ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/337050086

Guía y herramientas para el diseño responsable y la implementación de sistemas de IA | Guide and tools for responsible design and implementation of AI systems

Preprint · November 2019

DOI: 10.13140/RG.2.2.15303.24486

CITATIONS

0

1 author:

Juan Antonio Lloret Egea

FORMAEMPLEO

3 PUBLICATIONS **0** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Los estados de la Inteligencia Artificial | The states of Artificial Intelligence View project

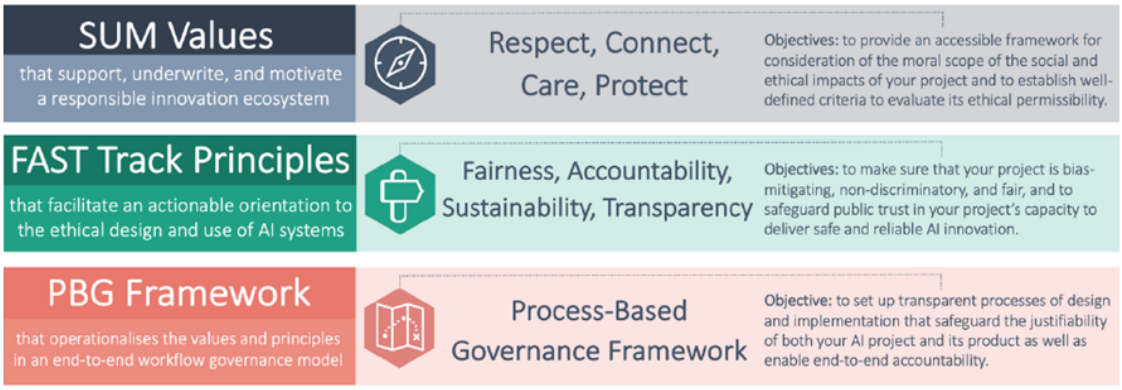
All content following this page was uploaded by Juan Antonio Lloret Egea on 06 November 2019.

The user has requested enhancement of the downloaded file.

Capítulo 9-2º | Chapter 9-2º

Los 7 mandamientos de la IA | The 7 commandments of AI

Guía y herramientas para el diseño responsable y la implementación de sistemas de IA | Guide and tools for responsible design and implementation of AI systems



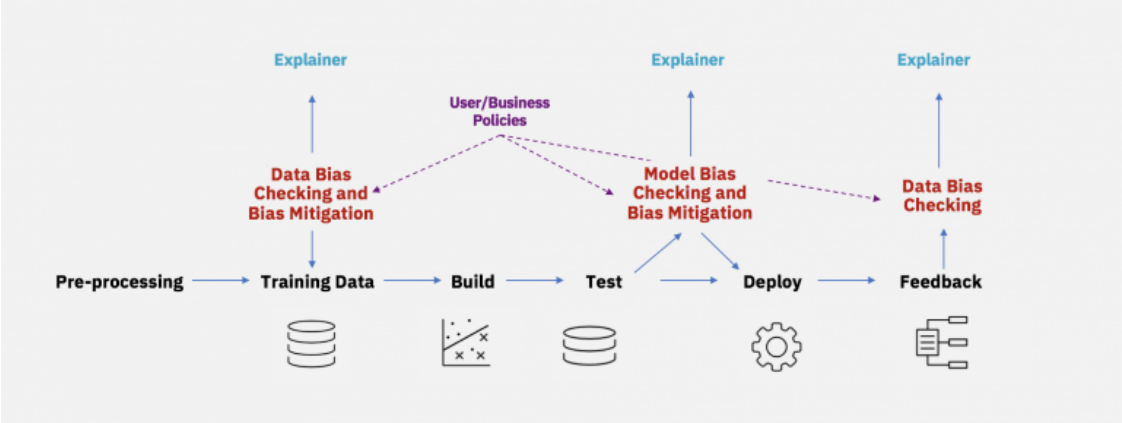


Fig. I A. 9.2.1- Plataforma ética para la entrega responsable de un proyecto de IA **&** Mitigando el sesgo a lo largo del ciclo de vida de AI | Ethical Platform for the responsible delivery of an AI project **&** Mitigating bias throughout the AI lifecycle. Crédito imag. (C9.1-1, C9.1-2). URL: https://doi.org/10.5281/zenodo.3240529 **&** https://www.ibm.com/blogs/research/wp-content/uploads/2018/09/IBM-Research-AI-Fairness-360-768x291.png

---

Autor / Author*:* Juan Antonio Lloret Egea | Miembro de la Alianza Europea para la IA / Member to the European AI Alliance | https://orcid.org/0000-0002-6634-3351|

© 2019. Licencia de uso y distribución / License for use and distribution: [Los estados de la inteligencia artificial (IA) | The states of artificial intelligence (AI)] creative commons CC BY-NC-ND |

ISSN 2695-3803|| Escrito / Writed: 23/10/2019. Actualizado / Updated: 05/11/2019 |

9.2-1.- Introducción | Introducction

Habitualmente encontramos abundante material teórico sobre la descripción, modelación, desarrollo e implementación de determinados modelos u objetivos. Menos frecuente es, sin embargo, encontrar a su vez los modelos prácticos sobre cómo llevar a cabo estos modelos teóricos. Por lo que, a juicio nuestro, este capítulo quizá sea uno de los más determinantes operacionalmente y documentalmente en el campo de la inteligencia artificial a la hora de su implementación teorico-práctica.

Entre sus contenidos se encuentra un extracto con carácter educativo e ilustrativo del estudio: *Comprender la ética y la seguridad de la inteligencia artificial: una guía para el diseño responsable y la implementación de sistemas de inteligencia artificial en el sector público*. Su utilidad en nuestra opinión resulta incuestionable como **eje directriz** (1ª sección). También se incluyen herramientas para la **explicabilidad**, **equidad** y **trazabilidad** de los sistemas IA para hacer más asequible su desarrollo real (2ª sección). Y por último las **conclusiones y recomendaciones** o pautas guiadas para obtener prototipos iniciales sujetos a prescripciones suficientemente garantes como para pensar que los nuevos desarrollos futuros de IA son fiables para el uso y bienestar del ser humano (3ª sección).

Por lo expuesto anteriormente, este capítulo se divide en tres secciones:

[English]

Usually we find abundant theoretical material about the description, modeling, development and implementation of certain models or objectives. Less frequent, however, is to find practical models on how to carry out these theoretical models. Therefore, in our opinion, this chapter may be one of the most operational and documentary determinants in the field of artificial intelligence at the time of its theoretical and practical implementation.

Among its contents is an educational and illustrative extract of the study: *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. Its usefulness in our opinion is unquestionable as a **guiding axis** (1st section). Tools for the **explainability**, **equity** and **traceability** of AI systems are also included to make their real development more affordable (2nd section). And finally, the **conclusions and recommendations** or guided guidelines for obtaining initial prototypes subject to sufficiently guarantor prescriptions to think that the new future developments of AI are reliable for the use and well-being of the human being (3rd section).

Based on the foregoing, this chapter is divided into three sections:

PRIMERA SECCIÓN | FIRST SECTION:

Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector

Ésta es la guía más completa sobre el tema de la ética y la seguridad de la IA en el sector público hasta la fecha, (*Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*), en palabras de sus autores. El documento proporciona elementos integrales sobre cómo aplicar los principios de ética y seguridad de la IA al diseño e implementación de sistemas algorítmicos en el sector público. Sus autores enuncian que en breve publicarán un libro de trabajo para dar vida a las recomendaciones hechas.

La convergencia de la disponibilidad cada vez mayor de *Bigdata*, la velocidad vertiginosa y la expansión de las plataformas de computación en la nube, y el avance de algoritmos de aprendizaje automático cada vez más sofisticados, han dado paso a un momento notable de promesa humana. Las innovaciones en IA ya están dejando una huella en el gobierno, al mejorar la provisión de bienes y servicios sociales esenciales, desde la atención médica, la educación y el transporte hasta el suministro de alimentos, la energía y la gestión ambiental. Es probable que estas recompensas sean solo el comienzo. La posibilidad de que el progreso en IA ayude al gobierno a enfrentar algunos de sus desafíos más urgentes es emocionante, pero abundan las preocupaciones legítimas. Al igual que con cualquier tecnología nueva y en rápida evolución, una curva de aprendizaje abrupta significa que se cometerán errores y errores de cálculo y que se producirán impactos imprevistos y nocivos.

Identifica los daños potenciales causados por los sistemas de IA y propone medidas concretas y operables para contrarrestarlos. Se enfatiza que las organizaciones del sector público pueden anticipar y prevenir estos daños potenciales al administrar una cultura de innovación responsable y al establecer procesos de gobierno que respalden el diseño e implementación de sistemas de IA éticos, justos y seguros. Es relevante para todos los involucrados en el diseño, producción y despliegue de un proyecto de IA del sector público: desde científicos de datos e ingenieros de datos hasta expertos en dominios, gerentes de entrega y líderes departamentales.[C9.1-1].

[English]

This is the most complete guide on the topic of ethics and AI security in the public sector to date, (*Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*), in the words of its authors. The document provides comprehensive elements on how to apply the principles of ethics and security of AI to the design and implementation of algorithmic systems in the public sector. Its authors state that they will shortly publish a workbook to give life to the recommendations made.

The convergence of the increasing availability of big data, the dizzying speed and expansion of cloud computing platforms, and the progress of increasingly sophisticated machine learning algorithms, have given way to a remarkable moment of human promise . The innovations in AI are already leaving a mark on the government, by improving the provision of essential social goods and services,

from medical care, education and transportation to food supply, energy and environmental management. It is likely that these rewards are only the beginning. The possibility that progress in AI helps the government face some of its most urgent challenges is exciting, but legitimate concerns abound. As with any new and rapidly evolving technology, an abrupt learning curve means that errors and miscalculations will be made and that unforeseen and harmful impacts will occur.

It identifies the potential damages caused by AI systems and proposes concrete and operable measures to counteract them. It is emphasized that public sector organizations can anticipate and prevent these potential damages by administering a culture of responsible innovation and by establishing governance processes that support the design and implementation of ethical, fair and safe AI systems. It is relevant for everyone involved in the design, production and deployment of a public sector AI project: from data scientists and data engineers to domain experts, delivery managers and departmental leaders[C9.1-1].

9.2-2.- Tres bloques de construcción de un ecosistema responsable de entrega de proyectos de IA | Three building-blocks of a responsible AI project delivery ecosystem

---

Establecer una plataforma ética para la entrega responsable de proyectos de IA implica no solo construir desde la base cultural; implica proporcionar a su equipo los medios para lograr los objetivos de establecer la permisividad ética, la equidad, la confiabilidad y la justificación de su proyecto. Se necesitarán tres bloques de construcción para hacer posible una plataforma tan ética:

1. En el nivel más básico, es necesario que obtenga un conocimiento práctico de un marco de valores éticos que respalde, suscriba y motive un ecosistema de diseño y uso de datos responsable. Estos se denominarán valores SUM, y estarán compuestos por cuatro nociones clave: **Respetar**, **Conectar**, **Cuidar** y **Proteger**. Los objetivos de estos Valores SUM son (1) proporcionarle un marco accesible para comenzar a pensar sobre el alcance moral de los impactos sociales y éticos de su proyecto y (2) establecer criterios bien definidos para evaluar su permisividad ética.
2. En un segundo nivel más concreto, una plataforma ética para la entrega responsable de proyectos de IA requiere un conjunto de principios accionables que faciliten una orientación hacia el diseño y uso responsable de los sistemas de IA. Estos se llamarán Principios de Rastreo RÁPIDO, y estarán compuestos por cuatro nociones clave: **Justicia**, **Responsabilidad**, **Sostenibilidad** y **Transparencia**. Los objetivos de estos Principios de FAST Track son proporcionarle las herramientas morales y prácticas (1) para asegurarse de que su proyecto mitigue los prejuicios, no sea discriminatorio y justo, y (2) para salvaguardar la confianza pública en la capacidad de su proyecto para ofrecer innovación de IA segura y confiable.
3. En un tercer y más concreto nivel, una plataforma ética para la entrega responsable de

proyectos de IA requiere un marco de gobernanza basado en procesos (**PBG Framework**) que operacionaliza los Valores SUM y los Principios de Rastreo RÁPIDO en todo el flujo de trabajo de entrega de proyectos de IA. El objetivo de este Marco PBG es establecer procesos transparentes de diseño e implementación que salvaguarden y permitan la justificación tanto de su proyecto de IA como de su producto[C9.1-1].

[English]

Setting up an ethical platform for responsible AI project delivery involves not only building from the cultural ground up; it involves providing your team with the means to accomplish the goals of establishing the ethical permissibility, fairness, trustworthiness, and justifiability of your project. It will take three building-blocks to make such an ethical platform possible:

1. At the most basic level, it necessitates that you gain a working knowledge of a framework of ethical values that Support, Underwrite, and Motivate a responsible data design and use ecosystem. These will be called SUM Values, and they will be composed of four key notions: **Respect**, **Connect**, **Care**, and **Protect**. The objectives of these SUM Values are (1) to provide you with an accessible framework to start thinking about the moral scope of the societal and ethical impacts of your project and (2) to establish well-defined criteria to evaluate its ethical permissibility.
2. At a second and more concrete level, an ethical platform for responsible AI project delivery requires a set of actionable principles that facilitate an orientation to the responsible design and use of AI systems. These will be called FAST Track Principles, and they will be composed of four key notions: **Fairness**, **Accountability**, **Sustainability**, and **Transparency**. The objectives of these FAST Track Principles are to provide you with the moral and practical tools (1) to make sure that your project is bias-mitigating, non-discriminatory, and fair, and (2) to safeguard public trust in your project's capacity to deliver safe and reliable AI innovation.
3. At a third and most concrete level, an ethical platform for responsible AI project delivery requires a Process-Based Governance framework (**PBG Framework**) that operationalises the SUM Values and the FAST Track Principles across the entire AI project delivery workflow. The objective of this PBG Framework is to set up transparent processes of design and implementation that safeguard and enable the justifiability of both your AI project and its product[C9.1-1].

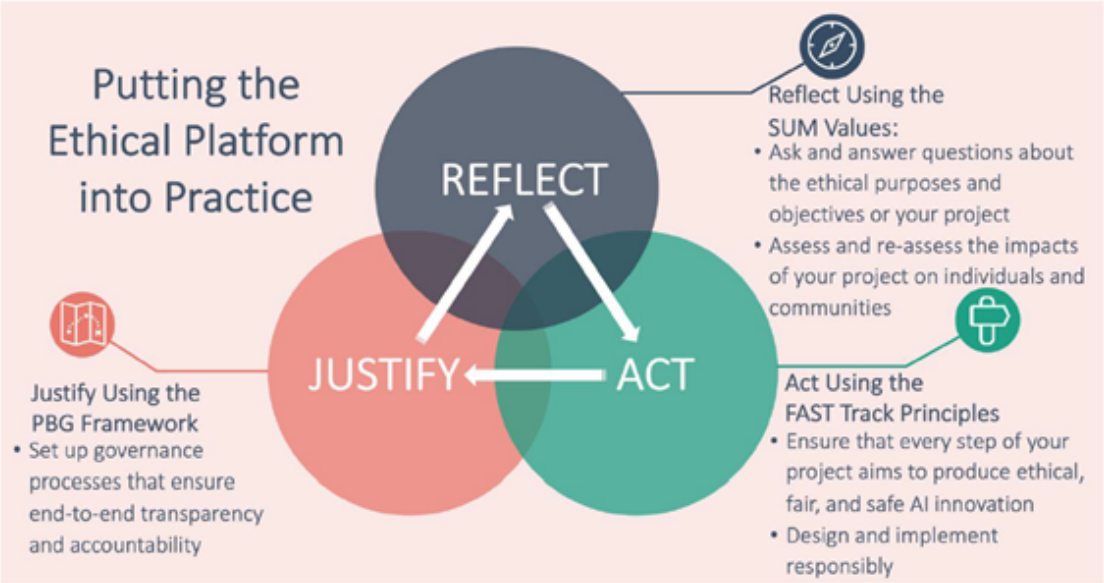


Fig. I A. 9.2.2- Poner en práctica la plataforma ética | Putting the Ethical Platform into Practice. Crédito imag. (C9.1-1). URL: https://doi.org/10.5281/zenodo.3240529

9.2-3.- Algunos de los principios de seguimiento FAST | Some of the FAST tracking principles

- **Justicia**: poner el principio de no-discriminación daños en acción | **Fairness**: putting the principle of discriminatory non-harm into action

Cuando esté considerando cómo poner en práctica el principio de no daño discriminatorio, debe reunirse con todos los gerentes del equipo del proyecto para mapear la participación de los miembros del equipo en cada etapa de la tubería del proyecto de IA desde alfa hasta beta. Tener en cuenta un diseño y una implementación que sean justos desde una perspectiva de flujo de trabajo le permitirá, como equipo, concretar y hacer caminos explícitos de responsabilidad de principio a fin de una manera clara y revisable por pares. Esto es esencial para establecer un marco de responsabilidad sólido. Aquí hay una representación esquemática del flujo de trabajo consciente de la equidad. Tendrá que completar la fila final[C9.1-1].

[English]

When you are considering how to put the principle of discriminatory non-harm into action, you should come together with all the managers on the project team to map out team member involvement at each stage of the AI project pipeline from alpha through beta. Considering fairness aware design and implementation from a workflow perspective will allow you, as a team, to concretise and make explicit end-to-end paths of accountability in a clear and peer-reviewable manner. This is essential for

establishing a robust accountability framework. Here is a schematic representation of the fairness aware workflow. You will have to complete the final row[C9.1-1].

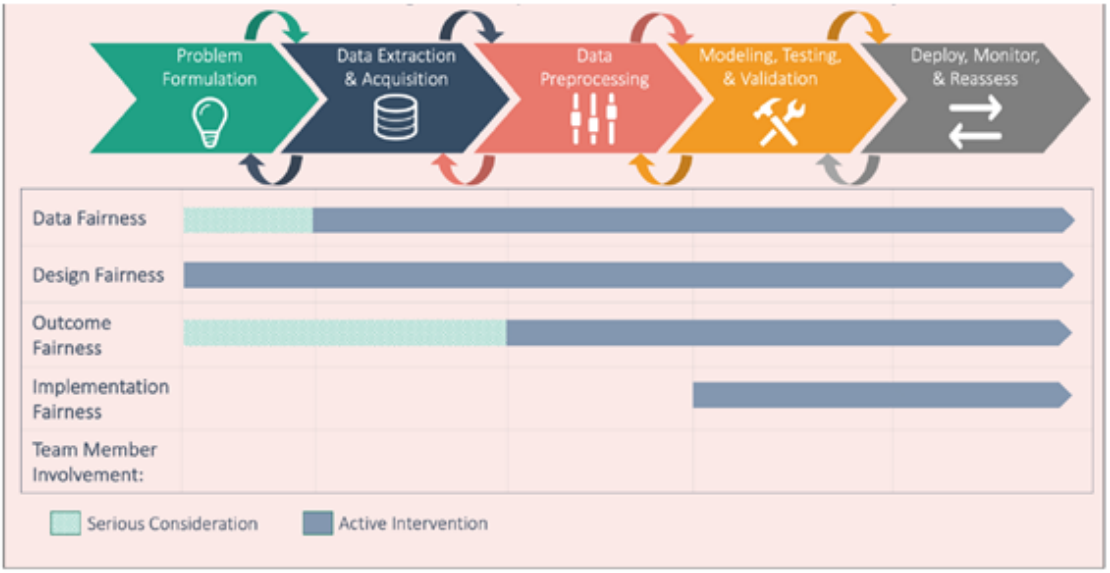


Fig. I A. 9.2.3- Flujo de trabajo de diseño e implementación justo para proyectos de IA | Fairness Aware Design and Implementation Workflow for AI Projects. Crédito imag. (C9.1-1). URL: https://doi.org/10.5281/zenodo.3240529

- **Responsabilidad**: la responsabilidad merece consideración en todo el flujo de trabajo de diseño e implementación | **Accountability**: accountability deserves consideration across the entire design and implementation workflow

Como práctica recomendada, debe considerar activamente las diferentes demandas que la responsabilidad del diseño le impone antes y después del lanzamiento de su proyecto de IA. Nos referiremos al proceso de garantizar la responsabilidad durante las etapas de diseño y desarrollo de su proyecto de IA como 'responsabilidad anticipada'. Esto se debe a que está anticipando las necesidades de responsabilidad de su proyecto de IA antes de que se complete. Siguiendo una lógica similar, nos referiremos al proceso de abordar la responsabilidad después del inicio del despliegue de su proyecto de IA como 'responsabilidad correctiva'. Esto se debe a que después de la implementación inicial de su sistema, está solucionando cualquiera de los problemas que pueden ser elevado por sus efectos y potenciales externalidades. Estos dos subtipos de responsabilidad a veces se denominan responsabilidad *ex ante* (o antes del evento) y responsabilidad *ex post* (después del evento) respectivamente.

Responsabilidad anticipatoria: el tratamiento de la responsabilidad como un principio anticipatorio implica que tome como importancia primordial las decisiones tomadas y las acciones tomadas por su

equipo de entrega del proyecto antes del resultado de un proceso de decisión respaldado algorítmicamente.

Este tipo de rendición de cuentas *ex ante* debe priorizarse sobre la rendición de cuentas correctiva, que se centra en cambio en las medidas correctivas o justificativas que se pueden tomar después de que se haya completado ese proceso de automatización.

Al garantizar que los procesos de entrega de proyectos de IA sean responsables antes de la aplicación real del sistema en el mundo, reforzará la solidez de los procesos de diseño e implementación y, por lo tanto, evitará de manera más efectiva posibles daños al bienestar individual y al bienestar público.

Del mismo modo, al establecer fuertes regímenes de responsabilidad anticipatoria y al hacer que el proceso de diseño y entrega sea lo más abierto y accesible al público posible, colocará a las partes interesadas afectadas en una posición para tomar decisiones mejor informadas y más informadas sobre su participación en estos sistemas antes de impactos potencialmente dañinos. Al hacerlo, también fortalecerá la narrativa pública y ayudará a proteger el proyecto del daño a la reputación[C9.1.1].

[English]

As a best practice, you should actively consider the different demands that accountability by design places on you before and after the roll out of your AI project. We will refer to the process of ensuring accountability during the design and development stages of your AI project as 'anticipatory accountability.' This is because you are anticipating your AI project's accountability needs prior to it being completed. Following a similar logic, we will refer to the process of addressing accountability after the start of the deployment of your AI project as 'remedial accountability.' This is because after the initial implementation of your system, you are remedying any of the issues that may be raised by its effects and potential externalities. These two subtypes of accountability are sometimes referred to as ex-ante (or before-the-event) accountability and ex-post (after-the-event) accountability respectively.

Anticipatory Accountability: Treating accountability as an anticipatory principle entails that you take as of primary importance the decisions made and actions taken by your project delivery team prior to the outcome of an algorithmically supported decision process.

This kind of ex ante accountability should be prioritised over remedial accountability, which focuses instead on the corrective or justificatory measures that can be taken after that automation supported process had been completed.

By ensuring the AI project delivery processes are accountable prior to the actual application of the system in the world, you will bolster the soundness of design and implementation processes and thereby more effectively pre-empt possible harms to individual wellbeing and public welfare.

Likewise, by establishing strong regimes of anticipatory accountability and by making the design and delivery process as open and publicly accessible as possible, you will put affected stakeholders in a

position to make better informed and more knowledgeable decisions about their involvement with these systems in advance of potentially harmful impacts. In doing so, you will also strengthen the public narrative and help to safeguard the project from reputational harm[C9.1-1].

- **Sustentabilidad**: evaluación de impacto de las partes interesadas | Sustainability: stakeholder Impact Assessment

---

Usted y su equipo de proyecto deben unirse para evaluar el impacto social y la sostenibilidad de su proyecto de IA a través de una Evaluación de Impacto de las Partes Interesadas (**SIA**), ya sea que el proyecto de IA se esté utilizando para prestar un servicio público o en una capacidad administrativa de *back-office*. Cuando nos referimos a '**partes interesadas**', nos referimos principalmente a personas individuales afectadas, pero el término también puede extenderse a grupos y organizaciones en el sentido de que los miembros individuales de estos colectivos también pueden verse afectados como tales por el diseño y la implementación de sistemas de IA. Debe prestarse la debida consideración a las partes interesadas en ambos niveles.

El propósito de llevar a cabo un SIA es multidimensional. Las SIA pueden servir para varios propósitos, algunos de los cuales incluyen:

1. Ayuda a generar confianza pública en que el diseño y la implementación del sistema de IA por parte de la agencia del sector público se ha realizado de manera responsable.
2. Facilitar y fortalecer su marco de responsabilidad.
3. Sacar a la luz riesgos invisibles que amenazan con afectar a las personas y al bien público.
4. Asegurar una toma de decisiones bien informada y prácticas de innovación transparentes.
5. Demostrar previsión y debida diligencia no sólo dentro de su organización sino también al público en general.

Su equipo debe reunirse para evaluar el impacto social y la sostenibilidad de su proyecto de IA a través del SIA en tres puntos críticos en el ciclo de vida de entrega del proyecto:

1. Fase alfa (**Formulación del problema**): lleve a cabo una evaluación de impacto de las partes interesadas (SIA) inicial para determinar la admisibilidad ética del proyecto. Consulte los valores SUM como punto de partida para conocer las posibles consecuencias de su proyecto en el bienestar individual y el bienestar público. En los casos en que concluya que su proyecto de IA tendrá impactos éticos y sociales significativos, debe abrir su SIA inicial al público para que sus puntos de vista puedan considerarse adecuadamente. Esto reforzará la inclusión de una diversidad de voces y opiniones en el proceso de diseño y desarrollo a través de la participación de una gama más representativa de partes interesadas. También debe considerar consultar con las partes interesadas internas de la organización, cuyo aporte también fortalecerá la apertura, la inclusión y la diversidad de su proyecto.
2. De alfa a beta (**pre-implementación**): una vez que su modelo ha sido entrenado, probado y validado, usted y su equipo deben revisar su SIA inicial para confirmar que el sistema de IA

que se implementará aún está en línea con las evaluaciones y conclusiones de su evaluación original. Este *check-in* debe registrarse en la sección de preimplementación del SIA con cualquier cambio aplicable agregado y discutido. Antes del lanzamiento del sistema, este SIA debe hacerse público. En este punto, también debe establecer un marco de tiempo para la reevaluación una vez que el sistema está en funcionamiento, así como una consulta pública que antecede y proporciona información para esa reevaluación. Los plazos para estas reevaluaciones deben ser decididos por su equipo caso por caso, pero deben ser proporcionales a la escala del impacto potencial del sistema en las personas y comunidades que afectará.

3. Fase beta (**reevaluación**): después de que su sistema de IA se haya activado, su equipo debe revisar y reevaluar su SIA de manera intermitente. Estos registros deben registrarse en la sección de reevaluación del SIA con cualquier cambio aplicable agregado y discutido. La reevaluación debe centrarse tanto en evaluar el SIA existente contra los impactos del mundo real como en considerar cómo mitigar las consecuencias no intencionadas que pueden haber surgido a raíz del despliegue del sistema. Se debe realizar una consulta pública adicional para obtener información en la etapa beta antes de la reevaluación, de modo que la contribución de los interesados se pueda incluir en las deliberaciones de reevaluación.

Debe tener en cuenta que, en su enfoque específico en la sostenibilidad social y ética, su evaluación de impacto de las partes interesadas constituye sólo una parte de la plataforma de gobernanza para su proyecto de IA y debe ser un complemento de su marco de responsabilidad y otra documentación de auditoría y monitoreo de actividades.

Su SIA debe dividirse en cuatro secciones de preguntas y respuestas. En la primera sección debería haber preguntas generales sobre los posibles impactos sociales y éticos generales del uso del sistema de IA que planea construir. En la segunda sección su equipo debe formular en colaboración preguntas relevantes específicas del sector y utilizar casos específicos sobre el impacto del sistema de IA en las partes interesadas afectadas. La tercera sección debe proporcionar respuestas a las preguntas adicionales relevantes para la evaluación previa a la implementación. La cuarta sección debe brindar la oportunidad a los miembros de su equipo de reevaluar el sistema a la luz de sus impactos en el mundo real, la opinión pública y las posibles consecuencias no deseadas[C9.1-1].

[English]

You and your project team should come together to evaluate the social impact and sustainability of your AI project through a Stakeholder Impact Assessment (SIA), whether the AI project is being used to deliver a public service or in a back-office administrative capacity. When we refer to ‘stakeholders’ we are referring primarily to affected individual persons, but the term may also extend to groups and organisations in the sense that individual members of these collectives may also be impacted as such by the design and deployment of AI systems. Due consideration to stakeholders should be given at both of these levels.

The purpose of carrying out an SIA is multidimensional. SIAs can serve several purposes, some of which include:

1. Help to build public confidence that the design and deployment of the AI system by the public sector agency has been done responsibly.
2. Facilitate and strengthen your accountability framework.
3. Bring to light unseen risks that threaten to affect individuals and the public good 4) Underwrite well-informed decision-making and transparent innovation practices.
4. Underwrite well-informed decision-making and transparent innovation practices.
5. Demonstrate forethought and due diligence not only within your organisation but also to the wider public.

Your team should convene to evaluate the social impact and sustainability of your AI project through the SIA at three critical points in the project delivery lifecycle:

Your team should convene to evaluate the social impact and sustainability of your AI project through the SIA at three critical points in the project delivery lifecycle:

1. Alpha Phase (**Problem Formulation**): Carry out an initial Stakeholder Impact Assessment (SIA) to determine the ethical permissibility of the project. Refer to the SUM Values as a starting point for the considerations of the possible effects of your project on individual wellbeing and public welfare. In cases where you conclude that your AI project will have significant ethical and social impacts, you should open your initial SIA to the public so that their views can be properly considered. This will bolster the inclusion of a diversity of voices and opinions into the design and development process through the participation of a more representative range of stakeholders. You should also consider consulting with internal organisational stakeholders, whose input will likewise strengthen the openness, inclusivity, and diversity of your project.
2. From Alpha to Beta (**Pre-Implementation**): Once your model has been trained, tested, and validated, you and your team should revisit your initial SIA to confirm that the AI system to be implemented is still in line with the evaluations and conclusions of your original assessment. This check-in should be logged on the pre-implementation section of the SIA with any applicable changes added and discussed. Before the launch of the system, this SIA should be made publicly available. At this point you must also set a timeframe for re-assessment once the system is in operation as well as a public consultation which predates and provides input for that re-assessment. Timeframes for these re-assessments should be decided by your team on a case-by-case basis but should be proportional to the scale of the potential impact of the system on the individuals and communities it will affect.
3. Beta Phase (**Re-Assessment**): After your AI system has gone live, your team should intermittently revisit and re-evaluate your SIA. These check-ins should be logged on the re-assessment section of the SIA with any applicable changes added and discussed. Re-assessment should focus both on evaluating the existing SIA against real world impacts and on considering how to mitigate the unintended consequences that may have ensued in the wake of the deployment of the system. Further public consultation for input at the beta stage should be undertaken before the re-assessment so that stakeholder input can be

included in re-assessment deliberations.

You should keep in mind that, in its specific focus on social and ethical sustainability, your Stakeholder Impact Assessment constitutes just one part of the governance platform for your AI project and should be a complement to your accountability framework and other auditing and activity-monitoring documentation.

Your SIA should be broken down into four sections of questions and responses. In the 1st section, there should be general questions about the possible big-picture social and ethical impacts of the use of the AI system you plan to build. In the 2nd section, your team should collaboratively formulate relevant sector-specific and use case-specific questions about the impact of the AI system on affected stakeholders. The 3rd section should provide answers to the additional questions relevant to preimplementation evaluation. The 4th section should provide the opportunity for members of your team to reassess the system in light of its realworld impacts, public input, and possible unintended consequences[C9.1-1].

- **Transparencia:** definición de IA transparente | **Transparency**: Defining transparent AI

Es importante recordar que la transparencia como principio de la ética de la IA difiere un poco del significado del uso diario del término. La comprensión común del diccionario de la transparencia lo define como (1) la calidad que tiene un objeto cuando uno puede ver claramente a través de él o (2) la calidad de una situación o proceso que puede justificarse y explicarse claramente porque está abierto a inspección y libre de secretos.

La transparencia como principio de la ética de la IA abarca ambos significados:

Por un lado, la IA transparente implica la interpretabilidad de un sistema de IA dado, es decir, la capacidad de saber cómo y por qué un modelo se desempeñó y cómo lo hizo en un contexto específico y, por lo tanto, de comprender la razón detrás de su decisión o comportamiento. Con frecuencia se hace referencia a este tipo de transparencia mediante la metáfora de '**abrir la caja negra**' de la IA. Implica clarificación de contenido e inteligibilidad o explicabilidad.

Por otro lado, la IA transparente implica la justificación tanto de los procesos que intervienen en su diseño e implementación como de su resultado. Por lo tanto, implica la solidez de la justificación de su uso. En este sentido más normativo, la IA transparente es prácticamente justificable de manera irrestricta si se puede demostrar que tanto los procesos de diseño e implementación que han entrado en la decisión o comportamiento particular de un sistema como la decisión o comportamiento en sí son éticamente permisibles, no discriminatorio / justo y digno de confianza pública / seguridad[C9.1-1].

[English]

It is important to remember that transparency as a principle of AI ethics differs a bit in meaning from the everyday use of the term. The common dictionary understanding of transparency defines it as

either (1) the quality an object has when one can see clearly through it or (2) the quality of a situation or process that can be clearly justified and explained because it is open to inspection and free from secrets.

Transparency as a principle of AI ethics encompasses both of these meanings:

On the one hand, transparent AI involves the interpretability of a given AI system, i.e. the ability to know how and why a model performed the way it did in a specific context and therefore to understand the rationale behind its decision or behaviour. This sort of transparency is often referred to by way of the metaphor of 'opening the black box' of AI. It involves content clarification and intelligibility or explicability.

On the other hand, transparent AI involves the justifiability both of the processes that go into its design and implementation and of its outcome. It therefore involves the soundness of the justification of its use. In this more normative meaning, transparent AI is practically justifiable in an unrestricted way if one can demonstrate that both the design and implementation processes that have gone into the particular decision or behaviour of a system and the decision or behaviour itself are ethically permissible, non-discriminatory/fair, and worthy of public trust/safety-securing[C9.1-1].

- **Transparencia:** tres tareas críticas para diseñar e implementar IA transparente | **Transparency**: Three critical tasks for designing and implementing transparent AI

Esta definición de transparencia en dos frentes como principio de la ética de la IA le pide que piense en la IA transparente tanto en términos del proceso que está detrás (las prácticas de diseño e implementación que conducen a un resultado respaldado algorítmicamente) como en términos de su producto (el contenido y la justificación de ese resultado). Tal distinción de proceso/producto es crucial, porque aclara las tres tareas de las que su equipo será responsable de salvaguardar la transparencia de su proyecto de IA:

- Proceso de Transparencia, Tarea 1: justificar proceso. Al ofrecer una explicación a las partes interesadas afectadas, debe poder demostrar que las consideraciones de permisibilidad ética, no discriminación / equidad y seguridad / confiabilidad pública fueron operativas de extremo a extremo en los procesos de diseño e implementación que conducen a una decisión automatizada o comportamiento. Esta tarea se respaldará siguiendo las mejores prácticas descritas en este documento a lo largo del ciclo de vida del proyecto de IA y poniendo en práctica medidas de auditabilidad sólidas a través de un marco de responsabilidad por diseño.
- Transparencia de resultados, tarea 2: aclarar el contenido y explicar los resultados. Al ofrecer una explicación a las partes interesadas afectadas, debe ser capaz de mostrar en un lenguaje claro que sea comprensible para los no especialistas cómo y por qué un modelo se desempeñó de la misma manera en un contexto específico de toma de decisiones o de comportamiento. Por lo tanto, debe poder aclarar y comunicar la justificación de su decisión o comportamiento. Esta explicación debería ser socialmente significativa en el sentido de que los términos y la lógica de

la explicación no deberían simplemente reproducir las características formales o los significados técnicos y la lógica del modelo matemático, sino que deberían traducirse al lenguaje cotidiano de las prácticas humanas y, por lo tanto, ser comprensibles en términos de los factores sociales y las relaciones que implica la decisión o el comportamiento.

- Transparencia del resultado, Tarea 3: justificar el resultado. Al ofrecer una explicación a las partes interesadas afectadas, debe poder demostrar que una decisión o comportamiento específico de su sistema es éticamente permisible, no discriminatorio / justo y digno de confianza pública / seguridad. Esta justificación del resultado debe tomar la aclaración del contenido / el resultado explicado de la tarea 2 como su punto de partida y comparar esa explicación con los criterios de justificación a los que se ha adherido a lo largo del diseño y el uso: permisibilidad ética, no discriminación / equidad y seguridad / confiabilidad pública. Adoptar un enfoque óptimo para procesar la transparencia desde el principio debe apoyar y salvaguardar esta demanda de explicación normativa y justificación de resultados[C9.1-1].

[English]

This two-pronged definition of transparency as a principle of AI ethics asks that you to think about transparent AI both in terms of the process behind it (the design and implementation practices that lead to an algorithmically supported outcome) and in terms of its product (the content and justification of that outcome). Such a process/product distinction is crucial, because it clarifies the three tasks that your team will be responsible for in safeguarding the transparency of your AI project:

- Process Transparency, Task 1: Justify Process. In offering an explanation to affected stakeholders, you should be able to demonstrate that considerations of ethical permissibility, non-discrimination/fairness, and safety/public trustworthiness were operative end-to-end in the design and implementation processes that lead to an automated decision or behaviour. This task will be supported both by following the best practices outlined herein throughout the AI project lifecycle and by putting into place robust auditability measures through an accountability-by-design framework.
- Outcome Transparency, Task 2: Clarify Content and Explain Outcome. In offering an explanation to affected stakeholders, you should be able to show in plain language that is understandable to non-specialists how and why a model performed the way it did in a specific decision-making or behavioural context. You should therefore be able to clarify and communicate the rationale behind its decision or behaviour. This explanation should be socially meaningful in the sense that the terms and logic of the explanation should not simply reproduce the formal characteristics or the technical meanings and rationale of the mathematical model but should rather be translated into the everyday language of human practices and therefore be understandable in terms of the societal factors and relationships that the decision or behaviour implicates.
- Outcome Transparency, Task 3: Justify Outcome. In offering an explanation to affected stakeholders, you should be able to demonstrate that a specific decision or behaviour of your system is ethically permissible, non-discriminatory/fair, and worthy of public trust/safety-securing. This outcome justification should take the content clarification/explicated outcome

from task 2 as its starting point and weigh that explanation against the justifiability criteria adhered to throughout the design and use pipeline: ethical permissibility, non-discrimination/fairness, and safety/public trustworthiness. Undertaking an optimal approach to process transparency from the start should support and safeguard this demand for normative explanation and outcome justification[C9.1-1].

- **Transparencia:** Mapeo de transparencia AI | **Transparency**: Mapping AI transparency

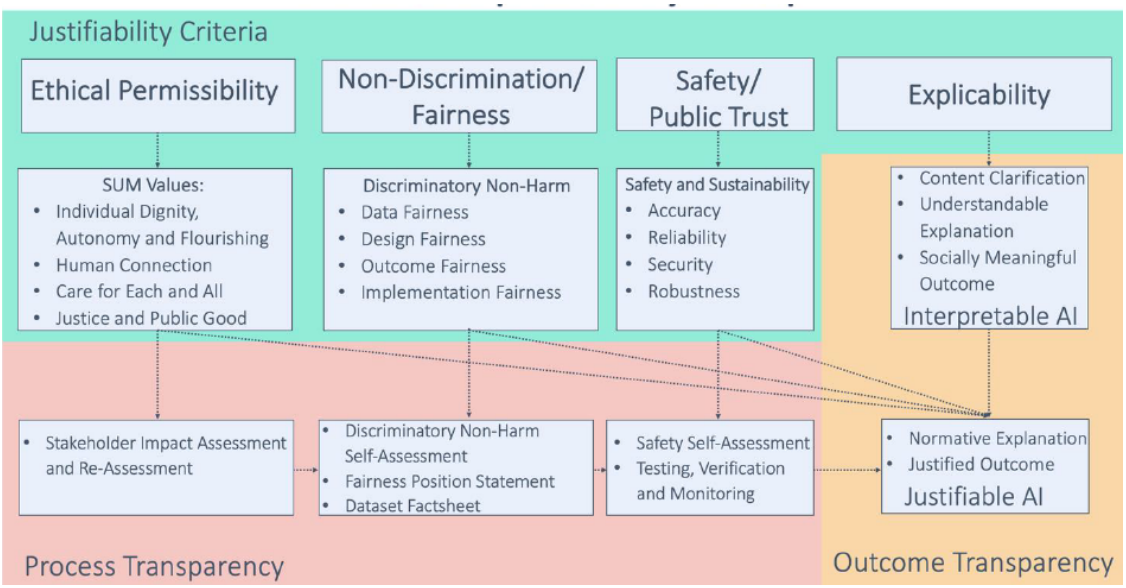


Fig. I A. 9.2.4- Mapa de transparencia AI | AI transparency map. Crédito imag. (C9.1-1). URL: https://doi.org/10.5281/zenodo.3240529

9.2-4.-Transparencia de resultados: explicando los resultados, aclarando el contenido, implementando responsablemente | Outcome transparency: Explaining outcome and clarifying content

- Aspectos técnicos de elegir, diseñar y usar un sistema de inteligencia artificial interpretable | Technical aspects of choosing, designing, and using an interpretable AI system

Tenga en cuenta que, si bien, a primera vista, la tarea de elegir entre los numerosos algoritmos de inteligencia artificial y aprendizaje automático puede parecer desalentador, no tiene por qué ser así. Al apegarse a la prioridad de la transparencia de resultados, usted y su equipo podrán seguir algunas pautas sencillas y directas para seleccionar técnicas algorítmicas suficientemente interpretables pero de rendimiento óptimo. Antes de explorar estas pautas, es necesario proporcionarle información

básica para ayudarlo a comprender mejor qué facetas de explicación están realmente involucradas en la IA técnicamente interpretable. Una buena comprensión de lo que realmente se necesita de tal explicación le permitirá enfocarse efectivamente en las necesidades de interpretación de su proyecto de IA.

Facetas de explicación en IA técnicamente interpretable: un buen punto de partida para comprender cómo funciona la dimensión técnica de la explicación en sistemas de IA interpretables es recordar que estos sistemas son en gran medida modelos matemáticos que realizan cálculos paso a paso en conjuntos transformadores de interacción estadística o entradas independientes en conjuntos de salidas objetivo. El aprendizaje automático es, en el fondo, solo estadística aplicada y teoría de la probabilidad fortificada con varias otras técnicas matemáticas. Como tal, está sujeto a los mismos requisitos metodológicamente rigurosos de validación lógica que otras ciencias matemáticas.

Tal demanda de rigor informa la faceta de la explicación formal y lógica de los sistemas de IA que a veces se llama la 'caja de cristal matemática'. Esta caracterización se refiere a la transparencia de la explicación estrictamente formal: no importa cuan complicado sea (incluso en el caso de una red neuronal profunda con cien millones de parámetros), un modelo algorítmico es un sistema cerrado de operaciones efectivamente computables donde las reglas y las transformaciones son aplicados mecánicamente a las entradas para determinar las salidas. En este sentido restringido, todos los modelos de inteligencia artificial y aprendizaje automático son totalmente inteligibles y matemáticamente transparentes, aunque sólo sea formal y lógicamente.

Ésta es una característica importante de los sistemas de IA, ya que hace posible que los enfoques computacionales suplementarios y eminentemente interpretables modelen, aproximen y simplifiquen incluso los más complejos y de alta dimensión entre ellos. De hecho, tal posibilidad alimenta algunos de los enfoques técnicos para la IA interpretable que pronto serán explorados.

Sin embargo, esta forma formal de comprender la explicación técnica de la IA y los sistemas de aprendizaje automático tiene limitaciones inmediatas. Puede decirnos que un modelo es matemáticamente inteligible porque funciona de acuerdo con una colección de operaciones y parámetros fijos, pero no puede decirnos mucho acerca de cómo o por qué los componentes del modelo transformaron un grupo específico de entradas en sus salidas correspondientes. No puede decirnos nada sobre la lógica detrás de la generación algorítmica de un resultado dado.

Esta segunda dimensión de la explicación técnica tiene que ver con la faceta semántica de la IA interpretable. Una explicación semántica ofrece una interpretación de las funciones de las partes individuales del sistema algorítmico en la generación de su salida. Mientras que la explicación formal y lógica presenta una explicación de la aplicación gradual de los procedimientos y reglas que comprenden el marco formal del sistema algorítmico, la explicación semántica nos ayuda a comprender el significado de esos procedimientos y reglas en términos de su propósito en la entrada-salida operación de mapeo del sistema, es decir, qué papel juegan en la determinación del resultado del cálculo del modelo.

Las dificultades que rodean la interpretabilidad de las decisiones y los comportamientos algorítmicos surgen en esta dimensión semántica de la explicación técnica[C9.1-1].

[English]

Keep in mind that, while, on the face of it, the task of choosing between the numerous AI and machine learning algorithms may seem daunting, it need not be so. By sticking to the priority of outcome transparency, you and your team will be able to follow some straightforward and simple guidelines for selecting sufficiently interpretable but optimally performing algorithmic techniques. Before exploring these guidelines, it is necessary to provide you with some background information to help you better understand what facets of explanation are actually involved in technically interpretable AI. A good grasp of what is actually needed from such an explanation will enable you to effectively target the interpretability needs of your AI project.

Facets of explanation in technically interpretable AI: A good starting point for understanding how the technical dimension of explanation works in interpretable AI systems is to remember that these systems are largely mathematical models that carry out step-by-step computations in transforming sets of statistically interacting or independent inputs into sets of target outputs. Machine learning is, at bottom, just applied statistics and probability theory fortified with several other mathematical techniques. As such, it is subject to same methodologically rigorous requirements of logical validation as other mathematical sciences.

Such a demand for rigour informs the facet of formal and logical explanation of AI systems that is sometimes called the 'mathematical glass box'. This characterisation refers to the transparency of strictly formal explanation: No matter how complicated it is (even in the case of a deep neural net with a hundred million parameters), an algorithmic model is a closed system of effectively computable operations where rules and transformations are mechanically applied to inputs to determine outputs. In this restricted sense, all AI and machine learning models are fully intelligible and mathematically transparent if only formally and logically so.

This is an important characteristic of AI systems, because it makes it possible for supplemental and eminently interpretable computational approaches to model, approximate, and simplify even the most complex and high dimensional among them. In fact, such a possibility fuels some of the technical approaches to interpretable AI that will soon be explored.

This formal way of understanding the technical explanation of AI and machine learning systems, however, has immediate limitations. It can tell us that a model is mathematically intelligible because it operates according to a collection of fixed operations and parameters, but it cannot tell us much about how or why the components of the model transformed a specified group of inputs into their corresponding outputs. It cannot tell us anything about the rationale behind the algorithmic generation of a given outcome.

This second dimension of technical explanation has to do with the semantic facet of interpretable AI. A semantic explanation offers an interpretation of the functions of the individual parts of the

algorithmic system in the generation of its output. Whereas formal and logical explanation presents an account of the stepwise application of the procedures and rules that comprise the formal framework of the algorithmic system, semantic explanation helps us to understand the meaning of those procedures and rules in terms of their purpose in the input-output mapping operation of the system, i.e. what role they play in determining the outcome of the model's computation.

The difficulties surrounding the interpretability of algorithmic decisions and behaviours arise in this semantic dimension of technical explanation[C9.1-1].

- Pautas para diseñar y entregar un sistema de IA suficientemente interpretable | Guidelines for designing and delivering a sufficiently interpretable AI system

Directriz 2: recurrir a técnicas estándar interpretables cuando sea posible | **Guideline 2**: Draw on standard interpretable techniques when posible

Para integrar activamente el objetivo de la suficiente capacidad de interpretación en su proyecto de IA, su equipo debe abordar el proceso de selección y desarrollo de modelos con el objetivo de encontrar el ajuste adecuado entre (1) riesgos y necesidades específicos del dominio, (2) recursos de datos disponibles y conocimiento del dominio, y (3) técnicas de aprendizaje automático apropiadas para la tarea. La asimilación efectiva de estos tres aspectos de su caso de uso requiere una actitud abierta y práctica.

A menudo, puede darse el caso de que entornos de alto impacto, críticos para la seguridad u otros entornos potencialmente sensibles aumenten las demandas de una responsabilidad y transparencia exhaustivas de los proyectos de IA. En algunos de estos casos, tales demandas pueden hacer que la elección de técnicas estándar pero sofisticadas no opacas sea una prioridad primordial. Estas técnicas pueden incluir árboles de decisiones, regresión lineal y sus extensiones como modelos aditivos generalizados, listas de decisiones / reglas, razonamiento basado en casos o regresión logística. En muchos casos, alcanzar el modelo de 'caja negra' primero puede no ser apropiado e incluso puede conducir a ineficiencias en el desarrollo del proyecto, porque los modelos más interpretables, que funcionan muy bien pero no requieren herramientas y técnicas complementarias para facilitar resultados interpretables, son también disponibles.

Nuevamente, el conocimiento de dominio sólido y la conciencia del contexto son componentes clave aquí. En los casos de uso en los que los recursos de datos se prestan a representaciones significativas bien estructuradas y la experiencia en el dominio se pueden incorporar a las arquitecturas modelo, las técnicas interpretables a menudo pueden ser más deseables que las opacas. El cuidadoso procesamiento previo de datos y el desarrollo iterativo de modelos pueden, en estos casos, perfeccionar la precisión de dichos sistemas interpretables de manera que las ventajas obtenidas por la combinación de su rendimiento y transparencia superen los beneficios de los enfoques semánticamente más transparentes.

Sin embargo, en otros casos de uso, las necesidades de procesamiento de datos pueden descalificar

el despliegue de este tipo de sistemas interpretables sencillos. Por ejemplo, cuando se buscan aplicaciones de inteligencia artificial para clasificar imágenes, reconocer el habla o detectar anomalías en las imágenes de video, los enfoques de aprendizaje automático más efectivos probablemente serán opacos. Los espacios de características de este tipo de sistemas de IA crecen exponencialmente a cientos de miles o incluso millones de dimensiones. A esta escala de complejidad, los métodos convencionales de interpretación ya no se aplican. De hecho, es la inevitabilidad de golpear ese muro de interpretabilidad para ciertas aplicaciones importantes de aprendizaje supervisado, no supervisado y de refuerzo lo que ha dado lugar a un subcampo completo de investigación de aprendizaje automático que se centra en proporcionar herramientas técnicas para facilitar una IA interpretable y explicable[C9.1-1].

[English]

In order to actively integrate the aim of sufficient interpretability into your AI project, your team should approach the model selection and development process with the goal of finding the right fit between (1) domain-specific risks and needs, (2) available data resources and domain knowledge, and (3) task appropriate machine learning techniques. Effectively assimilating these three aspects of your use case requires open-mindedness and practicality.

Often times, it may be the case that high-impact, safety-critical, or other potentially sensitive environments heighten demands for the thoroughgoing accountability and transparency of AI projects. In some of these instances, such demands may make choosing standard but sophisticated non-opaque techniques an overriding priority. These techniques may include decisions trees, linear regression and its extensions like generalised additive models, decision/rule lists, case-based reasoning, or logistic regression. In many cases, reaching for the ‘black box’ model first may not be appropriate and may even lead to inefficiencies in project development, because more interpretable models, which perform very well but do not require supplemental tools and techniques for facilitating interpretable outcomes, are also available.

Again, solid domain knowledge and context awareness are key components here . In use cases where data resources lend to well-structured, meaningful representations and domain expertise can be incorporated into model architectures, interpretable techniques may often be more desirable than opaque ones. Careful data pre-processing and iterative model development can, in these cases, hone the accuracy of such interpretable systems in ways that may make the advantages gained by the combination of their performance and transparency outweigh the benefits of more semantically intransparent approaches.

In other use cases, however, data processing needs may disqualify the deployment of these sorts of straightforward interpretable systems. For instance, when AI applications are sought for classifying images, recognising speech, or detecting anomalies in video footage, the most effective machine learning approaches will likely be opaque. The feature spaces of these kinds of AI systems grow exponentially to hundreds of thousands or even millions of dimensions. At this scale of complexity, conventional methods of interpretation no longer apply. Indeed, it is the unavoidability of hitting such an interpretability wall for certain important applications of supervised, unsupervised, and

reinforcement learning that has given rise to an entire subfield of machine learning research which focuses on providing technical tools to facilitate interpretable and explainable AI[C9.1-1].

9.2-5.-Pasos para garantizar procesos de implementación centrados en el ser humano | Steps to ensuring human-centred implementation processes

Paso 1: considere los aspectos del tipo de aplicación y el contexto del dominio para definir roles | **Step 1**: Consider aspects of application type and domain context to define roles.

Paso 2: definir las relaciones de entrega y los procesos de entrega de mapas | **Step 2**: Define delivery relations and map delivery processes.

Paso 3: construir una plataforma de implementación ética | **Step 3**: Build an ethical implementation platform.

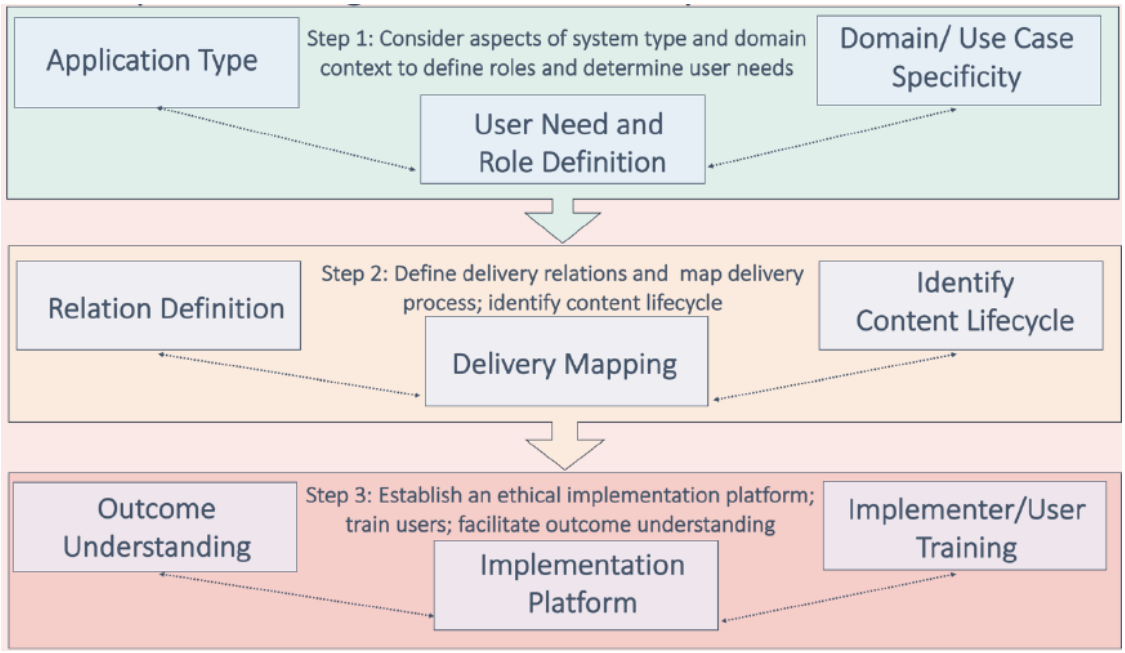


Fig. I A. 9.2.5- Pasos para garantizar procesos de implementación centrados en el ser humano | Steps to ensuring human-centred implementation processes. Crédito imag. (C9.1-1). URL: https://doi.org/10.5281/zenodo.3240529

Fig. I A. 9.2.6- *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. *The Alan Turing Institute*.

https://doi.org/10.5281/zenodo.3240529

SEGUNDA SECCIÓN | SECOND SECTION: Herramientas | Tools

9.2-6.- Herramientas para el diseño responsable y la implementación de sistemas de IA | Tools for responsible design and implementation of AI systems

> IBM entendemos que se sitúa como la pionera en la creación de herramientas de código abierto para hacer viable el diseño y la implementación de sistemas IA. Exponemos ahora una sección ya descrita en este libro en 4.5.4º.- IA confiable de IBM. | IBM we understand that it is positioned as the pioneer in the creation of open source tools to make the design and implementation of AI systems viable. We now expose a section already described in this book in 4.5.4º.- IBM Reliable AI

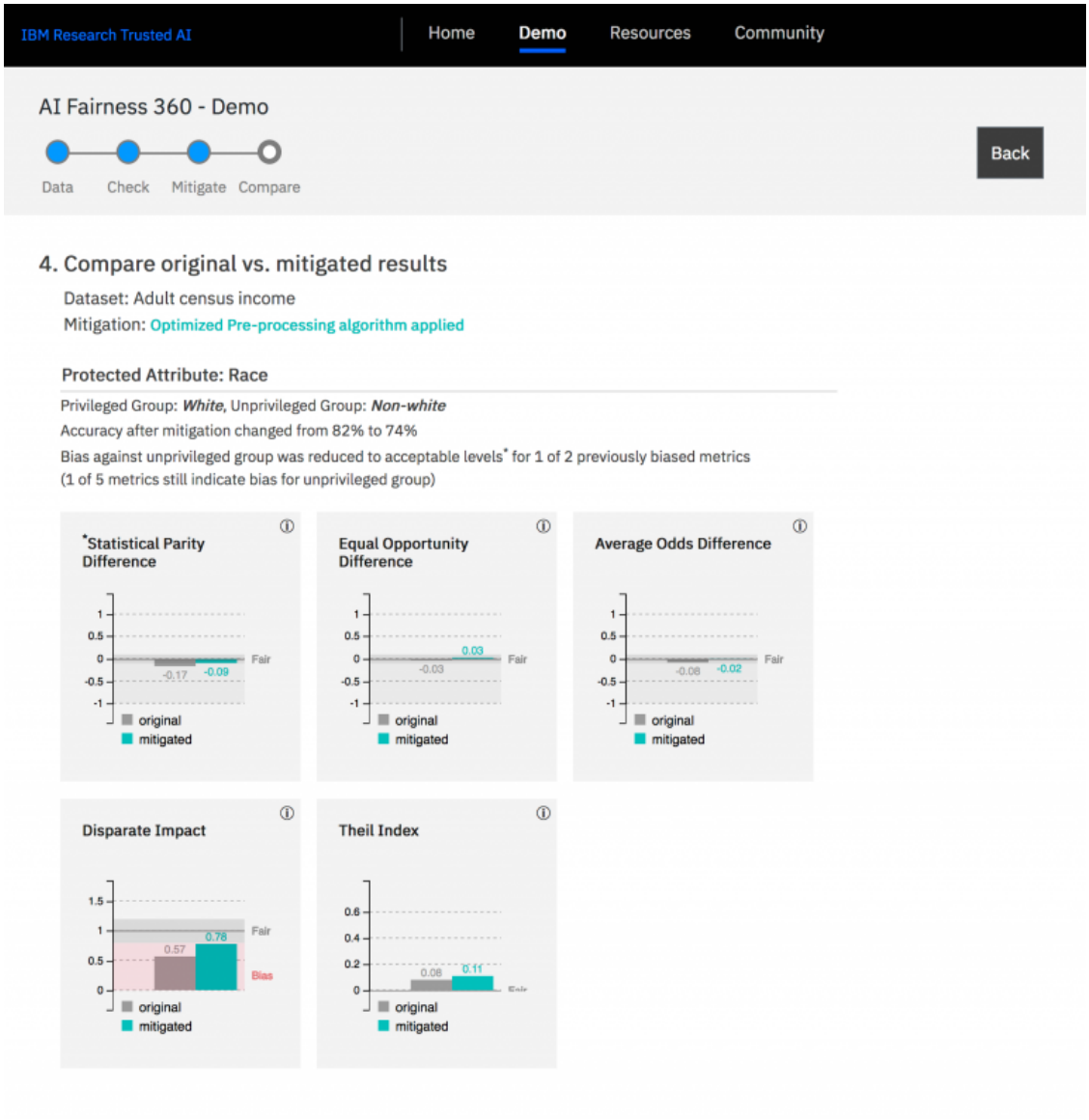


Fig. I A. 9.2.7-Experiencia interactiva AI Fairness 360 | AI Fairness 360 interactive experience. Crédito imag. (IBM Research Blog). URL: https://www.ibm.com/blogs/research/wp-content/uploads/2018/09/PosterPage-768x792.png

AI Fairness 360

RESUMEN. La equidad es una preocupación cada vez más importante, ya que los modelos de aprendizaje automático se utilizan para apoyar la toma de decisiones en aplicaciones de alto riesgo, como préstamos hipotecarios, contratación y sentencias de prisión. Este documento presenta un nuevo kit de herramientas Python de código abierto para la equidad algorítmica, AI Fairness 360 (AIF360), publicado bajo una licencia Apache v2.0 (https://github.com/ibm/aif360). Los objetivos principales de este conjunto de herramientas son ayudar a facilitar la transición de los algoritmos de investigación de equidad para usar en un entorno industrial y proporcionar un marco común para

que los investigadores de equidad compartan y evalúen algoritmos. El paquete incluye un conjunto integral de métricas de equidad para conjuntos de datos y modelos, explicaciones para estas métricas y algoritmos para mitigar el sesgo en conjuntos de datos y modelos. También incluye una experiencia web interactiva (https://aif360.mybluemix.net) que proporciona una introducción suave a los conceptos y capacidades para los usuarios de la línea de negocios, así como una amplia documentación, orientación de uso y tutoriales específicos de la industria para habilitar científicos de datos y profesionales para incorporar la herramienta más adecuada para su problema en sus productos de trabajo. La arquitectura del paquete ha sido diseñada para ajustarse a un paradigma estándar utilizado en ciencia de datos, mejorando así la usabilidad para los profesionales. Tal diseño arquitectónico y abstracciones permiten a los investigadores y desarrolladores ampliar el *ki* de herramientas con sus nuevos algoritmos y mejoras, y usarlo para la evaluación comparativa del rendimiento. Una infraestructura de prueba incorporada mantiene la calidad del código.

[English]

ABSTRACT. Fairness is an increasingly important concern as machine learning models are used to support decision makingin high-stakes applications such as mortgage lending, hiring, and prison sentencing. This paper introduces a newopen source Python toolkit for algorithmic fairness, AI Fairness 360 (AIF360), released under an Apache v2.0license (https://github.com/ibm/aif360). The main objectives of this toolkit are to help facilitate thetransition of fairness research algorithms to use in an industrial setting and to provide a common framework forfairness researchers to share and evaluate algorithms.The package includes a comprehensive set of fairness metrics for datasets and models, explanations for thesemetrics, and algorithms to mitigate bias in datasets and models. It also includes an interactive Web experience(https://aif360.mybluemix.net) that provides a gentle introduction to the concepts and capabilitiesfor line-of-business users, as well as extensive documentation, usage guidance, and industry-specific tutorials toenable data scientists and practitioners to incorporate the most appropriate tool for their problem into their workproducts. The architecture of the package has been engineered to conform to a standard paradigm used in datascience, thereby further improving usability for practitioners. Such architectural design and abstractions enableresearchers and developers to extend the toolkit with their new algorithms and improvements, and to use it forperformance benchmarking. A built-in testing infrastructure maintains code quality.

- AI Explainability 360 Kit de herramientas de código abierto | AI Explainability 360 Open Source Toolkit

---

Este kit de herramientas **de código abierto** extensible puede ayudar a comprender (explicabilidad) cómo los modelos de aprendizaje automático predicen por diversos medios a lo largo del ciclo de vida de la aplicación AI. Con ocho algoritmos de vanguardia para el aprendizaje automático interpretable,

así como métricas para la explicabilidad, está diseñado para traducir la investigación algorítmica del laboratorio en la práctica real de dominios tan amplios como finanzas, gestión de capital humano, atención médica, y educación.

[English]

This extensible **open source** toolkit can help you comprehend (explainability) how machine learning models predict labels by various means throughout the AI application lifecycle. Containing eight state-of-the-art algorithms for interpretable machine learning as well as metrics for explainability, it is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education.

- AI imparcialidad 360 Toolkit | AI Fairness 360 Toolkit

Este kit de herramientas **de código abierto** extensible puede ayudar a examinar, informar y mitigar la discriminación y el sesgo (justicia o igualdad) en los modelos de aprendizaje automático a lo largo del ciclo de vida de la aplicación AI. Con más de 70 métricas de equidad y 10 algoritmos de mitigación de sesgos de última generación desarrollados por la comunidad investigadora, está diseñado para traducir la investigación algorítmica del laboratorio en la práctica real de dominios tan amplios como finanzas, gestión de capital humano, salud y educación.

[English]

This extensible **open source** toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. Containing over 70 fairness metrics and 10 state-of-the-art bias mitigation algorithms developed by the research community, it is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education.

- Caja de herramientas Adversarial Robustness 360 | Adversarial Robustness 360 Toolbox

---

Adversarial Robustness Toolbox está diseñado para ayudar a los investigadores y desarrolladores a crear nuevas técnicas de defensa, así como a desplegar defensas prácticas de sistemas de IA del mundo real. Los investigadores pueden usar la caja de herramientas de robustez adversaria para comparar las defensas novedosas contra el estado del arte. Para los desarrolladores, la biblioteca proporciona interfaces que admiten la composición de sistemas de defensa integrales utilizando métodos individuales como bloques de construcción.

[English]

The Adversarial Robustness Toolbox is designed to support researchers and developers in creating novel defense techniques, as well as in deploying practical defenses of real-world AI systems. Researchers can use the Adversarial Robustness Toolbox to benchmark novel defenses against the state-of-the-art. For developers, the library provides interfaces which support the composition of comprehensive defense systems using individual methods as building blocks.

- IBM Watson OpenScale

Gestione la IA de producción, con confianza en los resultados. Vea cómo IBM Watson® OpenScale ™ rastrea y mide los resultados de la IA a lo largo de su ciclo de vida, y adapta y gobierna la IA a las situaciones comerciales cambiantes, para modelos construidos y funcionando en cualquier lugar.

[English]

Manage the production AI, with confidence in the results. See how IBM Watson® OpenScale ™ tracks and measures AI results throughout its life cycle, and adapts and governs AI to changing business situations, for models built and running anywhere.

TERCERA SECCIÓN | **THIRD SECTION**: Conclusiones y recomendaciones | conclusions and recommendations

---

Para definir de forma adecuada esta sección, tomaremos como referencia el modelo '**Esquema Europeo de Certificación de Tecnologías de Productos y Servicios de IA**' propuesto por el PhD. Carlos Galán (ver capítulo 9.1.3). Y siguiendo este modelo vamos a efectuar nuestras conclusiones y recomendaciones de forma pautada. (Estas conclusiones y recomendaciones pudieran ser tomadas como borrador para discusión).

A este apartado se me ocurre denominarlo Los 7 mandamientos de la IA.

[English]

To properly define this section, we will take as reference the model '**European Certification Scheme of AI Technologies, Products and Services**' proposed by the PhD. Carlos Galán (see chapter 9.1.3). And following this model we will make our conclusions and recommendations in a prescribed manner. (These conclusions and recommendations could be taken as a draft for discussion).

I think of this section as The 7 Commandments of AI.

Fig. I A. 9.1.4-**A** [ES]- Modelo de certificación propuesto en el documento *La certificación como mecanismo de control de la inteligencia artificial en Europa*[C9.1-5]. URL: http://www.ieee.es/Galerias/fichero/docs_opinion/2019/DIEEEO46_2019CARGAL-InteligenciaArtificial.pdf

---

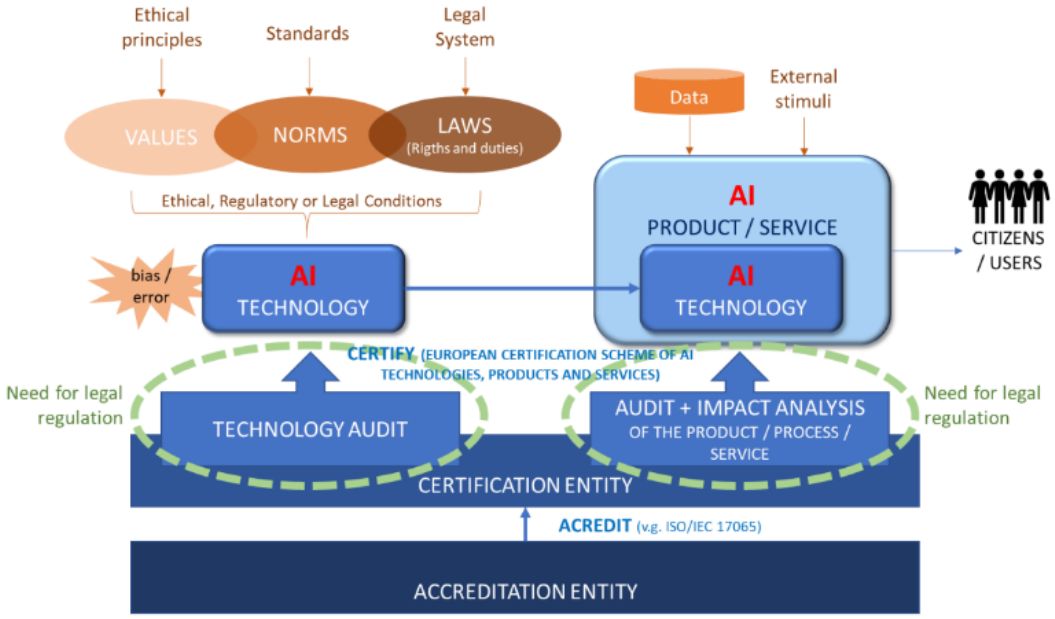


Fig. I A. 9.1.4-**B** [EN]- *Certification model proposed in the document Certification as a mechanism for controlling artificial intelligence in Europe* [C9.1-5].

1. El principio más universal es que la inteligencia artificial es un asunto a nivel mundial. Y como tal debe ser tratado.
2. Las pautas iniciales y mayormente importantes para referenciarse en una ética confiable pueden ser extraídas de 'La ética de la IA en Europa'. URL: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai que se puede ver complementada o mejorada por sus homólogas descritas aquí: [4.5.1.1.- Estrategias nacionales e internacionales de IA].
3. Los principales estándares para una ética confiable pueden ser extraídos del IEEE (familia de P7000) y del ISO/IEC ISO/IEC JTC1/SC42 . (Estos estándares debieran ser abiertos, conocidos, auditables en caso de necesidad y con un formato de funcionamiento similar a los desarrollos del tipo Open Source).
4. El sistema legal aún está por concretarse de manera adecuada. La regulación debiera establecerse de forma universal (mundial). Un buen punto de partida sería la ONU como elemento troncal (en consonancia con la Declaración Universal de Derechos Humanos y La protección de los derechos fundamentales en la Unión Europea). De tal forma que no existan variaciones de normativa al respecto entre países. (Quizá muy probablemente esas leyes debieran ser votadas y aprobadas de forma también universal para establecer los valores que afectan a nivel mundial, procurando evitar al máximo lagunas legislativas o agujeros legales).

5. Los datos de alto nivel (o bancos de datos) manejados por la inteligencia artificial debieran poder ser sometidos a auditoría mundial. Por lo que un organismo internacional debiera tener esa responsabilidad con el fin de evitar que los datos tengan distintos sesgos en función de dónde sean utilizados. Y aún a mayor responsabilidad, garantizar y evitar el mal uso de estos datos en detrimento de la población mundial. Este apartado también ha de comprometer a los datos o estímulos externos que reciban las IAs para que también puedan ser auditados públicamente. (En consonancia con el Reglamento Europeo de Protección de Datos, RGPD).
6. Una norma como la ISO/IEC 17065 (o similar mejorada) debe ser implantada con el objeto y fin de certificar y garantizar los servicios, y productos, y a las personas que participen en su desarrollo.
7. Con carácter bianual u otro el organismo encargado de auditar los datos (o banco de datos) deberá entregar un informe de impacto a cada país o gobierno que lo solicite con el fin de mantener y garantizar el aprovechamiento humano pretendido por la sociedad.

[English]

1. The most universal principle is that artificial intelligence is a worldwide issue. And as such it must be treated.
2. The initial and mostly important guidelines for referencing in a reliable ethic can be extracted from 'The ethics of AI in Europe'. URL:https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai that can be seen complemented or improved by their homologues described here: [4.5.1.1.- National strategies and international AI].
3. The main standards for reliable ethics can be extracted from the IEEE (P7000 family) and the ISO / IEC ISO/IEC JTC1/SC42. (These standards should be open, known, auditable if necessary and with an operating format similar to the Open Source type developments).
4. The legal system is still to be concretized properly. Regulation should be established universally (worldwide). A good starting point would be the UN as a core element (in line with the Universal Declaration of Human Rights and the Protection of Fundamental Rights in the European Union). In such a way that there are no variations in this respect between countries. (Perhaps these laws should probably be voted and passed universally to establish the values that affect the world, trying to avoid legislative loopholes or legal holes as much as possible).

5. High-level data (or data banks harmful) managed by artificial intelligence should be subject to worldwide auditing. Therefore, an international organization should have that responsibility in order to prevent the data from having different biases depending on where they are used. And even at greater responsibility, guarantee and avoid the misuse of these data to the detriment of the world population. This section must also commit to external data or stimuli received by AIs so that they can also be audited publicly. (In line with the European Data Protection Regulation, GDPR).
6. A standard such as ISO / IEC 17065 (or similar improved) must be implemented in order to certify and guarantee the services, products, and people involved in its development.
7. On a biannual or other basis, the agency in charge of auditing the data (or data bank) must deliver an impact report to each country or government that requests it in order to maintain and guarantee the human exploitation intended by society.

- Versión vigente de este documento: V.1 | Current version of this document: V.1
- Fecha | Date: 21/10/2019

Bibliografía | Bibliography

[C9.1-1]. Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. *The Alan Turing Institute*. https://doi.org/10.5281/zenodo.3240529

[C9.1-2]. Introducing AI Fairness 360. Website. [Recuperado 23/10/2019 de: https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/]

© 2019. Licencia de uso y distribución / License for use and distribution: [Los estados de la inteligencia artificial (IA) | The states of artificial intelligence (AI)] creative commons CC BY-NC-ND |
ISSN 2695-3803|

- Notas legales / Legal notes
- Página web de Formaempleo / Formaempleo website
- Formulario de contacto / Contact Form

Revision #27

Created Wed, Oct 23, 2019 7:26 PM by Juan Antonio Lloret Egea

Updated Wed, Nov 6, 2019 8:20 AM by Auto Backup Review

View publication stats