Conformity assessment under the EU AI act general approach

Thelisson, Eva; Verma, H.

DOI
10.1007/s43681-023-00402-5

Publication date
2024

Document Version
Final published version

Published in
AI and Ethics

AI and Ethics (2024) 4:113–121
https://doi.org/10.1007/s43681-023-00402-5

ORIGINAL RESEARCH

Conformity assessment under the EU AI act general approach

Eva Thelisson[1] · **Himanshu Verma**[2]

Accepted: 22 November 2023 / Published online: 3 January 2024

Abstract
The European Commission proposed harmonised rules on artificial intelligence (AI) on the 21st of April 2021 (namely the EU AI Act). Following a consultative process with the European Council and many amendments, a General Approach of the EU AI Act was published on the 25th of November 2022. The EU Parliament approved the initial draft in May 2023. Trilogue meetings took place in June, July, September and October 2023, with the aim for the European Parliament, the Council of the European Union and the European Commission to adopt a final version early 2024. This is the first attempt to build a legally binding legal instrument on Artificial Intelligence in the European Union (EU). In a similar way as the General Data Protection Regulation (GDPR), the EU AI Act has an extraterritorial effect. It has, therefore, the potential to become a global gold standard for AI regulation. It may also contribute to developing a global consensus on AI Trustworthiness because AI providers must conduct conformity assessments for high-risk AI systems prior to entry into the EU market. As the AI Act contains limited guidelines on how to conduct conformity assessments and ex-post monitoring in practice, there is a need for consensus building on this topic. This paper aims at studying the governance structure proposed by the EU AI Act, as approved by the European Council in November 2022, and proposes tools to conduct conformity assessments of AI systems.

Keywords EU AI Act · Conformity assessment · Assessment tools · AI governance · AI Trustworthiness

1 Introduction

In a silent and disruptive way, Artificial Intelligence (AI) systems transform many sectors of activities—both public and private. This is due to data-driven business models as well as the availability of a massive volume of data, computational power, and machine learning algorithms. AI applications increase efficiency while reducing costs. It also raises new challenges from a technical, legal, and ethical perspective. The large-scale deployment of AI applications, based on data fusion, data sharing and the automation of decision-making processes may lead to specific risks and harmful practices or misuse.

Following the entry into force of the GDPR in May 2018, aiming at strengthening the free flow of personal data in the single market while simultaneously reinforcing the control of the data subject on personal data, the EU Commission is now willing to bring legal certainty in the field of AI to reinforce the Single Market. It wants to build governance mechanisms to create safeguards regarding the lawful, safe, and trustworthy use of High-Risk Systems (HRS), that must conform with European values and human rights (art. 2, Lisbon Treaty). The EU AI Act proposes a risk-based approach. Instead of offering public and private enforcement mechanisms ex-post only, as is the case with the GDPR, the EU AI Act creates ex-ante governance mechanisms for high-risk AI applications, with the aim to prevent AI from causing harm to its users. Due to its extraterritorial effect, it has the potential to export European values abroad. This phenomenon is called the Brussels Effect by Prof. Bradford [7] from Columbia University in 2020 in her book "How the EU rules the world". Therefore, the EU AI Act may contribute to a global consensus on AI Trustworthiness. This global consensus must be reinforced by the new European Artificial Intelligence Board which may play a key role in collecting and sharing best practices among member states while issuing specific recommendations to strengthen the Single Market with appropriate safeguards.

✉ Eva Thelisson
eva@aitransparencyinstitute.com

Himanshu Verma
H.Verma@tudelft.nl

[1] AI Transparency Institute, Lausanne, Switzerland

[2] Knowledge and Intelligence Design Group at TU Delft, Delft, The Netherlands

2 A risk-based approach

The EU AI Act classifies AI systems in 4 categories based on the level of risks: low or minimal risk, limited risk, high risk, or unacceptable risk. For example, common AI applications like spam filters or video games belong to the low-risk category. For each risk level, the obligations of the AI provider are defined. Low risk systems do not have any obligations to fulfil. Systems with limited risk are those that (i) interact with humans, (ii) detect humans or determine a person's categorisation based on biometric data, or (iii) produce manipulable content. Chatbots and deep fakes are in this category. Organisations using systems with limited risk must comply with transparency obligations, i.e., users must be informed that they are interacting with an AI system, an AI system will be used to infer their characteristics or emotions, and/or the content they are interacting with has been generated using AI. The third category relates to high-risk systems. These systems can have a significant impact on the life of a user. These AI systems must comply with some requirements ex-ante, i.e., before any deployment on the EU market. The last category relates to unacceptable risk. These AI systems are banned from sale on the EU Market. In this category, we can find AI systems able to manipulate behaviour in a way that may result in physical or psychological harm, exploit the vulnerabilities of a group based on their age, physical or mental disability, or socioeconomic status. Also, AI systems used for social scoring by governments or for real-time biometric monitoring in a public area by law enforcement or on their behalf (except for those that meet strict criteria) are prohibited. Germany proposed banning biometric recognition technology and favours banning real-time biometric identification in public spaces while allowing ex-post identification. Furthermore, Germany advocates prohibiting any AI application that substitutes human judges in legal assessments of an individual's risk of committing a crime or repeat offending.

3 High-risks use cases

The EU AI Act identifies 8 use cases for high-risk systems. Specific requirements must be fulfilled ex-ante for these AI systems. Providers of AI systems must ensure that high-quality data have been used, that appropriate documentation is in place, that transparency practices are fulfilled, that adequate human oversight can be demonstrated, as well as testing processes for accuracy and robustness. The first use case deals with biometric identification systems used for real-time and post-remote identification of people without their agreement. The second use case relates to systems for critical infrastructure and protection of the environment, including those used to manage pollution. The third use case relates to education and vocational training systems used to evaluate or influence the learning process of individuals. The fourth use case deals with employment, talent management and access to self-employment. The fifth high-risk system relates to the access and use of private and public services and benefits, including those used in insurance. The sixth use case deals with AI systems used in or on behalf of law enforcement. The seventh use of AI systems deals with migration, asylum, and border control, including systems used on behalf of public authority. Finally, the last use case deals with AI systems used in the administration of justice and democratic processes, including systems used on behalf of the judicial authority. In the General Approach discussed at the EU Parliament, AI systems used for purposes of health and life insurance constitute high-risk. HRS are also classified as high-risk if they are, or are part of, the safety component for products covered by EU harmonisation legislation, or if they fall within the categories and use cases listed in Annex III. Members of the European Parliament recently added an additional requirement that the high-risk list only refers to systems with an intended purpose. Therefore, general-purpose AI will be treated separately pending further discussions. For applications listed in Annex III, a system will only be considered high-risk if it receives personal or biometric data as inputs or is intended to make or assist decisions affecting individuals' health, safety, or fundamental rights. The European Commission revised the conditions for assessing new risks. To remove used cases the EU Commission should consult with the EU AI Office, based on a procedure in a delegated act.

4 general-purpose AI systems

The EU AI Act plans to regulate generative models, namely "general-purpose AI" systems. These models can be used for many different applications and process different sources of data. Trade secrets protect companies to explain how applications derived from these models are built and how algorithms have been trained. Today, it is difficult to interpret how exactly the models generate harmful content or biassed outcomes, or how to mitigate those problems. The exact way in which these models will be regulated in the AI Act is still under debate, but in any case, creators of general-purpose AI models will likely need to be more open about how their models are built and trained [15]. The content of the EU AI Act depends on the stakeholders influencing the European legislative and standardisation process. Therefore, it is important to present these double processes. The

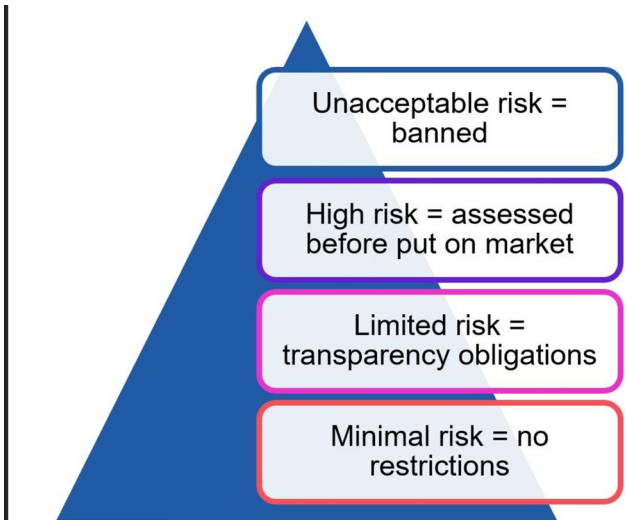

Fig. 1 The categorized risk levels in the EU AI Act

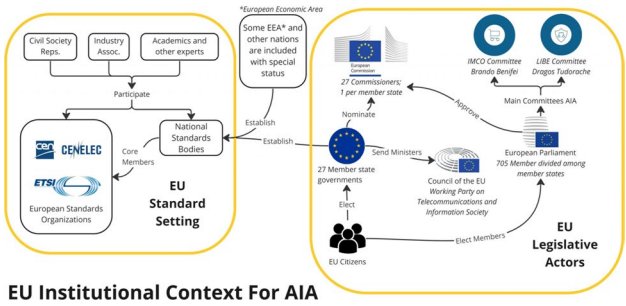

Fig. 2 EU Institutional Context for AI Act (Source: The Future of Life Institute)

figure below provides an overview of this double process (Figs. 1,2).

5 New legislative framework (NLF)

The EU adopted the New Legislative Framework (NLF) in 2008 to improve the internal market for goods, enhance the conditions for placing a wide range of products on the EU market, and increase the quality of conformity assessments and market surveillance. The NLF also clarifies the use of CE marking. The EU AI Act, which is based on this framework, sets out "essential requirements' that AI systems must meet to access the EU market. Companies can fulfil these legal obligations in their own way, based on "harmonised standards" that address the essential requirements. As described by [11], compliance with these standards creates a presumption of conformity with the relevant essential requirements. Following publication in the Official Journal of the EU, these standards became "harmonised standards." The EU AI Act requires organisations to adopt "suitable risk management measures," without specifying what is considered "suitable," which will be clarified during the standardisation process. Companies can demonstrate that their solution is at least equivalent to the standard if they choose not to implement the harmonised standards.

Once the EU AI Act is approved by the EU Parliament and the Council, harmonised standards, known as "European Standards," will be developed based on a draft developed by JTC1 (ISO and IEC Joint Technical Committee). However, European Standards will not be explicitly aligned with the essential requirements outlined in the EU AI Act until they are developed. In December 2022, the European Commission published a draft standardisation request for the EU AI Act in support of safe and trustworthy AI. According to EURACTIV [6], the new draft of the standardisation request for the EU AI Act includes significant changes compared to the earlier version from May 2022. The EU Commission has removed the European Telecommunications Standards Institute (ETSI)—one of the three European standardisation organisations—and instead addressed the request to the European Committee for Standardisation (CEN) and the European Committee for Electrotechnical Standardization (CENELEC). The European standardisation bodies will be required to have experts on fundamental rights, which means that standards must protect the principles, rights and values from the Charter of EU Fundamental Rights [European Union, [2]] and not be limited to technical standards. The request also calls for alignment with internationally recognized terminology and a reference to the EU-US AI Roadmap, a commitment to develop a shared understanding around critical concepts such as risk and trustworthiness, with specific metrics to measure them. The request also defines the state of the art as "technical capability based on consolidated technology and scientific findings" and excludes experimental stage research and immature research. The notion of accuracy is replaced by statistical accuracy and the impact for SMEs is taken into account by requesting CEN-CENELEC to consider the cost for quality management system and conformity assessment for SMEs. To ensure that European Standards are in place before the AI Act is enforced, CEN CENELEC IJTC 21 must complete and submit the deliverables to the EU Commission by January 31st, 2025. Once the AI Act is adopted, harmonised standards will be developed based on European Standards, which may necessitate a sector-specific approach.

6 Scope of the EU AI Act

6.1 Material scope

The definition of AI is still unclear and it would be consistent to align it with international frameworks like the

OECD and NIST. Parliament wants to impose an obligation on providers of foundation models to ensure a robust protection of fundamental rights, health, safety, the environment, democracy and the rule of law. Furthermore, generative foundation AI models that use large language models to generate art, music and other content would be subject to stringent transparency obligations. Finally, all foundation models should provide all necessary information for downstream providers to be able to comply with their obligations under the AI Act. Transparency requirements for HRS are proposed like standardised information about the model in the form of model cards and data sheets, access to the data used to train and fine-tune the model for the purpose of auditing, and explanation of the model's behaviour [12]. What is important is to create the conditions for users to comply with the obligations of the AI Act for high-risk uses of the model. Details of compliance assessment for general-purpose AI should be worked out by the European Commission after the Act enters into force, which creates a legal uncertainty. Parliament agrees that research activities and the development of free and open-source AI components would be largely exempted from compliance with the AI Act rules, to support innovation. European Council representatives also consider that law enforcement requires some of the AI tools that the European Parliament representatives want to ban like facial recognition and remote biometric identification systems. The use of these systems by law enforcement and intelligence agencies is challenged by civil society representatives.

6.2 Territorial scope

If providers of AI systems established in the EU must comply with the EU AI Act, those based in third countries that place AI systems on the market in the EU, and those located in the EU that use AI systems, will also be obliged to implement the EU AI Act. Providers and users based in third countries will also have to respect the EU AI Act, if the output of the system is used within the EU. However, if AI systems are used for military purposes in the EU or by public authorities in third countries, the EU AI Act won't be applicable.

6.3 Enforcement

National authorities competences have been strengthened and Parliament proposes also to establish an AI Office, a new EU body to support the harmonized application of the AI Act, provide guidance and coordinate joint cross-border investigations. The AI Office will also supervise the implementation of the tiered approach to foundational models.

6.4 6.4 Conformity assessment

Conformity assessments (CA) is a legal obligation, which must be fulfilled before a high-risk AI system is placed on the market. It aims at fostering accountability of AI providers. The EU AI Act defines a conformity assessment as the process of verifying whether the requirements set out in Title III, chapter 2 of the regulation relating to an AI system have been fulfilled, while Title III offers provisions for high-risks systems only. If a product is determined to meet all the relevant requirements, a declaration of conformity is issued, and a "CE" symbol is applied to the product. Companies must document the assessment to prove it was conducted correctly and may also delegate this responsibility to conformity assessment bodies, known as "notified bodies," which are accredited by "notifying authorities" established by member states. Public authorities can also serve as notified bodies. The EU AI Act may allow for self-assessment for some high-risk applications and require notified bodies for others, such as biometrics applications. The EU AI Act specifies the types of data that market surveillance authorities should have access to, such as documentation, datasets, and source code, and under what conditions. It also details how authorities should coordinate with the Commission, notified bodies, or authorities in other countries. Each member state must ensure market surveillance, which includes removing products that do not comply with EU legislation or have been found to be too dangerous, regardless of compliance.

6.5 Conformity assessment in practice

According to Article 3 (2) of the EU AI Act, the Conformity Assessment (CA) can be performed by the AI provider, the product manufacturer, the distributor, the importer of the high-risk system (HRS) or a third party. The product manufacturer is considered competent to perform the CA if the laws of Annex II section A of the EU AI Act apply, and in that case, the AI system is placed on the market and under the name of the product manufacturer (Article 24, recital 55). If the HRS system is put on the market or put into service under the name or trademark of the distributor, importer or any third party, or if they modify the intended purpose of HRS as determined by the provider, then they must perform the CA. In that case, the initial provider is no longer considered as the provider. If the initial provider makes a substantial modification to the HRS, then they are also no longer considered as the provider (Article 28).

6.6 Internal conformity assessment

The Conformity Assessment (CA) can be performed by the provider (or the distributor, importer or third party) who are well equipped and have the expertise to assess the compliance of AI systems (recital 64). However, they may have a conflict of interest in being transparent and may prefer not to disclose everything [Demetzou, [1]]. The CA includes a verification that the Quality Management System (QMS) complies with Article 17 and that the information in the technical documentation meets the legal requirements of the EU AI Act. It also includes a verification that the design and development process of the AI system and its post-market monitoring (Article 61) are consistent with the technical documentation. Following this verification, an EU Declaration of conformity for each AI system can be issued (Article 19(1)) and must be kept for 10 years and a copy should be provided to the national competent authority upon request. The responsible entity must affix a visible CE marking of conformity (Article 49, Article 30). It remains questionable that the industry itself will determine whether their own AI systems are high risk or not depending on various scenarios deemed high risk. The certification regime for high-risk AI systems is intended for use in these scenarios. However, AI system won't be classified as high risk, if it only performs "purely accessory" tasks. It will be the case if the AI system perform a narrow procedural task, or detect deviations from decision-making "patterns", or does not influence a decision, such as whether to provide a loan or to make a job offer or only improve the quality of work, such as a smart grammar checker. It would be important to clarify who bears the burden of proof.

6.7 External conformity assessment

In some use cases, the Conformity Assessment (CA) is performed by an external body, specifically a notified body, which is a CA body designated by the national notifying authority and must fulfil specific requirements (Article 33). If there are any major changes to the system such as retraining the model on new data or removing some features from the model, the system must undergo additional conformity assessments to ensure that the requirements are still being met before being re-certified and registered in the database.

External CA is only required for AI systems intended to be used for real-time and post-remote biometric identification for people that are not using the harmonised standards or common specifications of Article 41. If the high-risk system (HRS) is a safety component of a product and specific laws enumerated in Annex II, section A apply to it, the provider must follow the type of CA process stipulated in the relevant legal act. The notified body is responsible for assessing the quality management system (QMS) and the technical documentation (Annex VII) which must be included in the provider's application.

The notified body will issue an EU technical documentation assessment certificate (Article 44) which has a limited time validity and can be suspended or withdrawn by the notified body. The provider is responsible for creating the EU declaration of conformity and applying the CE marking of conformity. They must also prepare an EU declaration form which includes a description of the conformity assessment procedure that was performed (Article 19(1)). Article 45 grants the provider or any actor with a legitimate interest the right to appeal if the notified body considers the assessment of a high-risk system (HRS) as not being in conformity with the requirements for HRS. The decision and reasoning must be communicated in detail. Conformity Assessment (CA) is a continuous process aimed at evaluating the ongoing compliance of AI systems with the EU AI Act requirements for HRS. A post-market monitoring system must also be established, documented, and can be part of the technical documentation or product plan. Periodic audits must be carried out by the notified body to ensure that the Quality Management System (QMS) is maintained and applied. As the EU AI Act requires member states to designate a competent authority to supervise its implementation, ex-ante conformity assessment as well as post-market monitoring systems may be under supervision. This double mechanism is consistent with the deployment of a central EU database to increase transparency of HRS.

6.8 A comparative approach between the data protection impact assessment under GDPR and the conformity assessment under the EU AI Act

The General Data Protection Regulation (GDPR) requires a Data Protection Impact Assessment (DPIA) to be conducted if the processing of personal data is likely to result in a high risk to the rights and freedoms of individuals. The documentation of a DPIA enables the data controller to demonstrate that they acted in a diligent and responsible manner before processing the data in the event of damage or a lawsuit. Some data processing activities can be associated with an AI system as defined under the EU AI Act such as automated decision-making or making a significant contribution to such decision making. If an AI system falls under the scope of the EU AI Act and qualifies as a HRS, a Conformity Assessment (CA) must be carried out. The organisation responsible for the CA does not have discretion over whether to conduct the CA, regardless of whether personal data is processed. If personal data are processed, a Data Protection Impact Assessment (DPIA) may also be required in addition to the CA, and both can be conducted by the same organisation, such as the data controller. The scope of a DPIA and a CA is

different, with the DPIA assessing the nature, scope, context and purposes of processing, its necessity and proportionality to its aim by the data controller, while a CA verifies that the high-risk system (HRS) has been designed and developed according to the specific requirements of the EU AI Act for HRS. However, the CA can overlap with a DPIA as EU AI Act imposes requirements for high-risk training, validation, and testing data, data governance, and management practices in a similar way as the GDPR for sensitive data (Article 9 GDPR). These CA requirements aim to operationalize the principles of transparency and human oversight and ensure compliance with key GDPR principles such as lawfulness, fairness, purpose limitation, and accuracy principle. According to the GDPR, the data controller is required to identify risks to rights and freedoms, assess the severity and likelihood of those risks being materialised, and identify safeguards to mitigate them. A CA requires examining whether an AI system meets specific requirements set by the EU AI Act such as ensuring that training, validation, and testing data meet the quality criteria referred to in Article 10 of the EU AI Act, having the required technical documentation in place, enabling automatic recording of events (log files) when the HRS is operating, ensuring transparency of the system operation to allow the user to interpret the system's output and use it appropriately, enabling human oversight, and guaranteeing accuracy, robustness, and cybersecurity.

The provider of a HRS is responsible for conducting the CA. Those processing personal data are more likely to have the role of data processors under the GDPR in relation to the user of an AI system, as outlined in the joint opinion of the European Data Protection Supervisor (EDPS) and the European Data Protection Board (EDPB). Only data controllers have the obligation to conduct a DPIA and they may be users of AI systems under the EU AI Act. Data processors will then have to assess datasets against bias for accuracy and if they have the relevant characteristics for a specific geographical, behavioral, and functional context as part of the CA process. Due to the complexity of the AI supply chain, identifying the actors and their relevant responsibilities is central. The CA aims to guarantee compliance with certain legal requirements ex-ante, while a DPIA serves to demonstrate the diligence of the data controller ex-post. Mitigation measures for HRS (recitals 42 and 43) must be demonstrated ex-ante with a CA, thus the EU AI Act aims to increase legal certainty and to avoid fragmentation of actors involved in the AI supply chain.

6.9 Fundamental rights impact assessments and obligations for users of high-risk systems

According to EURACTIV [6] the European Parliament's co-rapporteurs circulated in January 2023 new compromise amendments to the AI Act proposing how to carry out fundamental rights impact assessments [EU [3]] and other obligations for users of high-risk systems. The co-rapporteurs want to include a requirement for all users of high-risk AI systems, both public bodies and private entities, to carry out a fundamental rights impact assessment, listing several minimum elements the assessment should include.

The EU AI Act requires users of AI systems to consider a range of factors, such as the intended purpose, geographical and temporal scope of use, categories of individuals and groups affected, specific risks for marginalised groups, and the potential environmental impact, when assessing the impact of the system on fundamental rights. This includes compliance with EU and national legislation, potential negative impacts on EU values, and considerations for public authorities such as democracy, the rule of law, and public funding. Similar to a Data Protection Impact Assessment (DPIA), users are required to draft a plan on how to mitigate any negative impact on fundamental rights and, in the absence of such a plan, they must inform the AI provider and relevant national authority without delay. Public bodies are also required to publish the results of the impact assessment as part of the registration to the EU register, and the logic of risk mitigation and safeguards documentation is similar to the GDPR. Users of AI systems considered at high-risk must ensure that they have the appropriate robustness and cybersecurity measures in place and that these measures are regularly updated. Moreover, "to the extent the user exercises control over the high-risk AI system," users would have to assess the risks related to the potential adverse effects of use and the respective mitigation measures. If the users become aware that using the high-risk system according to the instructions entails a risk to the health, safety, or protection of fundamental rights, they would have to immediately inform the AI provider or distributor and the competent national authority.

The users would have to ensure human oversight in all the instances required by the AI regulation and ensure that the people in charge have the necessary competencies, training, and resources for adequate supervision. High-risk AI users would also have to maintain the automatic logs generated by the system to ensure compliance with the AI Act, auditing any foreseeable malfunctioning or incidents and monitoring the systems throughout their lifecycle. Before a high-risk AI system is implemented in a workplace, the users should consult with worker representatives and inform and obtain the employees' consent. This raises the problem of the validity of the informed consent due to the hierarchical relationship between employees and employers. In addition, the users would have to inform the individuals affected by the high-risk system, notably concerning the type of AI being used, its intended purpose and the type of decision it makes. Distributors, importers, users, and any other third party would be considered providers of a high-risk system, with relative

obligations, under some specific circumstances (e.g., if they modify the intended purpose or make any substantial modification that makes an AI a high-risk application). This will also be the case if the high-risk system was put into services under their name or trademark unless a contractual arrangement assigns the obligations differently. When these third parties become a new AI provider, the original provider should cooperate closely with them to comply with the regulation's obligations. In a nutshell, this national authority should not conduct any conformity assessments (CA) itself. It will act as a notifying authority that assesses, designates, and notifies third party organisations. These organisations will conduct CA of providers of high-risk systems. In addition to ex-ante CA, providers of HRS will also have to establish and document post-market monitoring systems to study the behaviour and performance of HRS throughout their lifetime. They are expected to report any serious incidents or malfunctioning that constitute a breach of EU Law. They must take immediate and corrective actions to bring the AI systems under conformity or withdraw it from the market. This combination of ex-ante and ex-post controls offer a coordinated and robust approach for enforcing regulation. However, it remains unclear how to conduct conformity assessments and ex-post monitoring in practice. The AI Act contains limited guidelines on this aspect.

6.10 A comprehensive and holistic innovative approach

The AI Transparency Institute[1] (AITI) explored an innovative approach based on the research of International Institute for Management Development (IMD) in Lausanne [Bouquet, [8] and developed a methodology as well as a series of three online tools (careAI) with a SaaS platform for the upload of documentation to assess the trust and trustworthiness of high-risk AI systems.

The first tool is based on the recommendations of the EU Commission's High-Level Expert Group for Trustworthy AI from 2019. It takes into account several dimensions, as elaborated by Zicari et al. [18], and includes assessments along (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) societal and environmental wellbeing, and (7) accountability. The second tool evaluates the quality management system for HRS while considering all stakeholders (employees, clients, investors, environment, society). It has a focus on organizational aspects of the industrial process of the design, development, and continuous maintenance of HRS during the complete value chain. When we built this methodology in 2019, we took inspiration from various fields ranging from organizational theory (e.g., corporate governance), law (General Data Protection Regulation, EU Charter on Fundamental Rights, Environmental Law), standards (ISO norms), and digital ethics. This methodology was tested at EPFL (Swiss Federal Institute of Technology, Lausanne) with several complex research projects on AI involving cross borders data sharing among several jurisdictions, vulnerable participants, and the monitoring of public space at large scale. This methodology is not limited to an ethics-based auditing of AI systems as described by Mökander and Floridi [10], it is encompassing all relevant questions related to the management of AI risks within an organization. Many approaches aim at measuring and operationalizing abstract ethical principles into real-world applications, as this is the case of the Bertelsmann Foundation and VDE, or capAI which was developed by the University of Oxford. However, careAI is unique in that it provides organisations with a framework for verifying not only that ethical principles can be translated into verifiable criteria that help shape the design, development, deployment and use of ethical AI, but also that cybersecurity, privacy, human rights as recognized in the Charter of Fundamental Rights of the European Union [2], safety and environmental aspects (e.g., energy efficiency of data centres) have been considered at the governance level, as well as the impact on all stakeholders. The focus, here, is on internal organisational processes. The third tool deals with Corporate Digital responsibility. As many organizations are using AI as part of a larger ecosystem of digital technologies, it was necessary to develop a third online application with an overview of the impact of the design and deployment of digital products and services on its stakeholders.

Consisting of a comprehensive set of questionnaires, our toolkit provides a 360-degree view of the quality management system in place to mitigate the risks of each component of the responsible design, deployment, and post-market monitoring of AI as part of an industrial process that may impact different types of stakeholders. careAI goes beyond the guideline for Trustworthy AI published by Fraunhofer IAIS. It encompasses, likewise, questions on data governance, safety and security, robustness, Transparency, Human Agency and Oversight, Fairness, Post-Market monitoring but also evaluates the risks of the use of AI on human rights infringement and the environmental impact of HRS by asking questions, for example, on the policy in place regarding the energy efficiency of data centers. In a similar way as ISO norm 42,001, careAI offers self-assessment mechanisms to assist organizations in implementing a dedicated strategy on the responsible use of AI, assess its AI Policy, procedures, and guidelines. It provides a Quality Management System (QMS) that complements the current IT Management Systems Governance framework, full data lifecycle

[1] AI Transparency Institute's website: https://aitransparencyinstitute.com/

management, verification of internal controls for AI program, and offers specific guidelines for AI procurements. Training and education strategy are also important part of the conformity assessment methodology. careAI is also aligned with IEEE P7000 series which addresses specific issues at the intersection of technological and ethical considerations. In terms of safety, it requires organisations to document what policies and processes are in place to ensure that the behaviour implemented by the AI component is safe under all conditions. It will be checked that the organization has a clear understanding of what constitutes a safe behavior of an AI function, including under which conditions the component will provide which service. Unfortunately, the description and justification of the tool can't be further developed due to intellectual property restrictions (trade secrets). The set of holistic and detailed questionnaires we developed are based on laws (GDPR, Charter of fundamental rights), ethics (OECD AI principles[2] [17]) and technical standards (cybersecurity, QMS). They build upon the aforementioned indicators and generate different indices (or scores) along these dimensions and also a cumulative overall score. Each question within the questionnaire is assigned a weight. Moreover, the scores along each dimension and the cumulative score are computed based on a weighted mean. Higher scores signify a higher conformity, whereas a lower score corresponds to a lower conformity. They provide a pathway for the certification of AI systems. They can also be used as internal control mechanisms by companies or marketing tool to communicate on the responsible use of AI systems as part of a good corporate governance. These questionnaires also provide an extensive overview of organisation's (either the one using AI applications, or the one contributing to its development) performance across the different dimensions, in the form of a spider chart, i.e., a holistic overview highlighting the strengths as well as areas of improvements. Finally, our conformity assessment tools provide a cumulative score, in the form of a grade (highly conforming organisations are awarded with an 'A' and low conforming organisations are awarded a lower grade). These questionnaires, although provide self-assessment tools that can facilitate conformity of diverse actors involved in the development and deployment of AI systems, but simultaneously, can also empower regulatory and auditing bodies to scrutinise organisations'conformity with the EU AI Act. These conformity assessment tools are developed as webapps, which are hosted on the website of the AI Transparency Institute. This set of tools enable companies to put in place an appropriate governance of AI applications during the full value chain. It also ensures that the design and development and post-market monitoring of an AI system are trustworthy from a legal, ethical, and technical perspective—and thus compliant with the EU AI Act.

[2] The OECD AI Principles identified some key principles for the evaluation of AI trustworthiness in 2019. The main principles are fairness, transparency, contestability, and accountability.

7 Conclusion

The EU AI Act will be a legally binding instrument that aims to create ex-ante safeguards for high-risk AI systems (HRS) to foster the development and uptake of safe and lawful AI that respects fundamental rights. Following the Trilogue agreement, the European Parliament and the European Council would then formally approve the AI Act early 2024. The law will take effect two years later. It will be mandatory for all EU member states to comply with the EU AI Act. Due to the fast-paced nature of AI technology, it is likely that additional changes will be made to the EU AI Act after it takes effect, through implementing or delegated acts [EU Commission, n.d.]. The AI Act may become a gold standard due to its extraterritorial effect and the export of EU values in non-EU member states. The AI Act presents specific challenges for organisations, requiring AI design and development to be part of a specific industrial process and conformity assessment to be part of a corporate governance and risk management process. The AI Act aims to mitigate risks by evaluating the quality of HRS prior to its entry into the market and holds providers of AI systems liable for damages caused by defective products. Innovative tools have been developed to support organisations to comply with this EU Regulation. Our conformity assessment process assimilates the principles of corporate governance and digital ethics to provide detailed scores and guidelines for improving conformity to the EU AI Act. It ensures that the design and development of an AI system are trustworthy from a legal, ethical, and technical perspective.

References

1. Demetzou Katerina, Introduction to the conformity assessment under the draft of the EU AI Act, and how it compares to DPIA, Future of Privacy Forum, 12 Aug 2022. https://fpf.org/blog/introduction-to-the-conformity-assessment-under-the-draft-eu-ai-act-and-how-it-compares-to-dpias/
2. European Union, Charter of fundamental rights, 2000/C, 364/01, Official Journal of the European Communities, 18 Dec 2020. https://www.europarl.europa.eu/charter/pdf/text_en.pdf
3. EU Commission, Impact assessment of the EU AI Act, 21 Apr 2021. https://digital-strategy.ec.europa.eu/en/library/impact-assessment-regulation-artificial-intelligence
4. EU Commission, Implementing and delegating acts. https://commission.europa.eu/law/law-making-process/adopting-eu-law/implementing-and-delegated-acts_en. Accessed 28 Dec 2023
5. EU Parliament, Legislative Train Schedule, Artificial intelligence act, 20 Oct 2023. https://www.europarl.europa.eu/legislative-train/

theme-a-europe-fit-for-the-digital-age/file-regulation-on-artificial-intelligence
6. EURACTIV, Commission leaves European standardisation body out of AI standard-setting, Luca Bertuzzi, 7 Dec 2022. https://www.euractiv.com/section/artificial-intelligence/news/commission-leaves-european-standardisation-body-out-of-ai-standard-setting/
7. Bradford, A.: The Brussels effect: how the European Union rules the world. Oxford University Press, USA (2020)
8. Bouquet, C., Barsoux, J.L., Wade, M.: ALIEN Thinking: the unconventional path to breakthrough ideas. Hachette, UK (2021)
9. Floridi, L., Holweg, M., Taddeo, M., Amaya Silva, J., Mökander, J. and Wen, Y.: capAI-A procedure for conducting conformity assessment of ai systems in line with the EU artificial intelligence act. Available at SSRN 4064091 (2022)
10. Mökander, J., Floridi, L.: Ethics-based auditing to develop trustworthy AI. Mind. Mach. **31**(2), 323–327 (2021)
11. Pouget, A. https://artificialintelligenceact.eu/context/. Accessed 28 Dec 2023
12. Dandl, S., Molnar, C., Binder, M., Bischl, B.: Multi-objective counterfactual explanations. In: International conference on parallel problem solving from nature, pp. 448–469. Springer, Cham (2020)
13. Molnar, C., Casalicchio, G., Bischl, B.: Interpretable machine learning–a brief history, state-of-the-art and challenges. In: ECML PKDD 2020 Workshops: workshops of the European conference on machine learning and knowledge discovery in databases (ECML PKDD 2020), pp. 417–431. Springer International Publishing, Cham (2021)
14. NIST, EU-US AI Roadmap, 4 Dec 2022. https://www.nist.gov/system/files/documents/2022/12/04/Joint_TTC_Roadmap_Dec2022_Final.pdf
15. Veale, M., Borgesius, F.Z.: Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. Comp Law Rev Int **22**(4), 97–112 (2021)
16. Stiftung Bertelsmann (Eds.), Hallensleben, S., Hustedt, C., et al.: From Principles to Practice: An Interdisciplinary Framework to Operationalise AI Ethics. 2020. https://www.bertelsmannstiftung.de/en/publications/publication/did/from-principles-to-practice-wie-wir-ki-ethik-messbar-machenkoennen. Accessed 29 Dec 2023
17. Yeung, K.: Recommendation of the council on artificial intelligence (OECD). Int. Leg. Mater. **59**(1), 27–34 (2020)
18. Zicari, R.V., Amann, J., Bruneault, F., Coffee, M., Düdder, B., Hickman, E., Gallucci, A., Gilbert, T.K., Hagendorff, T., van Halem, I. and Hildt, E., 2022. How to assess trustworthy AI in practice. arXiv preprint arXiv:2206.09887