Banco Interamericano de Desarrollo

fAIr LAC

Adopción ética y responsable de la Inteligencia Artificial en América Latina y el Caribe

Marcelo Cabrol
Natalia González A.
Cristina Pombo
Roberto Sánchez A.

Sector Social

NOTA TÉCNICA Nº
IDB-TN-1839

Enero 2020

fAIr LAC

Adopción ética y responsable de la Inteligencia Artificial en América Latina y el Caribe

Marcelo Cabrol
Natalia González A.
Cristina Pombo
Roberto Sánchez A.

Enero 2020

BID
Banco Interamericano de Desarrollo

Catalogación en la fuente proporcionada por la
Biblioteca Felipe Herrera del
Banco Interamericano de Desarrollo
fAIr LAC: adopción ética y responsable de la inteligencia artificial en América Latina y el Caribe / Marcelo Cabrol, Natalia González Alarcón, Cristina Pombo, Roberto Sánchez Ávalos.
p. cm. — (Nota técnica del BID ; 1839)
Incluye referencias bibliográficas.
1. Artificial intelligence-Social aspects-Latin America. 2. Artificial intelligence-Social aspects-Caribbean Area. 3. Artificial intelligence-Moral and ethical aspects-Latin America. 4. Artificial intelligence-Moral and ethical aspects-Caribbean Area. 5. Social service-Technological innovations-Latin America. 6. Social service-Technological innovations-Caribbean Area. I. Cabrol, Marcelo. II. González Alarcón, Natalia. III. Pombo, Cristina. IV. Sánchez Ávalos, Roberto. V. Banco Interamericano de Desarrollo. Sector Social. VI. Serie.
IDB-TN-1839

Códigos JEL: O300, O330, O350
Palabras Claves: Inteligencia artificial, Aprendizaje automático, Tecnología, Ética, Transformación digital, Mercados laborales, Educación, Salud, Privacidad, Gobernanza de datos

http://www.iadb.org

BID
Banco Interamericano de Desarrollo

fAIr LAC

ADOPCIÓN ÉTICA Y RESPONSABLE DE LA INTELIGENCIA ARTIFICIAL EN AMÉRICA LATINA Y EL CARIBE

Marcelo Cabrol, Natalia González A.,
Cristina Pombo, Roberto Sánchez A.

Resumen

La inteligencia artificial (IA) ofrece oportunidades únicas para **promover la igualdad de oportunidades y mejorar la calidad de vida de todas las personas** de la región. Más allá de las posibilidades tecnológicas, su uso responsable y centrado en los individuos es esencial, además de que supone grandes desafíos.

El Banco Interamericano de Desarrollo (BID) aboga por construir un entendimiento común de lo que es la IA, sus oportunidades y sus aplicaciones, pero también de sus riesgos y posibles medidas para mitigarlos.

Para ello el BID, en colaboración con socios y aliados estratégicos, lidera una iniciativa denominada fAIr LAC mediante la cual busca promover la adopción responsable de la IA para mejorar la prestación de servicios sociales (principalmente en los sectores de educación, salud, protección social, mercados laborales y temas asociados con género y diversidad) y crear oportunidades de desarrollo en aras de reducir las brechas y atenuar la creciente desigualdad social. Trabajando en conjunto con los sectores público y privado, la sociedad civil y la academia, la iniciativa fAIr LAC liderará la ejecución de experimentos y proyectos piloto de sistemas de IA. Asimismo, creará modelos de evaluación ética y otras herramientas para que los gobiernos, los emprendedores y la sociedad civil puedan profundizar su conocimiento en la materia, contar con guías y marcos para la adopción responsable de la IA e incidir tanto en la política pública como en el ecosistema emprendedor.

En este documento se describen algunos de los retos y oportunidades que la IA presenta a la sociedad, junto con las líneas de acción que la iniciativa fAIr LAC propone para América Latina y el Caribe.

Agradecimientos

Por su tiempo y valiosos aportes, expresamos un agradecimiento especial al grupo consultivo conformado por Carolina Aguerre, Constanza Gómez-Mont, Daniel Korn, Elkin Echeverri, Fabrizio Scrollini, Gemma Galdon Clavell, Hugo Morales, Javier Barreiro, Joana Varón, Mario Arauz, Norberto Andrade, Renata Ávila, Ricardo Baeza-Yates, Ana Lucía Lenis, Ulises Cortés y Virginia Pardo.

Agradecemos igualmente el apoyo prestado y los comentarios recibidos de Arturo Muente, Jaime Granados, César Buenadicha, Elena Arias, Manuel Urquidi, David Rosas, Marcelo Perez, Luis Tejerina, Jennifer Nelson, Tamar Colodenco, Rodrigo Galindo, Alejandro Noriega, Carlos Amunategui, Daniel Castaño, Joan Manuel López y Lucía Camacho.

Índice

01. Introducción

Bajo la promesa de cambiar la forma en que vivimos y trabajamos, durante la última década la inteligencia artificial (IA) ha ganado protagonismo en los debates en múltiples esferas. Aunque existe cierto consenso en que aún faltan muchos años para llegar a una inteligencia artificial general, las aplicaciones de IA débil o acotada –donde el aprendizaje automático se especializa en la ejecución de una sola labor– han superado ya el nivel de precisión humana en algunas tareas específicas de reconocimiento visual y análisis de lenguaje natural en sectores como la medicina y el derecho (Ardila, et al., 2019 y Wood, 2018).

Al igual que en otras regiones, en América Latina y el Caribe (ALC) la IA promete mejorar la eficiencia en la prestación de servicios sociales[1] y la transparencia de la toma de decisiones públicas, así como incentivar la economía mediante aumentos en la productividad.

La IA tiene el potencial de ayudar a la sociedad a superar algunos de sus desafíos más importantes.

Desde reducir la pobreza y lograr avances en la educación, hasta mejorar la prestación de servicios de salud y la erradicación de enfermedades, pasando por incrementar la producción de alimentos de modo que satisfaga las necesidades de la población mundial.

Surgen, sin embargo, algunas preguntas: ¿Qué tan lejos está ALC de realizar ese potencial? ¿Cómo garantizar que en la búsqueda de esos beneficios no se incurra en costos sociales mayores que generen una sociedad más desigual? ¿Cómo debe ser la IA para que su uso resulte confiable? ¿Cómo asegurar que el uso de IA sea responsable?

La región debe prepararse para aprovechar los beneficios de una IA confiable donde el **ser humano esté en el centro de las decisiones, identificar los desafíos éticos y de privacidad que esta supone, y contar con mecanismos y estándares para el manejo y la mitigación de riesgos.**

[1] Se hace referencia al conjunto de servicios y actuaciones orientados a mejorar el bienestar social de la ciudadanía mediante el suministro de información, atención y apoyo, en particular para los sectores de educación, salud, protección social, mercados laborales, seguridad social y temas asociados con género y diversidad.

Si bien en este empeño es importante promover la innovación y la eficiencia, se debe favorecer especialmente el acceso. Propender por crear valor económico es loable, pero más lo es favorecer la distribución de la riqueza. El uso de la tecnología ha de ser proactivo en lo que se refiere a promover valores sociales como la integridad, la tolerancia y la diversidad, y evitar generar un mayor consumo que agudice el problema de los desechos, la contaminación y el cambio climático. Además, en el ámbito de la organización social misma, se ha de evitar que el uso masificado de la tecnología sea una solución para los pobres y el trato personalizado sea un privilegio solo para los ricos.

El Banco Interamericano de Desarrollo (BID) se propone lograr que exista un entendimiento común acerca de lo que es la IA en la actualidad y de lo que podría ser en el futuro cercano, así como sobre las oportunidades que ofrece, sus aplicaciones sectoriales, los riesgos que presenta y los posibles mecanismos para mitigarlos.

Para ello el BID, en colaboración con varios aliados estratégicos, lidera una iniciativa llamada fAIr LAC cuyo objetivo es **promover un desarrollo y aplicación responsables de la IA para mejorar la prestación de servicios sociales** –reduciendo con ello las brechas que existen– y eventualmente atenuar la creciente desigualdad.

La iniciativa fAIr LAC se apalanca en los sectores público y privado, así como en la sociedad civil, para incidir tanto en la política pública como en el ecosistema emprendedor. En este documento se introducen la definición estratégica y las líneas de acción que fAIr LAC propone para afrontar los retos y aprovechar las oportunidades que la inteligencia artificial presenta a la sociedad en América Latina y el Caribe.

El objetivo de fAIr LAC es promover un desarrollo y aplicación responsables de la IA para mejorar la prestación de servicios sociales y atenuar la creciente desigualdad.

02. Inteligencia artificial

La idea de las máquinas inteligentes fue planteada por Alan Turing (1950) en su documento teórico "Computing Machinery and Intelligence". Allí explo-raba la posibilidad de que una computadora pudiera simular la inteligencia humana y aprender.

El término "inteligencia artificial" aparecería unos años más tarde, en 1956, acuñado por el científico cognitivo John McCarthy en el marco de la conferencia "Dartmouth Summer Research Project on AI"[2] , reconocida como el hito fundacional de la IA como campo de estudio.

Hasta el día de hoy no existe una definición universalmente aceptada de IA, pues se trata de un área científica muy dinámica que ha ido evolucionando y dando lugar a múltiples tecnologías. Sin embargo, se puede afirmar que la IA es un campo de estudio que se enfoca en el desarrollo de capacidades en sistemas computacionales para realizar tareas tradicionalmente pensadas como exclusivas de la "inteligencia" humana. El problema entonces radica en definir qué es inteligencia.

En 1950, Turing (1950) propuso realizar la siguiente evaluación conversacional para definir inteli-gencia: una máquina se consideraría inteligente si al conversar con un humano lo hiciera de una forma tan natural que aquel no fuera capaz de distinguir que su interlocutor era una máquina. Esta evaluación ha tenido un importante número de objeciones como prueba para determinar inteligencia.

Casi cien años antes, en sus memorias sobre el motor analítico de Babbage, la matemática Ada Lovelace[3] había señalado que no se podía considerar inteligente a una máquina que solo podía hacer lo que se le ordenara (Epstein, 2008). Este argumento implica que la inteligencia requiere a su vez autonomía y capacidad de innovar. En el mismo sentido, John Searle (1980) argumentó que la prueba de Turing era inadecuada y podía superarse mediante el uso de reglas sintácticas sin requerir un entendimiento semántico real, por lo que no sería una muestra de inteligencia humana[4].

A lo largo de los años se han realizado también actividades y juegos a manera de métrica para

2 El objetivo de este evento era llevar a cabo una serie de talleres en los cuales se partía de la conjetura de que "cada aspecto del aprendizaje o cualquier otra característica de la inteligencia podía, en principio, describirse con tanta precisión que se podría lograr que una máquina la simulara".
3 Ada Lovelace (1815-1852) fue una matemática y escritora inglesa reconocida por algunos como la primera programadora del mundo, pues escribió el primer programa informático.
4 Por ejemplo, un equipo liderado por Vladimir Veselov diseñó un chatbot denominado Eugene que se hacía pasar por un chico ucraniano de 13 años. En 2014 este logró hacer creer a la mayor parte de los jueces de un concurso organizado por la Universidad de Reading que se trataba de un ser humano real. Esta victoria es, no obstante, controvertida, toda vez que el chatbot podía siempre excusar su incapacidad para expresarse correctamente en que no dominaba lo suficiente el inglés.

comparar las capacidades de inteligencia de una computadora y de un ser humano, desde una partida de ajedrez hasta un partido de *Go*. Estas pruebas también son debatibles bajo el mismo parámetro, dado que las máquinas son incapaces de explicar su estrategia.

La autora Pamela McCorduck (2004) se refiere al dinamismo anteriormente aludido como la paradoja de la IA. Según ella, existen innovaciones que en un cierto momento se pensaron como señales de inteligencia casi humana pero poco a poco se han convertido en pruebas superficiales hasta perder el privilegio de ser categorizadas como IA. Simultáneamente van surgiendo nuevas tecnologías que comienzan a asumir este papel (McCorduck, 2004 y Cristianini, 2014).

Existen innovaciones que en un cierto momento se pensaron como señales de inteligencia casi humana pero poco a poco se han convertido en pruebas superficiales hasta perder el privilegio de ser categorizadas como IA. Simultáneamente van surgiendo nuevas tecnologías que comienzan a asumir este papel"

PAMELA MCCORDUCK, 2004 Y CRISTIANINI, 2014

Este afán de medir la inteligencia centrándose en una sola labor ha llevado a que se separe el alcance y la concepción de la IA en tres objetivos:

- El de la **inteligencia débil o acotada**, con la cual se busca que el proceso de aprendizaje se especialice en la ejecución de una sola tarea, logrando una precisión igual o superior a la de un ser humano (este tipo de IA es la única que se ha logrado hasta el momento).

- El de la **inteligencia general o fuerte**, con la cual se aspira a que la IA logre generalizar el proceso de aprendizaje aplicándolo a distintas tareas con creatividad y conciencia propia.

- El de la **"súper inteligencia",** un tipo de IA que lograría sobrepasar la capacidad cognitiva del ser humano en todos los aspectos (Bostrom, 2014).

Hasta la fecha, todas las aplicaciones de IA son ejemplos de IA débil o acotada. Y aunque existen opiniones encontradas, la mayoría de expertos consideran que faltan muchos años para lograr una IA general o incluso que nunca se llegará a ese punto.

A finales de la década que acaba de concluir, en *MIT Technology Review* (Hao, 2019) se analizaron 16.625 productos de conocimiento sobre "inteligencia artificial" con el fin de examinar la evolución de la terminología y las técnicas utilizadas en ese campo de la ciencia[5]. Desde el nacimiento del término "inteligencia artificial" en los años cincuenta, cada década ha sido testigo de diferentes técnicas que han liderado el desarrollo de la IA (Hao, 2019). Dos de las corrientes que hicieron presencia en esa primera conferencia de Dartmouth en 1956, y que han modelado el avance de la IA hasta la fecha, son las de la IA simbólica y la IA subsimbólica o conectivista (Anexo II).

Esta última consta de un conjunto de técnicas de ajuste con componentes estadísticos que permiten que un sistema aprenda en forma automatizada a través de la extracción de patrones e inferencias, sin necesidad de recibir instrucciones explícitas de un ser humano.

5 Para ello se utilizaron las bases de datos de código abierto más grandes de artículos científicos, arXiv. Este análisis cubrió documentos publicados desde 1993 hasta el 18 de noviembre de 2018.

En las décadas de los años cincuenta y sesenta empezaron a desarrollarse las bases matemáticas de lo que hoy se conoce como redes neuronales[6] (Rosenblatt, 1958) pero que, aunque fueron ampliamente utilizadas, no contaban con la capacidad de procesamiento y/o con la cantidad de información necesaria para ajustar un modelo de forma exitosa.

Por esta razón, desde mediados de los años setenta hasta mediados de la década de los años noventa los sistemas de IA simbólica –también conocidos como **"sistemas basados en el conocimiento"** (*Knowledge-based Systems*)– dominaron el desarrollo de la IA. Fundamentados en la generación de acciones por deducción de reglas lógicas o axiomas, estos sistemas encuentran decisiones óptimas mediante reglas predefinidas dentro de un dominio específico (Cristianini, 2014). Una de sus aplicaciones más conocidas se produjo en 1996, cuando la computadora Deep Blue de IBM venció al campeón mundial de ajedrez, Garry Kasparov. Sin embargo, cada vez se hicieron más evidentes las limitaciones de los sistemas expertos, en particular por su falta de capacidad para adquirir conocimiento de manera autónoma y también por la complejidad de su construcción: había demasiadas reglas que debían codificarse para crear un sistema "inteligente", lo cual aumentaba los costos y los tiempos de la construcción de un sistema.

En consecuencia, el aprendizaje automático, más conocido como "*machine learning*" (un tipo de IA subsimbólica), recuperó su popularidad a finales de los años noventa y principios de la década de los años 2000 con el surgimiento de técnicas como las redes bayesianas[7] y las máquinas de soporte vectorial[8]; sin embargo, desde la década de los años 2010 las redes neuronales cobraron importancia nuevamente y han reinado a partir de entonces (Hao, 2019).

En este contexto histórico, tres avances hicieron que la IA subsimbólica renaciera, predominara y revolucionara el campo de la IA:

- **La mejora de los algoritmos de redes neuronales** a partir de los esfuerzos realizados en 1986 por los científicos Rumerhalt, Hinton y Williams (1986) mediante la propuesta de ajuste de retropropagación[9].

- **El acceso y recolección masiva de información**; y, más recientemente,

- **El aumento de la capacidad de procesamiento** mediante el desarrollo de las llamadas unidades gráficas de procesamiento (GPU por sus siglas en inglés).

6 El modelo matemático en que se basaban las redes neuronales o de perceptrones –como se las conocía en la época– no permitía el diseño de modelos multicapa, por lo que, en la práctica, no se podían aplicar en ámbitos distintos a los que contenían problemas lineales de decisión positiva o negativa. Lo anterior redundó en la imposibilidad de reconocer la función de disyunción exclusiva, de donde surgió la crítica realizada por Marvin Minsky y Seymour Papert (1969) a las primeras redes neuronales; ello provocó el abandono virtual del estudio de tales redes durante casi 20 años.

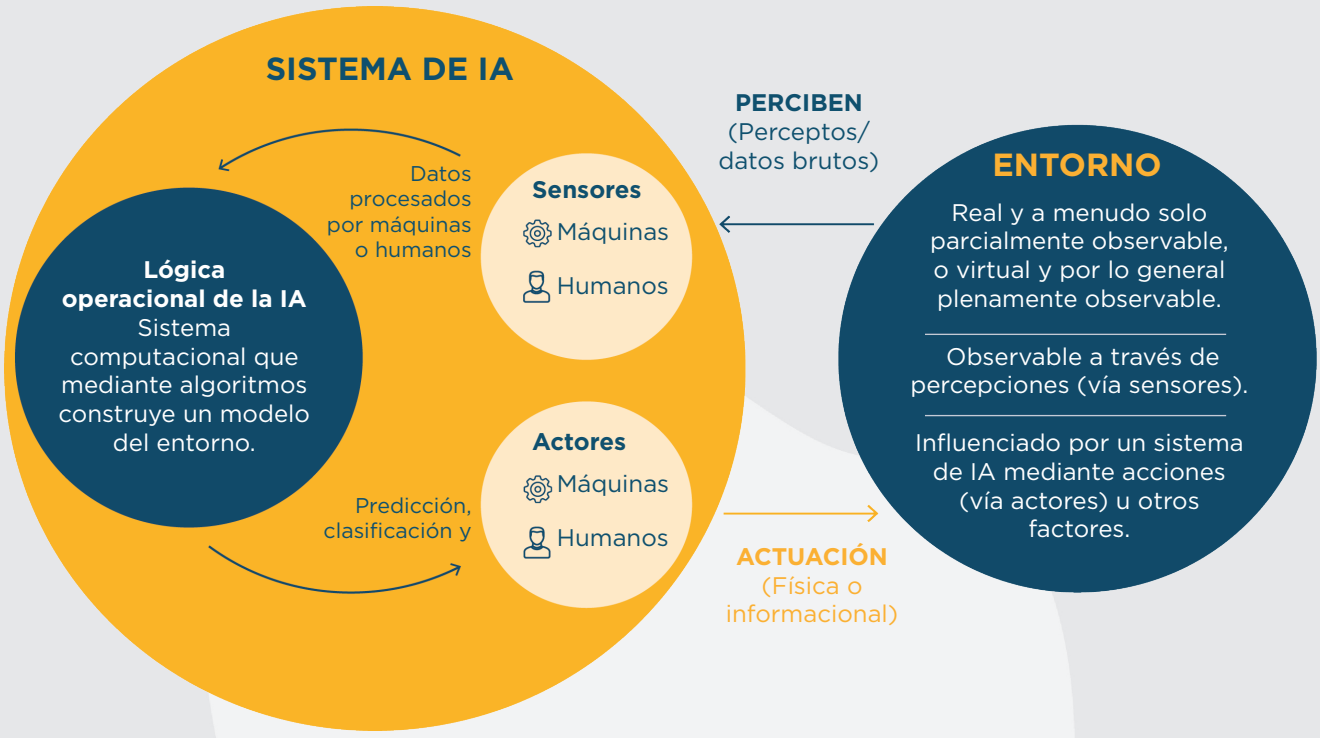
7 Es un modelo probabilístico multivariado que relaciona un conjunto de variables aleatorias mediante un grafo dirigido que indica explícitamente influencia causal.

8 Son sistemas de aprendizaje donde el algoritmo se enfoca en aprender a discriminar entre resultados positivos y negativos de una clase de vectores de n-dimensiones previamente establecida.

Lo anterior no implica que la IA simbólica se haya vuelto obsoleta, pues su uso sigue siendo amplio en aplicaciones de optimización y planificación. Más aún, en los últimos años se han desarrollado aplicaciones que combinan ambos paradigmas, como por ejemplo "los algoritmos de procesamiento del lenguaje natural [que] a menudo combinan enfoques estadísticos (basados en grandes cantidades de datos), y [los] enfoques simbólicos, que consideran temas como las reglas gramaticales" (OECD, 2019). Igualmente existe una convicción extendida acerca de que, para continuar con el desarrollo de la IA general, la IA subsimbólica no será suficiente y que por lo tanto se deberán encontrar mejores formas de combinar ambas corrientes.

La OCDE (OECD, 2019) incluye ambos paradigmas al describir a la IA como "**un sistema computacional que puede, para un determinado conjunto de objetivos definidos por humanos, hacer predicciones y recomendaciones o tomar decisiones que influyen en entornos reales o virtuales. Los sistemas de IA están diseñados para operar con distintos niveles de autonomía**" (OECD, 2019). Esta definición abarca los sistemas que pueden utilizar información o entradas suministradas por seres humanos o por máquinas, bien sea a través de análisis mediante algoritmos o en forma manual utilizando la inferencia estadística para formular opciones de información o actuación. Asimismo, se enfoca en el impacto e influencia que pueda tener la tecnología en el entorno social.

VISIÓN CONCEPTUAL DE ALTO NIVEL DE LA INTELIGENCIA ARTIFICIAL (gráfico 1)



Fuente: OECD (2019)

9 La retropropagación (del inglés backpropagation) es un método iterativo de aprendizaje supervisado que se emplea para entrenar redes neuronales artificiales mediante el ajuste del error de un nodo dentro de la red.

La iniciativa fAIr LAC adopta esta definición de IA de la OECD y sus líneas de acción se enfocan en cualquier sistema que pueda crear información para la toma de decisiones autónomas (A/TDA)[10] en materia de prestación de servicios sociales. Este enfoque es independiente del tipo de aprendizaje y del tipo de algoritmo, puesto que los resultados de los modelos de IA constituyen un proceso de creación de información que puede interactuar de distintas formas con las personas responsables de la toma de decisiones para afectar el entorno; esto ya sea que su resultado se adopte como una recomendación o que el sistema esté habilitado para iniciar una acción intermedia o final.

No obstante el hecho de que fAIr LAC utiliza una definición amplia sobre lo que significa la IA, esta iniciativa se enfocará principalmente en el análisis de riesgos y en la creación de mecanismos de mitigación de la IA subsimbólica, y específicamente de los algoritmos de aprendizaje de máquina (Anexo II), dado que estos se han popularizado durante las últimas décadas debido al potencial que tienen para mejorar el suministro de servicios sociales.

10 O también ADM/S por las siglas en inglés de automated decision-making or -support system.

02.1 Potencial de la Inteligencia Artificial para mejorar el bienestar social

Desde que se comenzó a hablar de tecnología en favor del bienestar social, han sido numerosos los enfoques para su uso. Las primeras aplicaciones se centraban en el gobierno electrónico y en mejorar los procesos de gobernanza. Recientemente, el enfoque se ha volcado a intentar solucionar problemas sociales de gran envergadura como la educación, la salud, la pobreza y la desigualdad, entre otros. En la medida en que la IA va adquiriendo carácter de tecnología accesible y de propósito general en la vida cotidiana, sus impactos serán mucho mayores en la existencia misma de las personas.

Las aplicaciones de IA son variadas y su crecimiento se nota en distintas áreas de la vida donde es posible detectar patrones a partir de grandes volúmenes de datos y de modelos complejos, así como en la disponibilidad de sistemas interdependientes que mejoren la toma de decisiones y generen políticas más igualitarias y eficientes. Los campos de investigación de IA como el procesamiento de lenguaje natural, la visión computacional, y los modelos de clasificación y predicción tienen un potencial muy significativo para incidir en el bienestar de la sociedad. En particular, tales aplicaciones se están utilizando en cuatro grandes ámbitos donde sus posibilidades de uso y alcance son elevadas: salud y pobreza, educación, equidad e inclusión social, y seguridad y justicia (McKinsey Global Institute, 2018a).

Algunos ejemplos ilustran este último punto. Según la Organización Mundial de la Salud (OMS), en el mundo existen cerca de 138 millones de pacientes que resultan perjudicados cada año por errores médicos, de los cuales 2,6 millones fallecen. Durante 2015, los errores médicos relacionados con el uso incorrecto de fármacos y fallas de diagnóstico representaron la tercera causa de muerte en Estados Unidos y el 10% de todas las muertes (GE Healthcare y UCSF, 2016). En tal sentido, el fortalecimiento de los sistemas informáticos de salud ayudará a que posteriormente se puedan implementar algoritmos preventivos que hagan más segura y eficiente este tipo de atención.

En el ámbito de la educación, por su parte, la IA puede adaptar el contenido de los cursos en función del progreso y aprendizaje de cada alumno (McKinsey Global Institute, 2018a). En materia de empleabilidad, la IA puede hacer más eficiente, justa e inclusiva la intermediación laboral. Hoy, la intermediación y el acoplamiento entre candidatos y puestos de trabajo ya no son las únicas áreas donde la IA puede incidir en los Servicios Públicos de Empleo. Con el uso de la IA, ahora es posible brindar también servicios integrales tanto a quienes buscan ocupaciones como a los departamentos de recursos humanos de las empresas, a los centros de capacitación y a la ciudadanía en general proporcionándoles la información laboral pertinente.

02.2 Inteligencia Artificial ética y responsable

La búsqueda y definición de una IA ética es aún materia de discusión. En los últimos años, organizaciones nacionales e internacionales han creado comités de expertos para debatir estos retos y han publicado una serie de principios (Jobin, Ienca y Vayena, 2019 y Mittelstadt, 2019). Los principios éticos proporcionan una orientación de alto nivel sobre lo que debe o no hacerse en el desarrollo y despliegue de sistemas de IA por parte de todos aquellos que tienen funciones activas en el ciclo de vida de un sistema de IA[11] (incluyendo su desarrollo, despliegue, operación, mantenimiento, etc.).

La iniciativa fAIr LAC utilizará como guía de desarrollo los principios éticos elaborados en mayo de 2019 por la OCDE (Recuadro 1) que fueron adoptados por todos los países miembros y seis países no miembros, y posteriormente por el G20 en junio de 2019 (OCDE, 2019).

Si bien seis países de la región (Argentina, Brasil, Colombia, Costa Rica, México y Perú) han adoptado los principios de la OCDE (2019), aún resulta incipiente la aplicación de recomendaciones específicas al respecto. Con el propósito de adoptar una IA responsable en la región, la iniciativa fAIr LAC ha optado por concentrarse en la definición de los retos que surgen al intentar operacionalizar los principios éticos para, a partir de allí, crear estrategias de implementación adecuadas desde las problemáticas y perspectivas latinoamericanas y caribeñas.

Lista de principios éticos de la OCDE (Recuadro 1)

Crecimiento inclusivo, desarrollo sostenible y bienestar

Las partes interesadas deberán participar activamente en la gestión responsable de una IA que esté pensada para alcanzar resultados beneficiosos para las personas y el planeta. Con el uso adecuado de la IA se podrá promover el aumento de las capacidades humanas y de la creatividad, la inclusión de poblaciones minoritarias, la reducción de las desigualdades económicas y sociales, así como la protección de entornos naturales, estimulando para ello el crecimiento inclusivo, el desarrollo sostenible y el bienestar.

Valores centrados en el ser humano y la equidad

Los actores del ecosistema de IA deben respetar el estado de derecho, los derechos humanos y los valores democráticos a lo largo de todo su ciclo de vida. Entre estos últimos sobresalen la libertad, la dignidad y la autonomía, la privacidad y la protección de los datos, la no discriminación y la igualdad, la diversidad, la equidad, la justicia social y los derechos laborales internacionalmente reconocidos. Con este fin, los actores de la IA deben implementar mecanismos y salvaguardias de protección de derechos como el de la autodeterminación de los individuos. Estos deben ajustarse al contexto y ser consistentes con el estado del arte.

Transparencia y explicabilidad

Los actores del ecosistema de IA deberán comprometerse con la transparencia y la divulgación responsable de los sistemas relacionados. Deberán proporcionar información relevante que se ajuste al contexto y sea coherente con el estado del arte. Con lo anterior se busca: (i) fomentar una comprensión general de los sistemas de IA, (ii) procurar que las partes interesadas tomen plena conciencia de sus interacciones con los sistemas de IA, (iii) asegurarse de que los afectados por un sistema de IA entiendan el resultado, y (iv) permitir que las personas afectadas adversamente por un sistema de IA impugnen sus resultados basándose en información clara y fácil de entender sobre los factores y la lógica que sirvieron de base para la predicción, recomendación o decisión que se busca refutar.

11 La OCDE (2019) describe el ciclo de vida de un sistema de IA como: (i) diseño, datos y modelación, (ii) verificación y validación, (iii) despliegue o publicación, y (iv) operación y monitoreo.

Robustez, seguridad y protección

La robustez, la seguridad y la protección son elementos esenciales de todo sistema de IA por las siguientes razones:

- Los sistemas de IA deben ser robustos, seguros y protegidos durante todo su ciclo para que, en condiciones de uso normal, uso previsible, uso incorrecto u otras condiciones adversas, funcionen adecuadamente y no supongan un riesgo irrazonable para la seguridad.

- Para ello, los actores de la IA deben garantizar la trazabilidad permanente, incluso en relación con los conjuntos de datos, procesos y decisiones tomadas durante el ciclo de vida del sistema de IA. Así será posible analizar correctamente, y en consonancia con el estado del arte, sus resultados y respuestas a las preguntas que se les formulen.

- En función de sus labores, del contexto y de su capacidad de actuación, los actores de la IA deben aplicar continuamente un enfoque sistemático de gestión de riesgos en cada fase del ciclo de vida del sistema para abordarlos de la mejor manera, incluyendo los relativos a la privacidad, seguridad digital y sesgos.

Rendición de cuentas

Los actores de la IA deben ser responsables del buen funcionamiento de los sistemas de IA y del respeto por los principios antes mencionados, en función de sus deberes, del contexto y del estado del arte.

Fuente: Traducción propia con base en OCDE (2019)[12]

12 Al momento de edición de este documento no existía una traducción oficial de los principios. La redacción tiene adaptaciones para facilitar su comprensión. Una vez la OCDE los publique en español se harán las modificaciones del caso.

RETOS DE LA INTELIGENCIA ARTIFICIAL EN LA REGIÓN

RETOS INTRÍNSECOS A LOS DATOS

Discriminación pasada

Desbalance de clase y validez externa

RETOS TÉCNICOS DE DESARROLLO

Fuga de información

Cambios temporales – retrolimentación
Sobreajuste y subajuste

RETOS DE INFRAESTRUCTURA Y SEGURIDAD

Ciberataques y falta de protocolos

Penetración ilegal de la IA y ataques adversarios

RETOS DE USO DURANTE LA EJECUCIÓN

Sesgo algorítmico injusto (output)

Impacto desigual (outcome)

RETOS DE ENTENDIMIENTO YPLANTEAMIENTO DE POLÍTICA

Necesidad y proporcionalidad

Definición correcta del problema y de la respuesta de política pública

RETOS DE RENDICIÓN DE CUENTAS

Redefinición de límites de privacidad

Rendición de cuentas y responsabilidad
Interpretabilidad y explicabilidad.

RETOS DE TRANSFORMACIÓN DIGITAL Y CAPACIDADES TÉCNICAS

Disponibilidad, interoperabilidad y calidad de datos

Falta de capacidades técnicas

RETOS DE IMPACTO SISTÉMICO

Interacción seres humanos-AI
Impactos económicos, educativos y laborales

Competencia y consolidación monopólica de la información

Fuente: Elaboración propia.

02.3 Retos de la Inteligencia Artificial

A medida que la IA se va expandiendo, de la mano del reconocimiento de sus beneficios y posibilidades también han surgido distintas voces que advierten sobre los retos que implica su adopción responsable y sobre las consecuencias indeseables que puede traer consigo su incorporación generalizada a los procesos y decisiones de la sociedad (Gráfico 2). Estas preocupaciones son muy diversas y no solo se centran en los sistemas de inteligencia artificial, sino que además consideran su impacto en el ecosistema de IA y en la forma en que el ser humano debe interactuar con aquellos y sus decisiones.

Para crear herramientas de mitigación exitosas que permitan diseñar sistemas cuyo impacto en los individuos y en la sociedad sea positivo, es necesario entender los tipos de retos y sus diferencias. La siguiente lista fue elaborada a partir de las discusiones que tuvieron lugar en mesas de trabajo multidisciplinarias con especialistas[13] en la región y contiene algunos de los retos más importantes en los cuales fAIr LAC estará enfocando sus líneas de acción.

Retos de transformación digital y capacidades técnicas

Los siguientes son algunos de los desafíos que conlleva la transformación digital y el desarrollo de capacidades técnicas que ello exige:

- **Disponibilidad, interoperabilidad y calidad de datos:** Entre los principales requisitos para ajustar los sistemas de IA con los modelos de aprendizaje de máquina es la disponibilidad de datos. Si bien es cierto que en los últimos años los países de ALC han avanzado en la digitalización de los servicios sociales, aún existe una importante fragmentación entre estos sistemas. Además, en la región se detecta un sesgo en la baja o nula disponibilidad de datos para zonas rurales con altos niveles de marginación. Es necesario que se sigan desarrollando sistemas digitales interoperables para que la información relevante "se encuentre disponible para el personal autorizado y contenga datos exactos, completos, necesarios y suficientes. Todo lo anterior dentro de un marco jurídico y regulatorio adecuado que respete las normas de privacidad, la ética, las leyes y las regulaciones vigentes" (Pombo et al., 2019).

- **Falta de capacidades técnicas:** En 2017, PricewaterhouseCoopers (PwC, 2017) estimó que la IA puede llegar a generar USD 15,7 billones a la economía mundial, lo cual representaría un aumento de 14% del PIB global. Los mayores beneficios económicos de la IA se registrarán en China (26% de aumento del PIB para 2030) y América del Norte (14,5% de aumento en el mismo lapso). Según este informe, si se mantiene la tendencia actual, América Latina capturaría solamente el 5,4% de esta suma, con lo cual quedaría en una

13 Especialistas e integrantes del grupo consultivo de fAIr LAC. Para más información, véase https://www.iadb.org/es/fairlac.

situación de desventaja competitiva, aumentando así su rezago frente a otras partes del mundo. Para aumentar el ritmo de adopción de la IA en la región es necesario crear capacidades técnicas e incentivar su uso responsable. Para ello, ALC deberá hacer esfuerzos encaminados a actualizar las competencias disponibles relacionadas con la IA, no solo para formar suficientes especialistas en este campo del conocimiento, sino también para permitir que un gran número de personas pueden convivir y trabajar con los sistemas de IA. La formación de capacidades en términos de capital humano es fundamental para absorber y aprovechar estas nuevas tecnologías (McKinsey Global Institute, 2018b).

Retos de entendimiento y planteamiento de política pública

En materia de asimilación de la IA para formular los planteamientos pertinentes de política pública existen los siguientes desafíos:

- **Necesidad y proporcionalidad:** Aunque la IA tiene el potencial suficiente para mejorar procesos y reducir desigualdades, también se debe tener en cuenta que adolece de una serie de limitaciones para no caer en lo que Morozov (2014) denomina el "solucionismo tecnológico" [14] , es decir, creer que la tecnología tiene la capacidad de resolver por sí sola todos los problemas sociales sin que paralelamente existan las políticas públicas adecuadas. Así pues, será necesario separar las capacidades reales de la IA como sistema generador de información, de la responsabilidad que compete a los encargados de la formulación de la política pública en cuanto a diseñar intervenciones para resolver las problemáticas sociales. Lo anterior implica aceptar que existen temas sociales para los cuales la incidencia que pueda tener la creación de un sistema de IA sea limitada o no sea adecuada. Esto exige que, en cada instancia en que se considere su aplicación, se deberá hacer un análisis de necesidad y proporcionalidad, para lo cual se requerirá considerar el riesgo de afectación de los derechos de las personas, la cantidad de personas que puedan resultar afectadas y su nivel de vulnerabilidad.

- **Definición correcta del problema y de la respuesta de política pública**: En consonancia con el punto anterior, existe el riesgo adicional de plantear los proyectos de IA desde la tecnología y no desde el problema social particular. Incluso un proyecto de IA necesario y funcional puede presentar riesgos si no se plantea la acción de política pública correcta.

14 Morozov identifica al "solucionismo tecnológico" como una ideología endémica que reformula fenómenos sociales complejos como la política, la salud pública, la educación y la aplicación de la ley como "problemas perfectamente definidos con soluciones definidas y computables o como procesos transparentes y evidentes que pueden optimizarse fácilmente, icon solo implementar los algoritmos correctos!".

Existe el riesgo adicional de plantear los proyectos de IA desde la tecnología y no desde el problema social particular. Un proyecto de IA necesario y funcional puede presentar riesgos si no se plantea la acción de política pública correcta.

Retos de infraestructura y seguridad

Uno de los principales temas que surgen en el campo de la IA se relaciona con la seguridad de su infraestructura. Aquí los desafíos para tener en cuenta son los siguientes:

- **Ciberataques y falta de protocolos:** Los riesgos de información están relacionados con la gobernanza y los protocolos de seguridad que se empleen a lo largo del ciclo de vida del proyecto. En los últimos años, en la región de ALC se ha registrado un incremento en casos de filtración de datos personales. En ocasiones los riesgos se generan por errores humanos originados por la falta de conocimiento y comprensión de los estándares de seguridad y buenas prácticas. Por ejemplo, en 2016 el padrón electoral del Instituto Nacional Electoral de México, que contenía información de cerca de 94 millones de personas, quedó expuesto sin contraseña en un servicio de almacenamiento debido a un error humano (Baraniuk, 2016). Aunque la falta de protocolos y el robo de información mediante ataques cibernéticos no es exclusivo de la IA, es cierto que la creciente expansión de sus aplicaciones en los servicios digitales aumenta la exposición de los datos personales de los ciudadanos. En los próximos años, en la región será cada vez más importante avanzar en la implantación de estándares y protocolos de seguridad de la información.

- **Penetración ilegal de la IA y ataques adversarios:** En diversos trabajos enfocados en visión computacional y clasificación de imágenes con redes neuronales se explora la manera de confundir a un algoritmo con ejemplos antagónicos. Szegedy et al. (2014) agregaron imágenes creadas sintéticamente con perturbaciones casi imperceptibles, cuyo propósito era hacer que el modelo las clasificara en forma errónea. Las aplicaciones de este tipo de ataques podrían ser muy diversas. Por ejemplo, en el caso del reconocimiento facial como medida de seguridad ampliamente utilizada en los dispositivos móviles, la penetración ilegal apuntaría a confundir al algoritmo haciendo que clasifique el rostro del atacante como si fuera el del dueño del dispositivo, con base en su información personal.

Retos intrínsecos a los datos (fuente de sesgo)

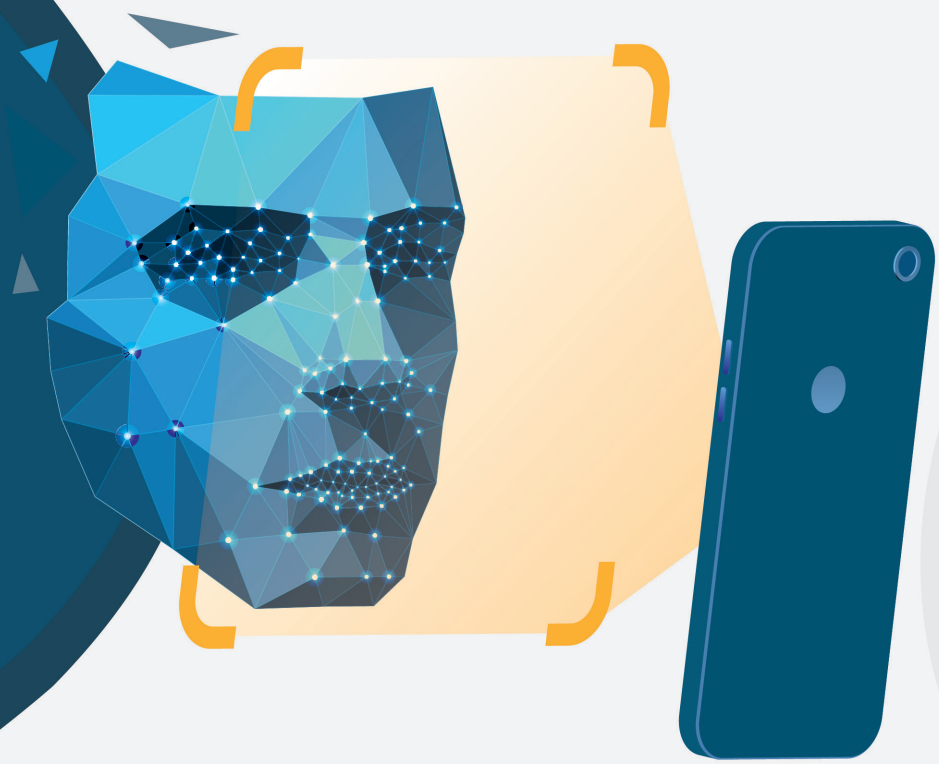
Aquí los retos se originan en problemas preexistentes en los datos que serán utilizados para ajustar el modelo. Son ocasionados por fenómenos intrínsecos a los datos de entrenamiento, como pueden ser la discriminación pasada y el desbalance de clase, los cuales se explican a continuación. Estos fenómenos pueden crear sesgos algorítmicos injustos y privilegiar a un grupo o perjudicar a otro.

- **Discriminación pasada:** Cuando se realiza un aprendizaje con información histórica o previamente etiquetada, los datos de entrenamiento pueden contener sesgos implícitos observados en la sociedad. Por ejemplo, en 2015 Amazon experimentó con un sistema de recomendación de recursos humanos a partir de técnicas de aprendizaje supervisado. El modelo entrenaba con los resultados históricos de procesos de selección de candidatos de los 10 años anteriores. Para cada currículum se contaba con una etiqueta binaria: 1 si el candidato había sido aceptado para la posición y 0 si había sido rechazado. Lo que el modelo no tomaba en cuenta es que la industria de la tecnología se ha caracterizado por ser predominantemente masculina[15], de modo que un algoritmo entrenado con esta información capturaba esos patrones de exclusión y terminaba recomendando una proporción mayor de hombres en forma consistente, reforzando así ladiscriminación de género (Dastin, 2018).

15 En los últimos años las empresas han puesto en marcha iniciativas para lograr la paridad de género.

- **Desbalance de clase y validez externa**: Este error se presenta cuando un modelo se entrena a partir de una base de datos que no es representativa de la población para la cual se quiere generalizar[16] o cuando no se tienen suficientes observaciones para las distintas subpoblaciones. Esto puede ocasionar que un modelo generalice con suficiente grado de precisión para la población en su conjunto, pero no para las subpoblaciones de forma particular (Guo, et al., 2008). Existen distintas metodologías para manejar el desbalance de clases, entre ellas, ajustar mediante submuestras o sobre un muestreo de los datos antes del entrenamiento.

En el ámbito de la inteligencia artificial, una de las aplicaciones en que se ha identificado este fenómeno de forma recurrente es la de visión computacional y reconocimiento facial. En un estudio publicado en 2018 (Buolamwini y Gebru, 2018) se señalaba que las bases de datos utilizadas para entrenar servicios comerciales de reconocimiento de rostros estaban compuestas mayoritariamente por sujetos cuyos tonos de piel eran más claros. Con este entrenamiento, la tasa de error para hombres blancos era del .8%, mientras que para mujeres con tonos oscuros de piel ascendía a 34,7% (Buolamwini y Gebru, 2018). **Dependiendo de la aplicación de un modelo con este tipo de errores, estas diferencias de equilibrio pueden tener un impacto importante en la vida de las personas que incremente la desigualdad entre la población.**



16 Existe una diferencia importante entre el error de validez externa, tal y como se entiende en los estudios econométricos, y los errores de desbalance de clase para modelos de aprendizaje automático. Buscando crear argumentos de causalidad, los primeros utilizan la representatividad para describir un fenómeno que afecta a la población. En cambio, los modelos de aprendizaje automático apuntan a minimizar el error en la predicción de observaciones externas al entrenamiento.

Retos técnicos de desarrollo e implementación

Son aquellos que se generan en el entrenamiento de los algoritmos, principalmente por errores metodológicos en la ejecución del proceso de entrenamiento de un modelo y en el manejo de la información. Estos errores también pueden crear sesgos algorítmicos injustos como privilegiar a un grupo o perjudicar a otro.

- **Fuga de información:** Este error es puramente metodológico y se produce cuando, durante el diseño del modelo, no se realiza una división apropiada entre el subconjunto de entrenamiento, prueba y validación. Esto conduce a que el modelo esté aprendiendo y evaluando con la misma información, lo que desemboca en un alto nivel de precisión que resulta ser muy poco realista. Un símil de esta situación en la vida real sería presentar un examen teniendo las respuestas correctas. Muy seguramente el estudiante obtendrá una calificación aprobatoria, lo cual no quiere decir que esté preparado para resolver esos problemas de forma independiente, dado que quizás solo se aprendió de memoria las respuestas.

- **Cambios temporales:** Este error se genera al seguir utilizando en forma indefinida un modelo que en algún momento haya tenido un buen nivel de precisión cuando el mundo real es complejo y se encuentra en cambio permanente. Por esto, la aplicación de tales modelos requiere que exista un equipo responsable que se pregunte constantemente si los supuestos, y la forma en la que el modelo se utiliza, siguen siendo útiles para la sociedad.

- **Sobreajuste y subajuste:** En modelos de aprendizaje de máquina entrenados con análisis supervisado, el objetivo del entrenamiento es generalizar ese aprendizaje. Esto quiere decir que las relaciones o los patrones aprendidos por el modelo mediante la observación de una porción de los datos le permiten tener información útil para clasificar ejemplos que aún no ha observado. Cuando se sobreajusta un algoritmo, este puede llegar a aprender perfectamente los datos de entrenamiento al nivel que impacta negativamente el rendimiento del modelo en información por fuera de ese subconjunto de aprendizaje. Es decir, es posible que no obtenga buenos resultados al clasificar información nueva. Por otro lado, el subajuste ocurre cuando el modelo es demasiado simple y no logra aprender lo suficiente como para permitirle crear una clasificación útil. Ambos fenómenos pueden presentarse para subpoblaciones que hacen parte de una población cuando se sobreajusta para grupos con características específicas o se subajusta para otras.

Retos de uso durante la ejecución

Son aquellos que se presentan cuando el algoritmo está siendo utilizado para tomar decisiones en una situación real. Se relacionan con el efecto que pueden ocasionar al segmentar o categorizar a la población y describen el riesgo que surge de la desconexión entre la toma de decisiones del algoritmo y el administrador del proceso.

- **Sesgo algorítmico injusto (*output*):** Se da cuando el sistema de IA comete errores sistemáticos que crean resultados injustos para subpoblaciones o individuos específicos. Como se vio en los dos tipos de retos anteriores (intrínsecos a los datos y técnicos de desarrollo e implementación), si bien estos sesgos se pueden crear por fenómenos como discriminación pasada, desbalance de clase, fuga de información, cambios temporales, y sobre y subajuste, entre otros, su impacto se manifiesta en cuanto se utilizan para tomar una decisión o acción de política pública. La forma en que se evalúa si el algoritmo está tomando una decisión injusta no es general para todos los problemas e incluso puede ser distinta para el mismo problema en diferentes contextos o países. Esto por cuanto lo que se entiende por "justicia" puede cambiar según la cultura y/o tradición de un grupo de población dado, o también en aquellas instancias en que una sociedad considerar necesario implementar políticas de discriminación positiva durante un periodo determinado para equilibrar una situación de discriminación previa (Cuadro 1). Es necesario propiciar discusiones abiertas al respecto, puesto que tener estas definiciones claras es la única forma de operacionalizar la mitigación del riesgo por sesgo tomando en cuenta las condiciones específicas de cada aplicación.

- **Impacto desigual (*outcome*):** El impacto desigual no requiere que exista un sesgo algorítmico en la recomendación (*output*), pues describe una situación en la cual el uso del producto o sistema impone una carga desproporcionada para los miembros de grupos específicos, creando la posibilidad de que su uso amplíe el estado de marginación o exclusión. Por ejemplo, las evaluaciones automatizadas de riesgos de reincidencia criminal, aunque estén bien calibradas mediante una definición de justicia como "paridad demográfica", pueden resultar en efectos acumulativos muy marcados para ciertos grupos, perpetuando así una condición de marginación.

ALGUNAS DEFINICIONES DE JUSTICIA[17] (cuadro 1)

Definición de justicia	Descripción
Justicia contrafactual	Esta medida considera que un predictor es "justo" si su resultado sigue siendo el mismo cuando se toma el atributo protegido y se invierte a su valor contrafactual (como por ejemplo introducir un cambio de raza, género u otra condición).
Paridad demográfica	Establece que la proporción de cada segmento de una clase protegida (por ejemplo, el género) debe obtener un resultado positivo en una misma proporción (como por ejemplo la asignación de becas escolares).
Igualdad de oportunidades	Esto implica que cada grupo debe obtener el resultado positivo en igualdad de condiciones, suponiendo que las personas cuentan con la misma calificación.

Fuente: Elaboración propia.

Retos de rendición de cuentas y cumplimiento

Uno de los principales desafíos de los sistemas de IA tiene que ver con el tipo de información que allí se recolecta sobre los individuos, y con la necesidad imperiosa de protegerla y asegurar que su uso será solo para aquellos fines que el titular haya aprobado. Aquí los desafíos son varios:

- **Redefinición de los límites de privacidad:** El ritmo acelerado de la innovación impide que los individuos, como sociedad, se planteen preguntas sobre los límites de privacidad que se deben establecer para estos sistemas. Además, es posible que cuando tales preguntas se formulen, ya sea demasiado tarde.

17A estas definiciones se puede agregar la de "justicia a través del desconocimiento": Se dice que un predictor logra la justicia a través del desconocimiento si los atributos protegidos no se utilizan explícitamente en el proceso de predicción. Esta medida tiene muchos problemas, ya que no permite evaluar el resultado (Gajane, 2018).

- **Rendición de cuentas y responsabilidad:** Este reto surge por la falta de claridad jurídica en materia de responsabilidad sobre las decisiones de un sistema. Por ejemplo, si un algoritmo decide el orden en que serán atendidos los pacientes que llegan a urgencias, y uno de los ellos falleciera por la falta de atención oportuna, ¿quién asume la responsabilidad de esa decisión? La necesidad de crear marcos de responsabilidad no puede subestimarse. Asimismo, se deben apropiar recursos de indemnización. Hoy en día, distintas empresas cuentan con modelos de negocios tipo Software como un Servicio (SaaS por sus siglas en inglés)[18]. Si un tercero contrata con una de estas plataformas y esto termina en un uso indebido o que cause algún daño cuantificable, ¿hasta qué punto es responsable la empresa proveedora del SaaS? Asimismo, ¿de qué forma se puede asegurar que las personas afectadas impugnen sus resultados?

- **Interpretabilidad y explicabilidad:** Estos dos términos relacionados se utilizan para describir el nivel de comprensión que se puede tener de los modelos. Por un lado, la interpretabilidad es la capacidad de observar bidireccionalmente en un sistema situaciones de causa y efecto. Esto implica tanto entender las razones por las cuales se ha realizado una predicción concreta, como predecir lo que sucederá dado un cambio en la entrada o en los parámetros algorítmicos. Entre tanto, la explicabilidad es un concepto más amplio que describe la capacidad de entender en términos humanos el funcionamiento de un modelo considerando sus entradas y salidas. En términos generales, se puede afirmar que existe una relación inversa entre explicabilidad e interpretabilidad.

Retos de impacto sistémico

Son aquellos que se presentan de forma indirecta en el sistema, entendido como el entorno social. Los principales desafíos son tres:

- **Interacción seres humanos-IA:** En esta instancia los retos abarcan las implicaciones indirectas del uso de la IA en los sectores sociales, incluyendo la definición del papel del usuario como receptor crítico de las recomendaciones de un sistema de IA para la toma de decisiones. Apoyarse en las nuevas tecnologías para ello es una práctica que se está extendiendo exponencialmente en el sector público. La forma óptima para que los responsables por la toma de decisiones (usuarios) utilicen las herramientas de IA es combinar los resultados de aquella con su propia intuición profesional. Sin embargo Nesta[19], partir de una extensa revisión de la literatura, entrevistas con servidores públicos y discusiones con expertos, identificó que algunos usuarios simplemente ignoran los resultados de tales herramientas (aversión algorítimica[20]), mientras otros recurren tanto a su conocimiento técnico como a su sentido común, generalmente sesgado, para informar el proceso de-

[18] El Software como un Servicio es un modelo donde el sistema o aplicación es administrado y mantenido por el proveedor; el cliente únicamente consume los resultados sin realizar cambios o ajustes.
[19] Nesta (National Endowment for Science, Technology and the Arts) es una fundación para la innovación radicada en el Reino Unido. El document se titula "Decision-making in the Age if the Algorithm: Three key principles to help public sector organisations make the most of AI tools" (Snow, 2019).
[20] Las personas son aversas a los algorítmicos predictivos después de verlos actuar, incluso cuando los ven superan una predicción humana. Esto se debe a que las personas pierden más rápido la confianza en el algoritmo que en las personas después de verlos cometer un error (Dietvorst et al., 2014).

cisorio. Esto implica que el sesgo perdura como una de las características de la toma de decisiones como actividad humana, no obstante la introducción de herramientas como la IA (Snow, 2019). Dadas las limitaciones existentes en los sistemas de IA en situaciones de alto riesgo, así como la aversión hacia la aplicación de los resultados por parte de algunos usuarios, se podría sugerir que la IA se emplee solo como soporte y como un insumo complementario de información para considerar. En su informe, Nesta destaca tres principios clave en la interacción seres humanos-IA que, de ser tenidos en cuenta, mejorarían la adopción y utilización de las de las recomendaciones de un sistema de IA: contexto, comprensión y agencia (Snow, 2019).

- **Impactos económicos,** educativos y laborales: Este desafío incluye los cambios que se están produciendo en los mercados de trabajo por concepto de la automatización masiva de funciones y tareas, si bien desde la perspectiva opuesta también comprende los efectos que tiene el surgimiento de empleos nuevos y la necesidad de que la ciudadanía adquiera las competencias que exigen estos trabajos emergentes.

- **Competencia y consolidación monopólica de la información:** En general se puede afirmar que un modelo de IA subsimbólica será más preciso si se entrena con mayor cantidad de información, siempre y cuando esta sea de buena calidad[21]. La posición privilegiada de, por ejemplo, gobiernos y empresas consolidadas en cuanto al acceso y capacidad de recolección de datos, puede propiciar el surgimiento de monopolios de información. Un caso puntual se puede ver cuando un producto de IA se convierte en líder en el mercado incrementando el número de usuarios de su servicio. A mayor cantidad de usuarios será mayor el volumen de información que alimente el sistema, lo que a su vez mejorará la calidad de los modelos. Esto genera un tipo de mercado en donde el ganador se lleva todo ("*winner-take-all market*"). Lo anterior significa que, gracias a su calidad ligeramente superior frente a la de sus competidores, un determinado producto obtendrá una proporción más amplia de usuarios e ingresos para esa clase de productos o servicios. Y al capturar un porcentaje mayor de la demanda consolidará su liderazgo. Es importante que en el desarrollo de soluciones dirigidas a prestar servicios sociales se encuentren mecanismos para disminuir tales efectos. Esto se puede lograr promoviendo proyectos de datos abiertos e iniciativas colaborativas como forma alternativa de construcción de IA, y también incentivando la creación de estructuras como los llamados *data trusts y data commons*[22].

[21] Existen casos en los cuales una mayor cantidad de datos puede no llevar a un mejor modelo o a aproximaciones metodológicas superiores (como la estadística bayesiana, por ejemplo), especialmente allí donde la falta información se puede compensar con conocimiento experto. Sin embargo, la relación entre calidad de datos y precisión es incuestionable.
[22] Data Trust se refiere a un marco repetible de términos y mecanismos para el manejo de la información. Data Commons se puede referir o bien a una plataforma tecnológica para almacenar y manipular conjuntos de datos o al conjunto de principios y estrategias de gobernanza para el uso de tales conjuntos de datos.

03. La iniciativa fAir LAC

Entre la década pasada y mediados de la presente, la región de América Latina y el Caribe (ALC) logró importantes avances en materia de reducción de pobreza y desigualdad. Sin embargo, la desigualdad social continúa siendo un problema sustancial en la región, considerada la más desigual del planeta.

Desde 2015 "se han registrado retrocesos, particularmente en materia de pobreza extrema, donde 167 millones de personas aún viven en esa situación" (CEPAL, 2019). Esta es una población que, en su mayoría, continúa siendo rural, con acceso insuficiente y de baja calidad a servicios básicos de salud, agua y saneamiento, y que arrastra un legado de discriminación de género, racial y de clase social.

ALC debe esforzarse por encontrar formas nuevas y mejores de reducir la pobreza, promover el crecimiento económico y favorecer la distribución de riqueza. La IA surge como una herramienta innovadora para lograr un mayor impacto socioeconómico.

El Government AI Readiness Index de 2019[23], producido con el apoyo de Oxford Insights y el Centro Internacional de Investigaciones para el Desarrollo (IDRC por sus siglas en inglés), muestra que los países de la región se enfrentan a tres desafíos cuando se trata de aprovechar el uso de la IA en favor del bien común: políticas, capacidad y recursos adecuados. En primer lugar, hasta la fecha ALC no cuenta con un enfoque coherente de **política** y tampoco con estándares éticos definidos. México, Colombia, Uruguay y Argentina están fijando actualmente políticas y estrategias de IA. Colombia, por ejemplo, por medio del documento CONPES 3975 definió su Política Nacional para la Transformación Digital e Inteligencia Artificial. Allí se identifican lineamientos concretos que, a través de su implementación, generarán un marco coherente de política para el desarrollo ético y responsable de la IA.

[23] El Government AI Readiness Index es el sistema de clasificación creado por Oxford Insights y el IDRC. El índice se obtiene mediante la suma de una normalización promedio de las métricas indexadas en una escala de 0 a 10 a partir de fuentes como la ONU, el WEF, el Global Open Data Index y el Banco Mundial, además de Gartner, Nesta y Crunchbase. Estas métricas se agrupan bajo cuatro temas de alto nivel: gobernanza, infraestructura y datos, habilidades y educación, y servicios públicos y gubernamentales.

En segundo lugar, la **capacidad** es un desafío para la región, y en particular para sus gobiernos. Aunque hay algunas empresas y académicos que trabajan en el campo de la IA, no existe un conocimiento generalizado en los sectores económicos, como tampoco una conexión clara sobre su aplicabilidad en el sector público. Por último, en comparación con países como Canadá, Estados Unidos y el Reino Unido, las naciones latinoamericanas aún no han logrado vincular los capitales público y privado con los recursos técnicos y académicos de los cuales dispone la región para establecer centrales de desarrollo de IA (*hubs*) (Oxford Insights e International Development Research Centre, 2019).

El potencial desaprovechado es significativo. Se estima que para 2035, el desarrollo de la IA podría agregar un punto porcentual al crecimiento económico anual de ALC (Ovanessoff y Plastino, 2017). Para no desperdiciar esta oportunidad, se necesita que tanto los responsables por la formulación de políticas como los emprendedores y la sociedad civil, vean la IA como una herramienta con el potencial de aprovechamiento para generar crecimiento económico y bienestar social en el largo plazo, y no simplemente como otro motor de la productividad. A su vez, el análisis de los riesgos actuales y emergentes de la IA y de su impacto en la fuerza laboral ayudará a crear una estrategia de mitigación de estos. Si bien hoy en día se tiende a considerar a la IA como una caja negra, es posible exigir transparencia, explicabilidad y trazabilidad en todo su desarrollo y ejecución.

Es así como el Banco Interamericano de Desarrollo (BID) está liderando fAIr LAC con el apoyo de varios socios estratégicos. Con esta iniciativa se busca promover una aplicación responsable de la IA para mejorar el suministro de servicios sociales y con ello atenuar la creciente desigualdad social en la región. fAIr LAC se apalanca en los sectores público y privado, así como en la sociedad civil, para lograr incidir tanto en la política pública como en el ecosistema emprendedor.

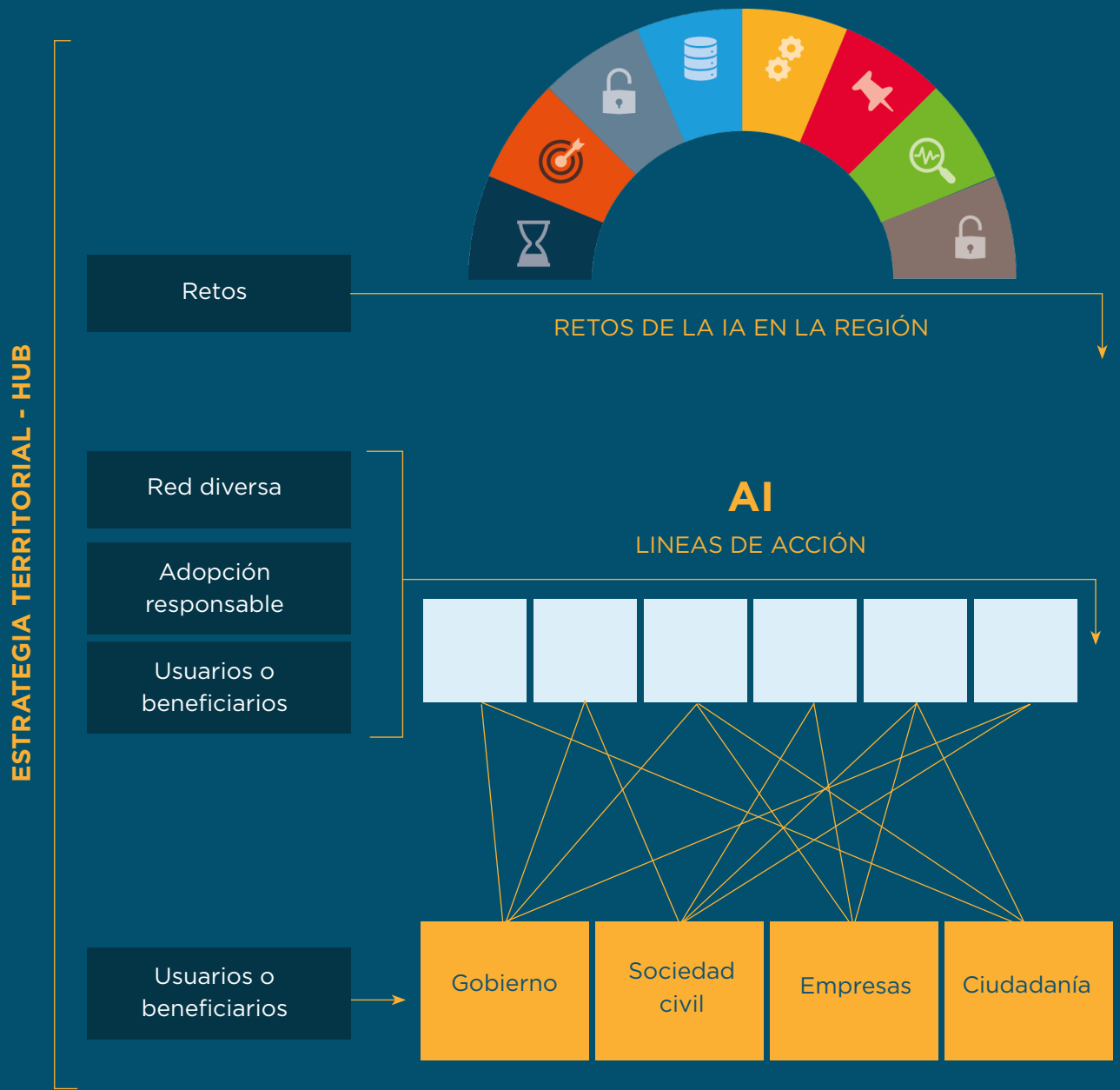
Dimensiones estratégicas y líneas de acción

Dados los retos identificados y el alcance determinado para fAIr LAC, se plantearon tres dimensiones que articulan estrategias para lograr el objetivo de promover un uso responsable de la IA en el campo de la prestación de servicios sociales. Las dimensiones estratégicas son tres: (i) desarrollo de una red diversa, (ii) formación de capacidades para una adopción responsable y (iii) promoción de calidad y mitigación de riesgos.

Dentro de cada dimensión estratégica se plantean líneas de acción con actividades e intervenciones específicas dirigidas a uno o varios de los usuarios o beneficiarios identificados, a saber, el gobierno, las organizaciones de la sociedad civil, las empresas y la ciudadanía (Gráfico 3)[24].

[24] La descripción detallada cada línea de acción supera el alcance de este trabajo. Esto será tema de documentos específicos posteriores.

Dimensiones estratégicas de fair lac (gráfico 3)



Fuente: Elaboración propia.

03.1 Dimensión 1: Desarrollo de una red diversa (visibilizar, difundir, construir y vincular)

Para garantizar que se escuche a una variedad de voces y se identifiquen los retos que afectan a la población de ALC, es necesario crear una red diversa que pueda contribuir a nutrir las discusiones sobre la IA, su potencial y sus implicaciones. Esto permitirá a ciudadanos, empresas y gobierno compartir conocimiento y generar vínculos entre el ecosistema y las iniciativas.

Aquí el objetivo es **crear espacios de encuentro, diálogo e intercambio** como los que se describen a continuación, donde la ciudadanía, las empresas y el gobierno puedan generar y compartir conocimiento.

- ***Grupo consultivo de personas expertas en IA.*** Esta es una red de profesionales y expertos que, desde la academia, el gobierno, la sociedad civil, la industria y el sector emprendedor asesora el desarrollo e implementación de las líneas de acción de fAIr LAC. El grupo promoverá una aplicación ética de la IA dentro de ALC, y propenderá por que se entiendan las implicaciones y particularidades del contexto latinoamericano en relación con discusiones globales. Asimismo, ayudará a guiar las iniciativas de los gobiernos para un uso responsable de la IA.

- ***Observatorio de casos de uso.*** Se creará una plataforma con fichas descriptivas de aquellos casos de uso de IA en la prestación de servicios sociales que se estén desarrollando en la región por parte de gobiernos, empresas, la academia y la sociedad civil. El objetivo es proveer un bien público regional para que tanto el sector público como el privado conozcan casos exitosos, identifiquen buenas prácticas y sistematicen aprendizajes de aplicaciones que se puedan emular en sus países y/o empresas, dando así visibilidad a las mejores prácticas de la región. El observatorio será alimentado y actualizado constantemente, buscando que se constituya en un referente documental sobre el estado de la IA en la región, incluyendo iniciativas no relacionadas con el BID y/o financiadas por este.

- ***Mecanismo de alianzas y alineación institucional.*** Este tiene por objeto identificar aquellas iniciativas que en materia de IA estén siendo implementadas simultáneamente por instituciones aliadas. Esto con el fin de explorar sinergias y oportunidades de colaboración, así como la existencia de una alineación institucional tanto en el ámbito regional como en el internacional.

- ***Sensibilización y comunicación.*** Si bien todo hace pensar que la región no se convertirá en un líder en el campo de la creación de inteligencia artificial en un futuro cercano, esta sí se aplicará, utilizará y desarrollará de forma generalizada en sus países.

El objetivo es proveer un bien público regional para que tanto el sector público como el privado conozcan casos exitosos, identifiquen buenas prácticas y sistematicen aprendizajes.

En tal sentido es clave que dirigentes, funcionarios públicos, empresarios y ciudadanos conozcan sus ventajas y riesgos. Para ello, fAIr LAC cuenta con una estrategia de sensibilización que incluye diseminación de conocimiento y campañas de comunicación en medios (Gráfico 4).

LINEAS DE ACCIÓN DE EJE DE RED DIVERSA (gráfico 4)

	Dimensiones estratégicas para fAIr LAC	Lineas de acción		Usuarios o beneficiarios
Modelo fAIr LAC	**Red diversa** Generación y difusión de conocimiento	**Grupo consultivo de personas expertas en IA** (una red de profesionales y expertos de la academia, el gobierno, la sociedad civil, la industria y el sector emprendedor)	**Mecanismo de alianzas y alineación institucional** (org. multilaterales, redes existentes)	Gobierno
		Sensibilización y comunicación (reportes, artículos de opinión y análisis, debates)	**Observatorio de casos de uso** (bien público regional para identificar casos exitosos y sistematizar buenas prácticas y aprendizajes)	Sector privado Ciudadanía

Fuente: Elaboración propia.

03.2 Dimensión 2: Formación de capacidades para una adopción responsable de la Inteligencia Artificial

Para asegurar que en la región se adopte la IA de manera responsable, es importante, primero, lograr que los encargados de la formulación de políticas, los ciudadanos y el ecosistema empresarial comprendan plenamente los retos y posibilidades que plantea la IA. Asimismo, la región debe incentivar la creación de espacios de capacitación que involucren a funcionarios públicos, representantes de la sociedad civil y emprendedores (Gráfico 5). Para esto se han considerado varias acciones y productos:

- ***Experimentos y proyectos piloto.*** Se crearán programas de apoyo para experimentos y casos de uso (asesoría, financiamiento, red de mentores). Estos tienen una doble función: acumular experiencia institucional en la aplicación de proyectos analíticos y prácticos, y sistematizar experiencias de aplicaciones donde la IA ayude a crear mayor impacto social, respetando los derechos humanos. Se priorizarán proyectos que puedan ser escalados y emulados en la región, considerando en todo momento la necesidad de adaptación de dominio por la multiplicidad de contextos y necesidades distintas en cada país, ciudad o municipalidad.

- ***Sistema de incentivos.*** Se propone crear modelos de incentivos (reconocimientos, acceso a financiamiento y recursos, respaldo institucional, retos) para funcionarios, empresas y ciudadanía.

- ***Herramientas diversas*** Aquí la idea es que funcionarios públicos y emprendedores tengan la posibilidad de acceder a formación educativa sobre IA, sus beneficios y riesgos, y también puedan profundizar su conocimiento en la materia a través del desarrollo y publicación de guías, marcos y otras herramientas metodológicas para la adopción responsable de la IA.

LINEAS DE ACCIÓN DE EJE DE ADOPCIÓN RESPONSABLE (gráfico 5)

	Dimensiones estratégicas para fAIr LAC	Lineas de acción		Usuarios o beneficiarios
Modelo fAIr LAC	Adopción responsable IA para servicios sociales	**Programa de apoyo para experimentos y casos de uso** (asesoría, financiamiento, red de mentores)	**Proyectos piloto** (con potencial de ser escalados en la región)	Gobierno
		Modelo de incentivos (reconocimientos, acceso a financiamiento y recursos, respaldo	**Herramientas diversas:** - Metodologías, guías, marcos de evaluación de impacto - Repositorios de modelos abiertos e índice de herramientas	Sector privado

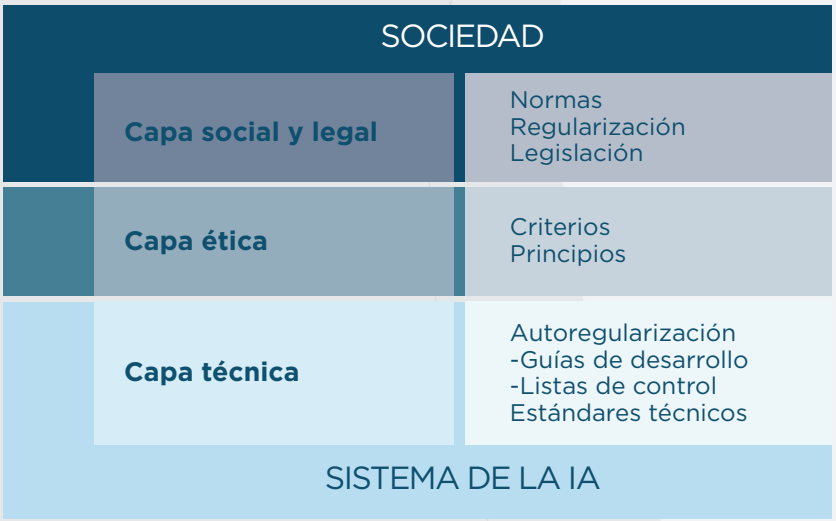
Fuente: Elaboración propia.

03.3 Dimensión 3: Promoción de calidad y mitigación de riesgos

A la luz de los beneficios transformadores de la IA, así como de sus riesgos, se debe priorizar el diseño de una política pública que los equilibre. Las sociedades deben establecer mecanismos de vigilancia a lo largo del ciclo de vida de los sistemas de IA. Esto incluye garantizar el desarrollo de algoritmos con la verificación de estándares mínimos, auditar la calidad de los datos y el proceso de diseño, establecer protocolos de seguridad y gobernanza de la información personal, y crear mecanismos jurídicos de responsabilidad y rendición de cuentas.

Es posible establecer mecanismos variados de mitigación de riesgos, entre los cuales figuran la autorregulación, la expedición de estándares, los procesos de auditoría y los marcos regulatorios oficiales (Gráfico 6). Cualquiera de ellos es deseable; su diferencia radica en si son o no jurídicamente vinculantes. Gasser y Almeida (2017) proponen un modelo de representación de gobernanza por capas para regular el uso de los sistemas de IA en su interacción con la sociedad. El modelo plantea distintos instrumentos que pueden ser desarrollados de forma paralela. La iniciativa fAIr LAC impulsará la elaboración de estos en el ámbito regional y territorial.

MODELO DE GOVERNANZA POR CAPAS (gráfico 6)



Fuente: Adaptado de Gasser y Almeida (2017)

- ***Guías y directrices.***
  La iniciativa fAIr LAC trabajará en la creación de herramientas que puedan guiar a desarrolladores y a los responsables por la formulación de políticas públicas sobre cómo abordar los retos de necesidad y proporcionalidad[25], así como los concernientes a desarrollo e infraestructura [26]. Estos documentos abarcarán campos de estudio que incluyan los siguientes temas:

  - **Evaluación de necesidad y proporcionalidad:** Con el apoyo de especialistas del sector social del Banco Interamericano de Desarrollo y de otros expertos de la red regional se evaluará la propuesta de sistemas de IA aplicados a los servicios sociales en los proyectos piloto; esto bajo un análisis integral de desarrollo que considere el efecto que puedan tener las propuestas de política públicas resultantes.

  - **Gobernanza de datos:** Aquí se describirán los estándares de seguridad de la información, protección de datos personales, procesos de gobernanza empresarial e interoperabilidad.

  - **Evaluación de equidad algorítmica:** Se abordarán criterios de no discriminación como son los de paridad demográfica, igualdad de oportunidades y umbrales de grupo, entre otros (Dignum, 2019 y Gajane, 2018) [27]. Dado que estos criterios dependerán del concepto de equidad culturalmente determinado, su análisis se realizará en los estudios de caso de los proyectos piloto implementados por fAIr LAC.

- ***Autorregulación.***
  Dado que, en general, la industria avanza a un ritmo más acelerado que la acción gubernamental y los esquemas regulatorios, incentivar la autorregulación de las empresas es un paso muy importante en la generación de buenas prácticas, pues sirve como instrumento para reflejar los principales problemas que encuentran las empresas que están desarrollando tecnología de punta. Las guías y listas de control figuran entre las herramientas de uso interno que establecen niveles mínimos de calidad en los procesos. Es importante mencionar que existen limitaciones a la autorregulación como herramienta de gobernanza, ya que las empresas tienen incentivos que pueden ir en dirección contraria al bienestar general. Aun así, la iniciativa fAIr LAC considera que la autorregulación es un paso inicial clave que por sus características y ritmo acelerado de implementación permite experimentar con normas y guías, y evaluarlas. Sin embargo, aquella deberá acompañarse de un esfuerzo encaminado a lograr estándares y marcos regulatorios oficiales. Por esta razón, fAIr LAC propende por crear un proceso de autoevaluación ética que pueda ser implementado por cualquier empresa y que le permita identificar sus principales riesgos y las acciones apropiadas para mitigarlos.

[25] Véase el apartado “Retos de entendimiento y planteamiento de política pública” en la sección II del presente documento.
[26] Véase el apartado “Retos técnicos de desarrollo e infraestructura” en la sección II del presente documento.
[27] Estos conceptos no se describen en este documento, pues superan su alcance. Se desarrollarán en materiales posteriores de fAIr LAC.

- ***Estándares.***
  Los estándares son un conjunto de buenas prácticas relacionadas con sistemas o herramientas específicas con los que se busca establecer un sistema de calidad. A diferencia de la autorregulación, para los estándares existen organismos independientes con comités técnicos que validan el compendio de normas. La formulación de estándares permite expedir certificaciones y realizar procesos de auditoría que generan incentivos de mercado, en la medida en que operan como factor diferenciador en un mundo competitivo. Aquí es fundamental la figura de "entidad de certificación", a saber, un tercero que realiza la verificación del cumplimiento de los procesos definidos en el estándar y expide la certificación. fAIr LAC trabajará con actores del ecosistema de IA en el desarrollo de estándares de calidad para proyectos de iniciativa pública y privada.

- ***Marco regulatorio oficial.***
  Es posible que entre los marcos regulatorios de los diversos países de la región existan diferencias. En este documento se caracterizan como estrategias de mitigación de largo plazo aquellas modificaciones introducidas en normas, reglamentos y leyes establecidas de forma oficial por un gobierno que sean jurídicamente vinculantes, es decir, que tengan carácter de obligatoriedad. Cuando una buena práctica o un estándar se convierte en norma o ley se vuelve legalmente exigible y las consecuencias de su incumplimiento deben ser claras. Se trabajará en el desarrollo de iniciativas de experimentación regulatoria (sandboxes regulatorios) y otros mecanismos que informen el desarrollo de marcos regulatorios locales (Gráfico 7).

LINEAS DE ACCIÓN DE EJE DE CALIDAD Y MITIGACIÓN DE RIESGOS (gráfico 7)

	Dimensiones estratégicas para fAIr LAC	Lineas de acción				Usuarios o beneficiarios
Modelo fAIr LAC	Calidad y mitigación de riesgos	Autoregulación, guías o directrices	Estándares y certificaciones	Iniciativas de experimentación regulatoria	Marcos normativos	Gobierno
		Fortalecer la participación de expertos y practicantes de ALC en esfuerzos internacionales (IEE, OCDE, etc) y acercar iniciativas internacionales a ALC				Sector privado

Fuente: Elaboración propia.

03.4 Estrategia regional y territorial

Para garantizar que produzca impacto social, además de tener enfoques sectoriales esta estrategia debe estar anclada a territorios y actores estratégicos en América Latina y el Caribe a través de centrales (*hubs*) locales de fAIr LAC que promuevan el uso responsable de la IA. Este anclaje territorial se basa en la adaptación a las particularidades locales de elementos como son la articulación de actores públicos y privados, el desarrollo de capacidades, el fortalecimiento y la curaduría de datos, los incentivos a la demanda de soluciones basadas en IA, y la promoción de casos de uso (desde la identificación hasta la implementación). Al mismo tiempo se debe buscar armonizar las iniciativas de la región como un todo para así evitar que la fragmentación normativa perjudique la posibilidad de establecer un ecosistema robusto de IA.

Una central o *hub* de IA es un ecosistema habilitador que presenta condiciones deseables para el desarrollo e implementación de la iniciativa fAIr LAC. Las dimensiones estratégicas se emularán en el ámbito local con el propósito de fomentar no solo la implementación de una IA responsable sino también alianzas entre instituciones públicas y privadas que promuevan a ALC como polo de innovación de IA para el impacto social (Gráfico 8).

La retroalimentación y aprendizaje entre la iniciativa regional y la territorial será bidireccional, ya que, **mediante la ejecución de los experimentos y casos de uso, el desarrollo de IA generará aprendizajes que puedan ser utilizados para producir conocimientos destinados a toda la región.**

A diciembre de 2019, fAIr LAC contaba con tres centrales o hubs regionales: Jalisco (México), Uruguay y Costa Rica. Estas se convertirán en referentes territoriales en el uso responsable de IA en Norteamérica, Centroamérica y el Cono Sur respectivamente.

MODELO REGIONAL Y TERRITORIAL (gráfico 8)

Fuente: Elaboración propia.

Conclusión

La importancia de la inteligencia artificial (IA) y otros desarrollos tecnológicos afines en el futuro de la humanidad es incuestionable. Estos avances están transformando radicalmente la manera como trabajamos y vivimos, y sus efectos – muchos positivos y otros no tanto – son materia de discusión en numerosas instancias de la sociedad en todos los países del mundo.

Como se ha visto a lo largo del presente documento, uno de los campos donde se espera lograr impactos significativos con el uso de la IA es el del **bienestar social de la ciudadanía a través de una prestación eficiente y efectiva de servicios sociales relacionados** principalmente con la educación, la salud y la pobreza, todo ello en aras de reducir la desigualdad.

Siendo estos los temas que ocupan un lugar privilegiado de la agenda pública en América Latina y el Caribe, la labor que logren desarrollar el BID y sus aliados de los sectores público y privado, la academia y las organizaciones de la sociedad civil a través de la iniciativa fAIr LAC será fundamental para abordar los retos que presenta el uso de a IA en la región.

Aunque los desafíos identificados son varios y comprenden asuntos de carácter técnico, regulatorio, de manejo de datos y de política pública, entre otros, los mayores esfuerzas deberán centrarse en garantizar un uso responsable y ético de la IA y en evitar la profundización de las desigualdades sociales neutralizando los efectos adversos de un aprendizaje automático sesgado en contra de los grupos menos favorecidos.

Los mayores esfuerzas deberán centrarse en garantizar un uso responsable y ético de la IA para evitar aumentar la desigualdad de nuestra region.

04. Referencias

- Baraniuk, C. (2016). Millions of Mexican Voter Records '*Were Accessible Online*'. Abril. Obtenida de BBC: https://www.bbc.com/news/technology-36128745.
- Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Nueva York: Springer-Verlag.
- Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press.
- Buolamwini, J. y T. Gebru. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Conference on Fairness, Accountability, and Transparency.
- CEPAL (Comisión Económica para América Latina y el Caribe). (2019). *Panorama Social de América Latina 2018*. Santiago: Naciones Unidas. Obtenido de LC/PUB.2019/3-P.
- Cristianini, N. (2014). On the Current Paradigm in Artificial Intelligence. *AI Communications 27*, No. 1. Obtenido de https://doi. org/10.3233/AIC-130582.
- Dastin, J. (2018). *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women.* Obtenido de Reuters: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.
- Dietvorst, B., J. Simmons y C. Massey. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General,* 1, 114–126. doi: 10.1037/xge0000033.
- Dignum, V. (2019). *Responsible Artificial Intelligence*. Springer. ISBN 978-3 -030-30371-6.
- Epstein, R., G. Roberts y G. Beber (2008). *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer.* Springer. p. 65. ISBN 978-1-4020-6710-5.
- Gajane, P. y M. Pechenizkiy. (2018). On Formalizing Fairness in Prediction with Machine Learning. Cornell University. Obtenido de https://arxiv.org/pdf/1710.03184.pdf
- Gasser, U. y V.A.F. Almeida. (2017). A Layered Model for *AI Governance. IEEE Internet Computing*, 21 (6) (Noviembre): 58–62. doi:10.1109/mic.2017.4180835.
- General Electric Healthcare y University of California-SF. (2016). Big Data, Analytics & Artificial Intelligence. Obtenido de http://newsroom.gehealthcare.com/wp-content/uploads/2016/12/GE-Healthcare-White-Paper_FINAL.pdf
- Guo, X., Y. Yilong, C. Dong, G. Yang y G. Zhou. (2008). On the Class Imbalance Problem. Fourth International Conference on Natural Computation.
- Hao, K.. 2019. *We Analyzed 16,625 Papers to Figure Out Where AI is Headed Next*. 25 de enero. Obtenido de https://www.technologyreview.com/s/612768/we-analyzed-16625-papers-to-figure-out-where-ai-is-headed-next/
- Jobin, A., M. Ienca y E.Vayena. (2019). The Global Landscape of AI Ethics Guidelines. *Nat Mach Intell* 1, 389–399. doi:10.1038/s42256-019-0088-2.
- cCarthy, J., M. L. Minsky, N. Rochester y C.E. Shannon. (1955). *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. Obtenido de http://jmc.stanford.edu/

articles/dartmouth/dartmouth.pd

- McCorduck, P. (2004). *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. A K Peters/CRC Press; 2a edición
- McKinsey Global Institute. (2018a). *Notes from the AI Frontier: Applying AI for Social Good*. McKinsey & Company.
- ----------. (2018b). *Notes from the AI Frontier: Modeling the Impact of AI on the World Economy.* McKinsey & Company.
- Minsky, M. y S. Papert. (1969). *Perceptrons: An Introduction to Computational Geometry. Cambridge*: MIT Press
- Mittelstadt, B. (2019) Principles Alone Cannot Guarantee Ethical AI. Obtenido de https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3391293.
- Morozov, E. (2014). T*o Save Everything, Click Here: The Folly of Technological Solutionism.* PUBLICAFFAIRS. ISBN 9781610391382
- OECD (Organizatión for Economic Co-operation and Development). (2019). *Artificial Intelligence in Society*. París: OECD Publishing.
- Ovanessoff, A. y E. Plastino. (2017). *How Artificial Intelligence Can Drive South America's Growth. Accenture.*
- Pombo, C., G. Ortega, F. Olmedo, M. Solalinde y A. Cubo. (2019) El ABC de la interoperabilidad de los servicios sociales: Marco conceptual y metodológico. Obtenido de https://publications.iadb.org/es/el-abc-de-la-interoperabilidad-de-los-servicios-sociales-marco-conceptual-y-metodologico.
- PwC (PricewaterhouseCoopers). (2017). *Sizing the Prize: What's the Real Value of AI for your Business and How Can you Capitalise?* Obtenido de https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf
- Rosenblatt, F. (1958) The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65(6), 386–408. https://doi.org/10.1037/h0042519
- Rumelhart, D., G. Hinton y R. Williams. (1986). Learning Representations by Back-propagating Errors. *Nature* 323: 533–536. doi:10.1038/323533a0
- Searle, J.R. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*. 3 (3): 417-457.
- Snow, T. (2019). *Decision-making in the Age if the Algorithm: Three Key Principles to Help Public Sector Organisations Make the Most of AI Tools. Nesta*, Londres: Nesta.
- Szegedy, C., Z. Wojciech, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow y R. Fergus. (2014). Intriguing Properties of Neuronal Networks. Cornell University. arXiv:1312.6199v4 [cs.CV]
- Turing, A. (1950). *Computing Machinery and Intelligence*. Obtenido de https://www.csee.umbc.edu/courses/471/papers/turing.pdf
- Wood, J. (2018). This AI Outperformed 20 Corporate Lawyers at Legal Work. Obtenido de https://www.weforum.org/agenda/2018/11/this-ai-outperformed-20-corporate-lawyers-at-legal-work/

05. Anexo I

MODELO Y LÍNEAS DE ACCIÓN DE FAIR LAC (gráfico A1.1)

Modelo fAIr LAC

Dimensiones estratégicas para fAIr LAC	Lineas de acción				Usuarios o beneficiarios
Red diversa Generación y difusión de conocimiento	**Grupo consultivo de personas expertas en IA** (una red de profesionales y expertos de la academia, el gobierno, la sociedad civil, la industria y el sector emprendedor)		**Mecanismo de alianzas y alineación institucional** (org. multilaterales, redes existentes)		Gobierno
	Sensibilización y comunicación (reportes, artículos de opinión y análisis, debates)		**Observatorio de casos de uso** (bien público regional para identificar casos exitosos y sistematizar buenas prácticas y aprendizajes)		Sector privado / Ciudadanía
Adopción responsable IA para servicios sociales	**Programa de apoyo para experimentos y casos de uso** (asesoría, financiamiento, red de mentores)		**Proyectos piloto** (con potencial de ser escalados en la región)		Gobierno
	Modelo de incentivos (reconocimientos, acceso a financiamiento y recursos, respaldo		**Herramientas diversas:** - Metodologías, guías, marcos de evaluación de impacto - Repositorios de modelos abiertos e índice de herramientas		Sector privado
Calidad y mitigación de riesgos	Autoregula-ción, guías o directrices	Estándares y certificaciones	Iniciativas de experimenta-ción regulatoria	Marcos normativos	Gobierno
	Fortalecer la participación de expertos y practicantes de ALC en esfuerzos internacionales (IEE, OCDE, etc) y acercar iniciativas internacionales a ALC				Sector privado

Bienestar y calidad de vida (a nivel personal)
Confianza + Impacto social * Equidad

Fuente: Elaboración propia.

06. Anexo II

Paradigmas de aprendizaje de la IA subsimbólica

La IA subsimbólica abarca distintas metodologías, entre las cuales se encuentran el aprendizaje automático que, hasta el momento, es la más utilizada (Gráfico A2.1).

INTELIGENCIA ARTIFICIAL SUBSIMBÓLICA (gráfico A2.1)

Fuente: OECD (2019)

El aprendizaje automático emplea tres paradigmas de aprendizaje: supervisado, no supervisado y por refuerzo.

- **Aprendizaje supervisado:** Se produce cuando se utiliza información donde el resultado deseado o "etiqueta" se conoce previamente. El algoritmo toma las variables relacionadas con el problema y aprende los patrones de relación entre aquellas y su resultado. El objetivo de este tipo de aprendizaje es que el modelo generalice y pueda realizar una predicción o clasificación con un determinado nivel de precisión para aquella información que no observó durante el entrenamiento. Algunos algoritmos que se utilizan en este paradigma son las redes neuronales, las máquinas de soporte vectorial y las regresiones logísticas, entre otros (Bishop, 2006).

La IA subsimbólica abarca distintas metodologías, entre las cuales se encuentran el aprendizaje automático que, hasta el momento, es la más utilizada (Gráfico A2.1).
El aprendizaje automático emplea tres paradigmas de aprendizaje: supervisado, no supervisado y por refuerzo.

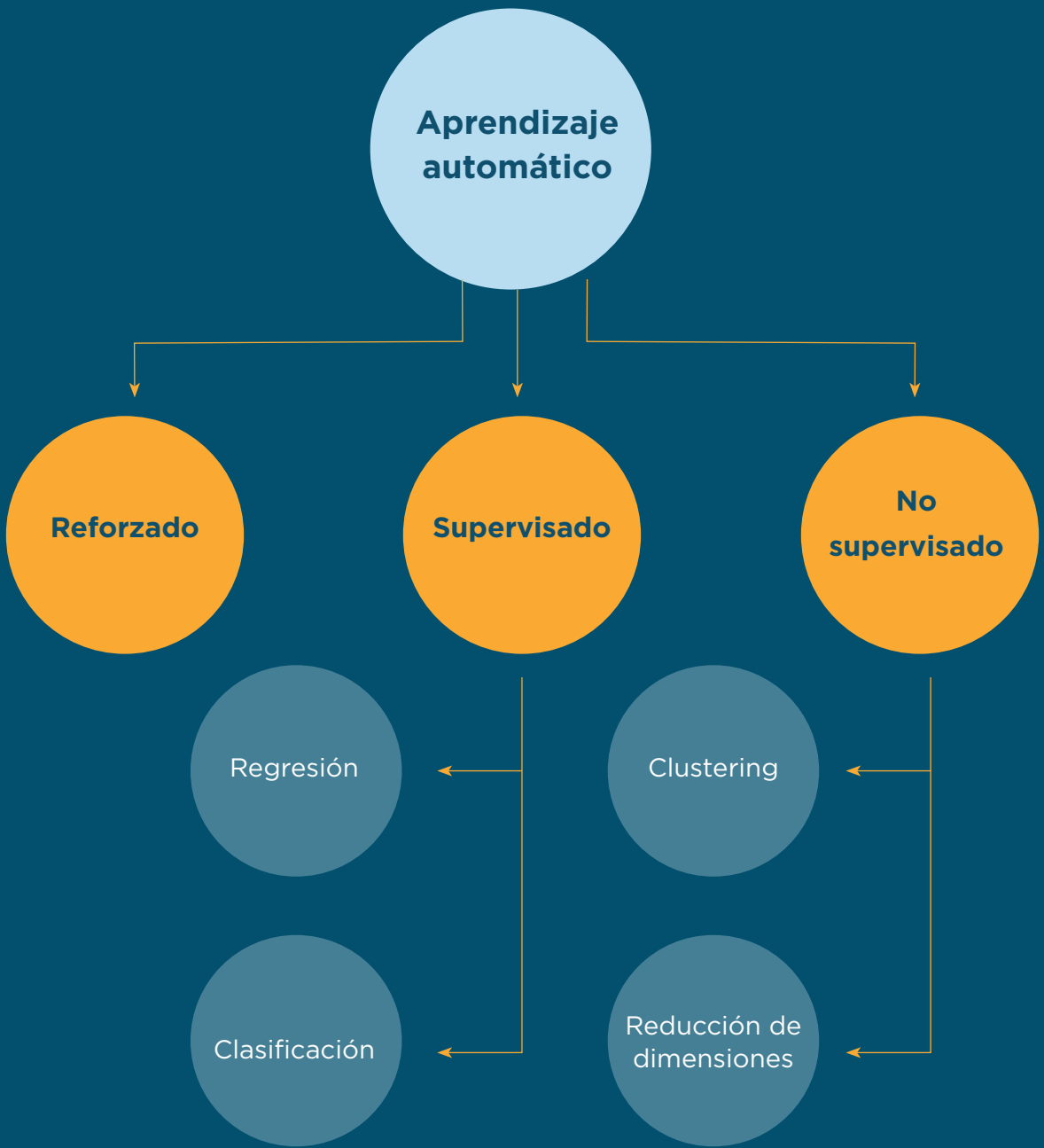
- **Aprendizaje supervisado:** Se produce cuando se utiliza información donde el resultado deseado o "etiqueta" se conoce previamente. El algoritmo toma las variables relacionadas con el problema y aprende los patrones de relación entre aquellas y su resultado. El objetivo de este tipo de aprendizaje es que el modelo generalice y pueda realizar una predicción o clasificación con un determinado nivel de precisión para aquella información que no observó durante el entrenamiento. Algunos algoritmos que se utilizan en este paradigma son las redes neuronales, las máquinas de soporte vectorial y las regresiones logísticas, entre otros (Bishop, 2006).

- **Aprendizaje no supervisado**: Se da cuando el algoritmo no observa un resultado deseado o etiqueta durante el entrenamiento. Este debe entonces encontrar patrones estructurales en la información que le permitan crear asociaciones, bien sea para descubrir grupos similares dentro de la información de entrenamiento (*clustering*) o para reducir el número de dimensiones. Algunos de los algoritmos que se utilizan en este paradigma son métodos de clusterización jerárquicos, *k-means* y modelos basados en árboles, entre otros (Bishop, 2006).

- **Aprendizaje por refuerzo:** En este caso al algoritmo no se le dan ejemplos de soluciones óptimas, sino que las tiene que descubrir mediante un proceso de ensayo y error. Esto se realiza creando un mecanismo de aprendizaje orientado por un sistema de reforzamiento. El algoritmo se convierte en un agente que debe interactuar con un entorno en el cual cada acción que realiza resulta en un premio o en un castigo. El agente, buscando maximizar sus ganancias, aprende guiado por este sistema sin intervención y/o conocimiento anterior. Dentro del paradigma de aprendizaje por refuerzo, las decisiones del agente se conocen como políticas, a saber, reglas matemáticas con las que el agente decide "explotar" [28] o explorar. QLearning y el gradiente de política determinista son dos de ellas (Bishop, 2006).

[28] La "explotación" ocurre cuando el agente repite una acción ya conocida, mientras que "exploración" se da cuando el agente experimenta con nuevas acciones.

PARADIGMAS DE APRENDIZAJE AUTOMÁTICO(gráfico A2.2)



Fuente: Elaboración propia.