sub simbólica, donde la representación del conocimiento se aparta del lenguaje usado en el dominio de aplicación y que tiene sus orígenes en las neurociencias (Hassabis, *et al.,* 2017).

El desarrollo de las capacidades de aprendizaje de los sistemas inteligentes ha creado nuevas posibilidades. Ya el sistema no solo depende del conocimiento aportado por los expertos humanos y codificado usando algún formalismo, ahora puede enriquecer ese conocimiento, y con ello ser más eficiente o eficaz en su desempeño. La combinación del aprendizaje profundo y el aprendizaje por reforzamiento ha generado una elevada capacidad de solución de problemas (Stone, *et al.,* 2016), como lo mostró el éxito de *AlphaGo,* programa desarrollado por *Google Deepmind,* que fue capaz de batir al campeón del juego de Go. Este programa fue entrenado inicialmente usando bases de datos de juegos de expertos humanos, pero fue refinando su desempeño a partir de jugar contra sí mismo y aplicando aprendizaje por reforzamiento.

Utilizando estas técnicas para la construcción de sistemas inteligentes se han desarrollado aplicaciones de la IA en prácticamente todos los aspectos de la vida socio-económica, causado por los avances obtenidos en las áreas de percepción, aprendizaje y capacidad de razonamiento. Algunos ejemplos de áreas de aplicación son, según Stone, *et al.,* (2016) y Bughin, *et al.,* (2017): transporte (autos inteligentes, vehículos autónomos, planificación del transporte); energía (ajustes de parámetros en sistemas, predicción de demanda, identificación de fallas, atención a clientes, contadores inteligentes); salud (asistentes clínicos, analítica de salud, robótica médica, cuidado al adulto mayor, optimización de la gestión hospitalaria); educación (robots-profesores, tutoriales inteligentes, analítica del aprendizaje); seguridad pública (ciberseguridad, detección de fraudes, prevención del crimen); y entretenimiento (producción de contenidos, juegos, interacción con sistemas basada en empatía y emoción). Aparejado a estas aplicaciones, se desatan numerosos retos para el empleo y las necesidades de capacitación y evaluación.

Un riesgo real con el empleo de las técnicas de IA es que con ellas se amplifiquen o refuercen los comportamientos sesgados de los humanos (Tommasi, *et al.,* 2017), (Li y Vasconcelos, 2019), (Mehrabi, *et al.,* 2019). El sesgo (bias) es una inclinación o prejuicio hacia o contra una persona, objeto, etc., puede ser bueno o malo, intencional o no. Esta problemática suce-

de cuando la salida inferida para un grupo particular es sistemáticamente diferente de otros grupos, y por eso un grupo es consistentemente tratado de forma distinta al resto. La calidad de los datos, y potencialmente el sesgo, es particularmente importante en la era del *Big Data,* cuando se generan en internet enormes volúmenes de datos, muy rápidamente y sin control de calidad; aun los mejores métodos de aprendizaje no pueden funcionar eficazmente usando datos de baja calidad.

Otra característica de algunas de estas aplicaciones (vehículos autónomos, cirugía robótica, control de redes eléctricas, armas inteligentes) es que algún fallo en su funcionamiento se puede convertir en errores catastróficos (Dietterich, 2017), por los que alguien debe responder.

Por otra parte, a pesar del universo diverso donde se pueden encontrar las técnicas de IA, todavía no existe una amplia confianza en las mismas; hay un problema de credibilidad como lo muestran algunos resultados presentados por Cannon (2019), Wang y Siau (2018) y Hengstler, *et al.,* (2016), donde todavía una parte importante de los encuestados manifiestan no sentirse cómodos interactuando con sistemas inteligentes en lugar de humanos, que el sesgo puede afectar las decisiones de estos sistemas, y que los problemas éticos y morales de los sistemas puede afectar su empleo pleno en la Sociedad.

Una cuestión interesante es analizar las razones que motivan esta falta de credibilidad en la IA a pesar de su alta presencia en la vida actual. Una respuesta pudiera estar en que los sistemas inteligentes son dinámicos, mediante su aprendizaje ellos se adaptan, modifican su comportamiento, y al hacer esto violan principios de usabilidad pues el usuario puede sentir que no tiene el control (Holliday, *et al.,* 2016). Otra razón por la que las personas tienden a asignar más credibilidad a los expertos humanos que a los sistemas computacionales es que se considera que los hombres son capaces de integrar más información y desde más fuentes (Alexander, *et al.,* 2018).

El objetivo de las aplicaciones de IA tiene que ser crear valores para la Sociedad. Para la adopción de esas aplicaciones es principal que los usuarios confíen en ellas, y para ello es necesario desarrollar estrategias que incrementen la capacidad para interpretar los sistemas de IA; promover la participación de los usuarios en su desarrollo y uso puede ayudar a creer y prevenir errores trágicos. Esto posibilitará el uso de las tecnologías de la IA de forma **ética, transparente** y **explicable.** Como se analizará en la siguiente sección, la búsqueda de este nuevo estado de desarrollo para la IA marcan algunas de las tendencias en su desarrollo; dando lugar a términos como IA fiable (*Trust AI)*, IA explicable (*Explainable AI, XAI)* e IA fuerte vs IA débil (*Strong AI* vs *Weak AI).*

ALGUNAS TENDENCIAS DE DESARROLLO DE LA IA

El uso creciente de las técnicas de IA plantea la cuestión de la confiabilidad de los usuarios en su empleo. Un aspecto relacionado con la fiabilidad de estos sistemas está relacionado con la comprensión que se tiene de su desempeño por lo que la interpretabilidad de los mismos es un tópico relevante. Por otra parte, la amplitud de áreas donde las técnicas de IA se emplean,

plantea la pregunta de hasta donde los sistemas inteligentes podrán desarrollarse, hasta donde su capacidad de solución de problemas será comparable con la del hombre.

IA FIABLE

La credibilidad del *software* está fundamentalmente basada en la naturaleza determinística de la mayoría de las aplicaciones en las cuales su comportamiento es únicamente determinado por el flujo del código, lo cual lo hace intrínsecamente predecible. La naturaleza no-determinística de los sistemas de IA dotados de aprendizaje rompen con este patrón tradicional del software e introduce nuevas dimensiones para dotar de credibilidad a los agentes de la IA. El enfoque tradicional es determinístico y predecible, el otro es no-determinístico y difícil de comprender (Rodriguez, 2018). Cuando los sistemas inteligentes están dotados de mecanismos para aprender a partir de su desempeño e interacción con el entorno (por ejemplo usando aprendizaje por reforzamiento), el mismo podría llegar a alcanzar un comportamiento no deseado ni previsto en su diseño inicial, que puede llegar a ser peligroso (Hibbard, 2012).

Si aceptamos que la IA será una parte relevante de nuestro futuro, es importante establecer los fundamentos de la credibilidad de los sistemas de IA. Hoy, regularmente descansamos en modelos de IA sin tener una clara comprensión de sus capacidades, conocimientos o procesos de entrenamiento. El concepto de Sistema de IA creíble (*trust AI*) se mantiene altamente subjetivo y no ha sido incorporado como parte de las plataformas populares de aprendizaje automático. Lograr la confianza de los usuarios en los productos y servicios informáticos que utilicen las técnicas de IA es hoy el propósito de diferentes instituciones, desde las academias, las empresas y los gobiernos. La credibilidad en los resultados de los sistemas inteligentes no solo se basa en la certidumbre de estos, sino también en la posibilidad de poder explicarlos a partir de la transparencia de su desempeño.

Según Rodriguez (2018), los pilares para lograr una mayor credibilidad de los sistemas basados en IA son la imparcialidad (estar libres de predisposición o sesgos); robustez (ser seguros y confiables, no comprometer los datos con que son entrenados); explicabilidad (proveer decisiones o sugerencias comprensibles por sus usuarios y desarrolladores); y genealogía (incluir detalles de su desarrollo, despliegue, y mantenimiento para poder ser auditados). La Unión Europea desarrolla iniciativas en esta dirección que enfocan el problema desde tres perspectivas: componente legal (los sistemas deben cumplir las leyes y regulaciones); componente ética (los sistemas deben satisfacer valores y principios éticos) y los sistemas deben ser robustos (EISMD, 2018).

Empresas, organizaciones y otras instituciones trabajan en el establecimiento de principios para el diseño, desarrollo y uso de los sistemas basados en IA, entre ellos están Asilomar (2017), IBM (IBM, 2018), *Google* (Pichai, 2018), IEEE (IEEE, 2018), y la Unión Europea (EISMD, 2018).

IA EXPLICABLE

Una de las direcciones seguidas en la búsqueda de incrementar la credibilidad en los sistemas inteligentes es que los usuarios puedan comprender su funcionamiento y las razones de las

decisiones que estos toman. Arribar a explicaciones llenas de significado sobre el conocimiento de los modelos de IA reduce la incertidumbre y ayuda a cuantificar su precisión. Mientras que la explicabilidad puede ser vista como un factor obvio para mejorar la credibilidad en los sistemas de IA, su implementación está lejos de ser trivial. La IA explicable (*Explainable AI, XAI)* (Barredo-Arrieta, *et al.,* 2020) está relacionada con la búsqueda del incremento de la credibilidad de los sistemas inteligentes y limitar el sesgo en el conocimiento utilizado. Esta perspectiva es especialmente relevante cuando el modelo computacional se ve como una caja negra cuyas interioridades no son conocidas o no son interpretables por los humanos.

El aprendizaje automático es una disciplina llena de fricciones y balances pero ninguna más importante que el balance entre precisión e interpretabilidad. En general, los modelos de IA altamente explicables tienden a ser muy simples y, por eso, no muy precisos. Desde esta perspectiva, establecer un correcto balance entre explicabilidad y precisión es esencial para mejorar la credibilidad de los modelos de IA.

El problema de la explicabilidad en IA no es nuevo, pero el crecimiento de los sistemas inteligentes autónomos (capaces de proponer una solución y ejecutarla) ha creado la necesidad de comprender cómo estos sistemas inteligentes obtienen una solución, hacen una predicción o una recomendación, o razonan para soportar una decisión, para incrementar la credibilidad de los usuarios en estos sistemas. Además, es importante que los sistemas de IA puedan responder a preguntas sobre por qué ocurre o no algo, cuándo se puede esperar éxito o fallo, cómo corregir un error (Lee, *et al.,* 2019). La importancia de estos aspectos puede no ser tan relevante para un sistema que recomiende películas a ver, pero sí puede serlo en gran medida en sistemas médicos o financieros. Por ejemplo, al sugerir una decisión sobre una solicitud de crédito o préstamo la explicación es muy importante, pues seguramente el usuario exigirá una explicación cuando su solicitud sea rechazada.

Hay algunos modelos de IA creados por algoritmos de aprendizaje automático (AA) como la regresión lineal, la regresión logística, el clasificador *Naive Bayes,* y los árboles de decisión que un humano puede fácilmente interpretarlos. Pero en modelos creados por algoritmos como las Máquinas de Soporte Vectorial (*Support Vector Machines), Random Forests, Gradient Boosted Trees* y métodos de aprendizaje profundo es un reto su interpretación aun para los investigadores en IA (LeCun, *et al.,* 2015), (Karmakar y Pal, 2018), (Montavon, *et al.,* 2018), y (Jay-Kuo, *et al.,* 2019). Incluso los modelos del primer grupo pueden ser difíciles de interpretar cuando sus dimensiones crecen; por ejemplo, cuando se tienen muchos rasgos, cuando se generan muchas reglas, o cuando la profundidad de un árbol es muy grande. En este contexto es que se desarrolla la denominada *XAI* (Barredo-Arrieta, *et al.,* 2020) y (Lamy, *et al.,* 2019).

En resumen, el propósito de la *XAI* es: (*i)* proveer una explicación de decisiones individuales; (*ii)* posibilitar la comprensión de fortalezas y debilidades; (*iii)* proveer una comprensión de cómo el sistema se comportará en el futuro; (*iv)* comunicar como corregir los errores del sistema.

Se han propuesto diferentes métodos para construir las explicaciones. En Abdollahi y Nasraoui (2018) se proponen tres: explicaciones basadas en ejemplos similares, explicaciones

donde se presentan los ítems que más incidieron en la solución encontrada, y la identificación de rasgos comunes entre palabras claves y el contenido. La mayoría de los métodos en *XAI* funcionan como una ingeniería inversa, se aprende el modelo computacional y luego se trabaja en mejorar su interpretabilidad (*post hoc interpretability)* (Guidotti, *et al.,* 2018); estos métodos se clasifican en modelos de explicación, cuando se trata de interpretar la lógica del modelo completo, explicación de la salida, cuando el objetivo es comprender las razones por la que se obtiene una solución particular (correlación entre los datos de entrada y la solución calculada), y de inspección, cuando el propósito es comprender como internamente el modelo (caja negra) se comporta cuando se varían los datos de entrada.

Una alternativa actual para resolver problemas en dominios de alta complejidad que incluyen la toma decisiones de alto nivel es la formación de equipos hombre-IA (con humanos y agentes basados en IA) y hay estudios que demuestran que el desempeño de tales equipos es mayor que el logrado por cada componente por separado (Tomsett, *et al.,* 2020), debido a que cada miembro del equipo es capaz de compensar las debilidades del otro.

Desarrollar formas más fuertes de interpretabilidad puede ofrecer varias ventajas, entre ellas incrementar la fiabilidad en los sistemas inteligentes, facilitar el análisis de posibles errores y ayudar a refinar el modelo. Para (Montavon, *et al.,* 2018), a pesar de los éxitos prácticos alcanzados hasta ahora, la interpretación de los sistemas inteligentes, especialmente los basados en redes profundas, se mantiene como un campo de investigación joven y emergente (Barredo-Arrieta, *et al.,* 2020).

IA débil versus IA fuerte

El gran desarrollo de la IA ha llevado a que resurja una interrogante relacionada con esta disciplina desde sus comienzos que se enfoca en decidir si hay una diferencia fundamental entre el hombre y la máquina, o si es solamente una diferencia en la potencia de cómputo.

Esto conduce a la pregunta sobre si las computadoras pudieran pensar. La forma de responder esta pregunta divide a la IA en dos grandes campos: la IA fuerte (*Strong AI*) y la IA débil (*Weak AI).* Los que se ubican en el primero plantean que no hay una diferencia fundamental entre el hombre y la máquina; mientras que los situados en el segundo defienden que solamente las personas pueden pensar, las máquinas no pueden. Estos últimos se enfocan en agregar rasgos "parecido a pensar" a las computadoras para hacerlas herramientas más útiles para resolver problemas en tareas específicas y no necesariamente en desarrollar una actividad mental o un modelo de comportamiento humano; mientras los primeros plantean que las computadoras puedan llegar a pensar a un nivel al menos igual que los humanos, que pueden ser conscientes y experimentar emociones, que pueden realizar actividad mental (como un ser humano). La IA fuerte también se ha denominado IA general o completa, y la IA débil como IA estrecha (*Narrow AI)* (Hengstler, *et al.,* 2016), (Lieto, *et al.,* 2018), y (Montes y Goertzel; 2019).

Los promotores de la IA fuerte sostienen que la inteligencia artificial puede igualar o superar la inteligencia humana, creen que las computadoras y el cerebro tienen una potencia

equivalente; que la inteligencia de una máquina puede exitosamente ejecutar cualquier tarea intelectual que un humano pueda hacer. Sin embargo, la visión de la IA fuerte es difícil de aceptar para muchos.

Han aparecido sistemas inteligentes como el sistema Watson de IBM con nuevas capacidades para resolver problemas mediante la combinación de diferentes técnicas para el procesamiento del lenguaje natural, el aprendizaje automático, el procesamiento de imágenes, etc. (Hossein; 2018). Un elemento importante para la IA general son las llamadas arquitecturas cognitivas, estas ofrecen una estructura tecnológica para la creación de sistemas inteligentes como *Icarus, Clarion* y *Soar* (Lieto, *et al.,* 2018).

Un rasgo importante que debe incluir la IA fuerte es la capacidad de transferir conocimiento de un dominio a otro. En esa dirección se ha desarrollado el aprendizaje continuo que tiene el propósito de aprender a partir de la experiencia del aprendizaje de varias tareas de una forma continua de manera que se explota el conocimiento y habilidades adquiridos previamente; un agente con aprendizaje continuo ideal debe ser capaz, entre otras funcionalidades, de resolver múltiples tareas, y exhibir sinergias cuando las tareas están relacionadas (Mankowitz, *et al.,* 2018). Según (Chen y Liu, 2018) se trata de aprender continuamente acumulando conocimiento pasado que se usará en procesos de descubrimiento conocimiento y solución de problemas en el futuro; este se contrasta con el paradigma aislado del aprendizaje automático tradicional, donde se aplica un método de aprendizaje a un conjunto de datos, se construye un modelo y este emplea en la solución de problemas.

En resumen, un sistema con IA general tiene el propósito de poder ejecutar la mayoría de las actividades que realizan los humanos; mientras que los sistemas con IA débil están diseñados para ejecutar una o pocas tareas específicas, en la actualidad la mayoría de los sistemas con IA son de IA débil. Aún están abiertos muchos retos éticos, científicos y tecnológicos para construir las capacidades que se requieren para alcanzar la IA general o fuerte (EISMD; 2018).

EN ESTE NÚMERO

De la explicación anterior se puede entender la amplitud y variedad de los retos que tiene hoy la IA, lo cual hace imposible incluir una representación de todas estas dimensiones en este número. Sin embargo, es notable el hecho de haber reunido en un mismo número contribuciones de entidades cubanas y extranjeras (España, Japón, Argentina) que van desde variantes de algoritmos concretos de aprendizaje, hasta el análisis de aspectos estratégicos de la gestión de la IA, pasando por aportes y experiencias en el uso de herramientas de optimización y simulación.

De esta manera el número contiene los siguientes trabajos.

Primero, aparecen un grupo de trabajos del tema de aprendizaje automático (incluyendo aprendizaje por reforzamiento y redes convolucionales, por ejemplo), a tono con la gran popularidad de estos métodos hoy. Es notable el enfoque aplicado de algunos de estos trabajos en áreas tan activas como la bioinformática (con el trabajo “Agrupamiento funcional de en-

zimas GH-70 utilizando aprendizaje semi-supervisado y *Apache Spark")* y el procesamiento de imágenes y videos (con los trabajos "Un estudio de la generalización en la clasificación de peatones" y "Exploración de Redes Neuronales Holográficas con cuantificación difusa para el monitoreo de conductores en Conducción Autónoma Condicional").

Como un reflejo de los sistemas que involucran varias ramas de la IA, en este número se incluyen dos trabajos donde se combinan los principios de Lógica Borrosa dentro de mecanismos para la representación y obtención del conocimiento (como son los trabajos "Algoritmo de aprendizaje reforzado para software de modelado basado en mapas cognitivos difusos" y "Tendencias en la sumarización lingüística de datos").

También se incluyen otros dos trabajos en que se combinan otras áreas de la IA como la simulación basada en agentes inteligentes, así como los métodos de búsqueda y optimización, donde se aprecia un enfoque hacia el desarrollo de herramientas de IA que puedan ser empleadas por un amplio espectro de usuarios finales (como son los casos de los trabajos "Herramienta de simulación para evaluar configuraciones semafóricas" y "Nodos *Knime* para ajustar modelos usando la biblioteca de clases Biciam").

Finalmente, aparecen dos trabajos que tratan temas muy importantes para lograr un mayor empleo de los algoritmos de IA y particularmente de Aprendizaje Automático. Por una parte, en el trabajo "Hacia la democratización del aprendizaje de máquinas usando *AutoGOAL"* se ilustra cómo puede aumentar el uso de este tipo de algoritmos a partir de simplificar su uso, lo cual los autores consideran que democratiza su uso. Por otra parte, en el trabajo "Marco de trabajo para la publicación de datos abiertos en Cuba" se analizan herramientas que pueden favorecer una mayor proliferación de datos para su uso ordenado y para aprender de ellos.

CONCLUSIONES

En este trabajo se ha presentado una apretada síntesis de algunos de los conceptos más relevantes hoy en el mundo de la Inteligencia Artificial. Sin entrar en los detalles de los métodos, el énfasis se ha puesto en los desafíos que entrañan estos desarrollos dentro del contexto de la transformación digital de las sociedades. También se enuncian algunas de las tendencias que están llamadas a conducir las investigaciones en esta temática en el futuro inmediato.

Finalmente, se hace una presentación de los trabajos del número, permitiendo mostrar la amplia gama de temáticas cubiertas.

REFERENCIAS

Abdollahi B. y Nasraoui, O. (2018). Transparency in Fair Machine Learning: the Case of Explainable Recommender Systems In *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent,* Jianlong Zhou and Fang Chen (Eds.). Springer International Publishing, Cham, 21–35

Alexander, V. et al. (2018). Why trust an algorithm? Performance, cognition, and neurophysiology. *Computers in Human Behavior 89:* 278-288

Asilomar (2017). AI Principles. Future of Life Institute, Recuperado de: https://futureoflife.org/ai-principles/

Axel-Montes, G. y Goertzel, B. (2019). Distributed, decentralized, and democratized artificial intelligence. *Technological forecasting and Social Sciences 141*:354-358

Baptiste-Lamy, J. et al. (2019). Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *AI in Medicine 94:*42-53.

Barredo-Arrieta, A., Díaz-Rodríguez, N., Del-Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatilaf, R. y Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58:* 82–115. doi:10.1016/j.inffus.2019.12.012

Bughin, J. et al. (2017). AI: the next digital frontier. Discussion paper. *McKinsey Global Institute.*

Cannon, J. (2019). Report shows consumers don't trust artificial intelligence, Recuperado de: https://www.fintechnews.org/report-shows-consumers-dont-trust-artificial-intelligence/

Chen, Z. y Liu, B. (2018). *Lifelong Machine Learning.* Second Edition, Morgan & Claypool. Print 1939-4608 Electronic 1939-4616. doi:10.2200/S00832ED1V01Y201802AIM037, p. 209.

Dietterich. T. G. (2017). Steps Toward Robust Artificial Intelligence. AI MAGAZINE, FALL.

EISMD (2018). AI4People's Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations" (november). Recuperado de: http://www.eismd.eu/wp-content/uploads/2018/11/ Ethical-Framework-for-a-good-AI-Society.pdf.

Samoili, S., López-Cobo, M., Gómez, E., De Prato, G., Martínez-Plumed, F.,y Delipetrev, B. (2020). AI Watch. Defining Artificial Intelligence. *Towards an operational definition and taxonomy of artificial intelligence,* EUR 30117 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-17045-7, doi:10.2760/382730, JRC118163.

Guidotti, R. et al. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv. 51(5)*:93 (august), doi: 10.1145/3236009.

Hassabis, D. et al. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron 95,* July(19): 245-258.

Hengstler, M. et al. (2016). Applied artificial intelligence and trust. The case of autonomous vehicles and medical assistance devices. *Technological Forecasting & Social Change 105:* 105–120

Hibbard, B. (2012). Avoiding Unintended AI Behaviors. *Lecture Notes in Artificial Intelligence, 7716:* 107-116

Holliday, D., et al. (2016). User trust in intelligent systems: A journey over time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces.* 164-168. New York, USA: ACM

Hossein, M. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons 61,* 577—586. Recuperado de: https://www.ibm.com/watson

IBM (2018). Trusted AI. Recuperado de: https://www.research.ibm.com/artificial-intelligence/trusted-ai/

IEEE (2018). IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. IEEE. Recuperado de: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_general_principles.pdf

Jay-Kuo, C.C. et al. (2019). Interpretable convolutional neural networks via feedforward design. *J. Vis. Commun. Image R. 60:* 346–359

Karmakar, B., y Pal, N.R. (2018). How to make a neural network say "Don't know". *Information Sciences* 430–431, 444–466

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature 521,* 436–444

Lee, Y. et al. (2019). Egoistic and altruistic motivation: How to induce users' willingness to help for imperfect AI. *Computers in Human Behavior 101,* 180-196

Li, Y. y Vasconcelos, N. (2019). REPAIR: Removing Representation Bias by Dataset Resampling. Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition,* 9572-9581. Recuperado de: arXiv: 1904.07911.

Lieto, A. et al. (2018). The role of cognitive architectures in general artificial intelligence. *Cognitive Systems Research 48,* 1–3

Luke, S. (2012). *Essentials of Metaheuristics.* ISBN: 978-0-557-14859-2. Recuperado de: http://cs.gmu.edu/ sean/book/metaheuristics/

Mankowitz, D.J. et al. (2018). Unicorn: Continual learning with a universal, off-policy agent. Recuperado de: arXiv: 1802.08294v2.

McCarthy, J., Minsky, M.L., Rochester, N. y Shannon, C.E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. August 31, 1955. Recuperado de: http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

Mehrabi, N. et al. (2019). A Survey on Bias and Fairness in Machine Learning. Recuperado de: arXiv: 1908.09635v2

Montavon, G. et al. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing 73,* 1–15

Pichai, S. (2018). AI at Google: our principles. June 7. Recuperado de: https://www.blog.google/technology/ai/ai-principles/

Rodriguez, J. (2018). Towards AI Transparency: Four Pillars Required to Build Trust in Artificial Intelligence Systems. Recuperado de: https://www.linkedin.com/pulse/towards-ai-transparency-four-pillars-required-build-trust-rodriguez

Rusell, S. y Norvig, P. (2010) *Artificial Intelligence: a modern approach.* Pearson Education

Silver, D., et al. (2016). Mastering the game of Go with deep neural networks and tree search. Nature 529, 484–489

Silver, D. et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science 362,* 1140–1144

Stone, P. et al. (2016). Artificial intelligence and Life in 2030, in *One Hundred Year Study on Artificial Intelligence (AI100).* Report of the 2015 Study Panel, Stanford University. Recuperado de: https://ai100.stanford.edu

Tommasi, T. et al. (2017). A deeper look at dataset bias. In *Domain Adaptation in Computer Vision Applications,* 37–55. Springer

Tomsett, R. et al. (2020). Rapid Trust Calibration through Interpretable and Uncertainty-Aware *AI. PATTER 1,* July: 1-9. doi:10.1016/j.patter.2020.100049

Wang, W., y Siau, K. (2018). Trusting Artificial Intelligence in Healthcare. *Twenty-fourth Americas Conference on Information Systems,* New Orleans