Gobernanza de Datos

Guía práctica para el desarrollo ético de sistemas basados en IA

SUBSECRETARÍA DE POLÍTICAS PÚBLICAS BASADAS EN EVIDENCIA
SECRETARÍA DE INNOVACIÓN Y TRANSFORMACIÓN DIGITAL

Buenos Aires Ciudad

Índice

1. Introducción

A la par de los progresos y la adopción masiva de las tecnologías de la información y comunicación (TIC), la era actual está marcada por un rápido progreso en el campo de la inteligencia artificial (IA), y su influencia en diversos sectores gubernamentales y de la sociedad es innegable. Los sistemas basados en IA se utilizan en los más diversos ámbitos y afectan desde la gestión de justicia hasta las preferencias de entretenimiento, o sea que 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 la implementación y el uso de estas tecnologías y que al mismo tiempo incorporen las normas éticas necesarias para minimizar los riesgos. En este contexto, es necesario que las áreas gubernamentales adopten estrategias de regulación y evaluación para guiar el desarrollo y el uso ético de estas herramientas con el fin de proteger a la ciudadanía de los posibles riesgos asociados a su implementación.

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 funcionarios 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 disciplinas con diversos enfoques. El conocimiento tecnológico se enriquece y 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 la sociedad.

Para abordar estas necesidades, el Área de Datos de Fundar y la Subsecretaría de Políticas Públicas Basadas en Evidencia (SSPPBE) del Gobierno de la Ciudad de Buenos Aires han trabajado 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, esta guía también puede servir como referencia en el sector privado.

2. Objetivos de la guía y a quién está dirigida

El principal objetivo de esta guía es disponibilizar herramientas de diagnóstico para lograr un desarrollo e implementación confiables de los sistemas basados en IA, y que estos minimicen los riesgos asociados a su uso. En particular, la presente guía viene a cubrir una carencia en el área de la ética en IA, en la que abundan las guías de principios éticos de

desarrollo de IA, pero que en la gran mayoría de ellas no se contemplan los conocimientos, los tiempos, ni los flujos de trabajo existentes de quienes desarrollan estas tecnologías. Es decir, que en su mayoría no tienen utilidad práctica, y entonces, según experiencias recolectadas de distintas fuentes, la ética en IA hoy en día se resuelve en general como un proceso ad hoc y a posteriori, donde los problemas se detectan de manera tardía (cuando ya ocurrieron), entonces las acciones necesarias para arreglar estos problemas se realizan en forma de parches, y una vez que parte del daño ya está hecho.

Esta guía está especialmente dirigida entonces a equipos de diseñadores/as, desarrolladores/as, y evaluadores/as de servicios basados en IA y sirve, a su vez, 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 permitir a las distintas áreas de gobierno minimizar los riesgos y el impacto potencial negativo de estas tecnologías.

Además, 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 sean afectadas por estas tecnologías, de modo que quienes estén implicados en el ciclo de vida de estos sistemas encuentren aquí herramientas para lograr un desarrollo seguro y robusto, y a la vez una evaluación efectiva de los mismos.

3. Modo de uso de la guía

El contenido se ha estructurado con el propósito de ser una herramienta práctica que se adapte a los ciclos de desarrollo que actualmente presentan las áreas que trabajan con tecnologías de **inteligencia artificial.** Sin embargo, el modo de utilizar esta guía depende de los objetivos que quien desarrolle, evalúe o utilice tenga. En primer lugar, se ofrece en la sección 4. Principios Éticos de IA 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.

Por otro lado, 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, se presenta en la sección 5. Guía de desarrollo ético de sistemas basados en IA la guía propiamente dicha que se nutre de los principios éticos descritos 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 logre abarcar 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, en la sección 6. Casos de uso se muestran dos ejemplos de aplicación de la guía de la sección 5 a la implementación de dos sistemas basados en IA de la Subsecretaría de Políticas Públicas Basadas en Evidencia (SSPPBE) del Gobierno de la Ciudad de Buenos Aires (GCBA). Se aplicó la guía 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 los mismos. Y por otro lado, 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 se desprenden además una serie de recomendaciones en vistas de mejorar la implementación de cada una de las iniciativas de uso de sistemas basados en IA en el gobierno.

4. 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 el uso de estas tecnologías puedan tener. Para lograr un desarrollo e implementación ética de sistemas basados en IA se presentan a continuación ciertos principios que proveen los conceptos mínimos de la ética en IA y que serán de utilidad a lo largo de la guía de desarrollo ético de la sección 5, a medida que sean requeridos en las distintas etapas de desarrollo de cada sistema basado en IA.

4.1. Propósito

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

Los algoritmos de IA se pueden clasificar por 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, o que detectan elementos específicos en imágenes, o aquellos que recomiendan películas basadas en los gustos de cada usuario, 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. Este tipo de algoritmos de IA, de propósito particular, no pueden ser utilizados con fines diferentes que para los que fueron programados y por lo tanto presentan un bajo riesgo ético desde el punto de vista del propósito con el que fueron creados.

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 que son independientes del dominio específico, y a la vez son versátiles

y adaptables a situaciones distintas. Por ejemplo, un algoritmo de propósito general puede imitar a un humano en una conversación de chat, de tal manera 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: estos crean imágenes, voces, videos o textos a partir de comandos de entrada, pero tienen un gran e indefinido grado de libertad a la hora de generar estos elementos. Estos algoritmos tienen entonces 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 ser utilizados no pueden en general ser previstas durante su diseño y desarrollo.

4.2. Colaboración con expertos en la materia de aplicación

Para todo sistema basado en IA existen desafíos específicos respectivos a cada área de aplicación de los mismos. Es necesario que se considere entonces el conocimiento experto de cada área de aplicación específica de estos sistemas, por ejemplo para asegurar que estos funcionan de manera acorde al propósito con el que fueron creados. Por ejemplo, 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 experto de geógrafos/as o expertos/as en las dinámicas terrestres a la hora de 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 los/as científicos/as de datos y aquellas personas que son expertas en la materia de aplicación de los mismos.

4.3. Seguridad

El principio de seguridad para los sistemas basados en IA debe integrarse al desarrollo de los mismos desde su diseño, especialmente para aplicaciones sensitivas. El principio de seguridad desde el diseño (SdB, por sus siglas en inglés, Security by Design) implica evaluar los potenciales riesgos que los sistemas puedan presentar, e implementar las prevenciones o soluciones desde el comienzo mismo de su desarrollo (desde la etapa de diseño). Esto permite minimizar o eliminar la necesidad de reacondicionar o establecer parches de seguridad en etapas posteriores del desarrollo de los mismos. Vale la pena aclarar que de ninguna manera se puede asegurar que los sistemas basados en IA (al igual que los sistemas informáticos en general) son 100% seguros, es decir que estos pueden presentar fallas no previstas o vías por las cuales estos sistemas pueden ser hackeados o vulnerados de maneras creativas.

Existen, además, una serie de problemas técnicos de seguridad que pueden arrastrarse desde el diseño o el entrenamiento de un algoritmo de IA, y que pueden generar respuestas erróneas o dar lugar a usos fraudulentos de los mismos. Estos son denominados “accidentes de IA” y pueden ser de varios tipos, de los cuales se mencionan tres ejemplos. Sin embargo, no se detallan aquí todas las posibles soluciones a estos problemas, ya que cada uno de estos accidentes requiere un estudio pormenorizado caso por caso de cómo evitarlos o revertirlos.

Box Seguridad 1 - Efectos secundarios negativos

Incluso cuando se declara una función objetivo precisa que el algoritmo de IA debe optimizar, pueden ocurrir efectos secundarios que consisten en 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 pasarles por encima o destruirlos para cumplir con su objetivo final. Una solución posible a este problema implica considerar los mecanismos de seguridad necesarios para que estos efectos secundarios no ocurran, y programar explícitamente al algoritmo para que éste tome las precauciones necesarias automáticamente.

Box Seguridad 2 - Hackeo de recompensas

El algoritmo puede lograr 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 éste observa en el suelo, pero en vez de limpiarla éste decidiera apagar sus sensores de visión, con lo cual no podrá observar la suciedad para limpiarla y entonces considerará en todo momento su objetivo cumplido. Esta es una forma en la cual un algoritmo puede encontrar una solución al problema que tiene que resolver no prevista por el humano que lo programa. Existen posibles soluciones a este problema, por ejemplo, determinar más específicamente la función objetivo de tal manera que no deje 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 Seguridad 3 - Robustez al cambio de distribución

El algoritmo de IA se puede encontrar durante su aplicación 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 aún 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, entonces 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 utilizada. Nuevamente, una solución general a este problema es difícil de obtener, por lo que es necesario prever de antemano los posibles contextos en los que esa IA va a ser utilizada.

4.4. 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 los datos utilizados durante todo el proceso de desarrollo. Desde el punto de partida, el propósito del sistema, junto a la arquitectura del algoritmo y diseño deben estar debidamente documentados. La recolección de datos debe ser trazable, su fuente debe estar documentada o los datos recolectados deben estar almacenados y disponibles. Si los datos se toman mediante un programa automático, éste debe formar parte de la documentación y su uso debe poder ser replicado independientemente. Luego, el preprocesamiento que se le realiza a los datos, desde la anonimización de los mismos hasta los procesos que sean necesarios para transformar los datos crudos en datos de entrada para la IA, deben estar documentados y poder ser replicados independientemente. Asimismo, el entrenamiento de los algoritmos de IA con esos datos debe ser replicable, y deben estar documentados los hiper-parámetros que fueron probados (estos son los parámetros determinados manualmente por el programador). Si bien el entrenamiento de un algoritmo de IA tiene en general componentes o parámetros aleatorios que hacen que éste no pueda ser replicado 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, se deben documentar todas las pruebas posteriores que se realizan sobre el sistema y los datos con los cuales se realizan estas pruebas deben estar almacenados y 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, para permitir la comprensión y replicación completa del desarrollo.

El principio de trazabilidad permite que un proyecto de IA sea tomado por un equipo de desarrolladores/as ajenos al proyecto, y que éstos puedan volver a crear el mismo algoritmo en las mismas condiciones con que fue creado en primer lugar. Esto permite, entre otros beneficios, poder identificar de manera precisa los motivos detrás de cualquier problema que surja posteriormente con ese sistema, y poder así corregirlos de manera efectiva. Además, 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 programadores/as como por los reguladores/as de estas tecnologías, y también por los usuarios o personas objetivo en general. El tema de la transparencia es abordado con mayor detalle en la sección 4.7, junto al concepto de explicabilidad de los algoritmos de IA.

4.5. 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 la información no sea utilizada para fines distintos a los que se han informado al/la 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 tales como el consentimiento informado —por el cual las empresas deben obtener el consentimiento informado del/la titular de los datos para recolectar, almacenar, usar y compartir sus datos personales—; la limitación de la finalidad —por el que las empresas solo deben recolectar y usar los datos personales para fines específicos y legítimos, y no deben utilizarlos para otros fines sin el consentimiento del/la titular—; la minimización de datos —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 las/os titulares de los datos tienen a acceder a sus datos personales y solicitar su corrección o eliminación si son inexactos o innecesarios—.

En el aspecto estrictamente normativo, en nuestro país existe la *Ley Nacional de Protección de Datos Personales* (Ley N° 25.326), que data del año 2000. El órgano de control establecido por dicha ley, la *Agencia de Acceso a la Información Pública* se encuentra —a la fecha de elaboración de esta guía— inmersa en un proceso de modificación de esa Ley. Asimismo, el país cuenta con normativa nacional más específica y dispersa en el cuerpo jurídico, entre las cuales de destacan normas como la *Reglamentación de la Ley de Protección de Datos Personales* (Decreto N° 1558/2001), que establece las pautas para la aplicación de la *Ley de Protección de Datos Personales*; o la *Ley de Protección de Datos Personales en el Ámbito Telefónico* (Ley N° 26.951), 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 se puedan recuperar o extraer 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, se deben implementar 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 sean informadas sobre qué datos se recopilan, cómo se utilizan y con quién se comparten, y 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, ya que al respetar la privacidad y proteger los mismos 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 los mismos.

Box - 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 aquéllas/os, respetando los principios del secreto profesional.

4.6. 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 éstos estén presentes en los datos utilizados para entrenar a estos algoritmos.

Los algoritmos de IA son entrenados con datos, y no solo aprenden los sesgos presentes en los mismos sino que pueden amplificarlos. Por ejemplo, imaginemos que los trabajadores 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 tal manera que se podría observar 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, trato desigual, o dar lugar a un aumento en las brechas y la exclusión de ciertos grupos sociales, es decir que pueden contribuir al aumento de la injusticia en general. Es por ello que las iniciativas orientadas a solucionar estos problemas se denominan de justicia algorítmica.

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 a esta pregunta es afirmativa, y esto puede depender del tipo de algoritmo utilizado. En este sentido, es crucial no solo 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, pasando por el prototipado del algoritmo y, finalmente, en el algoritmo en su estado de producción.

Volviendo al ejemplo previamente mencionado 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, se hace necesario incorporar la variable de género en el proceso de entrenamiento del algoritmo, lo que permitirá que, a través de su interacción con otras variables, se pueda corregir 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, mientras que si se le pide a los algoritmos 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 evaluación rigurosa de los conjuntos de datos utilizados en el entrenamiento de los algoritmos, de manera tal que se garantice el acceso a datos de calidad, con un volumen y variedad que los haga equitativamente representativos de diferentes grupos de personas, industrias o segmentos a los que se aplican los algoritmos de IA en cada caso. Además, se debe probar y testear el algoritmo una vez entrenado para que no contenga sesgos en ninguna de las dimensiones relevantes (ver Box - Detección de sesgos en la práctica). Sin embargo, la inexistencia de sesgos nunca se podrá asegurar en un 100%, por lo que se deben llevar a cabo la mayor cantidad de pruebas que permitan lograr los resultados más justos e inclusivos posibles.

Es importante tener en cuenta que la justicia no solo 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 de manera que contribuyan a un trato justo y equitativo para todas las personas, sin importar su origen, género, raza, orientación sexual u otras características protegidas.

Box - Detección de sesgos en la práctica

Se deben tener en cuenta las características específicas de los diferentes grupos de población objetivo de cada algoritmo de IA para determinar si existen sesgos a la hora de aplicarlos. Los posibles sesgos que pueden ser considerados son por:

- grupo etario
- grupo social
- lugar geográfico de vivienda
- nivel de educación
- nacionalidad
- sistema cultural
- etnia
- grupo lingüístico
- género
- niña/o menor de edad
- discapacidad
- grupo desfavorecido, marginado o en situación de vulnerabilidad
- toda otra característica específica del contexto de aplicación de cada algoritmo

4.7. 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 es que codifican la información necesaria para resolver los problemas que les son planteados, ni 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. Es decir, la transparencia para estos algoritmos de IA 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, 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, se debe informar que éste puede ser utilizado de maneras novedosas y creativas, no necesariamente previstas por los y las diseñadores y programadores en primer lugar. 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.

4.8. 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 los 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 en la cual tiene que decidir si priorizar el bienestar de los peatones o el de los pasajeros del vehículo. Este tipo de dilemas no tienen respuestas definitivas, y poseen la dificultad adicional de que los sistemas basados en IA pueden tomar decisiones mucho más rápido que lo que pueden reaccionar los 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, y por ello es fundamental que existan mecanismos de rendición de cuentas y sistemas de responsabilidad legal claros para garantizar que se tomen las medidas adecuadas en caso de problemas. 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, es cada vez 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 hacer a las personas y organizaciones que las crean responsables de responder por ellas.

5. Guía de desarrollo ético de sistemas basados en IA

A continuación se presenta la guía de desarrollo ético de sistemas basados en IA propiamente dicha, que brinda a los equipos de trabajo, diseñadores, desarrolladores y evaluadores de proyectos basados en IA una referencia para generar sistemas robustos y confiables con respecto a la ética del desarrollo de estas tecnologías. El desarrollo de sistemas basados en IA se puede dividir de forma general en las siguientes cinco etapas, para cada una de las cuales son relevantes distintas partes de la guía, y que son indicadas en cada caso:

1. Diseño:
    - Es un bosquejo del propósito del algoritmo de IA, sus funciones, estructura y características generales. Ver sección 5.1. Diseño del algoritmo de IA.
2. Prototipo:

- Implica el desarrollo de un prototipo inicial, que permita demostrar el correcto funcionamiento del algoritmo de IA. Requiere una recolección inicial de datos y su debido preprocesamiento y anonimización. Ver secciones 5.2. Recolección de datos e integración de fuentes a 5.5. Testeo del sistema de IA y pruebas de justicia algorítmica inclusive.

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 requiere hacer una recolección y preprocesamiento exhaustivo de los datos de entrenamiento, así como el testeo final del rendimiento del algoritmo. Ver secciones 5.2. Recolección de datos e integración de fuentes a 5.5. Testeo del sistema de IA y pruebas de justicia algorítmica inclusive.
4. Implementación en producción:
    - A partir de la salida del sistema a su despliegue y uso en el mundo real. Ver sección 5.6. Implementación del sistema de IA en producción.
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 5.7. Desarrollo continuo del sistema de IA.

5.1. Diseño del algoritmo de IA

Definir el propósito del algoritmo de IA (ver principio ético de 4.1. Propósito). Dar un diagnóstico del contexto en el cual va a ser aplicado. Definir posibles usos y aplicaciones del algoritmo, y el contexto de su aplicación (e.g. población objetivo, qué servicio ofrece, etc).	
¿Se consultó con especialistas en el/las área/s de aplicación del algoritmo de IA? ¿Cuáles fueron sus recomendaciones y/o advertencias? (Ver principio ético de 4.2. Colaboración con expertos en la materia de aplicación).	
¿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 4.3. Seguridad).	

5.2. Recolección de datos e integración de fuentes

¿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 4.4. Trazabilidad).	
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 4.5. Protección de datos y privacidad).	
¿Cuál es la calidad de los datos? ¿Hay datos faltantes? ¿Contienen errores o inconsistencias? (e.g. por carga manual). ¿Hay algún procedimiento para detectar esto?	
¿Cubren los datos los casos de uso intencionados? ¿Se integran datos de distintas fuentes? ¿Sobran datos? ¿Por qué?	
¿Se actualizan los datos? ¿Con qué frecuencia? Si no se actualizan, ¿son datos fijos que no pueden cambiar?	
¿Se almacenan los datos? ¿De manera segura? ¿Hay consentimiento y aviso en caso de ser datos personales o privados? ¿Pueden los usuarios acceder a los datos personales sobre ellos mismos que estén almacenados? ¿Puede un usuario decidir que sus datos personales sean borrados de la base de datos?	
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 4.6. Sesgos y justicia algorítmica).	

5.3. Preprocesamiento y etiquetado de datos

¿Es necesario almacenar todas las variables que actualmente se almacenan? ¿Son todas necesarias para el funcionamiento de los algoritmos?	
¿Se anonimizan los datos? ¿Cómo? ¿Se pueden re-identificar posteriormente a los individuos? Por ejemplo, cruzando distintas fuentes de datos externas que no sean parte del presente desarrollo (ver principio ético de 4.5. Protección de datos y privacidad).	
¿Cómo se realiza el etiquetado de los datos? ¿El etiquetado de los datos está alineado con el propósito del algoritmo de IA? Si el etiquetado proviene de fuentes diversas, ¿son coherentes las distintas fuentes?	

5.4. Definición del algoritmo, función objetivo y entrenamiento

Definición de los algoritmos de optimización que van a ser probados para su entrenamiento. ¿Son explicables o de caja negra? (ver principio ético 4.7. Explicabilidad y transparencia).	
Definición de la función objetivo. ¿Esta se alinea totalmente con el propósito del sistema, o es aproximada? Considerar problemas de seguridad que esta función objetivo podría tener (ver principio ético 4.3. Seguridad).	
Definición de métricas de rendimiento del algoritmo de IA (e.g. precisión,	

exhaustividad, etc). ¿Éstas están alineadas a los objetivos del algoritmo?	
¿Por qué se seleccionó el algoritmo que quedó finalmente? ¿Qué hiper-parámetros se probaron? Documentar todos los resultados (ver principio ético 4.4. Trazabilidad).	
¿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?	

5.5. Testeo del sistema de IA y pruebas de justicia algorítmica

¿Cuál es el resultado obtenido de las métricas de rendimiento finales del algoritmo? ¿En qué datos se obtuvieron estos resultados?	
Realizar un análisis de sesgos en el algoritmo finalizado. ¿Qué criterios de sesgos se probaron? Definir valores aceptables de sesgos (ver principio ético 4.6. Sesgos y justicia algorítmica).	
Evaluación realizada por usuarios 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 es posible que presente fallas).	

5.6. Implementación del sistema de IA en producción

¿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?	
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 4.6. Sesgos y justicia algorítmica).	
Evaluación realizada por usuarios finales. Crear una base de datos centralizada de reportes de incidentes. Evaluar cómo resolver estos incidentes.	

5.7. Desarrollo continuo del sistema de IA

La mayoría de los algoritmos de IA se degradan con el tiempo luego de su implementación en producción, en función de cómo se desempeña frente a nuevos datos o contextos. Además, la variación en sus respuestas es cada vez más grande a medida que pasa el tiempo (“la IA envejece”). Es por ello que 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.

Monitoreo continuo del rendimiento del algoritmo de IA. En caso de degradación se debe volver a re-entrenarlo utilizando datos nuevos.	
En caso de agregar nuevas funcionalidades o modificar las ya existentes, se debe volver a iniciar el chequeo de la guía ética desde el comienzo 5.1. Diseño del algoritmo de IA.	

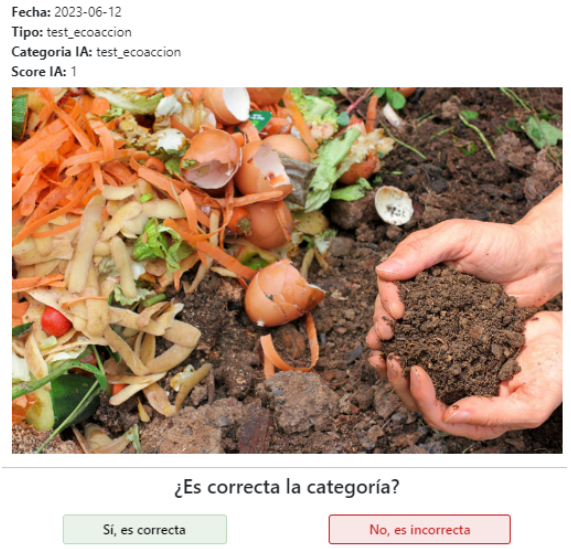
6. Casos de uso

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 se encuentran en desarrollo o fueron implementados por parte de la Subsecretaría de Políticas Públicas Basadas en Evidencia (SSPPBE) del Gobierno de la Ciudad de Buenos Aires (GCBA). El primero es un clasificador de residuos reciclables, por parte del proyecto de *Ecopuntos*, y el segundo es un recomendador de capacitaciones o cursos específicamente adaptados a cada usuario en particular, según su historial de cursos realizados o sus preferencias. Dado que la guía práctica fue desarrollada para poder ser aplicada 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, y por lo tanto solo se tomarán en cuenta para cada uno de los proyectos aquellas preguntas que sean relevantes para cada uno de ellos.

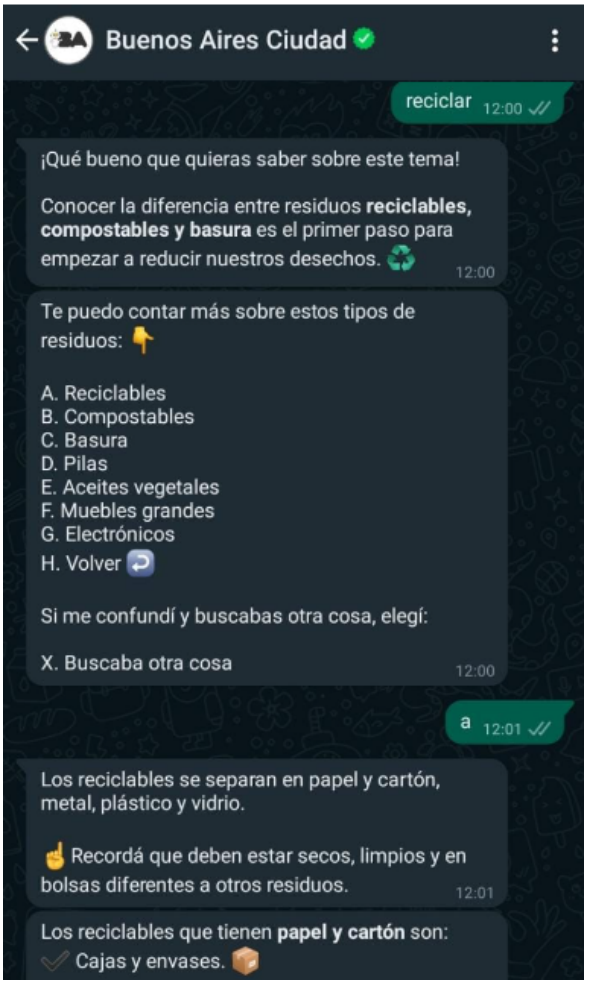
6.1 Ecopuntos

Diseño del algoritmo de IA

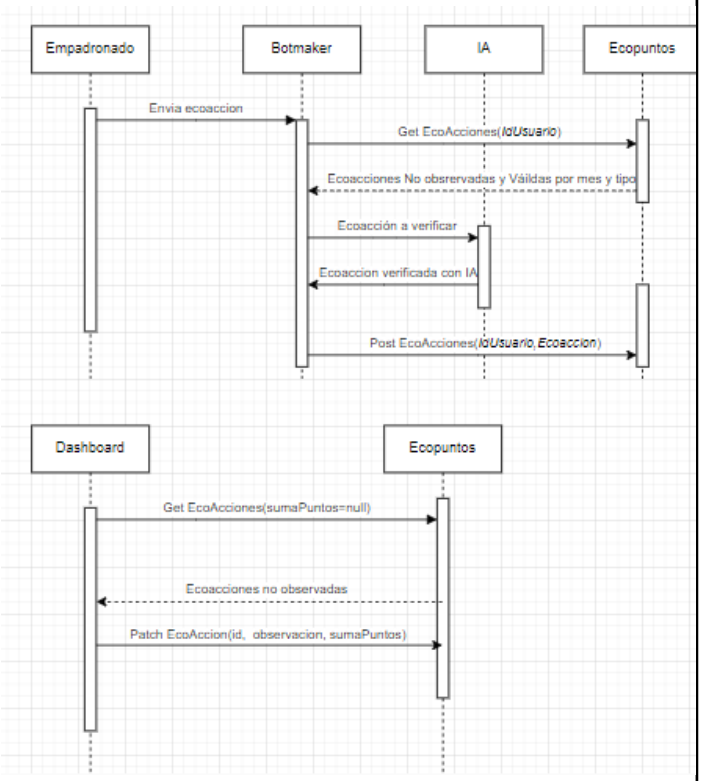
Definir el propósito del algoritmo de IA (ver principio ético de 4.1. Propósito). Dar un diagnóstico del contexto en el cual va a ser aplicado. Definir posibles usos y aplicaciones del algoritmo, y el contexto de su aplicación (e.g. población objetivo, qué servicio ofrece, etc).	El algoritmo de IA de Ecopuntos es de propósito particular y tiene como fin la categorización de fotos (proporcionadas por los usuarios) de materiales para el reciclado, con el objetivo de segmentar diferentes tipos de reciclables de manera automática. El algoritmo será utilizado en el programa Ecopuntos 2023, el cual se podrá acceder desde el chatbot de la Ciudad de Buenos Aires, Boti. El objetivo es mejorar la gestión de residuos por parte de los vecinos a través de la gamificación de los diferentes hábitos sustentables y dar premios a quienes más puntaje logran según distintas iniciativas que conforman el programa Ecopuntos. La presente versión del algoritmo de clasificación de Ecopuntos se entrenó con datos subidos por los usuarios durante el año 2021. Además, tiene el objetivo de descomprimir el trabajo del Observatorio de reciclables de Ecopuntos (que realizan una clasificación manual/visual de los residuos), y proveer una clasificación certera de los diferentes residuos que se depositan en la Ciudad de Buenos Aires, a fines de permitir la escalabilidad de las clasificaciones. El mismo sirve a su vez como una guía para los vecinos, a fines de aconsejar o sugerir formas de reciclar los diferentes tipos de desechos que sean consultados.

	Para más información se puede dirigir a la página del proyecto: https://ciudadverde.gob.ar/ecopuntos/.
¿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 4.3. Seguridad).	Podría existir el riesgo de que la IA haga predicciones equivocadas y que por esa razón ocurra que un usuario decida enviar un residuo a un destino equivocado (basado en una clasificación errónea que el algoritmo de IA de Ecopuntos haya realizado). Sin embargo, en la versión del programa de 2023 todavía funcionará un Observatorio que hace chequeos manuales/visuales (humanas) de las clasificaciones realizadas por el algoritmo de IA, es decir que la clasificación del algoritmo en ningún caso será tomada como la clasificación de residuos final en la versión del programa en 2023. El Observatorio de reciclables consiste en un validación o clasificación manual/visual de las fotos subidas a través de la siguiente interfaz gráfica:  Asimismo, podría ocurrir que un usuario no suba simplemente una foto de un residuo, sino que suba una selfie con su rostro al sistema de Boti, en cuyo caso será necesario tomar medidas para que dicha foto sea borrada o permanezca de manera confidencial en el sistema.

Recolección de datos e integración de fuentes

¿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 4.4. Trazabilidad).	Los datos (fotos) utilizados fueron subidos por los usuarios a través de Whatsapp, en una experiencia conversacional gamificada de Ecopuntos a la que se puede acceder a través de Boti, el chatbot del Gobierno de la Ciudad de Buenos Aires (GCBA). Los mismos 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, los cuales igualmente no hacían referencia explícita al uso de las fotos para ser utilizados como datos de entrenamiento de un algoritmo de IA, sino que hacía referencia al uso generalizado de estos datos. 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 que sea clasificado: 
Describir de manera general los datasets que se utilizan para entrenar	Para entrenar el algoritmo de IA se utilizaron fotos subidas por los usuarios de Boti en formato JPEG,

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 4.5. Protección de datos y privacidad).	cada una con alrededor de 200 KB de peso en promedio. Las mismas fueron clasificadas por los mismos usuarios en siete categorías posibles: aceite, secos, botellas, compost, electrónicos, orgánicos o baterías. Las fotos utilizadas se subieron durante los 4 meses de desarrollo del programa Ecopuntos en el año 2021, que contó con alrededor de 3000 usuarios que subieron fotos, y aproximadamente 18.000 fotos subidas en total. Con respecto a datos confidenciales o personales, existen fotos enviadas por los usuarios en donde se pueden ver sus rostros y/o diferentes espacios de sus hogares.
¿Cuál es la calidad de los datos? ¿Hay datos faltantes? ¿Contienen errores o inconsistencias? (e.g. 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 los usuarios, por lo que se recomienda realizar una validación manual/visual por parte del Observatorio de las clasificaciones realizadas por los usuarios que suben las fotos.
¿Cubren los datos 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 cumplen 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.
¿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 serán subidas durante la próxima versión del programa de Ecopuntos que se desarrollará durante 2023. Se contempla pedir a los usuarios que las fotos se tomen de maneras determinadas para facilitar el trabajo tanto del Observatorio como del algoritmo de IA.

¿Se almacenan los datos? ¿De manera segura? ¿Hay consentimiento y aviso en caso de ser datos personales o privados? ¿Pueden los usuarios acceder a los datos personales sobre ellos mismos que estén almacenados? ¿Puede un usuario decidir que sus datos personales sean borrados de la base de datos?	Los datos que se usaron 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, en donde se guardan los links a las fotos. Para la edición actual del programa el almacenamiento de Google almacena los links, los cuales se pueden revisar desde una nueva base de datos que estará alojada en la infraestructura de ASI. Se puede ver en el siguiente diagrama el flujo de los datos que corresponden al desarrollo del programa de Ecopuntos:  Actualmente, un usuario 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 que todo vecino tiene de realizar un pedido formal al GCBA para conocer los datos que se guardan sobre el mismo).
Realizar un análisis preliminar de sesgos y problemas de representatividad en los datos. ¿Se	Un sesgo de los datos es que las fotos corresponden a residuos de los ciudadanos de la Ciudad de Buenos Aires, por lo que el algoritmo

hacen suposiciones sobre los datos? ¿Qué criterios de sesgos y representatividad se probaron? Definir valores aceptables de sesgos (ver principio ético 4.6. Sesgos y justicia algorítmica).	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 para aquellos que 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 la Ciudad de Buenos Aires. Sin embargo, la aplicación en otros ambientes con otras lógicas de reciclado y/o con otros dispositivos para interactuar con el chatbot podrían modificar los resultados del algoritmo como está entrenando actualmente.

Implementación del sistema de IA en producción

¿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

Monitoreo continuo del rendimiento del algoritmo de IA. En caso de degradación se debe volver a re-entrenarlo utilizando datos nuevos.	Todavía no se implementó este proceso, pero uno de los objetivos es re-entrenar el algoritmo con las imágenes nuevas subidas por los usuarios que el algoritmo actual no haya podido clasificar correctamente.
En caso de agregar nuevas funcionalidades o modificar las ya existentes, se debe volver a iniciar el chequeo de la guía ética desde el comienzo 5.1. Diseño del algoritmo de IA.	Existe la posibilidad de agregar nuevas funcionalidades al algoritmo de clasificación de reciclables, como por ejemplo escalar la aplicación a diferentes latitudes (como por ejemplo a distintas provincias), agregar otras fuentes de datos (como videos), o agregar nuevas categorías de clasificación de reciclables.

6.2. Recomendador de capacitaciones

Diseño del algoritmo de IA

Definir el propósito del algoritmo de IA (ver principio ético de 4.1. Propósito). Dar un diagnóstico del contexto en el cual va a ser aplicado. Definir posibles usos y aplicaciones del algoritmo, y el contexto de su aplicación (e.g. población objetivo, qué servicio ofrece, etc).	El propósito del algoritmo de IA es poder recomendar a la ciudadanía un conjunto de capacitaciones o cursos ofrecidos por el Gobierno de la Ciudad de Buenos Aires (GCBA), con el fin de fomentar la formación de las personas y así aumentar la posibilidad de empleabilidad de las mismas. Para poder generar una recomendación el algoritmo tiene en cuenta tanto el historial de cursos realizados por las personas usuarias como así también sus intereses. De esta forma se puede realizar una recomendación personalizada para cada persona
¿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 4.3. Seguridad).	No hay un impacto negativo, ni existen riesgos para usuarios/as del Recomendador, en razón de que el algoritmo sólo analiza las variables de "cursos previamente realizados" e "intereses" para perfilar a las/os usuarias/os, por lo que 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 sirve para priorizar las capacitaciones y la oferta no se restringe.

Recolección de datos e integración de fuentes

¿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 4.4. 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. Si, hay consentimiento, ya que cada área firma un formulario en el cual establecen las condiciones del uso y tratamiento de los datos. En los casos de utilización de código, siempre se documentan.

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 4.5. Protección de datos y privacidad).	Se utilizan 3 tablas, las mismas son tomadas del datalake. Con la información de las tablas 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 las mismas. Conforme a lo detallado, no hay variables que contengan información personal ni confidencial.
¿Cuál es la calidad de los datos? ¿Hay datos faltantes? ¿Contienen errores o inconsistencias? (e.g. por carga manual). ¿Hay algún procedimiento para detectar esto?	Para evaluar la calidad de los datos, realizamos un procedimiento mediante el cual se utiliza un framework que evalúa la completitud y exactitud de los datos. No hay datos faltantes, ya que se utilizan los cursos realizados por cada persona. Tampoco hay errores o inconsistencias en la carga de datos.
¿Cubren los datos los casos de uso intencionados? ¿Se integran datos de distintas fuentes? ¿Sobran datos? ¿Por qué?	Los datos, si cubren los casos de uso intencionados. A través de los mismos se puede acceder al historial de cursos realizados por el usuario, luego generar la matriz cuadrada nombrada anteriormente, para luego generar la recomendación. No sobran datos, ya que se toman únicamente los datos que se necesitan.
¿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 1 mes se reentrena la matriz que calcula la similaridad entre cursos.
¿Se almacenan los datos? ¿De manera segura? ¿Hay consentimiento y aviso en caso de ser datos personales o privados? ¿Pueden los usuarios acceder a los datos personales sobre ellos mismos que estén almacenados? ¿Puede un usuario 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: https://portaldeoportunidades.buenosaires.gob.ar/home Si, pueden acceder a sus datos almacenados, como así también solicitar que sus datos sean eliminados, borrados o actualizado s conforme lo que establece 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.

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 4.6. Sesgos y justicia algorítmica).	Posibles sesgos que puede tener un algoritmo de recomendación de este estilo son: - Sesgo de muestreo: el dataset usado para entrenar el recomendador no es representativo de la población objetivo, por lo que podría hacer recomendaciones erradas. Lo mitigamos 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, lo que puede reforzar sesgos en las recomendaciones. Lo solucionamos 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. Lo solucionamos incorporando los intereses además del historial. - Sesgo de falta de transparencia: Los usuarios pueden no confiar en el algoritmo ni en sus recomendaciones porque no saben cómo funciona. Lo solucionaremos al momento de publicar el código del algoritmo, junto a un documento técnico explicando en detalle cómo funciona.

Preprocesamiento y etiquetado de datos

¿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 serán utilizadas por el algoritmo, es decir un identificador único para cada usuario/a, los nombre de los cursos que realizó y sus intereses. Es indispensable contar con estos datos para el buen funcionamiento del algoritmo.
¿Se anonimizan los datos? ¿Cómo? ¿Se pueden re-identificar	En este tipo de proyecto no se anonimizan los datos ya que es importante poder identificar a cada persona para

posteriormente a los individuos? Por ejemplo, cruzando distintas fuentes de datos externas que no sean parte del presente desarrollo (ver principio ético de 4.5. Protección de datos y privacidad).	hacer una recomendación personalizada, pero los datos personales no son utilizados para generar la recomendación, es decir, no los ve el algoritmo. No se puede re-identificar a los/as usuarios/as porque los IDs no se exponen, son únicamente para uso interno.

Definición del algoritmo, función objetivo y entrenamiento

Definición de los algoritmos de optimización que van a ser probados para su entrenamiento. ¿Son explicables o de caja negra? (ver principio ético 4.7. Explicabilidad y transparencia).	Se realizó y entrenó un algoritmo denominado “sistema de recomendación”, que se trata de generar una “matriz de cursos y capacitaciones” en la que esté contenida la información de “qué cursos son más similares entre sí que otros”. Esta similaridad entre cursos permitirá luego recomendar a los usuarios cursos similares (en varios sentidos) a los cursos que ya realizaron 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.
Definición de la función objetivo. ¿Esta se alinea totalmente con el propósito del sistema, o es aproximada? Considerar problemas de seguridad que esta función objetivo podría tener (ver principio ético 4.3. Seguridad).	La función objetivo busca determinar la similaridad entre los distintos cursos. Esto se logra al tener en cuenta como más similares a aquellos cursos que las personas realizan en conjunto con mayor frecuencia (se asume que estos cursos son más similares que otros cursos que usualmente no son realizados por una misma persona).
Definición de métricas de rendimiento del algoritmo de IA (e.g. precisión, exhaustividad, etc). ¿Éstas están alineadas a los objetivos del algoritmo?	Se utilizó como métrica el accuracy. Para cada usuario se extrajeron los cursos que realizó. Se quitó de la lista un curso al azar para cada usuario. Se corrió el algoritmo de recomendación sobre los usuarios, mostrándole todos los cursos realizados por cada uno excepto el eliminado. Se registró en cada caso si el curso eliminado estaba entre los recomendados. El accuracy representa la

	proporción de recomendaciones que incluía un curso eliminado.
¿Por qué se seleccionó el algoritmo que quedó finalmente? ¿Qué hiper-parámetros se probaron? Documentar todos los resultados (ver principio ético 4.4. Trazabilidad).	Dada la naturaleza del problema y la cantidad de cursos y usuarios disponibles, la solución más acorde es utilizar un algoritmo de sistema de recomendación (de tipo filtro colaborativo item-item, es decir, curso-curso).
¿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, con el que cuenta 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

¿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, en base a los datos usados para el desarrollo, previos al 31/12/2022. Lo cual indica que se encuentra dentro del margen aceptable. Como se describió anteriormente, el algoritmo necesita como información los cursos que realizó el usuario y sus intereses. La métrica evalúa principalmente la recomendación que se realiza en base al historial del usuario, la cual es la más importante que tiene el algoritmo.
Realizar un análisis de sesgos en el algoritmo finalizado. ¿Qué criterios de sesgos se probaron? Definir valores aceptables de sesgos (ver principio ético 4.6. Sesgos y justicia algorítmica).	En razón de las variables con las que trabaja el algoritmo, de su análisis resulta que no está sesgado y ni posee problemas de representatividad. No se hacen suposiciones, ya que toma la totalidad de los datos de acuerdo a los cursos e intereses que el/la usuario/a elige para realizar las capacitaciones.

Evaluación realizada por usuarios 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 usuario, como así también poder 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 es posible que presente 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: ¿Qué pasa si el usuario envía un DNI incorrecto? ¿Qué sucede si el usuario no envía sus intereses? De nuestro lado hicimos pruebas buscando algunos usuarios en la base de datos, les inventamos intereses y corroboramos que el algoritmo devolvía una recomendación.

Implementación del sistema de IA en producción

¿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.
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 4.6. Sesgos y justicia algorítmica).	En razón de las variables con las que trabaja el algoritmo, de su análisis resulta que no está sesgado y ni posee problemas de representatividad.

Evaluación realizada por usuarios finales. Crear una base de datos centralizada de reportes de incidentes. Evaluar cómo resolver estos incidentes.	No se obtuvieron comentarios o reacciones por parte de los usuarios que impliquen la necesidad de modificar una recomendación realizada.

Desarrollo continuo del algoritmo de IA

Monitoreo continuo del rendimiento del algoritmo de IA. En caso de degradación se debe volver a re-entrenarlo utilizando datos nuevos.	Vamos a descargar los datos actualizados para evaluar el algoritmo y reentrenarlo con cierta frecuencia.
En caso de agregar nuevas funcionalidades o modificar las ya existentes, se debe volver a iniciar el chequeo de la guía ética desde el comienzo 5.1. Diseño del algoritmo de IA.	No se hicieron nuevas modificaciones. En caso de hacerla sin duda volveremos a tomar los parámetros recomendados por la guia

7. Referencias

Varios autores (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.

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

Martínez, M. V. et al. (2022). Innovar con Ciencia de Datos en el sector público. Fundación Sadosky.

Aguerre, C. et al. (2020). Inteligencia Artificial en América Latina y el Caribe. Ética, Gobernanza y Políticas. CETyS (Universidad de San Andrés) y fAIr LAC (Banco Inter-americano de Desarrollo, BID).

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.

Varios autores (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. Banco Inter-americano de Desarrollo, BID.

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

Varios autores (2022). Microsoft Responsible AI Standard. Microsoft.

Amodei, et al. (2016). Concrete Problems in AI Safety. OpenAI.

Madaio, et al. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. ACM Digital Library.

Inioluwa, et al. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. ACM Digital Library.

Vela, et al. (2022). Temporal quality degradation in AI models. Nature Scientific Reports.

8. Anexo complementario

Consideraciones especiales para chatbots (y algoritmos generativos en general)

Los algoritmos de IA generativos son aquellos que generan contenido de manera flexible, como texto, imágenes o videos, a partir de comandos de texto de entrada flexibles, llamados *prompts*. Existen ya en la actualidad muchos riesgos éticos asociados al desarrollo, despliegue y uso de algoritmos generativos. Principalmente porque todos los algoritmos generativos son de propósito general, es decir que no son creados con el fin de satisfacer una aplicación concreta, sino que pueden ser utilizados de manera flexible y creativa por los usuarios.

Algunos de los riesgos éticos actuales asociados a los algoritmos generativos son la generación de voces artificiales o que simulen la voz de personas reales (usurpación de la voz), y que por lo tanto pueden ser utilizados con fines delictivos. También se pueden generar imágenes o videos artificiales que involucren personas específicas de la realidad, pero que muestran acontecimientos y acciones que nunca realizaron (*deepfake*). Por otro lado, también es posible crear algoritmos de IA que generen automáticamente noticias falsas, y al mismo tiempo es posible generar un gran número de opiniones automáticas

sobre distintos temas a partir de usuarios ficticios, de manera que es posible fabricar opinión pública para que ciertas miradas aparenten tener mayor envergadura que las que realmente tienen en la sociedad.

Un tipo especial de algoritmos generativos son aquellos que generan texto, y en particular aquellos que tienen la capacidad de interactuar mediante un chat con personas reales, denominados chatbots. Se presentan entonces algunas recomendaciones para un desarrollo confiable y ético de este tipo de algoritmos generativos, los chatbots.

Primero, al igual que para todo desarrollo de algoritmos de IA, es importante tener claro el propósito del chatbot, sus alcances y limitaciones. En especial se debe prestar mayor atención al uso de chatbots 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 actualmente puede reemplazar. A su vez, deben estar debidamente notificados los costos o riesgos de los errores que estos chatbots pudieran cometer, y los límites de su rendimiento.

Por otro lado, se debe mencionar claramente que (y cuando) las respuestas del chatbot son generadas por un algoritmo y no por un ser humano, ya que es probable que esto sea difícil o incluso imposible de diferenciar sin previo aviso, independientemente de la habilidad o experiencia de los usuarios.

El chatbot debe poder detectar ofensas, discursos de odio, o de temas controversiales por parte de los usuarios, de tal manera de poder responder apropiadamente ante esos ataques. Además, se debe asegurar la privacidad de los datos personales o confidenciales con los que el chatbot fue entrenado, y que por lo tanto están almacenados implícitamente en los algoritmos de respuesta.

Es importante notar además que de ninguna manera se puede asegurar en la actualidad 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. Es necesario entonces cerciorarse a través del conocimiento experto (humano) sobre todo en aplicaciones sensitivas de los mismos, como en las áreas de salud, empleo, finanzas, legales, etc.

Además, existen métodos para que los usuarios de un chatbot puedan hacer que éste diga cualquier cosa que los usuarios quieren que diga. Esto se denomina “prompt hacking”, y significa que un usuario puede fabricar discursos de tal manera que parezcan generados por los chatbots. A su vez, con estos métodos es también posible hacer que un chatbot responda con información confidencial con la que fue entrenado, incluso cuando el chatbot fue programado explícitamente para no revelar esta información. Esto es especialmente importante para la protección de datos y privacidad, ya que los chatbots están en general entrenados con una gran cantidad de datos que pueden estar almacenados en su algoritmo de forma implícita. Con lo cual si éstos fueron entrenados con datos privados o confidenciales, no se puede asegurar que estos datos no puedan ser recuperados por algunos de los usuarios.

Se recomienda entonces programar al chatbot para que detecte información privada o confidencial durante su conversación con usuarios, y que decida no almacenar esa información o, en caso de necesidad, almacenarla durante el menor tiempo posible. Se recomienda además proveer información detallada al usuario sobre la recolección de datos y pedir el consentimiento correspondiente de que estos sean recolectados (en caso de ser necesario recolectarlos).

Se recomienda además, para maximizar la seguridad del chatbot, agregar funciones que aporten a la privacidad de las personas y generen confianza en el mismo. Por ejemplo, se pueden agregar botones con funciones como “Dime todo lo que sabes sobre mí”, u “Olvida nuestra última interacción”, o “Borra todo lo que sabes sobre mí”, etc. En algunos casos estas funciones podrían ser incluso requeridas legalmente. También se debe dar un lugar para que los usuarios puedan hacer devoluciones, por ejemplo si los chatbots sirvieron con éxito para su finalidad prevista, y en caso de una devolución negativa, el chatbot debe proveer a los usuarios una forma de continuar la comunicación a través del contacto con humanos.

Todos los elementos anteriores deben estar debidamente documentados y estos deben ser de fácil acceso por los usuarios del chatbot, incluido un código de conducta para los usuarios entiendan los alcances del servicio, sus limitaciones y la forma de dirigirse al chatbot para su efectivo uso y funcionamiento.