adigital

A.I.: transparency and ethics certifications

State of the art

With the advice of

eticas

A.I.: transparency and ethics certifications

State of the art

Contents

Standardization rules

There are various organizations in the world that are responsible for creating domestic and international standards for performing processes, assessments, experiments and other tasks in a wide range of disciplines. Although these standards are not a certificate in themselves, complying with them ensures that a task or process conforms to the rules. With this in mind, various committees have started working on developing standards that cover different areas of the development and implementation process for Artificial Intelligence (AI) systems and *Machine Learning* (ML). There follows a list with various bodies and committees, together with a brief description of them prepared by Winter et al., 2021.

ISO/IEC: One of the most influential standardization organizations globally, ISO began working on the development of rules and standards for AI in May 2018, after establishing subcommittee 42 within the first technical committee (JTC 1/SC 42). Specifically, the most important ones are:

> ► ISO/IEC TR 24027:2021 — *Bias in AI systems and AI-aided decision-making*

> ► ISO/IEC TR 24028:2020 — *Overview of trustworthiness in artificial intelligence*

> ► ISO/IEC TR 24368 — *Overview of ethical and societal concerns*

ITU: The ITU-T group was set up in November 2017 and remained active until July 2020, preparing draft technical specifications for machine learning. At the same time, the ITU/WHO group (founded in 2018) works in partnership with the World Health Organization to standardized assessments of AI-based models for health, diagnosis, triage or treatment decisions.

IEEE: This global engineer association has published various reports on the evaluation and assessment of different systems that include AI. Specifically, in the IEEE P7000 series, it is developing standards to evaluate the interoperability, functionality and security of systems.

CEN and CENELEC: The CEN-CENELEC working group was established in April 2019 to address the need to standardize AI in Europe. In this connection, the group promotes the use of the ISO/IEC JTC 1/SC 42 standards, and works with the European Commission to identify the technical requirements for such use.

DIN: One of the most important organizations in Europe, the DIN information technology standards committee works to develop tools, standards and practices for processes and applications in the field of AI, taking into consideration social implications and opportunities. In this case, they also follow the guidelines in the ISO/IEC/JTC 1/SC 42 standards.

It should be noted that most standardization activities focus on issues relating to security, robustness, reliability, fairness, and human oversight. (Winter et al., 2021:18-9)

Fundamentals

The way in which Artificial Intelligence has to be developed and implemented and the potential biases of *machine learning* have become particularly important in recent years and, regardless of the advantages, disadvantages and challenges the situation poses, they are already part of our daily lives. Over the last few decades, computing power has increased exponentially. This, combined with global internet implementation and an increase in the capacity to create, store and manage data has facilitated the large-scale implementation of various AI or ML-based systems, thus promoting a low-regulation environment.

However, the regulatory proposals in this area have been prioritized at international, European and national levels and there are countless texts and initiatives aimed at establishing basic guidelines to detect and prevent biased or discriminatory decisions, and initiatives to define the ethical limits to the development and application of these technologies. Additionally, the transparency framework through which channels should be established for accessing and overseeing AI and the products that feature it without damaging intellectual property is emerging as another of the great and most pressing challenges in this field.

The General Data Protection Regulation (GDPR 2016/67) has already briefly covered automated decision-making and profiling based on specific categories of personal data, stating that *"it should be allowed only under specific conditions"*. In 2019, the first ethics guidelines for trustworthy AI were published in Europe, with the aim of shedding light on adherence to transparency and ethical principles[1] and, in April 2021, the European Commission published its proposal for an Artificial Intelligence Act[2], which is now being debated in the European Parliament and in the Council of the EU and is a priority file. Within the same framework, the European Commission has published its proposed Product Liability Directive adapted to the digital age and artificial intelligence[3], aimed at adapting public liability to the digital age, especially with regard to liability for damage caused by AI systems.

Furthermore, the Digital Services Act states that large digital platforms with over 45 million users in the EU must disclose the use of artificial intelligence to remove illegal content, and the content that has been removed, and they will have to publicly report on how they limit serious risks to society regarding freedom of speech, public health and elections, among other types of information.

In turn, from a national perspective, Spain is one of the most advanced jurisdictions promoting regulatory measures in AI. The Spanish Government defined the National Strategy for Artificial

---

1   https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

2   https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:52021PC0206

3   https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12979-Responsabilidad-civil-Adaptacion-de-las-normas-de-responsabilidad-a-la-era-digital-y-a-la-inteligencia-artificial_es

Intelligence structured through a variety of policies and regulatory actions.

The Act 12/2021 of September 28 is currently in force, amending the consolidated text of the Workers' Statute Act, approved by Royal Legislative Decree 2/2015, of October 23, to safeguard the employment rights of delivery workers in the field of digital platforms, introducing a new subsection d) in article 64.4, which reads as follows:

"d)   To be informed by the company of the parameters, rules and instructions on which the algorithms or artificial intelligence systems that affect decision-making are based where these may affect the working conditions, access to and retention of employment, including profiling."

In July 2022, Act 15/2022, of July 12, for equal treatment and non-discrimination also came into force, in which article 23 states:

"1. Within the framework of the National Artificial Intelligence Strategy, of the Charter of Digital Rights and of the European initiatives related to Artificial Intelligence, public sector bodies shall promote the implementation of mechanisms to ensure that the algorithms involved in decision-making that are used in public sector bodies follow criteria for bias minimization, transparency and accountability, wherever technically viable. These mechanisms shall include their design and training data, and address their potential discriminatory impact. To achieve this, the performance of impact assessments to determine possible discriminatory bias shall be promoted.

2. Within the scope of their competencies with regard to the algorithms involved in decision-making, public sector bodies shall prioritize transparency in the design, implementation and interpretability of the decisions that they make.

3. Public sector bodies and businesses shall promote the use of Artificial Intelligence that is ethical, trustworthy and respects fundamental rights, in particular following the recommendations made by the European Union in this regard.

4. An algorithm quality seal shall be promoted."

In turn, the country is going to create a Spanish Artificial Intelligence Supervisory Agency, which will be responsible for developing, overseeing and monitoring projects that fall within the National Artificial Intelligence Strategy (ENIA), and projects promoted by the EU, in particular those relating to the development of regulations on artificial intelligence and its possible uses. The Agency's specific function will be to minimize any significant risks that may be posed by the use of artificial intelligence systems to people's safety and health, and to their fundamental rights. In this regard, the text states that those measures "shall in themselves entail their own actions, actions coordinated with other competent authorities and actions to support private entities".

In fact, in part 16 of the Recovery Plan on the National Artificial Intelligence Strategy (ENIA), point 1.3 on the regulatory and ethical framework: AI Observatory and Seal stipulates that "*a certification architecture and trusted AI seal will be developed for AI products and services. This will include the creation of a collection of tools (toolkit) that guides the design of technologies according to the criteria recommended by the seal. (Demonstration project). This quality seal will be aligned and compatible with the European regulatory framework envisaged for March 2021. Spain is participating in European working groups*

in relation to this new regulation. The Spanish seal will also include references to Spain's strengths in AI such as respect for Spanish grammar in algorithms or alignment with the Green Algorithms program." In this connection, they will be put out to tender to make this certification a public-private partnership project.

Furthermore, in June, the Spanish Regulatory Sandbox on Artificial Intelligence was presented and it is likely to be launched in October. The Spanish sandbox has the following goals:

► to establish the conditions for the seamless implementation of future regulations on AI;

► to facilitate the testing of specific technical solutions and regulatory compliance and accountability procedures;

► to support businesses, especially SMEs, to avoid uncertainty and unnecessary burdens; and

► to provide practical experience and create guidelines, toolkits and good-practice materials for the development of harmonized European standards. With this initiative, the Commission is also seeking to create synergies with other national initiatives to develop a pan-European AI system of sandboxes.

It is also in the middle of creating an Observatory to monitor the social impact of algorithms, operating under the Spanish National Observatory for Telecommunications and the Information Society (ONTSI).

Finally, the Commission recently presented, on February 2, its proposal for a standardization strategy[4] outlining its approach to standards within the single market, as well as globally, in addition to a proposal to amend the Standardization Regulation; (1025/2012) a report[5] on its implementation, and an EU work program for European standardization for 2022.[6] Thierry Breton, Commissioner for the Internal Market, highlighted *"the strategic importance of technical standards for Europe"* and how *"Europe's technological sovereignty, ability to reduce dependencies and protection of EU values will reply on Europe's ability to become a global standard-setter"*. In the work program, the need to establish safe and trusted artificial intelligence systems is a key point and is aimed at ensuring that artificial intelligence systems can be safe and trustworthy and that they are properly monitored over their life cycles, respecting the fundamental values and human rights that are recognized by the EU and strengthening European competitiveness.

Others have also arisen in relation to market competition, regarding the use of algorithms and their transparency. In fact, in the White Paper published as part of the White Paper consultation on Artificial Intelligence launched by the European Commission, the National Commission of Markets and Competition (CNMC) and Catalan Competition Authority proposed adapting the regulations to allow competition authorities to use artificial intelligence to detect illegal activities so that, during investigations to monitor the code used in algorithms, they are able to access the information on computers or electronic media, databases or applications.[7]

---

4   https://ec.europa.eu/docsroom/documents/48598

5   Report from the Commission to the European Parliament and the Council on the implementation of Regulation (EU) No 1025/2012 from 2015 to 2020

6   The 2022 annual EU work program for European standardization

7   https://www.cnmc.es/prensa/cnmc-acco-contribucion-libro-blanco-inteligencia-artificial-CE-20200930

Existing certifications

USAII Certifications[8]

The United States Artificial Intelligence Institute offers three certifications for individuals which, in turn, certify the adequate development of AI systems. Therefore, it is not a certificate for a product or model, it is a credential to assess them. There are three different versions:

► Certified Artificial Intelligence Engineer (CAIE™): Credential that certifies that you have basic but adequate knowledge of the field.

► Certified Artificial Intelligence Consultant (CAIC™): Credential that certifies that you are able to orchestrate expertise on the deployment and management of AI systems.

► Certified Artificial Intelligence Scientist (CAIS™): Credential that certifies that you are able to lead complex projects that require the use of AI.

Equal AI Badge

The EqualAI Badge program, developed in collaboration with the World Economic Forum, is designed to equip professionals and executives with the necessary know-how to ensure that the practices within a business are responsible and inclusive. As above, this not a technical certification, it is a personal and introductory certification.

RAII Certification

The RAII consists of three main components: a white paper,[9] detailing how the certification conforms to the current regulations, with input from experts, a certification guide that facilitates the certification process and a sample for certification planning, together with the various points that must be

8   https://www.usaii.org/artificial-intelligence-certifications
9   https://drive.google.com/file/d/1272HcpXUbSkZWkK9MKPzzsUD3uKsfPXa/view

evaluated in the process. This certification has four areas of focus (financial services, healthcare, procurement and human resources), it is based on the OECD's guidelines and it details the principles adopted in the various aforementioned standards (IEEE, ISO, etc.).

2.4 AI Ethics Impact Group[10]

The AI Ethics Impact Group is an interdisciplinary consortium led by VDE Association for Electrical, Electronic & Information Technologies and Bertelsmann Stiftung. In view of the different AI technologies being adopted, the group has developed a framework to operationalize ethical principles and assess different systems according to those principles.

The framework is based on the VCIO model (i.e. values, criteria, indicators and observables) that allows different values to be operationalized and specified for different problems. This model has three main components: the VCIO model, an ethical *rating* (based on energy efficiency labels) and a two-dimensional matrix that is used to classify different application contexts for AI systems.

The idea behind this methodology is to obtain a label for any system that implements AI, based on six aspects: transparency, accountability, privacy, justice, reliability and sustainability. For this purpose, the risk matrix provides the minimum requirements for a system in a specific application context. Meanwhile, the VCIO model allows the model's consistency with a set of predetermined values to be measured. The final score is obtained by combining both aspects.

Figure 1. **Label**
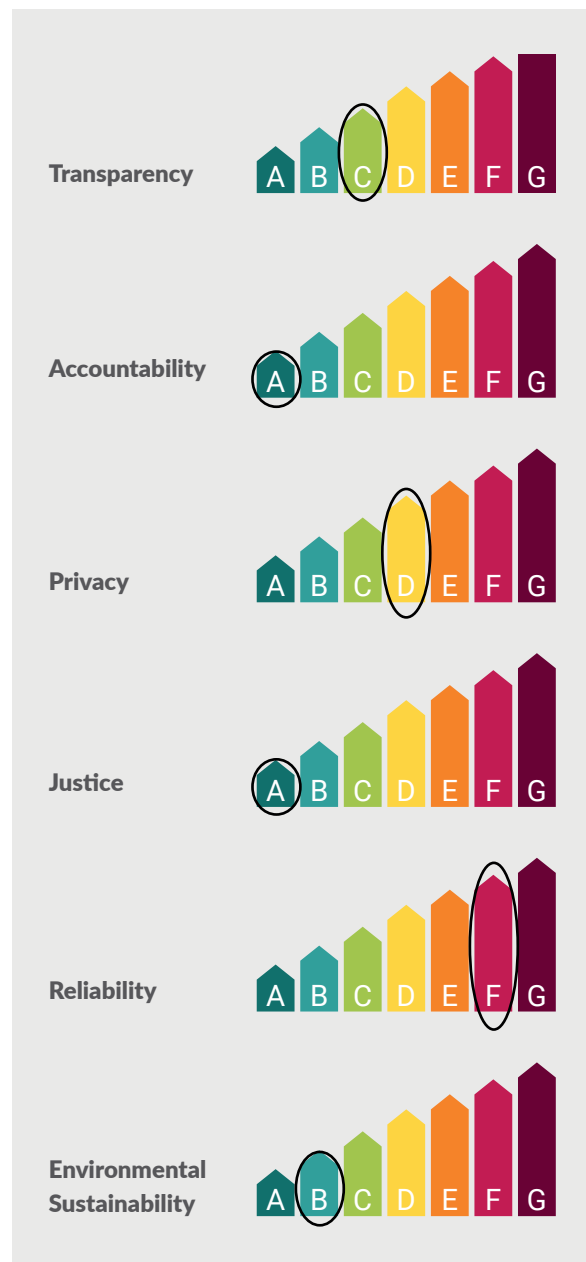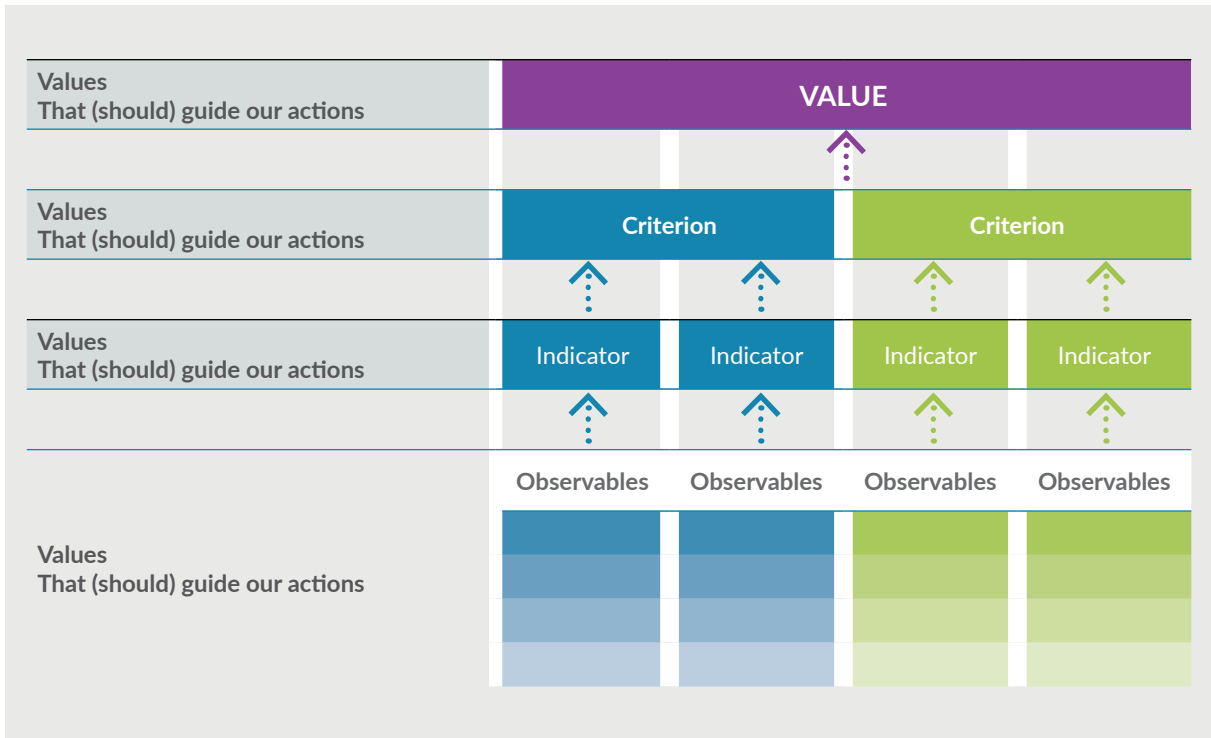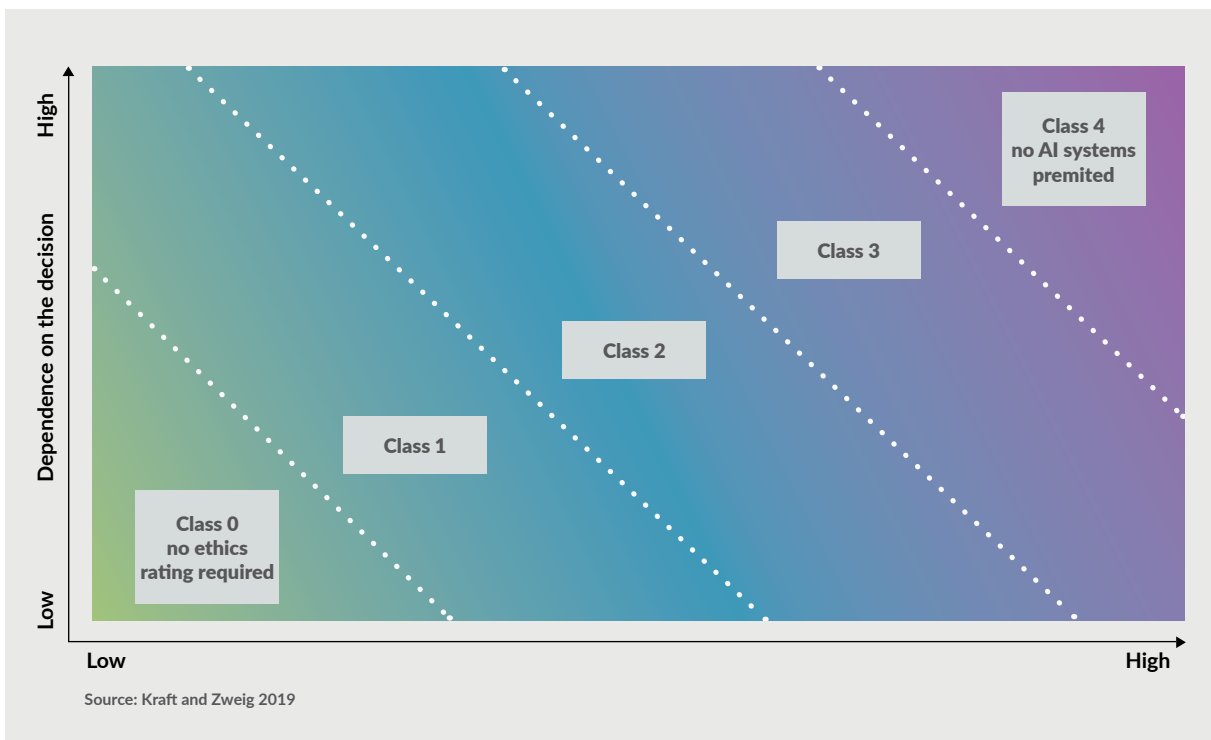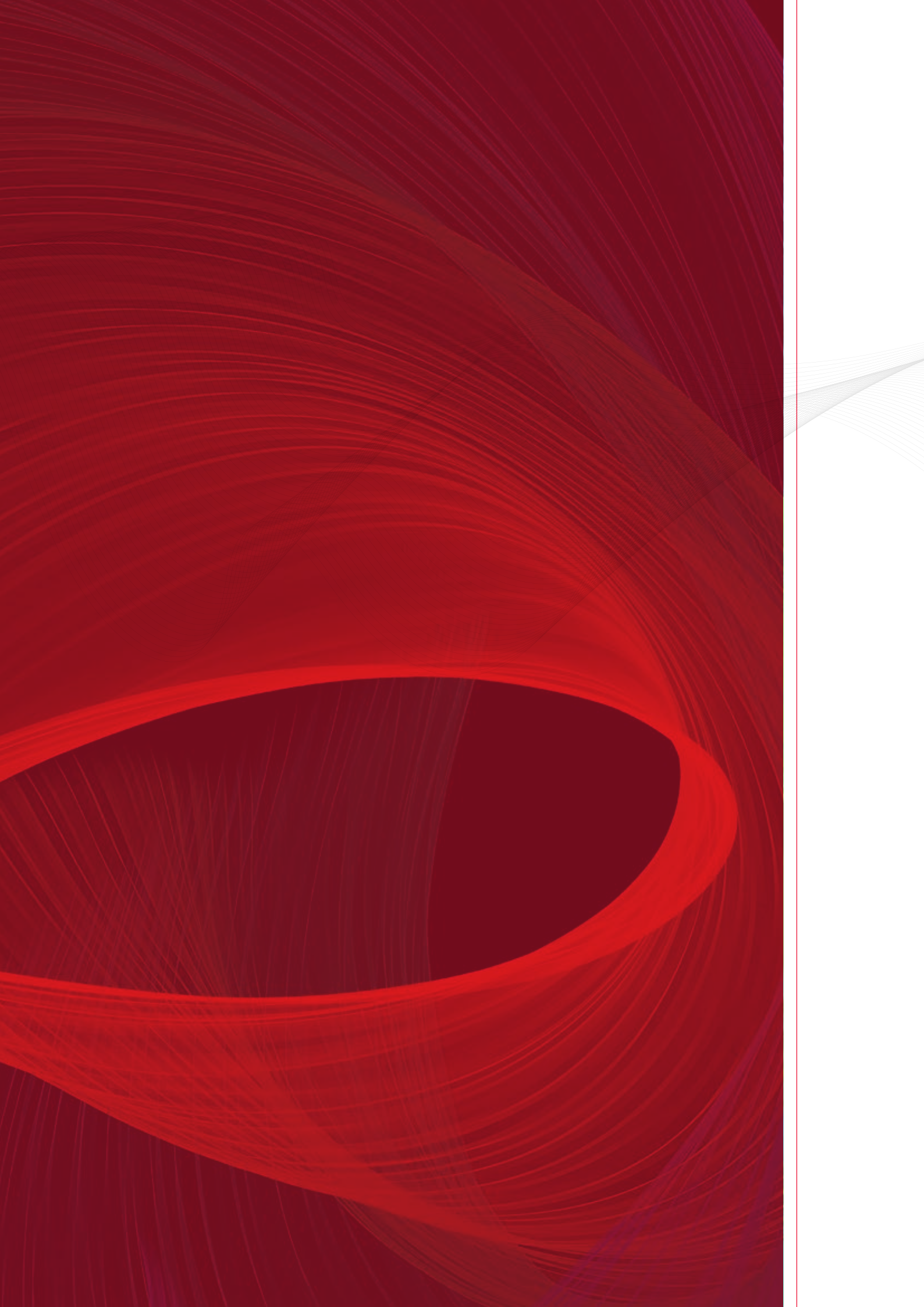
Figure 2. **VCIO model**

Figure 3. **Risk matrix with 5 classes of application areas with risk potential ranging from ´no ethics rating required´ in class 0 to the prohibition of AI systems in class 4**

Source: Kraft and Zweig 2019

Transparency

From the user's perspective, transparency in AI should be seen as a mechanism capable of increasing their confidence in complex, non-deterministic systems through their understanding - even when this is superficial - of both the structure and the information flows of those systems. However, it should be noted that transparency by itself does not guarantee the quality or fairness of the results produced by AI.

Thus, transparency should be seen as a key element for ensuring that different stakeholders are able to access sufficient information to make a decision about whether to adopt an algorithmic system and, if so, the risks that this entails. This, in turn, is only possible if the purpose, architecture, benefits and dangers, or corrective mechanisms of such systems are sufficiently clear. Ultimately, transparency should be seen as a precursor to ethical and impact assessments, as it promotes accountability and a critical approach to the outputs generated by these systems.

In this sense, transparency is usually seen as a mechanism that facilitates or ensures the explainability of artificial intelligence systems. By providing information about the technical aspects and principles of a given system, it makes it possible to explain the results that it provides or, at least, contextualize them more precisely from technical and social perspectives.

However, transparency should be approached as the complex exercise that it is: making the code freely accessible or publishing assessments of it, irrespective of the fact that the regulatory or societal context may be counterproductive. Some months ago, when Elon Musk declared his intention to acquire Twitter, he announced that he intended to provide full access to the platform's algorithms to increase transparency.[11] However, various experts identified fundamental problems in this understanding of transparency: firstly, they remarked that access to the source code does not provide information on how the learning models have been trained or evaluated. Secondly, they stated that the risks entailed in opening up Twitter's code (e.g. leading to misuse of the platform) could outweigh the potential benefits of such transparency.[12]

---

11 https://www.prnewswire.com/news-releases/elon-musk-to-acquire-twitter-301532245.html

12 https://www.technologyreview.com/2022/04/27/1051472/the-problems-with-elon-musks-plan-to-open-source-the-twitter-algorithm/

Finally, technical transparency (in the sense of allowing access to all or part of the code) can help to build trust among different stakeholders. However, given the status of AI systems as key elements in the operations of numerous technology companies, the limitations to disclosing certain elements of the models used by private companies are evident.

In turn, the discussion on transparency can be approached from different perspectives: on the one hand, it can be viewed as an exercise aimed at providing information about a single model within a service. On the other hand, it can be seen as information about a service as a whole, without delving into its component parts or the way data is collected and managed. Each of these perspectives provides information about different aspects at different levels of abstraction, which will be analyzed in greater detail below.

Data

The literature regarding the ethics of artificial intelligence has a plethora of proposals on how to monitor the fairness of results and proper database preparation. Regarding the latter point, domestic and international organizations have made numerous proposals (e.g. Europe's General Data Protection Regulation or GDPR, or the methods proposed in California[13] or Brazil[14]), and the well-known *Datasheets for Datasets. (Gebru et al., 2021)*

If we focus on the European case, while it is true that the GDPR covers data collection and management (emphasizing the criteria under which certain data can be collected or the requirements for data minimization and not selling data to third parties), it also proposes **transparency mechanisms in relation to data**. More specifically, articles 13 and 14

of the regulation not only cover the right to be informed when personal data is collected, used or accessed, but they also enshrine the principle of transparency, requiring any type of information or notification regarding the processing of personal data to be easy to understand.

Various initiatives covered in the GDPR seek to focus these principles in different areas, whether in the Internet of Things (IoT) in Wachter, 2018, or in the introduction of elements that reflect the child-related legal requirements for platform protocols in Milkaite & Lievenes, 2020.

However, transparency in relation to data may result in decontextualized information being provided which, in the event of poor practice when implementing database traceability, may lead to problems with interpretability, (Abiteboul & Stoyanovich, 2019:4) which may be exacerbated by the lack of unified criteria in the field.

While it is true that the information and transparency requirements for data have been regulated more extensively, the **development of regulatory proposals regarding artificial intelligence** (e.g. the Artificial Intelligence Act in Europe,[15] or the reform of the workers' statute in Spain to include transparency requirements for work involving AI[16]), **marks a new line of work, focused on technical aspects beyond data, which the various stakeholders must explore.**

Algorithmic transparency

In 2019, Google researchers developed the *Model Cards for Model Reporting* tool, (Mitchell et al., 2019) leading the development of socio-technical tools for the evaluation of artificial intelligence models. During the following

---

13 https://oag.ca.gov/privacy/ccpa

14 https://www.lgpdbrasil.com.br/

15 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

16 https://prensa.mites.gob.es/WebPrensa/noticias/ministro/detalle/4118

years, changes were made to this tool, such as the *Reward Reporting* proposal based on a reinforcement learning approach, (Gilbert et al., 2022) or various government proposals such as the one put forward by CIPEC and the Inter-American Development Bank, (Galdón & Lorente, 2022) or the ALTAI[17] list by the European Commission.

However, despite the potential offered by model assessment exercises, they still have certain limitations. Firstly, in business terms, they usually provide a lot of details about the way the models work, and the parameters used and certain design criteria, which could generate a conflict of interests or lead to certain competitive disadvantages if not adopted globally. Secondly, as they are developed from an academic and research perspective, these tools usually overlook the interactions between AI models and other elements that are crucial for the provision of a specific service, and propose solutions that cover a narrow range of technologies.

In this regard, cities like Amsterdam[18] and Helsinki[19] have implemented **algorithm registers** in which the reports on the artificial intelligence systems used by those city councils are contained and made available for users to consult. This allows the different stakeholders to learn about and understand the nature of those systems and their fields of application, being the first public databases available to consult and assess the implications of using AI.

Some of the limitations of this proposal are due to the context in which it has been implemented: products commissioned and implemented by city councils are likely to be compatible with pre-existing frameworks and supported by comprehensive evaluations, which allows us to conclude that in different city councils or

other regional organizations its use would be not only appropriate but also recommended. (Floridi, 2020:543) But it is also true that in an ecosystem with more diverse stakeholders and systems, the creation of a single useful register could be undermined. Furthermore, although **using registers** helps to build public trust in the AI elements used, **it is not enough**. To ensure effective involvement, it is necessary for the *feedback* to have tangible consequences (what is known in the field as accountability). (Ibid., 545)

System transparency

Although one of the main reasons why transparency exercises are promoted is to facilitate understanding of the nature of the various AI models used and to help eliminate "black box" type models, knowing the design parameters, the variables used, or the performance of a specific model may not be enough to determine the possible risks involved in a service in which it is employed. It is in this sense that the use of model ethical assessment tools comes up against the practical limitations of using models in business: to offer a specific service, more than one AI model may be used and they may be different in nature. In such an event, how can transparency be exercised?

Government proposals such as the algorithmic transparency tool[20] promoted by the Spanish Ministry of Labor achieve this by delimiting the domain of application. When focusing on the workplace, a single questionnaire is able to reflect the essential elements to assess the impact of the different models used despite their technical differences. However, when facing ecosystems with diverse activities with varied technical needs, **establishing a single questionnaire to exercise algorithmic transparency is more complex**.

---

17 https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence
18 https://algoritmeregister.amsterdam.nl/en/ai-register/
19 https://ai.hel.fi/en/ai-register/
20 https://prensa.mites.gob.es/WebPrensa/noticias/ministro/detalle/4118

In this sense, following the agile development methodology, there are proposals such as the TIRA toolbox, which make it possible to translate descriptions of APIs such as REST (representational state transfer) - covering both the interface and the description of a service - into **transparency requirements based on the General Data Protection Regulation. (Grünewald et al., 2021)**

Additionally, tools such as the responsible AI standard developed by Microsoft[21] establish various transparency goals, such as the intelligibility of the systems used for decision-making, communication with the various stakeholders and the need to disclose an AI interaction to a user. For this purpose, the proposal to implement those requirements consists of preparing semi-structured notes to provide information on the transparency of the systems, or evaluation mechanisms such as audits or impact assessments.
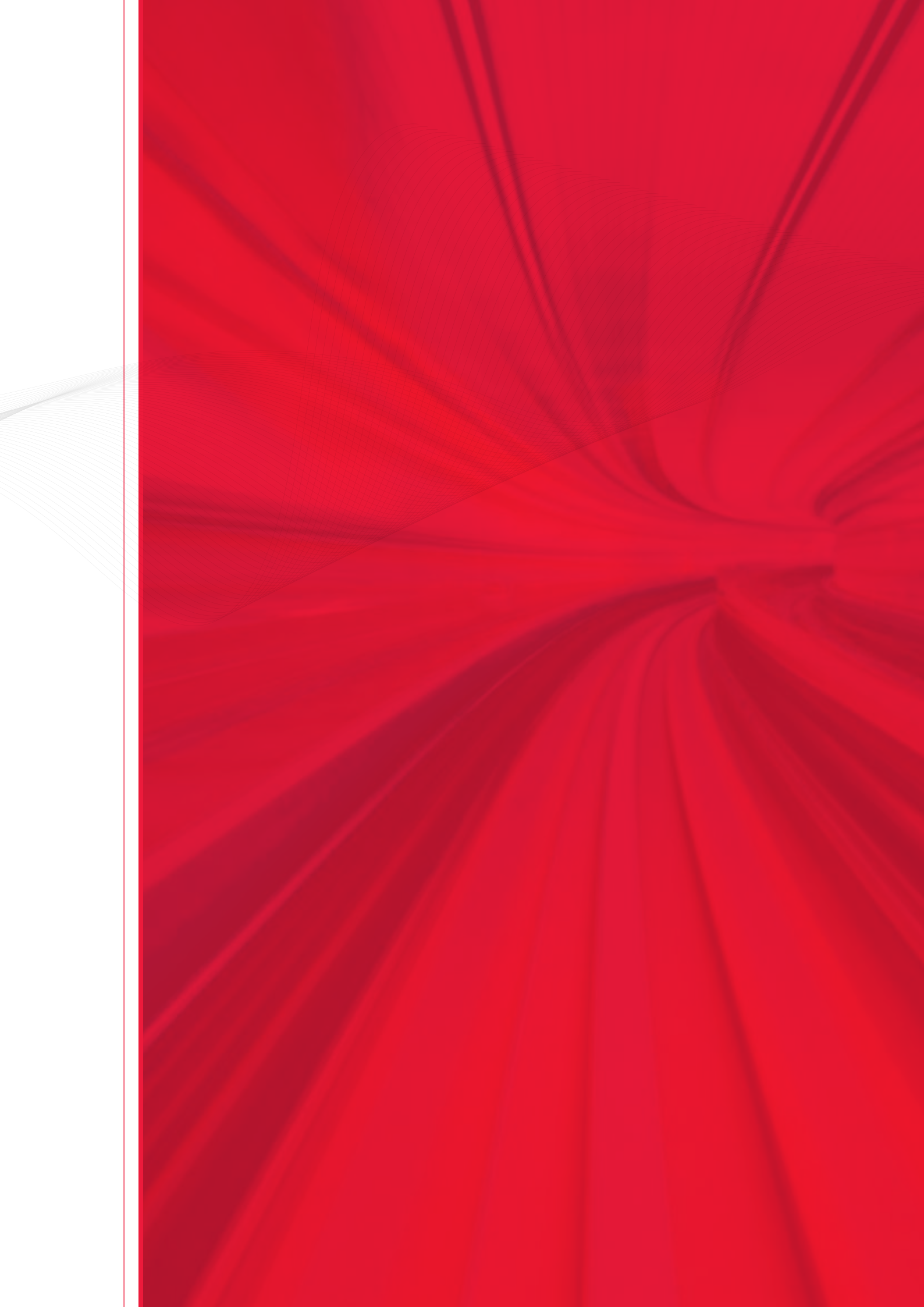
Summary and proposal

The complexity inherent in AI (both at a technical level and in the areas affected by it) means that there is a gap between the theoretical proposals to facilitate the flow of information from developers to different stakeholders and the need to supervise and organize the widespread use of such technology. However, there is a range of proposals aimed at bridging this gap: from transparency exercises for the collection and management of data to detailed reports on the nature, parameters and performance of different AI models, different stakeholders propose solutions to translate the AI principles of transparency and ethics into specific requirements.

Consequently, viewing transparency as a precursor to the ethical evaluation of the different models involved in providing a service

has various benefits. Firstly, revealing – to a greater or lesser extent – the architecture behind a service allows the companies that own the technology to retain their competitive advantage. The use of certain data points or parameters is necessary to assess the performance of the models, but it is not strictly necessary to carry out an initial evaluation of the potential risks that the technology entails.

Secondly, this mechanism is applicable to different technologies in any field of application. From a regulator's perspective, this makes it possible to implement a **single requirement for any company that uses AI**. Finally, producing documents that allow transparency to be implemented in the architecture makes it easier to identify the trouble spots in the system, enabling a more efficient and effective use of the auditing resources or model monitoring.

---

21 https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf

Bibliography

Abiteboul, S., & Stoyanovich, J. (2019). Transparency, fairness, data protection, neutrality: Data management challenges in the face of new regulation. *Journal of Data and Information Quality (JDIQ)*, *11*(3), 1-9.

Ala-Pietilä, P., Bonnet, Y., Bergmann, U., Bielikova, M., Bonefeld-Dahl, C., Bauer, W., ... & Van Wynsberghe, A. (2020). *The assessment list for trustworthy artificial intelligence (ALTAI)*. European Commission.

Floridi, L. (2020). Artificial intelligence as a public service: Learning from Amsterdam and Helsinki. *Philosophy & Technology*, *33*(4), 541-546.

Galdon Clavell, G., & Lorente Martínez, A. (2019) *Hacia un prospecto en el marco regulatorio laboral en Argentina: Análisis tecnológico, marco regulatorio y buenas prácticas*. IADB: Inter-American Development Bank.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86-92.

Gilbert, T. K., Dean, S., Zick, T., & Lambert, N. (2022). Choices, Risks, and Reward Reports: Charting Public Policy for Reinforcement Learning Systems. *arXiv preprint arXiv:2202.05716.*

Gobierno de España - Ministerio de Asuntos Económicos y Transformación Digital. Estrategia Nacional de Inteligencia Artificial

Grünewald, E., Wille, P., Pallas, F., Borges, M. C., & Ulbricht, M. R. (2021, September). TIRA: an OpenAPI extension and toolbox for GDPR transparency in RESTful architectures. In *2021 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* (pp. 312-319). IEEE.

Milkaite, I., & Lievens, E. (2020). Child-friendly transparency of data processing in the EU: from legal requirements to platform policies. *Journal of Children and Media, 14*(1), 5-21.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).

Ley 12/2021, de 28 de septiembre, por la que se modifica el texto refundido de la Ley del Estatuto de los Trabajadores, aprobado por el Real Decreto Legislativo 2/2015, de 23 de octubre, para garantizar los derechos laborales de las personas dedicadas al reparto en el ámbito de plataformas digitales.

Ley 22/2021, de 28 de diciembre, de Presupuestos Generales del Estado para el año 2022.

Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación.

Proposal for a Regulation amending Regulation (EU) No 1025/2012 as regards the decisions of European standardisation organisations concerning European standards and European standardisation deliverables

Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos - RGPD)