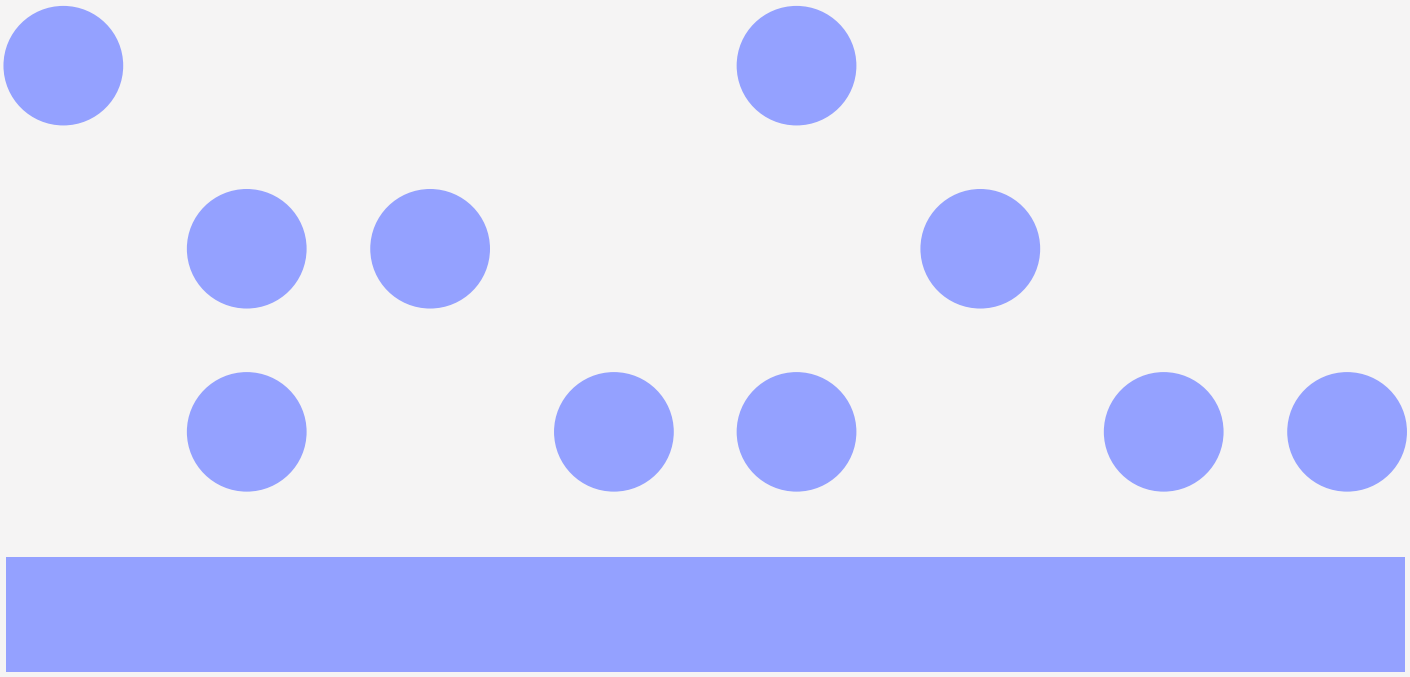
Guía práctica para el desarrollo ético de sistemas basados en IA



Datos

Marcos Feole
Juan Manuel Dias
Mariana Kunst
Zulma Carrizo
Germán Guido Lavalle

Noviembre 2023

Buenos Aires Ciudad

fundar

Guía práctica para el desarrollo ético de sistemas basados en IA

Marcos Feole
Juan Manuel Dias
Mariana Kunst
Zulma Carrizo
Germán Guido Lavalle

- Generar riqueza
- Promover el bienestar
- Transformar el Estado

Índice

Guía práctica para el desarrollo ético de sistemas basados en IA

fundar

Introducción

A la par de los progresos y de la masiva adopción de las tecnologías de la información y comunicación (TIC), la era actual está marcada por un rápido avance en el campo de la inteligencia artificial (IA). Su influencia en diversos sectores gubernamentales, en particular, y de la sociedad, en general, es innegable. Los sistemas basados en IA se usan en diversos ámbitos y afectan desde la gestión de justicia hasta las preferencias de entretenimiento, es decir, son transversales a toda la sociedad.

Estas tecnologías proveen aplicaciones beneficiosas en el ámbito de la administración pública, contribuyen a la eficiencia de los procesos gubernamentales y generan una amplia gama de servicios para la sociedad. Por lo tanto, es fundamental que los gobiernos diseñen estrategias que estimulen su implementación y uso y que, al mismo tiempo, incorporen las normas éticas, necesarias para minimizar los riesgos. En este contexto, es preciso que las áreas de gobierno adopten estrategias de regulación y evaluación para guiar el desarrollo y el uso ético de estas herramientas. Así podrán proteger a la ciudadanía de los posibles riesgos asociados a la implementación de estas tecnologías.

En la actualidad, el sector público se encuentra cada vez más inmerso en el empleo de este tipo de tecnologías para llevar a cabo una amplia variedad de tareas y ofrecer servicios esenciales. La mayoría de estos sistemas se basan en datos y se utilizan para proporcionar información crucial para la ciudadanía, optimizar la asignación de recursos, agilizar los procedimientos gubernamentales y respaldar a personas que se desempeñan como funcionarias en la toma de decisiones.

La creación de sistemas tecnológicamente eficientes, éticamente responsables, libres de sesgos, socialmente adecuados y legalmente conformes requiere la colaboración de profesionales y de disciplinas con diversos enfoques. El conocimiento tecnológico se enriquece y se complementa con otros saberes —como el jurídico, el social, el antropológico y muchas otras áreas—. Esta combinación de conocimientos es esencial para garantizar que los sistemas desarrollados sean equilibrados y satisfagan las necesidades de toda la sociedad.

Para abordar estas necesidades, Fundar y la **Subsecretaría de Políticas Públicas Basadas en Evidencia** (SSPPBE) del Gobierno de la Ciudad Autónoma de Buenos Aires trabajaron de manera conjunta en la creación de esta guía de ética algorítmica. Si bien su enfoque principal es la orientación en el desarrollo de algoritmos en el ámbito gubernamental, también puede servir de referencia para el sector privado.

Objetivos y público

Según experiencias recolectadas de distintas fuentes, hoy, la ética en IA se resuelve, en general, como un proceso *ad hoc* y *a posteriori*. En otras palabras, los problemas se detectan de manera tardía (cuando ya ocurrieron), las acciones necesarias para arreglar estos problemas se realizan en forma de parches y una vez que parte del daño ya está hecho.

Frente a ese escenario, el principal objetivo es disponibilizar herramientas de diagnóstico para un desarrollo e implementación confiables de los sistemas basados en IA, y que estos minimicen los riesgos asociados a su uso. En particular, esta guía busca cubrir una carencia en el área de la ética en IA, donde —si bien abundan las guías de principios éticos de desarrollo— la gran mayoría no contempla los conocimientos, los tiempos, ni los flujos de trabajo existentes de quienes desarrollan estas tecnologías; es decir, no tienen utilidad práctica.

Esta guía está especialmente dirigida a equipos de diseño, desarrollo y evaluación de servicios basados

en IA. También sirve para hacer un seguimiento de proyectos gubernamentales que contengan un componente de uso de esta tecnología (tanto de uso interno del gobierno como de cara a la sociedad), con el objetivo de que distintas áreas puedan minimizar los riesgos y el potencial impacto negativo de estas tecnologías.

Complementariamente, aporta un marco de comprensión respecto de la ética en IA a los equipos implicados en dichos procesos, e incluso a las personas usuarias o que, de alguna manera, se vean afectadas por estas tecnologías. Así todas las personas implicadas en el ciclo de vida de estos sistemas podrán encontrar herramientas para un desarrollo seguro y robusto y, a la vez, una evaluación efectiva de ellos.

Modo de uso

El contenido se ha estructurado para que sea una herramienta práctica que se adapte a los ciclos de desarrollo que actualmente presentan las áreas que trabajan con tecnologías de IA. Sin embargo, el modo de utilizar esta guía depende de los objetivos de quien desarrolle, evalúe o utilice estas tecnologías. En primer lugar, en la sección **Principios éticos de IA** se ofrece una serie de requerimientos y principios que son estándares éticos de IA y que son necesarios a la hora de pensar, diseñar, desarrollar y evaluar los sistemas basados en IA. Asimismo, esta sección es útil para acceder a un panorama sobre los puntos éticos claves a tener en cuenta durante la implementación, desarrollo o evaluación de estas tecnologías.

A continuación, en caso de querer implementar o llevar a la práctica el desarrollo de un sistema basado en IA que tenga en cuenta estos estándares éticos, la sección **Guía de desarrollo ético de sistemas basados en IA** consta de la guía propiamente dicha (que se nutre de los principios éticos descriptos en la sección anterior). El modo de uso de esta sección consiste en que el equipo de trabajo complete cada punto de la guía a lo largo del proceso de desarrollo de un proyecto de IA, de forma que este proceso abarque de manera robusta todos los aspectos de la ética en IA considerados, desde el comienzo del diseño del sistema hasta su implementación en producción.

Por último, la sección **Casos de uso** muestra dos ejemplos de aplicación de la guía a la implementación de dos sistemas basados en IA de la Subsecretaría de Políticas Públicas Basadas en Evidencia (SSPPBE) del Gobierno de la Ciudad Autónoma de Buenos Aires (GCBA). La guía se aplicó a dos proyectos basados en IA. Por un lado, el caso Ecopuntos que tiene como objetivo segmentar automáticamente los diferentes tipos de residuos reciclables, basándose en fotografías de ellos. Por otro, el caso Recomendador de capacitaciones, que tiene como propósito acercar a la ciudadanía oportunidades de capacitaciones o cursos que se adecúen a su perfil o a sus intereses específicos. De la aplicación de la guía ética de IA, también se desprende una serie de recomendaciones para mejorar la implementación de cada una de las iniciativas de uso de sistemas basados en IA en el gobierno.

Principios éticos de IA

La ética en IA se refiere a los principios y valores que deben guiar el desarrollo, despliegue y uso de sistemas de IA, con el fin de minimizar los riesgos potencialmente negativos que pueda tener el uso de estas tecnologías. Para un desarrollo e implementación ética de sistemas basados en IA, a continuación se presentan ciertos principios que proveen los conceptos mínimos de la ética en IA y que serán de utilidad a lo largo de la Guía práctica para el desarrollo ético de sistemas basados en IA, a medida que sean requeridos en las distintas etapas de desarrollo de cada sistema basado en IA.

Principios éticos de IA

Propósito

El primer principio ético se refiere al "propósito" con el cual se crea un sistema basado en IA. Aquí se parte entonces de la presunción de que la IA no se crea con fines maléficos o dañinos. Además se asume que, en términos generales, los beneficios proporcionados por la aplicación de la IA deben superar a sus potenciales riesgos negativos. Los sistemas deben diseñarse y utilizarse con un propósito claro y beneficioso para la sociedad y las personas. Esto implica que se desarrollen con objetivos éticos, que su implementación y uso no causen daño innecesario o injusto, y que se alineen con los valores y normas éticas de la sociedad.

Los algoritmos de IA pueden clasificarse, según su propósito, en dos grupos principales; los de propósito particular y los de propósito general. Entre los primeros, se encuentran aquellos algoritmos que clasifican imágenes en distintas categorías, que detectan elementos específicos en imágenes, que recomiendan películas basadas en los gustos de cada persona usuaria, etc. Estos tienen un fin específico, que queda codificado en la función objetivo a ser optimizada por el algoritmo de IA (donde, por ejemplo, buscan minimizar el error en sus predicciones). Estos algoritmos de IA no pueden usarse con fines diferentes que para los que fueron programados y, por lo tanto, tienen un bajo riesgo ético desde el punto de vista del propósito con el que fueron creados.

El segundo grupo —los algoritmos de IA de propósito general— son, como su nombre lo indica, los que están diseñados para resolver una amplia variedad de problemas en diferentes campos de aplicación. Es decir, ellos son independientes del dominio específico y, a la vez, son versátiles y adaptables a distintas situaciones. Por ejemplo, un algoritmo de propósito general puede imitar a un ser humano en una conversación de chat, de manera tal que resulte muy difícil —o casi imposible— distinguir si las respuestas del chat corresponden a un algoritmo o a una persona real. Los algoritmos de IA generativos son, en general, de propósito general: crean imágenes, voces, videos o textos a partir de comandos de entrada, pero tienen un gran e indefinido grado de libertad al generar estos elementos. Por lo tanto, tienen una mayor probabilidad de riesgo —asociado a la posibilidad de que sean usados de maneras creativas para fines dañinos—, ya que las formas en que pueden usarse no pueden, en general, ser previstas durante su diseño y desarrollo.

Colaboración con personas expertas en la materia de aplicación

Todo sistema basado en IA tiene desafíos específicos sobre su área de aplicación. Por eso es necesario contemplar el conocimiento experto de cada área de aplicación específica de estos sistemas; por ejemplo, para garantizar que funcionan de manera acorde al propósito con el que fueron creados. Así, si se desea desarrollar un algoritmo que detecta elementos o variaciones del suelo sobre imágenes satelitales, va a ocurrir que —por la alta variabilidad de los ambientes naturales, la dinámica de climas, las estaciones o los cambios en la vegetación— se requerirá del conocimiento de perfiles expertos en geografía o en las dinámicas terrestres para diseñar y desarrollar estos algoritmos. En este sentido, para asegurar que cada algoritmo de IA cumpla su propósito y funcione de manera correcta, será necesario establecer una colaboración dinámica entre aquellas personas que se desempeñan como científicas de datos y aquellas que son expertas en la materia de aplicación.

Seguridad

El principio de seguridad para los sistemas basados en IA debe integrarse a su desarrollo desde su diseño, especialmente para aplicaciones sensitivas. El principio de seguridad desde el diseño (SdB, por su nombre en inglés, Security by Design) implica evaluar los potenciales riesgos que los sistemas puedan presentar e implementar las prevenciones o soluciones desde el comienzo de su desarrollo (desde la etapa de diseño). Así podrá minimizarse o eliminarse la necesidad de reacondicionar o de establecer parches de seguridad en etapas posteriores de su desarrollo.

Principios éticos de IA

Vale la pena aclarar que, de ninguna manera, puede asegurarse que los sistemas basados en IA (al igual que los sistemas informáticos, en general) sean 100% seguros. Ellos pueden presentar fallas no previstas o vías en que pueden ser hackeados o vulnerados de maneras creativas.

También existe una serie de problemas técnicos de seguridad, denominados "accidentes de IA". Pueden arrastrarse desde el diseño o el entrenamiento de un algoritmo de IA y generar respuestas erróneas o dar lugar a usos fraudulentos. Estos son, y pueden ser, de varios tipos (a continuación, se mencionan tres ejemplos). Sin embargo, aquí no se detallan todas las posibles soluciones a estos problemas, ya que cada uno de estos accidentes requiere un estudio pormenorizado de cómo evitarlos o revertirlos.

Box 1

Efectos secundarios negativos

Incluso cuando se declara una precisa función objetivo que el algoritmo de IA debe optimizar, pueden ocurrir efectos secundarios —acciones negativas no previstas que le permiten al algoritmo llegar igualmente a cumplir su objetivo—. Por ejemplo, un robot programado para mover cajas de un lado a otro podría tener que atravesar materiales frágiles en su camino, y pasarlos por encima o destruirlos para cumplir con su objetivo final.

Una posible solución a este problema implica considerar los mecanismos de seguridad necesarios para que estos efectos secundarios no ocurran y programar explícitamente al algoritmo para que tome las precauciones necesarias de manera automática.

Box 2

Hackeo de recompensas

El algoritmo puede optimizar su función objetivo, pero tomando atajos o haciendo trampa, de tal manera que no termina de cumplir con el propósito previsto. Un ejemplo de esto puede ser un robot programado para limpiar la suciedad que observa en el suelo, pero, en vez de limpiarla, decidiera apagar sus sensores de visión. Así, no podrá observar la suciedad y entonces considerará, en todo momento, su objetivo cumplido. Esta es una forma donde un algoritmo encuentra una solución no prevista por el humano que lo programa al problema que tiene que resolver.

Entre las posibles soluciones a este problema, por ejemplo, puede determinarse específicamente la función objetivo, no dando lugar a tantas ambigüedades. Sin embargo, por su naturaleza de funcionamiento, siempre se deberá esperar que los algoritmos de IA lleguen a sus objetivos de maneras creativas o sorpresivas, incluso para las mismas personas que los programan.

Box 3

Robustez al cambio de distribución

Durante su aplicación, el algoritmo de IA puede encontrarse con un contexto distinto al que fue programado y no tener manera de detectar este cambio de contexto, en general. Entonces, puede asumir que está realizando un buen trabajo, aun cuando esté llevando a cabo tareas que no tienen sentido en ese nuevo contexto. Por ejemplo, si una IA está programada para detectar tumores en manchas de la piel, pero fue entrenada únicamente con personas de tez clara, cuando esa IA sea utilizada en personas de tez oscura es probable que detecte tumores donde no los haya (o que no los detecte cuando los haya) y, al mismo tiempo, asegure (erróneamente) una "alta confianza" en su solución. Este problema se debe puramente a que esa IA no fue entrenada para el contexto en el que es usada.

Nuevamente, es difícil obtener una solución general a este problema. Por ello es necesario prever de antemano los posibles contextos en que esa IA va a utilizarse.

Principios éticos de IA

Trazabilidad

La trazabilidad se refiere a la determinación de los procesos que forman parte de todo el ciclo de desarrollo de los sistemas basados en IA, y su debida documentación. Este principio implica entonces mantener un registro de las acciones y de los datos usados durante todo el proceso de desarrollo. El punto de partida, el propósito del sistema, la arquitectura del algoritmo y el diseño deben estar debidamente documentados. La recolección de datos debe ser trazable, su fuente debe estar documentada y los datos recolectados deben estar almacenados y disponibles. Si los datos se toman mediante un programa automático, este programa debe formar parte de la documentación y su uso debe replicarse independientemente. Luego, el preprocesamiento realizado a los datos —desde su anonimización hasta los procesos necesarios para transformar los datos crudos en datos de entrada para la IA— deben documentarse y replicarse independientemente. Asimismo, el entrenamiento de los algoritmos de IA con esos datos debe ser replicable y los hiper-parámetros probados (los parámetros determinados manualmente por el programador) deben estar documentados. Si bien el entrenamiento de un algoritmo de IA tiene, en general, componentes o parámetros aleatorios que hacen que este no pueda replicarse independientemente de manera idéntica a la original, todas las pruebas realizadas y los resultados obtenidos en cada caso deben estar debidamente documentados. Por último, el proceso de testeo de los sistemas de IA debe ser replicable; es decir, todas las pruebas posteriores que se realizan sobre el sistema y los datos con los que ellas se llevan a cabo deben documentarse, almacenarse y estar disponibles.

En definitiva, la trazabilidad en el desarrollo de los sistemas basados en IA implica documentar todos los procesos. Desde el primero hasta el último: propósito, diseño, arquitectura, datos, preprocesamiento, entrenamiento, pruebas y resultados. Así podrá haber comprensión y replicación completa del desarrollo.

El principio de trazabilidad permite que un proyecto de IA sea tomado por un equipo de desarrollo ajeno al proyecto y que puedan volver a crear el mismo algoritmo en las mismas condiciones con que fue creado. Esto permite, entre otros beneficios, identificar de manera precisa los motivos detrás de cualquier problema que surja posteriormente con ese sistema y corregirlos de manera efectiva. La trazabilidad es un aspecto clave de la transparencia detrás de estas tecnologías, ya que no permite que existan partes de los sistemas cuyo desarrollo sea desconocido o incierto tanto por otros equipos de programadores como por quienes regulan estas tecnologías, además de las personas usuarias o personas objetivo en general.

Protección de datos y privacidad

La protección de datos personales se refiere a la práctica de salvar la información que una persona posee y comparte con una empresa u organización. Esta información puede incluir nombres, direcciones, números de teléfono o cualquier otro dato que pueda identificar a una persona. La protección de datos personales implica garantizar que dicha información sea recolectada, almacenada, usada y compartida de forma segura y responsable. Esto implica que no sea utilizada para fines distintos a los informados a la persona titular de los datos, ni compartida con terceros sin su consentimiento. En este sentido, los principios de protección y privacidad de datos personales requieren que las empresas y organizaciones cumplan normas y prácticas para garantizar la seguridad y privacidad de los datos personales que recolectan, almacenan, usan y comparten.

Asimismo, estos principios involucran a otros. Entre ellos, el consentimiento informado (por el cual las empresas deben obtener el consentimiento informado de quien es titular de los datos para recolectar, almacenar, usar y compartir sus datos personales); la limitación de la finalidad (por el que las empresas sólo deben recolectar y usar los datos personales para fines específicos y legítimos y no deben utilizarlos para otros fines sin el consentimiento de la persona titular); la minimización de datos

Principios éticos de IA

(que exige que las empresas sólo deben recolectar los datos personales necesarios para cumplir con el propósito especificado y no deben recolectar más datos de los necesarios); la exactitud (mediante el cual las empresas deben tomar medidas razonables para garantizar que los datos personales sean precisos y estén actualizados); el acceso y rectificación (que implica el derecho de titulares de los datos a acceder a sus datos personales y solicitar su corrección o eliminación si son inexactos o innecesarios).

En cuanto a la normativa, en nuestro país rige la Ley 25.326 de Protección de Datos Personales del 2000. El órgano de control establecido por dicha ley, la Agencia de Acceso a la Información Pública, está —a la fecha de elaboración de esta guía— inmersa en un proceso de modificación de esa ley. Complementariamente, el país cuenta con normativa nacional más específica y dispersa en el cuerpo jurídico. Entre ellas destacan normas como la Reglamentación de la Ley de Protección de Datos Personales (decreto 1558/2001), que establece las pautas para la aplicación de la Ley de Protección de Datos Personales; la Ley 26.951 de Protección de Datos Personales en el Ámbito Telefónico, que regula la protección de datos personales en el ámbito telefónico; entre otras.

Como sabemos, los sistemas basados en IA necesitan una gran cantidad de datos para ser entrenados. Además, en algunos de estos sistemas, los datos de entrada quedan codificados de manera implícita dentro de los mismos algoritmos —por lo que, sobre todo para los algoritmos de IA generativos, es probable que puedan recuperarse o extraerse una parte de estos datos durante su uso—. Este principio se centra entonces en garantizar que tales sistemas respeten y salvaguarden la privacidad y la información de las personas u organizaciones contenidas en estos datos. Para ello, deben implementarse medidas adecuadas para proteger la información personal de accesos no autorizados, así como para garantizar la exactitud y la integridad de los datos; y que las personas estén informadas sobre qué datos se recopilan, cómo se utilizan y con quién se comparten, y que puedan tener el control y el consentimiento sobre el uso de sus datos en estos sistemas.

La privacidad y la protección de datos personales y confidenciales son aspectos esenciales para generar confianza en la IA. Al respetar la privacidad y proteger los datos se evita el riesgo de discriminación, abuso o violaciones de los derechos fundamentales de las personas. Entre las diversas formas de garantizar este principio, se sugiere limitar los datos de entrada con los que se alimentan los algoritmos de IA, eliminar datos confidenciales, personales o privados, y realizar una anonimización exhaustiva de ellos.

Box 4

Datos personales sensibles

Los datos personales sensibles son aquellos que revelan información especialmente delicada sobre una persona como su origen racial o étnico, opiniones políticas, creencias religiosas, afiliación sindical, orientación sexual, datos biométricos, datos de salud, etc. Estos datos requieren una protección adicional debido a su naturaleza sensible y el potencial riesgo de discriminación o violaciones a la intimidad que podrían derivar de su tratamiento inadecuado.

La protección legal de los datos personales sensibles puede variar según la legislación de cada país, pero también existen principios y normativas que son comunes. Por ejemplo, en la Unión Europea existe el Reglamento General de Protección de Datos (GDPR) que establece que el tratamiento de datos personales sensibles requiere un consentimiento explícito y específico del titular de los datos, a menos que exista una base legal específica para su procesamiento.

En Argentina los datos personales se encuentran especialmente protegidos. Su tratamiento puede ser realizado sólo cuando medien razones de interés general autorizadas por ley o cuando sean utilizados con finalidades estadísticas o científicas, de tal manera que no puedan ser identificados sus titulares. Asimismo se establece que los establecimientos sanitarios —públicos o privados— y las y los profesionales vinculados a las ciencias de la salud pueden recolectar y tratar los datos personales relativos a la salud física o mental de las y los pacientes que acudan a los mismos o que estén o hubieren estado bajo tratamiento de aquellas/os, respetando los principios del secreto profesional.

Principios éticos de IA

Sesgos y justicia algorítmica

Los sesgos se refieren a una incorrecta o injusta representación de una población o fenómeno por parte de los datos, por ejemplo, a través de una recolección parcial o incorrecta de los mismos, o también pueden deberse a sesgos ya existentes en la sociedad (a través de tratamientos injustos o tendenciosos de distintos grupos de personas sobre la base de ciertas características de los mismos). Además, pueden existir sesgos en las respuestas de los algoritmos de IA, incluso sin que estén presentes en los datos utilizados para entrenar a estos algoritmos.

Los algoritmos de IA son entrenados con datos. No sólo aprenden los sesgos presentes en ellos, sino que pueden amplificarlos. Por ejemplo, imaginemos que las personas trabajadoras de un banco tienen un sesgo de género a la hora de determinar si se le otorga o no un crédito a una persona. Entonces, si se entrena una IA con los datos de créditos otorgados por este banco, este sesgo se vería reflejado, y quizás amplificado, por el algoritmo. De esta manera, podría observarse una disparidad de género aún mayor en el otorgamiento de créditos por parte de la entidad que utilice este algoritmo que la que se observaba en el banco original.

Este tipo de problemas pueden traer aparejadas aplicaciones de IA que resulten en discriminación y en trato desigual, o que den lugar a un aumento de las brechas y la exclusión de ciertos grupos sociales. Es decir, pueden contribuir al aumento de la injusticia en general. Las iniciativas orientadas a solucionar estos problemas se denominan "justicia algorítmica".

En este escenario, una cuestión de gran relevancia es si los algoritmos pueden generar respuestas sesgadas a partir de datos que, en principio, no presentan sesgos. La respuesta (si bien puede depender del tipo de algoritmo utilizado) es afirmativa. En este sentido, es crucial no sólo investigar y comprender los sesgos presentes en los datos, sino también abordar y corregir los posibles sesgos inherentes a los algoritmos. Estos sesgos pueden manifestarse incluso cuando los datos de entrada no contienen sesgos evidentes. Por lo tanto, es fundamental abordar y corregir los sesgos en cada fase del desarrollo de un algoritmo de IA. Desde la recopilación inicial de datos, el prototipado del algoritmo y, finalmente, en el algoritmo en su estado de producción.

Volviendo al ejemplo sobre el sesgo en la asignación de créditos, es importante reconocer que puede existir un sesgo de género en el algoritmo final, incluso si la variable de género no se utiliza en el entrenamiento del algoritmo. En estos casos, es necesario incorporar la variable de género en el proceso de entrenamiento del algoritmo. Ello permitirá que, a través de su interacción con otras variables, pueda corregirse de manera manual el sesgo presente en el algoritmo final. De ahí, la importancia de identificar y abordar los sesgos en cada fase del desarrollo de los algoritmos de IA.

Otros ejemplos de sesgos respecto a la utilización de algoritmos de IA generativos pueden ser, por ejemplo, en aquellos algoritmos que generan imágenes a partir de texto. Si se le pide a muchos de estos algoritmos que generen imágenes de doctores, abogados o jueces, raramente generarán imágenes de mujeres. Si se le pide que generen imágenes de personas cometiendo crímenes, generarán mayormente imágenes de varones de tez oscura; si se le pide generar imágenes de trabajadores de cocina en cadenas de comida rápida, generarán mayormente imágenes de mujeres de tez oscura.

Entonces, para promover la justicia algorítmica y revertir la presencia de sesgos en los algoritmos de IA es fundamental adoptar medidas que incluyan una rigurosa evaluación de los conjuntos de datos utilizados en el entrenamiento de los algoritmos. Así podrá garantizarse el acceso a datos de calidad, con un volumen y una variedad que los haga equitativamente representativos de diferentes grupos de personas, industrias o segmentos. Además, una vez entrenado el algoritmo, debe probarse y testearse para que no contenga sesgos en ninguna de las dimensiones relevantes (Box 5: Detección de sesgos en la práctica). Sin embargo, la inexistencia de sesgos nunca se podrá asegurar por completo. Por ello debe llevarse a cabo la mayor cantidad de pruebas para lograr los resultados más justos e inclusivos posibles.

Principios éticos de IA

Es importante tener en cuenta que la justicia no sólo se refiere a la eliminación de sesgos y discriminaciones existentes, sino también a la promoción de la equidad y la igualdad de oportunidades. Los algoritmos de IA deben diseñarse y utilizarse para contribuir a un trato justo y equitativo para todas las personas, sin importar su origen, género, raza, orientación sexual u otras características protegidas.

Box 5

Detección de sesgos en la práctica

Para determinar si existen sesgos, se debe tener en cuenta los resultados específicos de aplicar los algoritmos de IA sobre cada uno de los diferentes grupos de población objetivo. Los posibles sesgos que pueden considerarse son según:

- Grupo etario
- Grupo social
- Lugar geográfico de vivienda
- Nivel de educación
- Nacionalidad
- Sistema cultural
- Etnia
- Grupo lingüístico
- Género
- Infancia menor de edad
- Discapacidad
- Grupo desfavorecido, marginado o en situación de vulnerabilidad
- Toda otra característica específica del contexto de aplicación de cada algoritmo

Explicabilidad y transparencia

Una característica importante de la mayoría de los algoritmos de IA es que son tan complejos y flexibles que muchas veces no podemos entender ni explicar —*a priori*— cómo codifican la información necesaria para resolver los problemas planteados. Tampoco podemos saber exactamente por qué obtuvieron resultados exitosos (o no). Estos algoritmos se denominan de "caja negra" y, aunque obtengan soluciones exitosas, pueden no existir formas intuitivas de interpretarlos ni de dar explicaciones razonables de cómo llegan a sus resultados. Para estos algoritmos de IA, la transparencia tiene límites duros, porque sus resultados, predicciones o decisiones no son necesariamente explicables ni siquiera por las personas que los diseñan o programan.

En este sentido, es importante asegurar la mayor transparencia posible en el diseño y el uso de los algoritmos de IA. Como primera medida, se debe ser transparente con el propósito de la creación de estos sistemas; es decir, si estos serán usados para realizar predicciones (y cuáles), o si serán utilizados para tomar decisiones automáticamente; cómo es que esas decisiones serán tomadas y quién será responsable en caso de existir conflictos. En particular, si el algoritmo de IA es de propósito general, debe informarse que puede usarse de maneras novedosas y creativas, no necesariamente previstas por quien, en primer lugar, diseña y programa. Además, es necesario que estos puedan reportar con qué variables de los datos disponibles se alimenta el entrenamiento del algoritmo y cuál es la fuente y alcance de los datos utilizados para su creación.

Autonomía y responsabilidad

Una característica común a muchos sistemas que utilizan IA es su capacidad para tomar decisiones de manera flexible en contextos complejos, siempre y cuando estén programados con ese fin. A medida que estos sistemas son usados cada vez más para la toma de decisiones, pueden surgir dudas sobre cuánta autonomía debe delegarse sobre ellos y quiénes son responsables de sus acciones, en particular si se utilizan en contextos sensibles para la ciudadanía.

Por ejemplo, un algoritmo de IA que conduce un vehículo de manera autónoma puede encontrarse en una situación crítica donde tiene que decidir si priorizar el bienestar o de peatones o de pasajeros del vehículo. Este tipo de dilemas no tiene respuestas definitivas y posee la dificultad adicional de que los sistemas basados en IA pueden tomar decisiones mucho más rápido que lo que pueden reaccionar los seres humanos para corregirlas.

En este sentido, este principio se refiere a la posibilidad de que los sistemas de IA actúen de manera autónoma y a la adjudicación de responsabilidades por dichas decisiones. Esto incluye la responsabilidad por los posibles daños causados por decisiones erróneas, sesgos, discriminación o violaciones de derechos fundamentales. Por ello es fundamental que haya mecanismos de rendición de cuentas y sistemas de responsabilidad legal claros para garantizar que, en caso de problemas, se tomen las medidas adecuadas. Si un algoritmo de IA toma una decisión que genera daños, el problema es que la organización que la desplegó fue descuidada o indiferente durante su diseño y prueba, además de que la tecnología no se reguló fehacientemente.

A medida que se le da a los algoritmos más poder y autonomía, cada vez es más importante evaluarlos, regularlos y asegurar que tengan el suficiente control humano. Es importante además entender cómo garantizar que las personas afectadas por estos sistemas sean amparadas, si los sistemas automatizados toman decisiones que los perjudican. La ética detrás de estas tecnologías implica la necesidad de que las personas y organizaciones que las crean sean responsables de responder por ellas.

Guía práctica para el desarrollo ético de sistemas basados en IA

El desarrollo de sistemas basados en IA puede organizarse, de forma general, en las siguientes cinco etapas. Para cada una son relevantes distintas partes de la guía y están indicadas en cada caso:

Paso 1: Diseño. Es un bosquejo del propósito del algoritmo de IA, sus funciones, estructura y características generales. Ver sección Diseño del algoritmo de IA.

Paso 2: Prototipo. Implica el desarrollo de un prototipo inicial, para demostrar el correcto funcionamiento del algoritmo de IA. Requiere una recolección inicial de datos y su debido preprocesamiento y anonimización. Ver desde sección Recolección de datos e integración de fuentes a Testeo del sistema de IA y pruebas de justicia algorítmica inclusive.

Paso 3: Algoritmo final. Abarca el desarrollo del algoritmo con su arquitectura y datos finales, tal como se espera que sea implementado en producción. Esta etapa precisa una recolección y preprocesamiento exhaustivo de los datos de entrenamiento, así como el testeo final del rendimiento del algoritmo. Ver desde sección Recolección de datos e integración de fuentes a Testeo del sistema de IA y pruebas de justicia algorítmica inclusive.

Paso 4: Implementación en producción. A partir de la salida del sistema a su despliegue y uso en el mundo real. Ver sección Implementación del sistema de IA en producción.

Paso 5: Desarrollo continuo. Un algoritmo de IA nunca está finalizado; su desarrollo y evolución es continuo. Los algoritmos “envejecen” (pierden rendimiento) o pueden dejar de ser relevantes en nuevos contextos. Ver sección Desarrollo continuo del sistema de IA.

Guía práctica para el desarrollo ético de sistemas basados en IA

Entidad:

Proyecto:

Diseño del algoritmo de IA

Propósito. Definir el propósito del algoritmo de IA . Dar un diagnóstico del contexto en el que será aplicado. Definir posibles usos y aplicaciones del algoritmo y el contexto de su aplicación (por ejemplo: población objetivo, qué servicio ofrece, etc.) (ver principio ético de Propósito).

Especialistas. ¿Se consultó con especialistas en áreas de aplicación del algoritmo de IA? ¿Cuáles fueron sus recomendaciones y/o advertencias? (ver principio ético de Colaboración con personas expertas en la materia de aplicación).

Riesgos. ¿Se puede prever algún impacto negativo de la herramienta? ¿Cuáles? ¿Qué riesgos pueden existir por la utilización del algoritmo? Definir si alguno de estos usos tiene un potencial dañino o su uso puede ser sensible o vulnerar derechos humanos (ver principio ético de Seguridad).

Recolección de datos e integración de fuentes

Fuente. ¿Cuál es la fuente de los datos? ¿Cómo se obtienen? ¿Hay consentimiento de la fuente, las personas o las organizaciones? En caso de utilizar código de programación para obtenerlos, documentar el código (ver principio ético de Trazabilidad).

Datasets. Describir, de manera general, los *datasets* que se utilizan para entrenar el algoritmo. ¿Cuál es su tamaño? ¿Qué variables contiene? Hacer una descripción general de cada variable y tipo de dato que contiene. ¿Hay variables que contengan información personal o confidencial? (ver principio ético de Protección de datos y privacidad).

Calidad. ¿Cuál es la calidad de los datos? ¿Hay datos faltantes? ¿Contienen errores o inconsistencias? (por ejemplo, por carga manual). ¿Hay algún procedimiento para detectar esto?

Integración. ¿Los datos cubren los casos de uso intencionados? ¿Se integran datos de distintas fuentes? ¿Sobran datos? ¿Por qué?

Actualización. ¿Se actualizan los datos? ¿Con qué frecuencia? Si no se actualizan, ¿son datos fijos que no pueden cambiar?

Almacenamiento. ¿Se almacenan los datos? ¿De manera segura? ¿Hay consentimiento y aviso en caso de ser datos personales o privados? ¿Pueden las personas usuarias acceder a los datos personales almacenados sobre ellos mismos? ¿Puede una persona usuaria decidir que sus datos personales sean borrados de la base de datos?

Sesgos I. Realizar un análisis preliminar de sesgos y problemas de representatividad en los datos. ¿Se hacen suposiciones sobre los datos? ¿Qué criterios de sesgos y representatividad se probaron? Definir valores aceptables de sesgos (ver principio ético de Sesgos y justicia algorítmica).

Preprocesamiento y etiquetado de datos

Variables. ¿Es necesario almacenar todas las variables que actualmente se almacenan? ¿Son todas necesarias para el funcionamiento de los algoritmos?

Anonimización. ¿Se anonimizan los datos? ¿Cómo? ¿Se pueden volver a identificar posteriormente a los individuos? Por ejemplo, cruzando distintas fuentes de datos externas que no sean parte del presente desarrollo (ver principio ético de Protección de datos y privacidad).

Etiquetado. ¿Cómo se realiza el etiquetado de los datos? ¿El etiquetado está alineado con el propósito del algoritmo de IA? Si el etiquetado proviene de fuentes diversas, ¿son coherentes las distintas fuentes?

Definición del algoritmo, función objetivo y entrenamiento

Explicabilidad. Definición de los algoritmos de optimización que van a ser probados para su entrenamiento. ¿Son explicables o de caja negra? (ver principio ético de Explicabilidad y transparencia).

Objetivo. Definición de la función objetivo. ¿Ella se alinea totalmente con el propósito del sistema, o es aproximada? Considerar los problemas de seguridad que esta función objetivo podría tener (ver principio ético de Seguridad).

Rendimiento I. Definición de métricas de rendimiento del algoritmo de IA (por ejemplo, precisión, exhaustividad, etc.). ¿Están alineadas a los objetivos del algoritmo?

Trazabilidad. ¿Por qué se seleccionó el algoritmo que finalmente quedó? ¿Qué hiperparámetros se probaron? Documentar todos los resultados (ver principio ético de Trazabilidad).

Sistema externo. ¿Se utilizó algún sistema o servicio externo para la creación o el entrenamiento del algoritmo de IA? ¿Cuáles? ¿El proveedor de servicios cuenta con estándares de seguridad y privacidad alineados a los del equipo desarrollador?

Testeo del sistema de IA y pruebas de justicia algorítmica

Rendimiento II. ¿Cuál es el resultado obtenido de las métricas de rendimiento finales del algoritmo? ¿En qué datos se obtuvieron estos resultados?

Sesgos II. Realizar un análisis de sesgos en el algoritmo finalizado. ¿Qué criterios de sesgos se probaron? Definir valores aceptables de sesgos (ver principio ético de Sesgos y justicia algorítmica).

Evaluación I. Evaluación realizada por personas usuarias objetivo o partes interesadas, y también por expertos del dominio de aplicación del algoritmo. Crear una base de datos centralizada de reportes de incidentes. Evaluar cómo resolver estos incidentes.

Diagnóstico manual.

- Probar casos de uso previstos. ¿La IA funciona bien donde se supone que debería funcionar bien?
- Probar casos de borde (donde su funcionamiento se supone que es ambiguo).
- Probar posibles fallas (casos en lo que la IA puede presentar fallas).

Implementación del sistema de IA en producción

Discrepancias. ¿Hay discrepancias del funcionamiento del algoritmo según el entorno de desarrollo y el entorno de implementación final en producción? ¿Cuáles? ¿Cambiaron las métricas de rendimiento finales del sistema?

Sesgos III. Realizar un análisis de sesgos en el algoritmo implementado en producción. ¿Qué criterios de sesgos se probaron? Definir valores aceptables de sesgos (ver principio ético de Sesgos y justicia algorítmica).

Evaluación II. Evaluación realizada por personas usuarias finales. Crear una base de datos centralizada de reportes de incidentes. Evaluar cómo resolver estos incidentes.

Desarrollo continuo del sistema de IA

La mayoría de los algoritmos de IA, una vez implementados en producción, se degradan con el tiempo y en función de cómo se desempeñan frente a nuevos datos o contextos. Además, a medida que pasa el tiempo, la variación en sus respuestas es cada vez más grande ("la IA envejece"). Por ello, es necesario contar con un monitoreo automático y continuo de su rendimiento y reentrenar el algoritmo, en caso de que este rendimiento caiga por debajo de cierto umbral.

Rendimiento III. Monitoreo continuo del rendimiento del algoritmo de IA. En caso de degradación, se debe volver a entrenarlo utilizando datos nuevos.

Reiniciar. En caso de agregar nuevas funcionalidades o modificar las existentes, se debe volver a iniciar el chequeo de la guía ética desde el comienzo (consultar Diseño del algoritmo de IA).

Casos de uso

Casos de uso

A continuación, se muestran **dos aplicaciones de la guía práctica de desarrollo ético de sistemas basados en IA** de la sección anterior a dos proyectos de IA que están en desarrollo o fueron implementados por parte de la Subsecretaría de Políticas Públicas Basadas en Evidencia (SSPPBE) del Gobierno de la Ciudad Autónoma de Buenos Aires (GCBA).

El primero es un clasificador de residuos reciclables, por parte del proyecto de Ecopuntos. El segundo es un recomendador de capacitaciones o cursos específicamente adaptados a cada persona usuaria en particular, según su historial de cursos realizados o sus preferencias. Dado que la guía práctica fue desarrollada para aplicarse, en general, a cualquier proyecto que involucre un algoritmo de IA, algunas de las preguntas de la sección anterior pueden no corresponder a todos los casos de aplicación que se consideren. Por lo tanto, sólo se tomarán en cuenta aquellas preguntas relevantes para cada uno de los proyectos.

Guía práctica para el desarrollo ético de sistemas basados en IA

Ecopuntos

Entidad: Subsecretaría de Políticas Públicas Basadas en Evidencia (SSPPBE) del Gobierno de la Ciudad Autónoma de Buenos Aires (GCBA).

Proyecto: Ecopuntos - seleccionador de basura para mejorar el sistema de reciclaje.

Diseño del algoritmo de IA

Propósito. Definir el propósito del algoritmo de IA . Dar un diagnóstico del contexto en el que será aplicado. Definir posibles usos y aplicaciones del algoritmo y el contexto de su aplicación (por ejemplo: población objetivo, qué servicio ofrece, etc.) (ver principio ético de Propósito).

El algoritmo de IA de Ecopuntos es de propósito particular y tiene como finalidad categorizar fotos de materiales (proporcionadas por personas usuarias) para el reciclado. Así podrán segmentar, de manera automática, diferentes tipos de reciclables. El algoritmo se utilizará en el programa Ecopuntos 2023, al que podrá accederse desde el chatbot de la Ciudad Autónoma de Buenos Aires (Boti). El objetivo es mejorar la gestión de residuos de la comunidad a través de la gamificación de los diferentes hábitos sustentables, premiando a quienes logren más puntaje según distintas iniciativas que conformen el programa Ecopuntos. La presente versión del algoritmo de clasificación de Ecopuntos se entrenó con datos subidos por las personas usuarias durante 2021. Además, tiene el objetivo de descomprimir el trabajo del Observatorio de reciclables de Ecopuntos (que realiza una clasificación manual y visual de los residuos) y proveer una clasificación certera de los diferentes residuos que se depositan en CABA, a fines de permitir la escalabilidad de las clasificaciones. A su vez, el algoritmo sirve como una guía para la comunidad, ya que aconseja o sugiere formas de reciclar los diferentes tipos de desechos se consulten.

Para más información, consultar el sitio web del proyecto Ecopuntos.

Riesgos. ¿Se puede prever algún impacto negativo de la herramienta? ¿Cuáles? ¿Qué riesgos pueden existir por la utilización del algoritmo? Definir si alguno de estos usos tiene un potencial dañino o su uso puede ser sensible o vulnerar derechos humanos (ver principio ético de Seguridad).

Podría existir el riesgo de que la IA haga predicciones equivocadas y que, por esa razón, ocurra que una persona usuaria decida enviar un residuo a un destino equivocado (basado en una clasificación errónea del algoritmo de IA de Ecopuntos). Sin embargo, en la versión del programa de 2023

todavía funcionará un observatorio que chequea manual o visualmente (humanas) las clasificaciones realizadas por el algoritmo de IA. Es decir, la clasificación del algoritmo en ningún caso se tomará como la clasificación de residuos final en la versión del programa en 2023.

El Observatorio de reciclables consiste en un validación o clasificación manual o visual de las fotos subidas a través de la siguiente interfaz gráfica:

Imagen 1

Asimismo, podría ocurrir que una persona usuaria no suba simplemente una foto de un residuo, sino que suba una *selfie* con su rostro al sistema de Boti. En ese caso, será necesario tomar medidas para que dicha foto se borre o permanezca, de manera confidencial, en el sistema.

Recolección de datos e integración de fuentes

Fuente. ¿Cuál es la fuente de los datos? ¿Cómo se obtienen? ¿Hay consentimiento de la fuente, las personas o las organizaciones? En caso de utilizar código de programación para obtenerlos, documentar el código (ver principio ético de Trazabilidad).

Los datos (fotos) utilizados fueron subidos por las personas usuarias a través de Whatsapp en una experiencia conversacional gamificada de Ecopuntos. Se puede acceder a través de Boti, el chatbot del GCBA. Los datos se obtuvieron durante el desarrollo del programa Ecopuntos en 2021 y se subieron, de manera manual, luego de dar consentimiento y de aceptar los términos y condiciones correspondientes del programa (que igualmente no hacían referencia explícita al uso de las fotos como datos de entrenamiento de un algoritmo de IA, sino que hacía referencia a su uso generalizado de ellos). A continuación, se muestra una captura de pantalla del chat de Boti, a través del que se puede subir una foto de un residuo para su clasificación.

Casos de uso

Ecopuntos

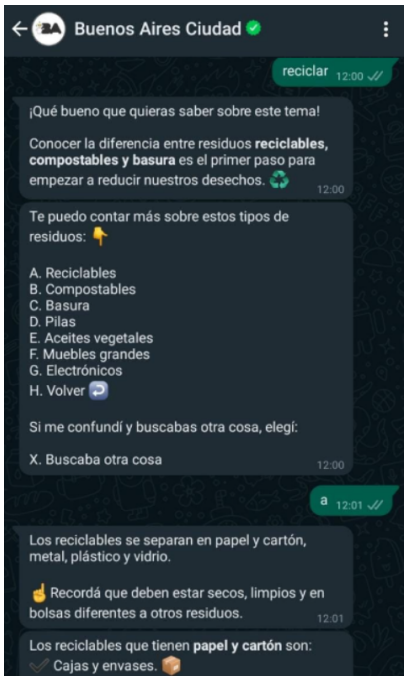


Imagen 2

Datasets. Describir, de manera general, los *datasets* que se utilizan para entrenar el algoritmo. ¿Cuál es su tamaño? ¿Qué variables contiene? Hacer una descripción general de cada variable y tipo de dato que contiene. ¿Hay variables que contengan información personal o confidencial? (ver principio ético de Protección de datos y privacidad).

Para entrenar el algoritmo de IA se utilizaron fotos subidas por personas usuarias de Boti en formato JPG (cada una con alrededor de 200 KB de peso en promedio). Estas fotos fueron clasificadas por personas usuarias en siete categorías posibles (aceite, secos, botellas, compost, electrónicos, orgánicos o baterías).

Las fotos se subieron durante los 4 meses de desarrollo del programa Ecopuntos en 2021, cuando contó con alrededor de 3000 perfiles usuarios que subieron fotos, y aproximadamente 18.000 fotos subidas en total.

Con respecto a datos confidenciales o personales, existen fotos enviadas por personas usuarias donde pueden verse sus rostros y/o diferentes espacios de sus hogares.

Calidad. ¿Cuál es la calidad de los datos? ¿Hay datos faltantes? ¿Contienen errores o inconsistencias? (por ejemplo, por carga manual). ¿Hay algún procedimiento para detectar esto?

No existen datos faltantes ya que las categorías se armaron sobre la base de las fotos que fueron subidas durante el funcionamiento del programa Ecopuntos. Sin embargo, es posible que haya fotos con residuos erróneamente clasificados por usuarios. Por ello se recomienda realizar una validación manual o visual por parte del Observatorio de las clasificaciones realizadas por perfiles usuarios que suben las fotos.

Casos de uso
Ecopuntos

Integración. ¿Los datos cubren los casos de uso intencionados? ¿Se integran datos de distintas fuentes? ¿Sobran datos? ¿Por qué?

Todos los datos corresponden a la misma fuente, por lo que no hubo necesidad de integrar datos de fuentes distintas. Hay fotos sobrantes que el Observatorio no consideró como válidas para sumar puntos, ya sea porque no cumplieron los requisitos explicados en la descripción del programa o porque no estaba claro el contenido de las fotos. Estas fotos ambiguas no se tomaron en cuenta para el entrenamiento del algoritmo de IA.

Actualización. ¿Se actualizan los datos? ¿Con qué frecuencia? Si no se actualizan, ¿son datos fijos que no pueden cambiar?

Se van a sumar nuevas fotos al entrenamiento futuro del algoritmo que se subirán durante la próxima versión del programa de Ecopuntos, a desarrollar durante 2023. Se contempla pedir a personas usuarios que las fotos se tomen de determinadas maneras, para facilitar el trabajo (tanto del Observatorio como del algoritmo de IA).

Almacenamiento. ¿Se almacenan los datos? ¿De manera segura? ¿Hay consentimiento y aviso en caso de ser datos personales o privados? ¿Pueden las personas usuarias acceder a los datos personales almacenados sobre ellos mismos? ¿Puede una persona usuaria decidir que sus datos personales sean borrados de la base de datos?

Los datos usados para entrenar el algoritmo de IA se guardan en el espacio de almacenamiento de Google que posee Botmaker (el sistema de creación de chatbots propietario de Google). Luego, se accede a estas fotos a través de un archivo de hoja de cálculo de Google, donde se guardan los enlaces a las fotos. Para la edición actual del programa, el espacio de Google almacena los enlaces, que pueden revisarse desde una nueva base de datos que estará alojada en la infraestructura de ASI. En el siguiente diagrama puede verse el flujo de datos que corresponde al desarrollo del programa de Ecopuntos:

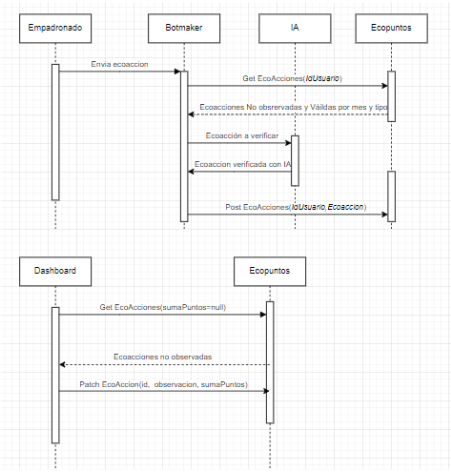


Imagen 3

Casos de uso
Ecopuntos

Actualmente, la persona usuaria no puede acceder a la información sobre los datos que fueron guardados y tampoco puede pedir que sus datos sean borrados (más allá de los mecanismos formales, como vecino, de realizar un pedido formal al GCBA para conocer los datos personales propios que se guardan).

Sesgos I. Realizar un análisis preliminar de sesgos y problemas de representatividad en los datos. ¿Se hacen suposiciones sobre los datos? ¿Qué criterios de sesgos y representatividad se probaron? Definir valores aceptables de sesgos (ver principio ético de Sesgos y justicia algorítmica).

Un sesgo de los datos es que las fotos corresponden a residuos de residentes de la CABA, por lo que el algoritmo presenta sesgos en relación con los reciclables de la Ciudad, que pueden ser distintos a los reciclables a nivel nacional.

Además, las fotos son subidas a través de un teléfono celular con acceso a internet y por parte de quienes reciben la comunicación de los distintos programas del GCBA.

Para este algoritmo y aplicación específicas estos sesgos no parecen ser un problema, ya que, por el momento, el algoritmo solo se aplica en CABA. Sin embargo, la aplicación en ambientes con lógicas de reciclado y/o dispositivos diferentes, para interactuar con el chatbot habría que modificar los resultados del algoritmo tal como está entrenando actualmente.

Preprocesamiento y etiquetado de datos

Variables. ¿Es necesario almacenar todas las variables que actualmente se almacenan? ¿Son todas necesarias para el funcionamiento de los algoritmos?

No es necesario. Los metadatos de las fotos, como las personas empadronadas que las subieron y la fecha en que fueron tomadas, no se utilizan para el entrenamiento del algoritmo.

Anonimización. ¿Se anonimizan los datos? ¿Cómo? ¿Se pueden volver a identificar posteriormente a los individuos? Por ejemplo, cruzando distintas fuentes de datos externas que no sean parte del presente desarrollo (ver principio ético de Protección de datos y privacidad).

La tabla con las personas empadronadas no está anonimizada ni se utiliza para entrenar el algoritmo. La información personal solamente se utiliza internamente. Para entrenar el algoritmo se usan únicamente las fotos y la categoría a la que cada una corresponde.

Casos de uso
Ecopuntos

Etiquetado. ¿Cómo se realiza el etiquetado de los datos? ¿El etiquetado está alineado con el propósito del algoritmo de IA? Si el etiquetado proviene de fuentes diversas, ¿son coherentes las distintas fuentes?

En 2021, las personas que subían la foto ponían la categoría del reciclable y luego el Observatorio la validaba (aceptando o rechazando esa categorización).

En 2023 sucederá lo mismo que en 2021, sumando que el algoritmo entrenado con los datos de 2021 estará en funcionamiento y aportará una clasificación automática y un puntaje asociado a dicha clasificación.

El etiquetado de las fotos responde entonces, específicamente, al propósito del algoritmo.

Definición del algoritmo, función objetivo y entrenamiento

Explicabilidad. Definición de los algoritmos de optimización que van a ser probados para su entrenamiento. ¿Son explicables o de caja negra? (ver principio ético de Explicabilidad y transparencia).

El algoritmo de entrenamiento de IA es un AutoML provisto por Vertex (*software* propietario de Google), que detecta un objeto en una imagen (busca el objeto que tiene mayor correspondencia con alguna categoría de las que predice el algoritmo). A continuación, el algoritmo devuelve una predicción (con su probabilidad) para cada objeto con que encuentra correspondencia con las categorías de reciclables establecidas. El algoritmo se queda entonces con aquel objeto que posee el puntaje más alto (la categoría más probable de todos los objetos detectados en cada foto). Este tipo de algoritmos son de caja negra.

Objetivo. Definición de la función objetivo. ¿Ella se alinea totalmente con el propósito del sistema, o es aproximada? Considerar los problemas de seguridad que esta función objetivo podría tener (ver principio ético de Seguridad).

La función objetivo busca minimizar el error en las predicciones del algoritmo. Esta función se alinea totalmente con el propósito de la IA .

Rendimiento I. Definición de métricas de rendimiento del algoritmo de IA (por ejemplo, precisión, exhaustividad, etc.). ¿Están alineadas a los objetivos del algoritmo?

La métrica de rendimiento del sistema utilizada es la precisión en la predicción para cada una de las siete categorías definidas.

Casos de uso
Ecopuntos

Sistema externo. ¿Se utilizó algún sistema o servicio externo para la creación o el entrenamiento del algoritmo de IA? ¿Cuáles? ¿El proveedor de servicios cuenta con estándares de seguridad y privacidad alineados a los del equipo desarrollador?

Se utilizó el sistema propietario de Google, llamado Botmaker.

Testeo del sistema de IA y pruebas de justicia algorítmica

Rendimiento II. ¿Cuál es el resultado obtenido de las métricas de rendimiento finales del algoritmo? ¿En qué datos se obtuvieron estos resultados?

La precisión en la clasificación de cada una de las categorías puede verse en la tabla a continuación:

Imagen 4

Etiqueta de confianza \ Etiqueta predicha	Aceite	Secos	Botellas	Compost	Electronicos	Organicos	Baterias
Aceite	89%	6%	0%	0%	6%	0%	0%
Secos	0%	98%	1%	0%	0%	0%	0%
Botellas	1%	5%	92%	0%	2%	0%	0%
Compost	0%	0%	0%	98%	0%	2%	0%
Electronicos	0%	6%	6%	0%	71%	3%	13%
Organicos	0%	1%	0%	2%	0%	97%	0%
Baterias	0%	8%	8%	0%	8%	0%	75%

Estas precisiones se obtuvieron para los datos de entrenamiento, es decir, para los mismos datos con los que el algoritmo fue entrenado. Queda pendiente la aplicación del algoritmo con datos nuevos (que no haya visto durante su entrenamiento) para definir su rendimiento real.

Sesgos II. Realizar un análisis de sesgos en el algoritmo finalizado. ¿Qué criterios de sesgos se probaron? Definir valores aceptables de sesgos (ver principio ético de Sesgos y justicia algorítmica).

Los sesgos previamente mencionados son razonables para el propósito del algoritmo, es decir, entran dentro de los valores aceptables para la aplicación del presente algoritmo.

Casos de uso

Ecopuntos

Evaluación I. Evaluación realizada por personas usuarias objetivo o partes interesadas, y también por expertos del dominio de aplicación del algoritmo. Crear una base de datos centralizada de reportes de incidentes. Evaluar cómo resolver estos incidentes.

Quien sube la foto define la categoría del reciclable. Luego, el Observatorio verifica la categoría proporcionada.

El Observatorio ya detectó categorías que el algoritmo no clasificó correctamente. En estos casos, se almacena un detalle de esas imágenes.

No hubo incidentes en producción, ya que el algoritmo todavía no fue implementado en este entorno.

Diagnóstico manual.

- Probar casos de uso previstos. ¿La IA funciona bien donde se supone que debería funcionar bien?
- Probar casos de borde (donde su funcionamiento se supone que es ambiguo).
- Probar posibles fallas (casos en lo que la IA puede presentar fallas).

Las métricas correspondientes a las predicciones con los datos de entrenamiento del algoritmo son positivas. El algoritmo todavía no ha sido probado en casos nuevos. Se utilizará durante la apertura del programa de Ecopuntos en 2023.

Implementación del sistema de IA en producción

Discrepancias. ¿Hay discrepancias del funcionamiento del algoritmo según el entorno de desarrollo y el entorno de implementación final en producción? ¿Cuáles? ¿Cambiaron las métricas de rendimiento finales del sistema?

No se implementó aún el algoritmo en producción. No se cuentan con métricas ni ningún tipo de información del rendimiento del algoritmo en este entorno.

Desarrollo continuo del sistema de IA

Rendimiento III. Monitoreo continuo del rendimiento del algoritmo de IA. En caso de degradación, se debe volver a entrenarlo utilizando datos nuevos.

Todavía no se implementó este proceso, pero uno de los objetivos es entrenar el algoritmo con nuevas imágenes —subidas por personas usuarias— que no haya clasificado correctamente.

Reiniciar. En caso de agregar nuevas funcionalidades o modificar las existentes, se debe volver a iniciar el chequeo de la guía ética desde el comienzo (consultar Diseño del algoritmo de IA).

Pueden agregarse nuevas funcionalidades al algoritmo de clasificación de reciclables. Como, por ejemplo, escalar la aplicación a diferentes latitudes (distintas provincias), agregar otras fuentes de datos (videos), o sumar nuevas categorías de clasificación de reciclables .

Guía práctica para el desarrollo ético de sistemas basados en IA

Recomendador de capacitaciones

Entidad: Subsecretaría de Políticas Públicas Basadas en Evidencia (SSPPBE) del Gobierno de la Ciudad Autónoma de Buenos Aires (GCBA).

Proyecto: Recomendador de capacitaciones o cursos, específicamente adaptados a cada usuario/a, según su historial de cursos realizados o sus preferencias.

Diseño del algoritmo de IA

Propósito. Definir el propósito del algoritmo de IA . Dar un diagnóstico del contexto en el que será aplicado. Definir posibles usos y aplicaciones del algoritmo y el contexto de su aplicación (por ejemplo: población objetivo, qué servicio ofrece, etc.) (ver principio ético de Propósito).

El propósito del algoritmo de IA es recomendar a la ciudadanía un conjunto de capacitaciones o cursos ofrecidos por el GCBA, para fomentar la formación de las personas y así aumentar su posibilidad de empleabilidad.

Para generar una recomendación, el algoritmo tiene en cuenta tanto el historial de cursos realizados por las personas usuarias, como sus intereses. De esta forma, puede recomendar personalizadamente.

Riesgos. ¿Se puede prever algún impacto negativo de la herramienta? ¿Cuáles? ¿Qué riesgos pueden existir por la utilización del algoritmo? Definir si alguno de estos usos tiene un potencial dañino o su uso puede ser sensible o vulnerar derechos humanos (ver principio ético de Seguridad).

No hay un impacto negativo, ni existen riesgos para quien usa el Recomendador, ya que el algoritmo sólo analiza las variables de “cursos previamente realizados” e “intereses” para perfilar personas usuarias. Por ese motivo, no utiliza datos como el género, la condición social, etc. Por lo tanto, no podemos afirmar, *a priori*, que tenga potencial dañino, ni vulnere los derechos humanos. Otro punto a destacar es que, más allá de las variables analizadas, el Recomendador puede priorizar las capacitaciones y la oferta no se restringe.

Casos de uso

Recomendador de capacitaciones

Recolección de datos e integración de fuentes

Fuente. ¿Cuál es la fuente de los datos? ¿Cómo se obtienen? ¿Hay consentimiento de la fuente, las personas o las organizaciones? En caso de utilizar código de programación para obtenerlos, documentar el código (ver principio ético de Trazabilidad).

Los datos de los cursos se obtienen de diferentes fuentes: SIENFO (Sistema de Información de Educación No Formal), GOET (Gerencia Operativa de Educación y Trabajo), SIU (Sistema de Información Universitaria), MOODLE (Modular Object-Oriented Dynamic Learning Environment) y CRMSL (Customer Relationship Manager SocioLaboral). Luego son ingestados y modelados en el *datalake*.

Sí, hay consentimiento, ya que cada área firma un formulario en el que establece las condiciones de uso y tratamiento de los datos.

Los casos de utilización de código siempre se documentan.

Datasets. Describir, de manera general, los *datasets* que se utilizan para entrenar el algoritmo. ¿Cuál es su tamaño? ¿Qué variables contiene? Hacer una descripción general de cada variable y tipo de dato que contiene. ¿Hay variables que contengan información personal o confidencial? (ver principio ético de Protección de datos y privacidad).

Se utilizan tres tablas tomadas del *datalake*. Con la información de ellas se conforma una matriz cuadrada (que contiene la misma cantidad de filas y de columnas [800x800]). Las filas y las columnas representan las capacitaciones y los valores que contiene la matriz son los valores de similaridad entre ellas.

Conforme a lo detallado, no hay variables que contengan información personal ni confidencial.

Calidad. ¿Cuál es la calidad de los datos? ¿Hay datos faltantes? ¿Contienen errores o inconsistencias? (por ejemplo, por carga manual). ¿Hay algún procedimiento para detectar esto?

Para evaluar la calidad de los datos, se realiza un procedimiento mediante el cual se utiliza un *framework* que evalúa su completitud y exactitud.

No hay datos faltantes, ya que se utilizan los cursos realizados por cada persona. Tampoco hay errores o inconsistencias en la carga de datos.

Casos de uso

Recomendador de capacitaciones

Integración. ¿Los datos cubren los casos de uso intencionados? ¿Se integran datos de distintas fuentes? ¿Sobran datos? ¿Por qué?

Los datos cubren los casos de uso intencionados. A través de ellos puede accederse al historial de cursos realizados por la persona, generar la matriz cuadrada (nombrada anteriormente), y luego efectuar la recomendación. No sobran datos, ya que se toman únicamente aquellos que se necesitan.

Actualización. ¿Se actualizan los datos? ¿Con qué frecuencia? Si no se actualizan, ¿son datos fijos que no pueden cambiar?

Se actualizan datos en el sentido de que se publican cursos nuevos y se pueden dar de baja cursos que estaban publicados. Cada un mes se rentrena la matriz que calcula la similaridad entre cursos.

Almacenamiento. ¿Se almacenan los datos? ¿De manera segura? ¿Hay consentimiento y aviso en caso de ser datos personales o privados? ¿Pueden las personas usuarias acceder a los datos personales almacenados sobre ellos mismos? ¿Puede una persona usuaria decidir que sus datos personales sean borrados de la base de datos?

Son almacenados de forma segura. Las personas usuarias se registran, previo consentimiento informado en el Portal de Oportunidades del GCBA. Sí, pueden acceder a sus datos almacenados, como así también solicitar que sus datos sean eliminados, borrados o actualizados conforme lo establecido por la Ley 25.326 de Protección de Datos Personales. Esto es fundamental, ya que permite verificar la exactitud de los datos y corregir cualquier error.

Sesgos I. Realizar un análisis preliminar de sesgos y problemas de representatividad en los datos. ¿Se hacen suposiciones sobre los datos? ¿Qué criterios de sesgos y representatividad se probaron? Definir valores aceptables de sesgos (ver principio ético de Sesgos y justicia algorítmica).

Los posibles sesgos que puede tener un algoritmo de recomendación son:

- Sesgo de muestreo: el *dataset* usado para entrenar el recomendador no es representativo de la población objetivo. Por ello podría hacer recomendaciones erradas. Se mitiga usando el total de la población de muestreo para entrenar el algoritmo.

- Sesgo de selección: la recomendación está influenciada, principalmente, por el historial de la persona. Ello puede reforzar sesgos en las recomendaciones. Se soluciona incorporando los intereses, además del historial.

- Sesgo de "comienzo en frío": hacer malas recomendaciones debido a que la persona no tiene un historial en el sistema. Se soluciona incorporando los intereses, además del historial.
- Sesgo de falta de transparencia: puede no confiarse en el algoritmo ni en sus recomendaciones, porque no saben cómo funciona. Se soluciona al momento de publicar el código del algoritmo, junto con un documento técnico explicando en detalle cómo funciona.

Preprocesamiento y etiquetado de datos

Variables. ¿Es necesario almacenar todas las variables que actualmente se almacenan? ¿Son todas necesarias para el funcionamiento de los algoritmos?

Se almacenan únicamente aquellas variables que el algoritmo utilizará, es decir, un identificador único para cada persona usuaria, los nombre de los cursos que realizó y sus intereses. Es indispensable contar con estos datos para el buen funcionamiento del algoritmo.

Anonimización. ¿Se anonimizan los datos? ¿Cómo? ¿Se pueden volver a identificar posteriormente a los individuos? Por ejemplo, cruzando distintas fuentes de datos externas que no sean parte del presente desarrollo (ver principio ético de Protección de datos y privacidad)

En este tipo de proyecto no se anonimizan los datos, ya que es importante identificar a cada persona para una recomendación personalizada. Pero los datos personales no son utilizados para generar la recomendación, es decir, no los ve el algoritmo.

No se puede volver identificar a las personas usuarias porque los ID no se exponen, son únicamente para uso interno.

Definición del algoritmo, función objetivo y entrenamiento

Explicabilidad. Definición de los algoritmos de optimización que van a ser probados para su entrenamiento. ¿Son explicables o de caja negra? (ver principio ético de Explicabilidad y transparencia).

Se realizó y entrenó un algoritmo denominado "sistema de recomendación". Se trata de generar una "matriz de cursos y capacitaciones" en que esté contenida la información de "qué cursos son más similares entre sí que otros". Esta similaridad entre cursos permitirá luego recomendar cursos similares (en varios sentidos) a los cursos realizados en el pasado.

El algoritmo es entrenado de forma transparente, en el sentido de que las razones por las que recomienda un curso son, en buena medida, explicables.

Casos de uso

Recomendador de capacitaciones

Objetivo. Definición de la función objetivo. ¿Ella se alinea totalmente con el propósito del sistema, o es aproximada? Considerar los problemas de seguridad que esta función objetivo podría tener (ver principio ético de Seguridad).

La función objetivo busca determinar la similaridad entre los distintos cursos. Esto se logra al tener en cuenta como más similares aquellos cursos que las personas realizan en conjunto con mayor frecuencia (se asume que ellos son más similares que otros cursos que usualmente no son realizados por una misma persona).

Rendimiento I. Definición de métricas de rendimiento del algoritmo de IA (por ejemplo, precisión, exhaustividad, etc.). ¿Están alineadas a los objetivos del algoritmo?

Se utilizó como métrica el *accuracy*. Para cada persona usuaria se extrajeron los cursos que realizó. Se quitó de la lista un curso al azar para cada usuaria. Se corrió el algoritmo de recomendación, mostrándole todos los cursos realizados por cada persona, excepto el eliminado. En cada caso, se registró si el curso eliminado estaba entre los recomendados. El *accuracy* representa la proporción de recomendaciones que incluía un curso eliminado.

Trazabilidad. ¿Por qué se seleccionó el algoritmo que finalmente quedó? ¿Qué hiperparámetros se probaron? Documentar todos los resultados (ver principio ético de Trazabilidad).

Dada la naturaleza del problema y la cantidad de cursos y personas usuarias disponibles, la solución más acorde es utilizar un algoritmo de sistema de recomendación (de tipo filtro colaborativo ítem-ítem, es decir, curso-curso).

Sistema externo. ¿Se utilizó algún sistema o servicio externo para la creación o el entrenamiento del algoritmo de IA? ¿Cuáles? ¿El proveedor de servicios cuenta con estándares de seguridad y privacidad alineados a los del equipo desarrollador?

Se utilizaron librerías Open Source (pandas, numpy, sklearn, time, boto3). El almacenamiento de datos se realizó en el *datalake* de la Subsecretaría de Políticas Públicas Basada en Evidencia. Siempre se evalúa quiénes serán los proveedores que cumplan con los estándares de privacidad, seguridad y protección. Se establecen convenios que garantizan el nivel de protección.

Testeo del sistema de IA y pruebas de justicia algorítmica

Rendimiento II. ¿Cuál es el resultado obtenido de las métricas de rendimiento finales del algoritmo? ¿En qué datos se obtuvieron estos resultados?

La precisión del algoritmo de recomendación es de 0.58, con base en los datos usados para el desarrollo, previos al 31 de diciembre de 2022. Ello indica que se encuentra dentro del margen aceptable. Como se describió anteriormente, el algoritmo necesita, como información, los cursos que realizó la persona usuario y sus intereses. La métrica evalúa principalmente la recomendación que se realiza con base en el historial usuario, lo más importante que tiene el algoritmo.

Sesgos II. Realizar un análisis de sesgos en el algoritmo finalizado. ¿Qué criterios de sesgos se probaron? Definir valores aceptables de sesgos (ver principio ético de Sesgos y justicia algorítmica).

En razón de las variables con las que trabaja el algoritmo, de su análisis resulta que no está sesgado y no posee problemas de representatividad.

No se hacen suposiciones, ya que toma la totalidad de los datos en función de los cursos e intereses que la persona elige para realizar las capacitaciones.

Evaluación I. Evaluación realizada por personas usuarias objetivo o partes interesadas, y también por expertos del dominio de aplicación del algoritmo. Crear una base de datos centralizada de reportes de incidentes. Evaluar cómo resolver estos incidentes.

Siempre es importante hacer una evaluación y testeo por persona usuaria, como también tener presente los posibles casos de incidentes que se reporten.

Diagnóstico manual.

- Probar casos de uso previstos. ¿La IA funciona bien donde se supone que debería funcionar bien?
- Probar casos de borde (donde su funcionamiento se supone que es ambiguo).
- Probar posibles fallas (casos en lo que la IA puede presentar fallas).

El algoritmo funciona bien. No hay casos de borde, donde el algoritmo puede ser ambiguo. Previo a la salida en producción del algoritmo, para chequear posibles fallas del estilo, se probó: ¿Qué pasa si la persona envía un DNI incorrecto? ¿Qué sucede si no especifica sus intereses? Se hicieron

pruebas buscando algunos usuarios en la base de datos, se les inventó intereses y se corroboró que el algoritmo devolvía una recomendación.

Implementación del sistema de IA en producción

Discrepancias. ¿Hay discrepancias del funcionamiento del algoritmo según el entorno de desarrollo y el entorno de implementación final en producción? ¿Cuáles? ¿Cambiaron las métricas de rendimiento finales del sistema?

No hay discrepancias del funcionamiento del algoritmo según el entorno. El algoritmo fue entrenado en desarrollo con datos productivos.

Sesgos III. Realizar un análisis de sesgos en el algoritmo implementado en producción. ¿Qué criterios de sesgos se probaron? Definir valores aceptables de sesgos (ver principio ético de Sesgos y justicia algorítmica).

En razón de las variables con las que trabaja el algoritmo, de su análisis resulta que no está sesgado y no posee problemas de representatividad.

Evaluación II. Evaluación realizada por personas usuarias finales. Crear una base de datos centralizada de reportes de incidentes. Evaluar cómo resolver estos incidentes.

No se obtuvieron comentarios o reacciones de las personas usuarias que impliquen la necesidad de modificar una recomendación realizada.

Desarrollo continuo del sistema de IA

Rendimiento III. Monitoreo continuo del rendimiento del algoritmo de IA. En caso de degradación, se debe volver a entrenarlo utilizando datos nuevos.

Se descargarán los datos actualizados para evaluar el algoritmo y volver a entrenarlo con cierta frecuencia.

Reiniciar. En caso de agregar nuevas funcionalidades o modificar las existentes, se debe volver a iniciar el chequeo de la guía ética desde el comienzo (consultar Diseño del algoritmo de IA).

No se hicieron nuevas modificaciones. En caso de hacerlas, volverán a tomar los parámetros recomendados por la Guía.

Anexo

Anexo

Consideraciones para chatbots (y algoritmos generativos en general)

Los algoritmos de IA generativos generan contenido de manera flexible —como texto, imágenes o videos—, a partir de comandos de texto de entrada flexibles (llamados *prompts)*. En la actualidad ya existen muchos riesgos éticos asociados al desarrollo, despliegue y uso de algoritmos generativos. Principalmente, porque todos los algoritmos generativos son de propósito general, es decir no se crean para satisfacer una aplicación concreta, sino que pueden usarse de manera flexible y creativa por las personas usuarias.

Entre los riesgos éticos actuales asociados a los algoritmos generativos, se halla la generación de voces artificiales o que simulen la voz de personas reales (usurpación de la voz) y que, por lo tanto, pueden usarse con fines delictivos. También pueden generar imágenes o videos artificiales que involucren personas específicas de la realidad, pero que muestran acontecimientos y acciones que nunca realizaron (denominado *deepfake*). Además, es posible crear algoritmos de IA que generen automáticamente noticias falsas y, al mismo tiempo, produzcan un gran número de opiniones automáticas sobre distintos temas a partir de perfiles usuarios ficticios (fabricando así opinión pública para que ciertas miradas aparenten mayor envergadura social que las que realmente tienen).

Un tipo especial de algoritmo generativo es aquel que genera texto. En particular, aquel que tiene la capacidad de interactuar mediante un chat con personas reales (denominados “chatbots”). A continuación, se presentan algunas recomendaciones para un desarrollo confiable y ético de este tipo de algoritmos generativos, los chatbots.

Tener claro el propósito del chatbot, sus alcances y limitaciones. En especial, prestar mayor atención a su uso para el ofrecimiento de servicios relacionados con datos sensibles (como de salud, educación, empleo, o financieros). Estos, probablemente, requieran la asistencia de criterios expertos o humanos que ningún algoritmo de IA hoy en día puede reemplazar. A su vez, deben estar debidamente notificados los costos o riesgos de los errores que estos chatbots pudieran cometer, junto con los límites de su rendimiento.

Mencionar explícitamente cuando las respuestas del chatbot sean generadas por un algoritmo, y no por un ser humano; ya que esto es difícil —o incluso imposible— de diferenciar sin previo aviso, independientemente de la habilidad o experiencia usuaria.

Permitir que el chatbot detecte ofensas, discursos de odio, o temas controversiales de personas usuarias, para responder apropiadamente ante esos ataques. Además, asegurar la privacidad de los datos personales o confidenciales con los que el chatbot fue entrenado; y que, por lo tanto, estén almacenados implícitamente en los algoritmos de respuesta.

Notar que, en la actualidad, de ninguna manera puede asegurarse la veracidad de las respuestas de un algoritmo generativo. Este principio se denomina “alucinación de una IA” y significa que tampoco pueden reemplazar el conocimiento experto. En aplicaciones sensibles, como las áreas de salud, empleo, finanzas y legales, es fundamental contar con la contribución y supervisión de expertos humanos, en lugar de depender únicamente de las respuestas generadas por la IA.

Además, es importante tener en cuenta que existen métodos, englobados bajo el nombre de ‘*prompt hacking*’, que permiten a las personas usuarios de un chatbot generar respuestas personalizadas a su conveniencia. Esto significa que es posible manipular las respuestas del chatbot para que parezcan haber sido generadas por la IA. Mediante estos métodos, es posible obtener respuestas que

involucren información confidencial con la que el chatbot fue entrenado, incluso si inicialmente se programó para no divulgar dicha información. Esto plantea preocupaciones significativas sobre la protección de datos y la privacidad, especialmente cuando los chatbots se entrenan con grandes cantidades de datos que pueden incluir información sensible. En consecuencia, no se puede garantizar que esta información no pueda ser recuperada o mal utilizada por los usuarios

Programar al chatbot para que detecte información privada o confidencial durante su conversación con personas usuarias y que decida no almacenar esa información —o, en caso de necesidad, hacerlo durante el menor tiempo posible—. Se sugiere, además, proveer información detallada a la persona usuaria sobre la recolección de datos y pedir el consentimiento correspondiente para su recolección (en caso de que sea necesario).

Maximizar la seguridad del chatbot, agregar funciones que aporten a la privacidad de las personas y que generen confianza en él. Por ejemplo, pueden agregarse botones con funciones como “Dime todo lo que sabes sobre mí”, u “Olvida nuestra última interacción”, o “Borra todo lo que sabes sobre mí”, entre otras. En algunos casos estas funciones podrían incluso requerirse legalmente. También se debe dar lugar a que las personas usuarias hagan devoluciones. Por ejemplo, si los chatbots sirvieron con éxito para la finalidad prevista, y en caso de una devolución negativa, que el chatbot provea a personas usuarias una forma de continuar la comunicación a través del contacto humano.

Todas las recomendaciones mencionadas deben documentarse debidamente y ser de fácil acceso por las personas usuarias del chatbot. En ese sentido, puede incluirse un código de conducta, para que la personas usuarias entiendan los alcances del servicio, sus limitaciones y la forma de dirigirse al chatbot para un uso y funcionamiento efectivos.

Bibliografía

• AA VV (2023). Declaración de Montevideo sobre Inteligencia Artificial y su impacto en América Latina. Varias instituciones.

• Feole, M. y Guaymás Canavire, A. (2022). Guía práctica para el uso de imágenes satelitales en la definición de políticas públicas. Fundar.

• Luvini, P. (2022). Guía práctica para la protección de datos. Fundar.

• López, S.; Alonso Alemany, L.; Dias, J.M.; Ación, L. y Xhardez, V. (2023). Guía práctica para la protección de datos personales en salud. Fundar.

• Yankelevich, D. (2021). ¿Sueñan los robots con el deber? Notas para una política activa sobre ética e inteligencia artificial. Fundar.

• Martínez, M. V., Dumas, V. G., Sarabia, M., y Feldfeber Kisilevsky, I. (2022). Innovar con Ciencia de Datos en el sector público. Fundación Sadosky.

• Aguerre, C., Amunátegui Perelló, C., Brathwaite, C., Castañeda, J. D., Castaño, D., Del Pozo, C., Flórez Rojas, L., Gómez Montt, C., Lara Gálvez, J. C., López, J., Madrid, R., Martín del Campo, A. V., Vargas Leal, J. (2020). Inteligencia Artificial en América Latina y el Caribe. Ética, Gobernanza y Políticas. CETyS (Universidad de San Andrés).

• Ortiz Freuler, J. e Iglesias, C. (2018). Algoritmos e Inteligencia Artificial en Latinoamérica: Un estudio de implementaciones por parte de gobiernos en Argentina y Uruguay. World Wide Web Foundation.

• AA VV (2020). Data Ethics Framework. Government Digital Service, United Kingdom.

• Rosales Torres, C. S., Buenadicha Sánchez, C. y Narita, T. (2021). Auto-evaluación ética de IA para actores del ecosistema emprendedor. BID.

• Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura [UNESCO] (2021). Recomendación sobre la ética de la inteligencia artificial.

• Microsoft (2022). Microsoft Responsible AI Standard.

• Amodei, D., Olah, C., Steinhardt, J., Christiano, P. F., Schulman, J., Man, D. (2016). Concrete Problems in AI Safety. OpenAI.

• Madaio, M., Stark, L., Wortman Vaughan, J. y Wallach, H. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. ACM Digital Library.

• Inioluwa, D. R., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. y Barnes, P. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. ACM Digital Library.

• Vela, D., Sharp, A., Zhang, R., Nguyen, T., Hoang, A., y PianykhVela, O. S. (2022). Temporal quality degradation in AI models. Nature Scientific Reports.

Acerca del equipo autoral

Marcos Feole

Científico de Datos de Fundar

Licenciado y magíster en Física por el Instituto Balseiro, y magíster en Estadística por la Universidad de Illinois en Urbana-Champaign. Trabajó en investigación, desarrollo de software, ciencia de datos y modelado matemático para el sector financiero.

Juan Manuel Dias

Científico de Datos de Fundar

Licenciado en Sociología por la UBA y maestrando en Estadística de la UNTREF. Es egresado de la carrera de ciencia de datos de la EANT y de la Diplomatura de Ciencias Sociales Computacionales de la UNSAM.

Mariana Kunst

Coordinadora de Datos de Fundar

Licenciada en Economía y maestranda en Métodos Cuantitativos para la Gestión y Análisis de Datos por la Universidad de Buenos Aires. Se desempeñó como coordinadora del Sistema de Información Cultural de la Argentina (SInCA). Actualmente es docente en la Universidad de Buenos Aires.

Zulma Carrizo

Responsable del Área de Gobernanza y Protección de Datos en la Subsecretaría de Políticas Públicas Basadas en Evidencia del Gobierno de la Ciudad Autónoma de Buenos Aires

Abogada y especialista en Innovación y nuevas tecnologías, Data & AI Strategy (Universidad de San Andrés) Gobernanza de datos e Inteligencia Artificial.(UBAIALB) Visualización y Data Mininggal (Universidad Nacional de Cuyo) Certificación Ciberseguridad- Academia Cisco AWS Technical Essentials (BOOTCAMP INSTITUTE) Data Analytics, Product Manager, Scrum Master. Metodologías Agiles.

German Guido Lavalle

Doctor en Ingeniería e ingeniero nuclear por el Instituto Balseiro

Fundador y Director de la consultora CANDOIT, empresa argentina de ingeniería y desarrollo de software. Miembro del Gabinete de la Secretaría de Innovación y Transformación Digital (SECITD) del Gobierno de la Ciudad Autónoma de Buenos Aires, donde se dedica a temas de Inteligencia Artificial. Dirigió más de 40 proyectos de desarrollo tecnológico, en temas de modelado, simulación por computadora e inteligencia artificial. Fue Rector del ITBA, Rector de UADE, Decano de su Facultad de Ingeniería, Gerente de Relaciones Internacionales de la Comisión Nacional de Energía Atómica, y Profesor e Investigador del Instituto Balseiro. Co-autor de dos libros. Recibió el Premio a las Iniciativas de Vinculación Tecnológica, del Ministerio de Cultura y Educación, y el Premio "Ernesto E. Galloni" de la Academia Nacional de Ciencias Exactas, Físicas y Naturales.

Dirección ejecutiva: Martín Reydó

Revisión Institucional: Juliana Arellano

Coordinación editorial: Gonzalo Fernández Rozas

Corrección: Victoria Inverga

Diseño: Jimena Zeitune

Modo de citar

Feole, M.; Dias, J. M.; Kunst, M.; Carrizo, Z. y Lavalle, G. G. (2023) Guía práctica para el desarrollo ético de sistemas basados en IA. Fundar. Disponible en https://www.fund.ar

Sobre Fundar

Fundar es un centro de estudios y diseño de políticas públicas que promueve una agenda de desarrollo sustentable e inclusivo para la Argentina. Para enriquecer el debate público es necesario tener un debate interno: por ello lo promovemos en el proceso de elaboración de cualquiera de nuestros documentos. Confiamos en que cada trabajo que publicamos expresa algo de lo que deseamos proyectar y construir para nuestro país. Fundar no es un logo: es una firma.

Trabajamos en tres misiones estratégicas para alcanzar el desarrollo inclusivo y sustentable de la Argentina:

Generar riqueza. La Argentina tiene el potencial de crecer y de elegir cómo hacerlo. Sin crecimiento, no hay horizonte de desarrollo, ni protección social sustentable, ni transformación del Estado. Por eso, nuestra misión es hacer aportes que definan cuál es la mejor manera de crecer para que la Argentina del siglo XXI pueda responder a esos desafíos.

Promover el bienestar. El Estado de Bienestar argentino ha sido un modelo de protección e inclusión social. Nuestra misión es preservar y actualizar ese legado, a través del diseño de políticas públicas inclusivas que sean sustentables. Proteger e incluir a futuro es la mejor manera de reivindicar el espíritu de movilidad social que define a nuestra sociedad.

Transformar el Estado. La mejora de las capacidades estatales es imprescindible para las transformaciones que la Argentina necesita en el camino al desarrollo. Nuestra misión es afrontar la tarea en algunos aspectos fundamentales: el gobierno de datos, el diseño de una nueva gobernanza estatal y la articulación de un derecho administrativo para el siglo XXI.

En Fundar creemos que el lenguaje es un territorio de disputa política y cultural. Por ello, sugerimos que se tengan en cuenta algunos recursos para evitar sesgos excluyentes en el discurso. No imponemos ningún uso en particular ni establecemos ninguna actitud normativa. Entendemos que el lenguaje inclusivo es una forma de ampliar el repertorio lingüístico, es decir una herramienta para que cada persona encuentre la forma más adecuada de expresar sus ideas.

Sobre la Subsecretaría de Políticas Públicas Basadas en Evidencia del Gobierno de la Ciudad Autónoma de Buenos Aires

La Subsecretaría de Políticas Públicas Basadas en Evidencia se creó en diciembre de 2019 con el objetivo de posicionar a Buenos Aires como ciudad líder en materia de gestión de datos, innovadora y éticamente responsable. Está formada por un equipo multidisciplinario que trabaja día a día para que la Ciudad de Buenos Aires sea una ciudad data-driven, es decir, gobernada por los datos.

La Subsecretaría aborda este desafío con una visión estratégica enfocada en cuatro ejes: gobernar, analizar y disponibilizar los datos, así como expandir la cultura de datos en todo el Gobierno de la Ciudad.

Guía práctica para el desarrollo ético de sistemas basados en IA / Marcos Feole ... [et al.]. - 1a ed. - Ciudad Autónoma de Buenos Aires : Fundar , 2023.
Libro digital, PDF

Archivo Digital: descarga y online
ISBN 978-631-90201-6-8

1. Ética. 2. Inteligencia Artificial. 3. Algoritmo. I. Feole, Marcos.
CDD 006.301

ISBN 978-631-90201-6-8
9 786319 020168