Legal Protection by Design in the AI Value Chain

What role for AI Metrics?

Gianmarco Gori

Postdoctoral Researcher at the Law, Science, Technology and Society Research Group (LSTS),

Vrije Universiteit Brussel (VUB)

HumanE AI Net

European Union's Horizon 2020 research and innovation programme

Grant Agreement No 952026

WP 5

Page intentionally left blank

Executive summary

T5.1: ‘Legal Protection by Design’ (LPbD) (VUB)

The goal of T.5.1. is to address the question of the incorporation of fundamental rights protection into the architecture of AI systems. The report “Legal Protection by Design in the AI Value Chain: What role for ELS Metrics?”, authored by LSTS-VUB, addresses this challenge by integrating the results of legal research with the findings achieved through the participation in the macro-project “Metrics for Ethics” carried out in the context of WP5. Participation in this macro-project has offered a very good vantage point from which to investigate the role of By Design practices in strengthening Legal Protection throughout the lifecycle of AI systems and models. The macro-project has i) explored, through the analysis of case studies, approaches that measure in an integrated way Ethical, Legal, Social aspects (ELS) such as fairness, bias, privacy, robustness, transparency; and ii) sought to implement an integrated prototype dashboard aimed to provide an interface with real-time metrics, visualizations, and contextual information. The disciplinary diversity within and reiterative dialogue with the macro-project team has also provided a hands-on perspective to explore the question of *whether* and, if so, *how* and *to what extent*, 'metrics' could contribute to ‘Legal Protection by Design’. The topic of AI Metrics has proven particularly apt to investigate the connection between the practices of designing, testing and documenting AI systems and models and the roles and requirements established by positive law. The report therefore takes the role of AI metrics in the AI pipeline as an emblematic case study to shed light on the delicate interplay between legal and technical requirements that underpins Legal Protection By Design. Based on the analysis of the relevant positive law, whilst treasuring the practical lessons learnt from the participation in the macro-project, this report analyses the issues emerging in the context of technical operationalisation of a set of core ethical and legal principles. The ethical and legal frameworks are analysed while having specific regard to their similarities and dissimilarities, focusing in particular on the distinctive effects that are characteristic of metrics that assume relevance under *legal* norms.

The report situates the relationship between Legal Protection By Design and AI Metrics in the context of the evolving EU positive law. During the course of the HumanE AI Net project, a set of new legal instruments have substantively redesigned the framework governing the design, development and deployment of AI systems and AI models. The report takes into account how the updated legal framework - in particular, the AI Act - (re)configures the AI value chain by introducing a new vocabulary and new legally relevant roles. The report pays particular attention to the relationship between AI metrics and the obligations of the relevant operators in the AI value chain, in particular under the AI Act.

The report examines the obligations of providers of General Purpose AI Models and providers of High-risk AI systems, focusing on the interplay between metrics and legal requirements in the context of providers’ practices of risk management and testing as well as with respect to the requirements of accuracy, robustness, cybersecurity, transparency, interpretability, human oversight.

Adopting the lenses offered by the concept of Agonistic Machine Learning, the report leverages the results of the legal analysis and the lessons learned in the microproject to illustrate how AI metrics can contribute to Legal Protection by Design.

Page intentionally left blank

Table of Contents

1. Introduction 1

2. Legal Protection by Design and AI Metrics 2

2.1. AI metrics at the intersection of AI practice, Ethics and Law 2

2.2. What Metrics measure and how they "count" in different normative frameworks: By Design approaches and Legal Protection by Design. 3

3. The relevance of AI metrics in EU Digital Law 7

3.1. AI Metrics, legal obligations and technical specifications 7

3.2. AI Metrics and the AI Act 9

3.2.1. The goal and architecture of the AI Act 10

3.2.2. Metrics and High-risk AI systems 17

3.2.3. Towards operationalisation 24

4. The macro-project "Metrics for Ethics": Lessons learned on how operationalise AI Metrics for Legal Protection by Design 27

4.1. From the case study to the "Ethics Dashboard" 27

4.2. Seeing AI Metrics through the lenses of an Agonistic approach to Machine Learning 29

5. Conclusions 31

References 32

Page intentionally left blank

List of abbreviations

AI Artificial Intelligence

AIA AI Act, Regulation

AIS AI system

DSA Digital Services Act, Regulation

GDPR General Data Protection Regulation

GPAIM General Purpose AI Model

HRAIS High-risk AI system

LbD Legal by Design

LPbD Legal Protection by Design

ML Machine Learning

Page intentionally left blank

1. Introduction

T.5.1. of the HumanE AI Net project asks the question of how to incorporate fundamental rights protection into the architecture of AI systems. The present report tackles this challenge by merging the results of legal research with the findings achieved through the participation in the macro-project "Metrics for Ethics" carried out in the context of WP5 of the HumanE AI Net project. The participation in this Macro-project has provided a unique standpoint to investigate the role of By Design practices in strengthening Legal Protection throughout the lifecycle of AI systems and models.

The macro-project "Metrics for Ethics" aims to contribute to the research on the technical approaches to monitor Ethical, Legal, and Social requirements in AI design, development and deployment. To this end, the participants in the macro-project have conducted a case study that, by addressing a use case classified as High-risk under the AI Act, i.e., creditworthiness evaluation, sought to develop a dashboard that allows its users to explore different aspects of the use of AI metrics to measure requirements such as bias, fairness, robustness, explainability and transparency.

The cross-disciplinary dialogues with the team members of the macro-project has offered a hands-on perspective on the question of *whether* and, if so, *how* and *to what extent*, AI metrics can contribute to Legal Protection by Design, a concept coined by Hildebrandt to characterise an approach aimed to articulate fundamental rights protection and the checks and balances of the Rule of Law in the design of digital technologies. In this perspective, the topic of AI Metrics has proved particularly apt to investigate the delicate interplay between legal and technical requirements in the practices of designing, testing and documenting AI systems and models.

Section 2 of this report introduces the role that AI metrics play at the intersection of AI practice, ethics and law. The section illustrates how the concept of Legal Protection by Design can help understand the normative relevance that metrics can assume within multiple normative frameworks. By showing the differences between Legal Protection by Design and other "by design" approaches, the report examines the role of AI metrics in the context of the technical operationalisation of legal requirements.

Section 3 aims to answer the question of what it means for AI metrics to be relevant under EU law. A set of preliminary considerations is offered that clarify *how* metrics can take on significance for the purposes of an assessment of compliance with the law. Section 3.2. is dedicated to the examination of the role of AI metrics in the context of the provisions of the AI Act on General-purpose AI models and High-risk AI systems.

Section 4 illustrates the findings of the research conducted in the context of the macro-project "Metrics for Ethics". Adopting the lenses offered by the concept of Agonistic Machine Learning, the report discusses how the integration of AI metrics in the design, development and deployment of AI systems can contribute to Legal Protection by Design.

2. Legal Protection by Design and AI Metrics

This section provides an overview of the debate in computer science, ethics and law around the topic of AI metrics and introduces the fundamental legal-theoretical concepts that inform the present report.

2.1. AI metrics at the intersection of AI practice, Ethics and Law

Metrics play a **constitutive role in AI**. As Mitchell points out in his seminal book on Machine Learning: "A well-defined learning problem requires a well-specified task, *performance metric*, and source of training experience"[1].

More broadly speaking, AI metrics measure the extent to which a certain technical specification is satisfied. As such, metrics have inherent normative relevance. By contributing to the determination of *what counts as success* within AI community, they inform the very goals of AI research.

Simultaneously, the topic of AI metrics has attracted growing attention within the scientific and institutional debate on AI Ethics[2]. Over the last few years, an increasing number of initiatives have been undertaken by academics, institutional actors and tech companies alike to develop and make metrics and methodologies available that are aimed at measuring the extent to which AI solutions respect ethical requirements[3].

In the context of the European Union, the "Ethics Guidelines for Trustworthy AI" issued in 2019 by the High-Level Expert Group on Artificial Intelligence appointed by the European Commission, identify metrics as an important technical tool that can contribute to translating ethical principles into the design and use of AI, in particular, the seven principles of human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being and accountability[4]. The Guidelines emphasise that AI metrics are key in the context of testing and validating "all components of an AI system, including data, pre-trained models, environments and the behaviour of the system as a whole throughout its entire life cycle"[5]. In this respect, the Guidelines exhort the AI community to develop multiple metrics in order to "cover the categories that are being tested for different perspectives"[6].

As will be discussed in more detail in Section 3, the importance of AI metrics has been further consolidated by recently adopted law, and especially the AI Act[7].

---

[1] Tom M Mitchell, *Machine Learning* (McGraw-Hill 1997), p. 17 (emphasis added). See also, at p. 2, "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E".

[2] For a systematic literature review: Guilherme Palumbo, Davide Carneiro and Victor Alves, 'Objective Metrics for Ethical AI: A Systematic Literature Review' [2024] International Journal of Data Science and Analytics <https://doi.org/10.1007/s41060-024-00541-w>.

[3] For instance, see OECD, Catalogue of Tools & Metrics for Trustworthy AI https://oecd.ai/en/catalogue/metrics; AI Fairness 360 (AIF360), https://github.com/Trusted-AI/AIF360; IBM, Everyday ethics for AI https://www.ibm.com/design/ai/ethics/everyday-ethics/

[4] High Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI', 2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, p. 21

[5] Ivi

[6] Ivi

[7] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), http://data.europa.eu/eli/reg/2024/1689/oj

A problem of coordination can arise from the fact that AI metrics can assume normative relevance within different normative frameworks, i.e., AI practice, AI Ethics and law. Even though the requirements that metrics are meant to measure can be simultaneously relevant for AI practice, ethics and law, each of these domains is **autonomous** and distinguished by different forms of normativity. For instance, the results of the use of the very same metric can be recognised as a great success within the AI community and, at once, lead to the conclusion that a certain AI system is not in compliance with the law. As this example shows, autonomy does not mean that practitioners in one domain can ignore the other domains.

This is especially relevant in light of the fact that the prominent role that metrics play in AI practice has recently been faced with growing discontent about the "unreasonable effectiveness of metric optimization"[8]. Thomas and Uminsky have illustrated how the use of metrics in AI practice is subject to the infamous Goodhart's law, i.e., "When a measure becomes a target, it ceases to be a good measure"[9]. As the Authors highlight, metric optimization can lead to "manipulation, gaming, a focus on short-term quantities (at the expense of longer-term concerns), and other undesirable consequences …"[10].

In order to discern between the risks posed and the opportunities offered by AI metrics, it is crucially important to correctly understand what the **effects** are of the use of metrics. Importantly, such effects will depend on the domain in which they operate. By addressing the questions of "*what metrics measure*" and "*how metrics count in different normative frameworks*", the next section further explores the relationship between legal, ethical and technical requirements.

2.2. What Metrics measure and how they "count" in different normative frameworks: By Design approaches and Legal Protection by Design.

This section illustrates how the conceptual assumptions informing this report help trace and clarify the different values that metrics can assume – and the effects that they can produce - within different normative frameworks.

This report builds on the perspective of **Legal Protection by Design (LPbD)**, a concept coined by Hildebrandt to characterise an approach aimed to articulate the legal protection provided by fundamental rights and the checks and balances of the Rule of Law in the design of digital technologies[11]. Legal Protection by Design **is different from other "by design approaches" in several important ways**.

The first distinctive element of LPbD concerns the kind of **requirements** that this approach aims to articulate for the design of technologies. In this respect, LPbD is distinguished from "**ethics by design"** approaches that aim to align the design of technology to ethical values. It is possible to acknowledge that, at a high level, many ethical requirements are characterised by an apparent overlap with legal requirements. For instance, the 7 principles established in the Ethics Guidelines for Trustworthy AI also cover values that are legally protected. However, the overlap between legal and ethical requirements is only partial and the

[8] Rachel L Thomas and David Uminsky, 'Reliance on Metrics Is a Fundamental Challenge for AI' (2022) 3 Patterns 100476.

[9] Charles Goodhart (2015). Goodhart's law. In The Encyclopedia of Central Banking, L. Rochon and S. Rossi, eds. (Edward Elgar Publishing), pp. 227–228

[10] Ivi

[11] M. Hildebrandt, 'Boundary Work between Computational 'Law' and 'Law-as-We-Know-it'', in Deirdre Curtin, and Mariavittoria Catanzariti (eds), *Data at the Boundaries of European Law*, Collected Courses of the Academy of European Law (Oxford, 2023; online edn, Oxford Academic, 23 Mar. 2023), https://doi.org/10.1093/oso/9780198874195.003.0002, accessed 24 July 2024; Mireille Hildebrandt, 'Three Framing Concepts', in Laurence Diver, Tatiana Duarte, Gianmarco Gori, Emilie van den Hoven and Mireille Hildebrandt, COHUBICOL Research Study on Text-Driven Law, https://publications.cohubicol.com/assets/uploads/cohubicol-research-study-on-text-driven-law-final.pdf, p. 21; Mireille Hildebrandt, *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology* (Edward Elgar Publishing 2015).Mireille Hildebrandt, 'A Vision of Ambient Law' in Roger Brownsword and Karen Yeung (eds), *Regulating Technologies* (Hart 2008).

two should not be conflated. As Hildebrandt put it "Ethics is both more and less than law"[12]. The law might not address concerns that are of ethical relevance, and vice versa, e.g. something that is morally neutral might be required under the law. The fact that a requirement has both ethical and legal relevance, e.g., human autonomy, does not imply that such requirement assume the same meaning respectively in law and ethics. Stronger still, the legal or ethical nature of a requirement has crucial differences. Legal requirements have "legal teeth", i.e., they can be backed up with the force of the law. The law ensures that an enforcement system is in place to guarantee that independent and impartial courts of law can have the final - and binding - word as to the correct interpretation to be given to a legal requirement. The differences between the ethical and legal framework implies that the metrics used to measure whether AI technologies comply with ethical requirements do not at the same time measure whether those technologies also comply with legal requirements.

Secondly, Legal Protection by Design is also different from "**Legal by Design**" (LbD) approaches. An apparent similarity exists between LPbD and LbD, as both focus on the implementation of **legal requirements** into the design of computational technologies. Also in this case, however, the similarity can be misleading, as LPbD and LbD are distinguished by radically different assumptions as to the nature of law. This, in turn, affects the goals pursued by these two by-design approaches. Legal by Design is a form of *techno-regulation* aimed to design technology with regulatory effects. This entails that LbD aims to **translate legal requirements directly into technical requirements** and design technologies that inhibit, preclude or enforce a certain behaviour or state of affairs in order to achieve automatic compliance with the law. Under this perspective, it is conceivable that AI metrics could be developed to automatically determine whether a certain AI system complies with the law and, potentially, AI systems could be designed to automatically adjust their parameters to ensure compliance with the law. The Legal Protection by Design paradigm highlights that the translation of legal requirements into technical requirements is as necessary as it is complex. It is *necessary* because failing to do so would imply designing technologies that do not incorporate the legal safeguards needed to avoid jeopardising fundamental rights. It is *complex* because the translation between legal and technical requirements implies interfacing "different animals": the law and digital technologies each have different affordances and "modes of existence"[13].

Any attempt to answer the question of which technical requirements can ensure that the relevant legal requirements are satisfied requires facing an "**operationalisation gap**". Legal by design approaches see this gap as an obstacle to be overcome by deploying more technical resources – more automation – and by normalising legal requirements, e.g., expressing them in a form that lends itself better to technical operationalisation, including by expressing legal norms in the form of code. The LPbD approach rejects the idea that this operationalisation gap between a legal requirement and a technical requirement can be filled by equating the former and the latter. Crucially, LPbD emphasises that any act of technical operationalisation is an *interpretation* of a legal requirement, not the *legal requirement itself* and that maintaining that there is equivalence between the two is fundamentally incompatible with the essential features of modern positive law.

Legal requirements are expressed in written text in natural language and, as such, they are characterised by open texture[14]. This open texture results in legal requirements being characterised by a form of constrained multi-interpretability. Legal texts can accommodate multiple interpretations, but this does not mean that anything goes. The requirements formulated into legal texts are part of a broader legal framework that must be interpreted in light of the *principle of integrity*. This implies that interpreters are always required

[12] Mireille Hildebrandt, *Law for Computer Scientists and Other Folk* (Oxford University Press 2020)

[13] Laurence Diver, Tatiana Duarte, Gianmarco Gori, Emilie van den Hoven and Mireille Hildebrandt, COHUBICOL Research Study on Text-Driven Law, https://publications.cohubicol.com/assets/uploads/cohubicol-research-study-on-text-driven-law-final.pdf,

[14] Ivi; Herbert Hart, *The Concept of Law* (Oxford University Press 1992)

to test whether their interpretation could "form part of a coherent theory justifying the network as a whole"[15]. In constitutional democracies, this "network" includes **fundamental rights** and the **Rule of law**. The law is constrained by the need to ensure the protection of fundamental rights, meaning that legal requirements must be given the interpretation that mostly promotes the enjoyment of fundamental rights and provides effective remedies against their violation. This also means that a legal requirement that negatively affects fundamental rights is to be considered invalid and, therefore, devoid of legal effects[16]. Any act of interpretation-operationalisation must take into account the fact that the meaning that a legal requirement assumes in a certain setting is affected by the whole legal system. The concrete circumstances in which a legal requirement is applied-operationalised might trigger something akin to a backpropagation and readjusting of the entire network on legal norms in such a way as to ensure an interpretation that is coherent with the protection of fundamental rights.

Moreover, the fact that modern positive law is grounded on the principle of the Rule of law means that no one - and *no thing*, including technical operationalisation of a legal requirement - is above the law. This affects the relation between legal requirements and their translation into technical requirements in the sense that it is always possible that an interpretation-operationalisation is invalidated by making reference to the legal requirement that the former aims to operationalise.

The same applies to the metrics used to measure and evaluate whether or not a certain behaviour or state of affairs is compliant with a certain interpretation-operationalisation of a legal requirement. This aspect can be expressed in a phrase that, admittedly, might even sound paradoxical in the context of a discussion of AI metrics, i.e., "ultimately, legal requirements are what measure, and not what is measured" [17]. For the result of a measurement-evaluation performed with AI metrics to produce the legal consequence, metrics have to be, as it were, assessed or "*measured*" in light of the legal criteria that establish what counts as a legally valid measuring-evaluation. In the end, only a *legal* interpretation can determine whether, both *in abstracto* and *in concreto*:

- using the metric Z is an adequate way of measuring whether a behaviour or state of affairs complies with Y, i.e., a certain operationalisation of the legal requirements X,

- Y is an adequate operationalisation of the legal requirement X.

The idea of Legal Protection by Design entails that adequate legal protection requires resisting the temptation of equating legal and technical requirements, because such an equation would oversimplify the complexity inherent in the operationalisation of legal requirements . LPbD takes the challenge posed by the need to interface legal and technical requirements seriously, noting that the relationship between law and technology can be characterised by certain tensions. These tensions cannot be solved, but must be sustained through attentive design strategies that valorise the potential of technology to contribute to ensuring compliance with the law, without degrading into technological solutionism. The perspective of Legal Protection by Design helps to understand metrics as an instruments that, although *per se* not resolutive, can provide essential information that contributes to the *legal* assessment of whether the design, development and deployment of AI is carried out in compliance with the law.

---

[15] Ronald Dworkin, *Law's Empire* (Belknap Press 1986), p. 225

[16] Unless the interference with a fundamental right can, in turn, be justified by another norm of constitutional rank, e.g., when necessary to protect another fundamental right.

[17] Gianmarco Gori, 'Legal and Computer Rules: An Overview Inspired by Wittgenstein's Remarks' in Alice Helliwell, Alessandro Rossi and Brian Ball (eds), *Wittgenstein and Artificial Intelligence*, vol II (Anthem Press 2024).

The next section will show how the approach of LPbD finds recognition in the legal provisions of EU law under which AI metrics assume relevance.

3. The relevance of AI metrics in EU Digital Law

In the last decade, the regulation of the Digital Market has become a priority of the EU Legislator and the scope of EU law on digital technologies has widened. This has made AI metrics a topic of growing legal relevance. This section examines **the relevance of AI metrics in the context of EU law** and, in particular, the **AI Act**.

Before focusing on the provisions of the AI Act, it is worth making some preliminary observations that will help clarify *how* metrics can take on significance for the purposes of an assessment of compliance with law. To this end, the approach taken in the AI Act will be compared with that taken in two other legal acts that have particular importance in EU digital law, i.e., the General Data Protection Regulation (GDPR)[18] and the Digital Service Act (DSA)[19].

3.1. AI Metrics, legal obligations and technical specifications

It is noteworthy that neither the AI Act nor any other relevant EU legal instrument contains a **legal *definition*** of **"AI metrics"**. Equally, no legal instrument establishes the use of a *specific* metric as legally mandatory.

In fact, the GDPR and the DSA do not contain any reference to metrics.

In this respect, as we will see more in detail below, the AI Act is an exception, as it makes **direct reference to AI metrics**. For instance, it requires providers to test their AI systems "against prior defined metrics and probabilistic thresholds that are appropriate to the intended purpose of the High-risk AI system"[20]. The presence of direct references to metrics in the AI Act can be partly explained by considering that, other than the GDPR and the DSA, the AI Act directly regulates **AI systems and models *as products***. Accordingly, in the architecture of the AI Act, technical aspects of AI technologies have a more prominent role.

The GDPR and the DSA, on the contrary, are examples of legal acts informed by the principle of technological neutrality, i.e., they do not directly regulate AI systems and models, nor any other specific technology. Yet, the GDPR and the DSA regulate *activities* that are likely to involve the deployment of AI models and AI systems, i.e., respectively, the processing of personal data and the providing of mere conduit, caching and hosting services, including Very Large Online Platform and Search Engines. AI systems and models are therefore indirectly regulated, that is, to the extent that it is relevant for the purposes of assessing compliance with the legal obligations established in those acts.

However, the fact that the GDPR and the DSA lack direct reference to metrics does not mean that metrics are not legally relevant. AI metrics can assume relevance *as a tool* to ensure and demonstrate compliance with the requirements established by the GDPR and the DSA. For what concerns the GDPR, AI Metrics can be used, for instance, to measure and document compliance with the (cyber)security obligations established under art. 32 GDPR. More broadly, AI metrics can be used as part of the "technical and organisational measures" that data controllers are required to adopt to ensure compliance with the GDPR[21]. For what concerns the DSA, AI metrics can play a role, for instance, in the identification and mitigation of

[18] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), http://data.europa.eu/eli/reg/2016/679/oj

[19] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), http://data.europa.eu/eli/reg/2022/2065/oj

[20] Article 9, § 8, AI Act

[21] Article 24 GDPR

the systemic risks that Very Large Online Platforms and Search Engines might pose to fundamental rights[22], e.g., in the testing and adapting of algorithmic systems.[23] such as those used for content moderation as well as recommender systems and advertising systems.[24]

This should clarify that metrics can be relevant whether or not a legal Act includes direct references to them. It is up to the involved actors to argue and concretely demonstrate that the use of certain metrics is relevant for the purpose of complying with their obligations.

Although the same logic applies in the case of the AI Act, it presents some differences that are worth discussing. The AI Act follows the so-called "New Approach" that has informed EU Product Legislation since 1985. [25] Under the New Approach, EU product legislation is defined by a two-layer structure.

1. EU legal acts (regulations, directives) establish the so-called **essential requirements** that the regulated products – in the case of the AI Act, AI systems and AI models - must meet for them to be lawfully placed on or put into service in the internal EU market. As we will see below, these essential requirements are generally high-level requirements. The high-level character of the essential requirements gives rise to the "operationalisation gap" discussed in the previous section.

2. Following the New Approach, the AI Act gives the relevant operators a choice as to *how* to close this operationalisation gap.

2.1. Firstly, providers can decide to autonomously operationalise the legal requirements and prove that their products are compliant with the essential requirements. Compliance with EU product legislation requires the identification and implementation of the technical specifications that correctly operationalise the essential legal requirements. This also implies selecting appropriate methodologies to effectively monitor and measure the extent to which products are – and keep being – in compliance with the law.

2.2. Alternatively, providers can adhere to the **technical specifications** provided in **European Harmonised Standards, i.e.,** standards that are developed by European Standardisation Organisations on request of the EU Commission[26]. Although it is not mandatory for providers to adhere to the technical specifications contained in EU harmonised standards, doing so triggers a **presumption of conformity** with the essential legal requirement that the standard is meant to operationalise. It is clear that, this presumption imbues the technical specifications provided in EU harmonised standards with a certain *gravitas*: adherence to harmonised standards allows the relevant operators to reduce part of their compliance burden and to carry out their design and development activities with a higher level of legal certainty. With respect to the AI Act, the Commission has requested the European Committee for

[22] Articles 34 and 35 DSA

[23] Article 35, § 1, d, DSA

[24] Article 35, § 1, c-d-e, DSA

[25] Council Resolution of 7 May 1985 on a **new approach to technical harmonization and standards**, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31985Y0604(01).

[26] At now, European Standardization is regulated by Regulation (EU) No 1025/2012 of the European Parliament and of the Council of 25 October 2012 on **European standardisation**, amending Council Directives 89/686/EEC and 93/15/EEC and Directives 94/9/EC, 94/25/EC, 95/16/EC, 97/23/EC, 98/34/EC, 2004/22/EC, 2007/23/EC, 2009/23/EC and 2009/105/EC of the European Parliament and of the Council and repealing Council Decision 87/95/EEC and Decision No 1673/2006/EC of the European Parliament and of the Council Text with EEA relevance, http://data.europa.eu/eli/reg/2012/1025/oj

Standardisation (CEN) and European Committee for Electrotechnical Standardisation (CENELEC) to adopt standards by April 2025[27].

These harmonised standards will represent a key normative instrument that will inform the operationalisation choices of providers, including choices as to which metrics to use to measure legal compliance. However, it is important to keep the legal requirements established in the AI Act and the specifications and metrics that will be provided in the harmonised standards analytically distinct. The importance of maintaining this distinction can be illustrated by clarifying the **scope of the effects of the presumption of conformity** triggered by adherence to harmonised standards.

Firstly, the presumption only covers the conformity with the essential requirements that are covered in the harmonised standards. Secondly, this presumption does not shield providers from civil liability. Providers can still be found liable for the damage and/or violations of fundamental rights that can be traced back to their actions or omissions. Thirdly, the presumption of conformity, as the name implies, is a *presumption*. As such, it can be rebutted, meaning that it can be demonstrated that, despite the adherence to harmonised standards, an AI product actually does not comply with the law. In essence, what providers must comply with, ultimately, are the legal requirements, not technical specifications.

These preliminary considerations should clarify that the role that EU law assigns to AI metrics – and, more broadly, to technical specifications - does not correspond to the "**Legal by design" paradigm**. The relationship between legal and technical requirements is not one of identity: technical specifications are an operationalisation-interpretation of the law. AI metrics ***inform, but do not settle once and for all,*** the legal assessment as to whether a legal requirement is complied with. The use of a certain metric does not *per se* entail compliance with the legal requirement for which such metrics assume relevance.

Simultaneously, these preliminary considerations should also make clear that, *as a tool that contributes to demonstrating compliance* with legal requirements (directly or indirectly related to AI technologies), AI metrics can always play a relevant role, also when they are not directly referred to by the letter of the law.

The place that EU law assigns to AI metrics is actually a particularly important one. AI metrics lie at the intersection of two crucial needs, the balancing of which is as challenging as necessary. On the one hand, a certain level of standardisation of technical specifications is crucial to facilitate compliance with legal requirements. In this respect, AI metrics are key in that they make possible a standardised measurability and monitoring of compliance with technical specifications. On the other hand, standardised tools to quantify and measure compliance with technical specifications have a complementary function and do not supplant the need to perform a case-by-case assessment of whether, in concrete circumstances, compliance with technical specifications ensures compliance with legal requirements.

The following subsection will examine how this difficult balance plays out in the provisions of the AI Act.

3.2. AI Metrics and the AI Act

The following subsection analyses the role of AI metrics in the context of the legal requirements established by the AI Act. Subsection 3.2.1. will provide a brief outline of the goals and architecture of the AI Act. This

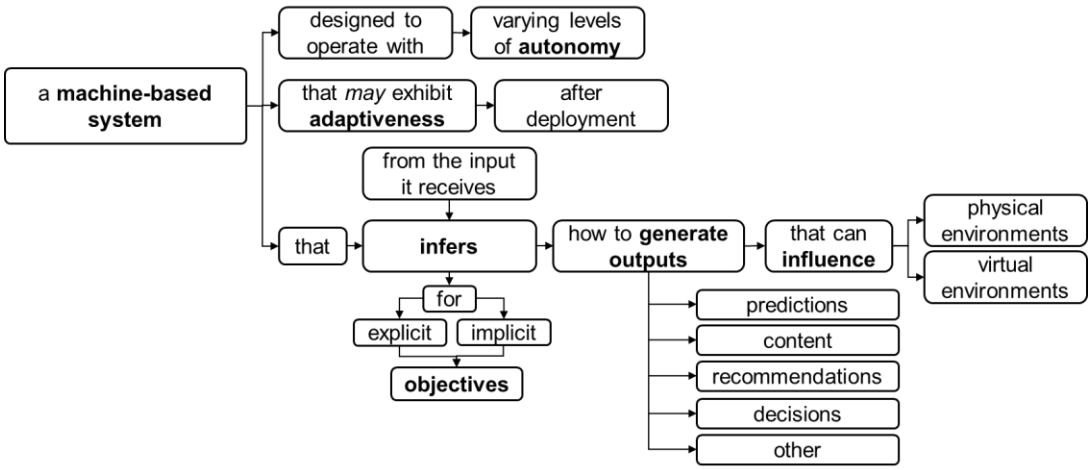
[27] Commission Implementing Decision on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on artificial intelligence Brussels, 22.5.2023 C(2023) 3215 final

will serve the purpose of better situating the relationship between AI metrics and the provisions of the AI Act that will be addressed in subsection 3.2.2.[28].

3.2.1. The goal and architecture of the AI Act

The AI Act is a complex legislative text. Its thirteen Chapters and thirteen Annexes introduce a layered body of rules aimed to regulate the whole AI value chain, from the development to the placing on the market, putting into service and deployment of AI systems (AISs) and, in the final version of the AI Act, also AI models that are classified as general-purpose (GPAIM).

Figure 1. The definition of AI system



The distinguishing feature of the products "AI systems" is their capacity to generate outputs that influence the environment. This differentiates AI systems from AI models. The AI Act considers AI models as components of AI systems, that despite being generally considered as essential lack the capacity to directly influence the environment. To have an impact "in the world", AI models require the addition of further components, such as for example a user interface[29]. Because of this, the initial version of the AI Act regulated AI models only indirectly, or better, only to the extent that they were integrated as components into an AI system.

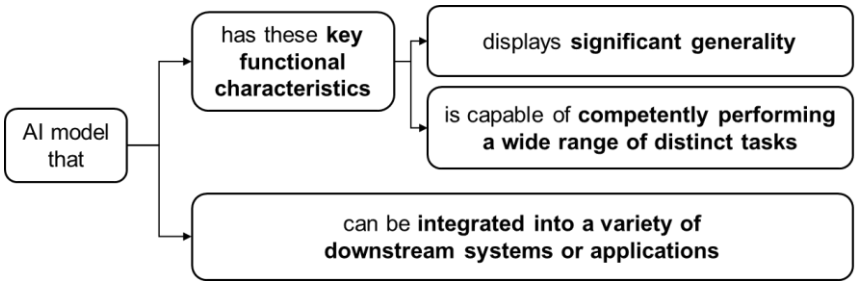
The final version of the AI act has introduced further rules that apply specifically to AI models that are classified as General purpose, irrespective of whether such models are integrated into an AI system or not. The legislative choice to include rules on General Purpose AI Models in the AI Act is motivated by the

[28] It is out of the scope of the present report to provide a thorough analysis of the provisions of the AI Act. The reader can find a more systematic discussion of the AI Act in the two tutorials given in the context of the HumanE AI Project: Mireille Hildebrandt, Tutorial on the proposal for an AI Act, HumanE AI Net, https://www.humane-ai.eu/event/tutorial-on-the-proposal-for-an-ai-act/; Mireille Hildebrandt and Gianmarco Gori, Tutorial on the final text of the AI Act, HumanE AI Net, https://www.humane-ai.eu/event/second-tutorial-on-the-ai-act/

[29] R(97) AI Act

consideration of the particular features of such models, i.e., the fact that they i) display a *significant generality* and ii) have the capability of *competently performing a wide range of distinct tasks* regardless of the way they are placed on the market[30]. Because of these features, the AI Act emphasises that providers of GPAIMs play a particular role with respect to the downstream systems that may integrate such models[31]. Their influential role makes it necessary to subject providers of GPAIMs to an adequate form of responsibility[32].

Figure 2. Definition of General Purpose AI Model



By regulating **AI systems** and **General purpose AI Models**, the AI Act aims to achieve the following **goals**:

- improve the functioning of the internal market, promote the uptake of human-centric and trustworthy AI and support innovation; and

- ensure a high level of protection of health, safety, fundamental rights enshrined in the Charter of Fundamental Rights, including democracy, the rule of law and environmental protection, against the harmful effects of artificial intelligence systems (AI systems) in the Union[33].

In pursuing such goals, the AI Act follows a **risk-based approach**[34], meaning that different rules are established depending on the level of risk that different AI systems, models and practices pose to fundamental rights, health and safety.

Subsection 3.2.1.1 will briefly illustrate how AI metrics assume relevance with respect to the legal requirements that the AI Act establishes for providers of GPAIMs (left in the figure below). Subsection 3.2.1.2. will then address the role of AI metrics with respect to AI systems (right in the figure below).

[30] Art. 3, § 1(63) AI Act. AI models that are used for research, development or prototyping activities before they are released on the market are out of the scope of the AI Act

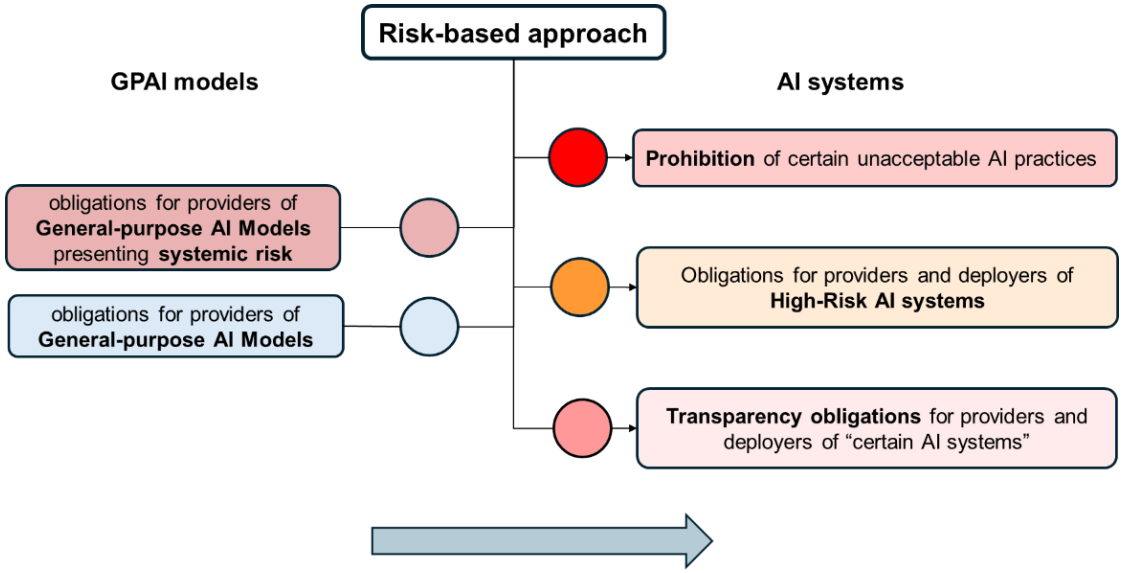
[31] R(101) AI Act

[32] Whenever a subject is both the provider of a GPAIM and provider of an AI system that integrates such model, such provider will have to comply cumulatively with the obligations concerning the model and those concerning the system. See R(97) AI Act

[33] Art. 1 AI Act

[34] R(26) AI Act

Figure 3. The risk-based approach in the AI Act



3.2.1.1. Rules on GPAI models

The rules on General Purpose AI Models introduced in Chapter V of the AI Act aim to ensure a thorough documentation of such models and an adequate management of the risks that the latter may pose. These rules are structured in two layers:

- a first set of obligations that applies to *all providers of GPAIMs*,

- an additional set of obligations that applies to providers of GPAIMs that have been designated as presenting *systemic risk*[35].

Among the obligations that apply to all providers of GPAIMs, one that is particularly relevant with respect to the topic of AI metrics – especially accuracy and fairness metrics - is the obligation to draw-up and keep up to date **technical documentation**[36]. Such technical documentation includes *inter alia* a detailed description[37] of

> the **design specifications** of the model and training process, including training methodologies and techniques, the key design choices including the rationale and assumptions made; **what the model is designed to optimise for** and the relevance of the different parameters, as applicable[38].

[35] For an illustration of the obligations of providers of GPAIMs and GPAIMs presenting systemic risk, see Mireille Hildebrandt and Gianmarco Gori, Tutorial on the final text of the AI Act, HumanE AI Net, https://www.humane-ai.eu/event/second-tutorial-on-the-ai-act/

[36] Art. 53, § 1(a); Annex XI, section 1

[37] Annex XI, § 2

[38] Annex XI**,** Section 1, n. 1(b), AI Act, my emphasis

Moreover, providers of GPAIMs must indicate in their technical documentation the measures and methods that they have adopted to "detect the unsuitability of data sources" and "identifiable biases"[39].

Providers of **GPAIMs presenting systemic risk** are subject to **additional obligations** that are functional to identifying and mitigating the systemic risk posed by their models. AI metrics assume relevance especially with respect to providers' obligations to perform and document **evaluation** and **adversarial testing** of their models[40].

Providers are required to conduct model evaluation "in accordance with standardised protocols and tools reflecting the state-of-the-art"[41]. The choices made in this respect by providers must be illustrated and justified in the technical documentation, which must contain "[a] detailed description of the evaluation strategies, including evaluation results, on the basis of available public evaluation protocols and tools or otherwise of other evaluation methodologies". Annex XI of the AI further specifies that "[e]valuation strategies shall include evaluation criteria, metrics and the methodology on the identification of limitations"[42].

The technical documentation of GPAIMs presenting systemic risk must also contain "a detailed description of the measures put in place for the purpose of conducting internal and/or external adversarial testing (e.g., red teaming), model adaptations, including alignment and fine-tuning"[43]. In this respect, providers of GPAIMs presenting systemic risk must ensure that their models have undergone testing procedures aimed at assessing their level of robustness and cybersecurity[44].

3.2.1.2. Rules on AI systems

According to the adopted risk-based approach (see figure 3), the rules on AI systems established in the AI Act are structured as follows:

- at one end of the spectrum, the AI Act identifies a set of AI practices that are **prohibited** because they are considered particularly harmful and abusive[45];

- at the other end of the spectrum, the AI Act introduces **transparency obligations** for providers and deployers of certain AI systems listed in art. 50., for instance, systems that are intended to interact directly with natural persons or generating synthetic audio, image, video or text content[46];

- the bulk of the provisions of the AI Act are dedicated to **AI systems classified as High-risk**.

As the rules on High-risk AI systems are also the most relevant for the topic of AI metrics, the rest of the analysis will be dedicated to these rules.

---

[39] Annex XI**,** Section 1, n. 1(c), AI Act

[40] Art. 55, § 1, a; Annex XI, section 2, n. 1, AI Act

[41] Ivi

[42] Annex XI**,** Section 2, n. 1, AI Act

[43] Annex XI, Section 2, n. 2, AI Act

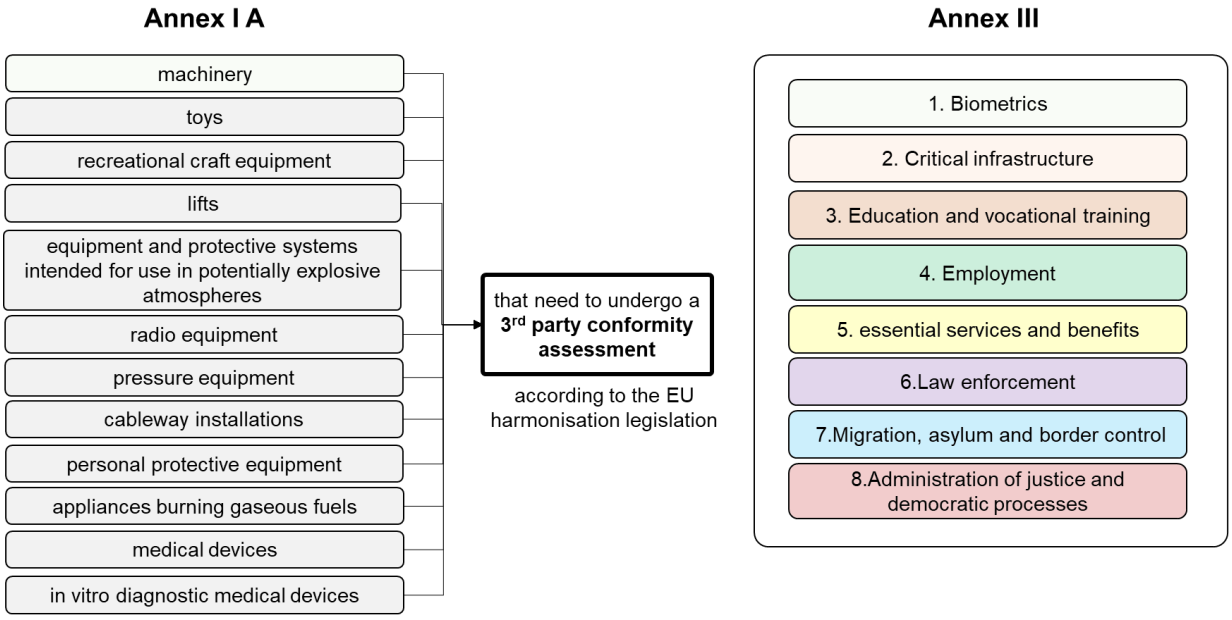
[44] *Mutatis mutandis*, similar requirements of robustness and cybersecurity apply to providers of High-risk AI systems. For this reason, we refer to the considerations that will be made in subsection 3.2.2.

[45] Art. 5, R(28) AI Act. See, see the Mireille Hildebrandt, Tutorial on the proposal for an AI Act, HumanE AI Net, https://www.humane-ai.eu/wp-content/uploads/2024/07/1.-AIA-General-defintions-and-prohibitions.pptx

[46] See, Mireille Hildebrandt, Tutorial on the proposal for an AI Act, HumanE AI Net, https://www.humane-ai.eu/wp-content/uploads/2024/07/11.-AIA-Transparency-for-medium-risk-systems.pptx

The category of High-risk AI systems is a *numerus clausus*, meaning that an AI system is considered High-risk only if it falls within the lists contained in Annex I or Annex III of the AI Act[47]. The figure below provides a synoptical visualisation of the AI systems currently classified as High-risk.

Figure 4. AI systems classified as High-risk in the AI Act



Annex I, Section A (on the left part of the figure 4 above), classifies as High-risk standalone systems and systems that are a safety component of products covered by EU product legislation and that, under such legislation, are required to undergo a third-party conformity assessment. This list includes, for instance, AI systems that are a component of certain medical devices or machines.

Annex III classifies as High-risk a set of AI systems that may pose a likely and severe harm to health, safety and fundamental rights because of their *specific intended purpose* and the *areas in which they are deployed*. This includes, for instance:

- AI systems intended to be used to make decisions affecting terms of work-related relationships, the promotion or termination of work-related contractual relationships[48];
- AI systems intended to be used by public authorities to evaluate the eligibility of natural persons for essential public assistance benefits and services[49];
- AI systems intended to be used by law enforcement authorities to evaluate the reliability of evidence in the course of the investigation or prosecution of criminal offences[50];
- AI systems to be used by competent public authorities to assess a risk of irregular migration posed by a natural person who intends to enter or who has entered into the territory of a Member State[51];

[47] As eventually amended by the Commission through the procedure set out in art. 6, § 6 of the AI Act

[48] Annex III, 4.a

[49] Annex III, 5.a.

[50] Annex III, 6.b.

[51] Annex III, 7.b.

- AI systems intended to be used by a judicial authority in researching and interpreting facts and the law and in applying the law to a concrete set of facts[52].

The AI Act regulates the whole lifecycle of AI systems that are classified as High-risk by setting out a set of mandatory requirements that these systems must meet; introducing a set of obligations for all the relevant operators in the value chain of High-risk systems, in particular providers and deployers and by establishing an institutional framework aimed to monitor and enforce the compliance with such obligations and requirements. Schematically, the lifecycle of a High-risk AI system can be subdivided into two main phases, i.e., the pre-market and post-market phases.

• **Pre-market phase and conformity assessment procedure**

This is the phase in which providers design and develop an AI system (potentially integrating a GPAIM in it). Providers who want to place on the market or put into service the HRAISs that they are developing (or have developed by others) are required to put in place technical and organisational measures to ensure that their systems comply with the requirements established in Articles 9 to 15 of the AI Act. Compliance with such requirements is necessary for a HRAIS to pass the conformity assessment procedure which, in turn, determines whether the provider can place on the market or put into service the HRAIS.

For almost all the High-risk AI systems listed in Annex III, the conformity assessment procedure is **internal**, that is, providers themselves verify that their systems comply with the mandated requirements through their quality management system. For some of the biometrics AI systems in Annex III, n. 1, and for the AI system listed in Annex I, Section A, the conformity assessment procedure involves a **third-party conformity assessment** body, the so-called notified bodies[53].

The pre-market phase ends when, having successfully passed the conformity assessment procedure, the provider draws up the declaration of conformity of the HRAIS, affixes the CE marking and places the system on the market or puts it into service.

• **Market phase: deployment and post-market monitoring**

Once a HRAIS is placed on the market or put into service, a set of obligations arises for both providers and deployers of the system. This phase is particularly relevant in that it requires meaningful cooperation between all the actors in the AI value chain, in particular, providers and deployers and, eventually, public authorities (e.g., market surveillance authorities, human rights authorities). The cooperation, reactive attitude and exchange of information between these actors is key to ensure that the deployment of a High-risk AI system does not pose risks to health, safety and fundamental rights. Providers are required to keep monitoring their HRAISs throughout their lifecycle and, if needed, adopt the measures necessary to keep their systems in compliance with the requirements set out in the AI Act. To this end, providers are required to establish a **post-market monitoring system** aimed to "actively and systematically collect, document and analyse relevant data … on the performance of HRAISs throughout their lifetime"[54]. Through the post-market monitoring system, the provider must "evaluate the **continuous compliance of AI systems with the essential requirements established under the AI Act. This includes** "an **analysis of the interaction with other AI systems**"[55].

[52] Annex III, 8.a.

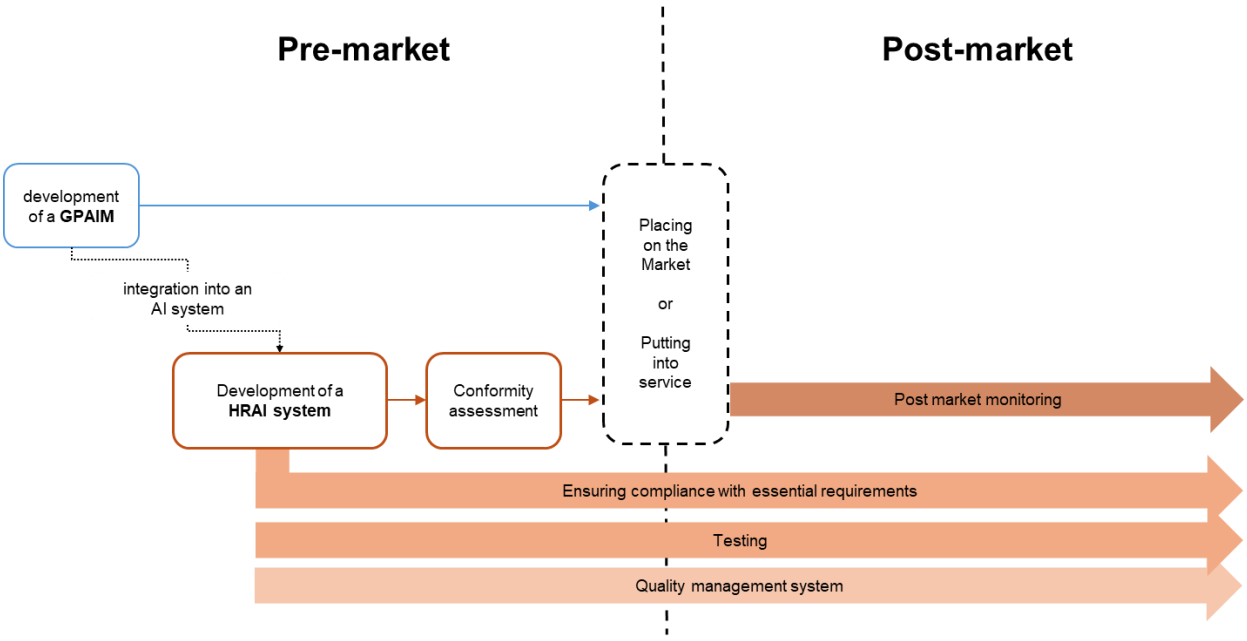
[53] Chapter III, Section IV, AI Act

[54] Art. 72 AI Act

[55] Ivi

First, deployers must take appropriate technical and organisational measures to ensure that the system is used in accordance with the instructions for use[56] and is not fed with input data that are not relevant and sufficiently representative in view of its intended purpose[57]. Moreover, deployers, as well as providers, are required to contribute to the monitoring of HRAISs and inform Market Surveillance Authorities in case a system causes a serious incident[58].

Figure 5. High-risk AI system lifecycle



In the next subsection, we will see more closely how AI metrics can provide both providers and deployers of HRAISs with a tool to **ensure** and **demonstrate** compliance with their respective obligations.

Before moving on, it is important to note a caveat. AI systems and AI models that do not fit in any of the categories mentioned above are not regulated by AI Act[59]. Nonetheless, it is worth noting that as "products" placed or put into service in the EU market, AI systems and models are still subject to legal requirements established by other legislative instruments, e.g., the General Product Safety Regulation[60]. Equally, it is

[56] Art. 26, § 1, AI Act

[57] Art. 26, § 4, AI Act

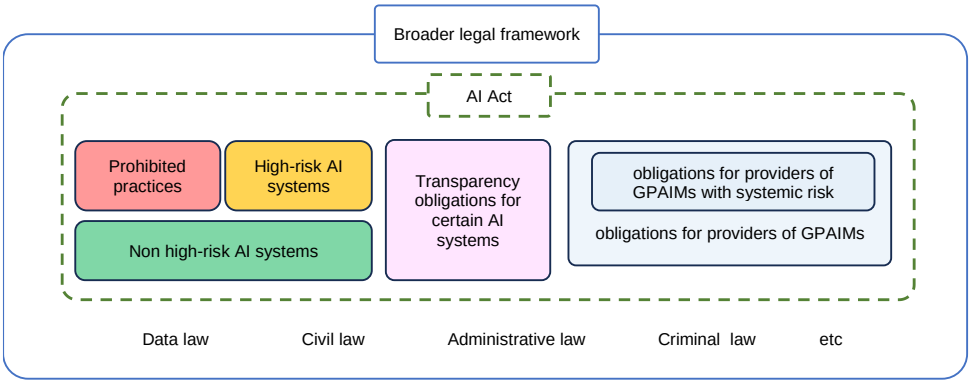
[58] Art. 73, AI Act

[59] Moreover, see R(165): "Providers of AI systems that are not high-risk should be encouraged to create codes of conduct, including related governance mechanisms, intended to foster the voluntary application of some or all of the mandatory requirements applicable to High-risk AI systems".

[60] Regulation (EU) 2023/988 of the European Parliament and of the Council of 10 May 2023 on general product safety, amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council and Directive (EU) 2020/1828 of the European Parliament and the Council, and repealing Directive 2001/95/EC of the European Parliament and of the Council and Council Directive 87/357/EEC, http://data.europa.eu/eli/reg/2023/988/oj

worth underlining that the action or omissions of providers and deployers might assume relevance under other applicable law, for instance, data protection law or law on information service providers. Indeed, metrics might also assume relevance in the context of the assessment of AI operators' administrative, criminal, civil liability and, in particular, product liability[61]. However, the following subsection will focus exclusively on the relevance that AI metrics assume directly under the AI Act.

Figure 6. The AI Act in the broader legal framework



3.2.2. Metrics and High-risk AI systems

This subsection illustrates the role played by AI metrics with respect to the requirements established by the AI Act *within* and *across* the legally relevant phases of the lifecycle of High-risk AI systems delineated in the previous subsection.

AI metrics are particularly relevant in the context of the operationalisation of the legal requirements of accuracy, robustness, cybersecurity, transparency, interpretability and oversight established for High-risk AI systems under the AI Act. Before examining such specific requirements, however, it is worth considering some general features that characterise AI metrics in the context of the broader architecture of the AI Act provisions on HRAISs, that will be relevant for the analysis of specific requirements. The *locus* of AI Metrics in the AI Act results from the overlapping of:

- three interlocking systems that providers of HRAISs are required to establish, i.e., the Quality Management System[62], the Risk Management System[63] and the Post Market Monitoring System[64];
- the information and documentation obligations of providers, in particular, the obligation to draw up technical documentation[65] and the obligation to provide clear a instruction for use to deployers of the HRAIS[66].

[61] Thiis is especially so with the proposed reform of the product liability directive, explicitly considering software as a product.

[62] Art. 17 AI Act

[63] Art. 9 AI Act

[64] Art. 72 AI Act

[65] Art. 11 and Annex IV of the AI Act

[66] Art. 13, AI Act

A systematic reading of these provisions makes it possible to identify some general features that characterise the role of AI metrics in the architecture of the AI Act.

The first feature is the strict relationship that the metrics in the AI Act are always to be considered in the light of the **intended purpose** of the HRAIS. This aspect is emphasised throughout the provisions on the quality management system and risk management system. The obligation to establish a **quality management system** has the purpose of ensuring that providers put in place a strategy to ensure regulatory compliance[67] and procedure for **"**the design, design control and design verification" as well as the "development, quality control and quality assurance" of their HRAISs[68]. This requires providers to carefully identify the technical means necessary to ensure that the HRAIS complies with the requirements set out in the AI Act, including harmonised standards, other relevant technical specifications[69] and AI metrics. These operationalisation choices inform the examination, testing and validation procedures that providers are required to carry out before, during and after the development of the HRAIS[70]. In this respect, the quality management system must include a **risk-management system** aimed to identify, assess and mitigate the risks that the HRAIS might pose to health, safety or fundamental rights[71]. The **testing** of the system is one of the most relevant strategies that providers are required to put in place to identify the most appropriate risk management measures[72]. Through testing, providers are required to make sure that their HRAISs **perform consistently** for their intended purpose and comply with the requirements established in articles 10-15 of the AI Act[73]. As per art. 9 of the Act AI Act, "testing shall be carried out against **prior defined metrics** and **probabilistic thresholds** that are ***appropriate to the intended purpose*** *of the HRAIS*".[74].

Secondly, the AI Act ascribes great importance to the thorough **documentation** and **justification** of providers' choices concerning AI metrics. The obligation to draw up and keep up to date **technical documentation**[75] requires providers, *inter alia*,

- to give a **detailed description** of the **metrics** used to measure the compliance of their AI systems with *relevant requirements established by the AIA[76], in particular accuracy* and *robustness*, as well as potentially discriminatory impacts[77].

- to justify why the metrics chosen are **appropriate** for the specific AI system under consideration[78].

Documentation is key also to **enable deployers to understand and correctly use the systems**.[79]

As said, in addition to the provisions on the Quality Management System, Risk Management System and documentation obligations, the AI Act contains a set of provisions that concern specific essential

---

[67] Art. 17, § 1, a, AI Act

[68] Art. 17, § 1, b and c, AI Act

[69] Art. 17, § 1, e, AI Act

[70] Art. 17, § 1, d, AI Act

[71] Art. 9 AI Act

[72] Art. 9, § 6, AI Act

[73] Cf, also, the provision concerning the post market monitoring system, discussed *supra*. AI Act, art. 72, § 2.

[74] Art. 9, § 8, AI Act (emphasis added)

[75] Art. 11 and Annex IV, AI Act

[76] Annex IV, § 2(g) and § 4, i.e., the requirements established in Chapter III, Section 2, AI Act

[77] Annex IV, § 2(g), AI Act

[78] Annex IV, § 4, AI Act

[79] See infra, pp. 21-24

requirements of High-risk AI systems, such as accuracy, robustness, cybersecurity, transparency, interpretability and human oversight. Although only some of these provisions contain direct references to metrics, they are all relevant for metrics in that they specify the requirements with respect to which the use of metrics can assume relevance. Put differently, by articulating the content of the essential requirements of HRAISs, the provisions that will be analysed in the remaining part of this subsection also determine which metrics are or are not appropriate to measure compliance with such requirements.

• ***Accuracy, robustness and cybersecurity***

Metrics play a key role in ensuring compliance with the requirements of **accuracy, robustness** and **cybersecurity** set out in art. 15 of the AI Act[80]. This provision requires providers to ensure that

- their High-risk AI systems achieve an "appropriate level" of accuracy, robustness and cybersecurity and

- such performance level is constant throughout the lifecycle of the systems.

It is worth noting that the AI Act does not provide a definition of accuracy, robustness and cybersecurity, nor does it identify specific metrics or methodologies to be used in their assessment. A more granular understanding of these requirements can nonetheless be achieved through a systematic reading of other relevant provisions of the AI Act. In particular, the provisions of the AI Act on technical documentation and data practices show that accuracy and robustness are closely connected to the legal requirement of non-discrimination.

First, as illustrated before, the provisions on technical documentation require providers to document, together with the metrics used to measure accuracy and robustness, also the metrics used to measure "potentially discriminatory impacts"[81].

Secondly, art. 10 of the AI Act highlights a strong relationship between the requirements of accuracy and robustness and the requirements concerning the **data** used to **train**, **validate** and **test**. Art. 10 demands providers to implement **data governance and management practices** and to ensure that their data meet a set of **quality criteria**. These mandatory practices and criteria request that providers duly consider the extent to which the datasets they employed for the training, validation and testing of HRASIs are *appropriate for the intended purpose* of the latter[82], i.e., that data are relevant, sufficiently representative, free of errors and complete[83]. In this respect, providers must assess the statistical properties of their datasets, taking into account "the persons or groups of persons in relation to whom the HRAIS is intended to be used"[84] and "the characteristics or elements that are particular to the specific geographical, contextual, behavioural or functional setting within which the HRAIS is intended to be used"[85]. These requirements are compounded by providers' obligations to

[80] See also R(74), AI Act

[81] Annex IV, § 2(g), AI Act

[82] Art, 10, §§ 2 to 5, AI Act

[83] Art. 10, § 3, AI Act

[84] Art. 10, § 3, AI Act. Those characteristics of the data sets may be met at the level of individual data sets or at the level of a combination thereof

[85] Art, 10, § 4, AI Act

- **examine possible biases** that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under EU law, especially in view of feedback loops[86];
- adopt appropriate measures to **prevent and mitigate such biases**[87];
- identification of relevant data gaps or shortcomings that prevent compliance with this Regulation, and how those gaps and shortcomings can be addressed[88].

Considered together, the provisions on accuracy, documentation, data governance and management practices and data quality criteria require providers to base their choices as to which metrics to use to measure accuracy on a careful consideration of the features of their datasets. This implies to duly consider the potential imbalances in the latter and to choose metrics accordingly, in order to avoid incurring in an accuracy paradox, i.e., a situation where a model achieves a high accuracy but fails to correctly classify instances of a class that is underrepresented in the dataset.

Furthermore, adopting accuracy metrics without duly considering potential discriminatory effects stemming from the selection and design of training datasets and inferred models would *per se* entail a lack of compliance with the requirement that demands such choices to be informed by the consideration of the intended purpose of High-risk AI systems. It is worth stressing again that the category of High-risk AI system incudes AI systems such as

- AI systems intended to be used to analyse and filter job applications and to evaluate candidates[89]
- AI systems intended to be used by competent public authorities for the examination of applications for asylum, visa or residence permits[90]
- AI systems intended to be used by public authorities to evaluate the eligibility of natural persons for essential public assistance benefits and services, including healthcare services, as well as to grant, reduce, revoke, or reclaim such benefits and services[91]
- AI systems intended to be used to evaluate the creditworthiness of natural persons or establish their credit score[92].

These are systems the deployment of which can result in a particularly high risk of discrimination. Because of this, providers are required to identify methods to measure accuracy that take into account – for instance through the use of adequate metrics – the potential biases of their systems.

Similar considerations can be made with respect to the requirement of **robustness**. The goal of the requirement of robustness is to avoid that "*erroneous decisions or wrong or biased outputs* generated by the AI system" lead to "safety impacts or negatively affect the fundamental rights"[93]. To this end, Article 15 mandates the adoption of measures to mitigate and address the risk of feedback loops with biased outputs[94] and metrics capable of measuring the resilience of the system with respect to "errors, faults and

[86] Art. 10, § 2(f), AI Act
[87] Art. 10, § 2(g), AI Act
[88] Art. 10, § 2(h), AI Act
[89] Annex III, § 4(a) AI Act
[90] Annex III, § 7(c) AI Act
[91] Annex III, § 5(a) AI Act
[92] Annex III, § 5(b) AI Act
[93] R(75) AI Act (emphasis added)
[94] Art. 15, § 4, AI Act

inconsistencies" that might be caused by the interaction of the system with natural persons or other systems[95].

Finally, with respect to robustness-cybersecurity, providers are required to put in place measures to ensure the resilience of their system against attempts by unauthorised third parties to alter the use, outputs or performance of HRAISs by exploiting system vulnerabilities[96], such as:

- data poisoning,
- model poisoning,
- adversarial examples
- model evasion,
- confidentiality attacks
- model flaws[97].

• ***Transparency, interpretability, human oversight***

Another important set of requirements established by the AI Act for HRAISs are the requirements of Transparency, Interpretability and Oversight (hereafter abbreviated as TIO). Understanding *what and how AI Metrics can measure* in order to contribute to the compliance with these requirements necessitates a close look at a set of interconnected provisions – art. 13, 14 and Annex IV - that address both providers and deployers of HRAISs[98].

Article 13 requires providers to design and develop their High-risk AI systems in such a way that the operations of their systems achieve a "**sufficient level of transparency"**[99]. What counts as a "sufficient" level of transparency is to be determined by "taking into account the needs and foreseeable knowledge of the target deployers"[100]. More specifically, for the level of transparency of a HRAIS to be deemed sufficient, it must be such as to:

- enable deployers to **interpret the system's output** and **use it appropriately**[101];
- enable deployers to **comply with their obligations**.

Article 14 requires providers to implement measures to enable deployers to exercise **meaningful oversight** on the system. A meaningful human oversight is that which, taking into account the level of autonomy and the context of use of the HRAIS[102], enables deployers to **prevent or minimise risks to health, safety or fundamental rights.**

The AI Act identifies a set of measures that providers are required to adopt to satisfy these TIO requirements.

First, providers are required to implement built-in oversight measures – e.g., appropriate human interfaces, measures to ensure that the system is responsive to the human operator, constraints that cannot be

[95] Ivi

[96] Art. 15, § 5, AI Act

[97] It is worth noting that similar requirements find applications also with respect to GPAIMs presenting systemic risks. *Supra*, Section 3.2.2.

[98] Art. 13, 14, art. 11 and Annex IV, especially, § 2, lett. b, e, and §§ 3-4, AI Act

[99] Art. 13, AI Act

[100] R(72), AI Act

[101] Ivi

[102] Art. 14, § 3, AI Act

overridden by the system itself – as well as establish measures to be implemented by deployers[103]. Secondly, and in close connection, providers are required to give deployers **instructions for use** containing complete, correct, clear, relevant, accessible and comprehensible information, including illustrative examples[104], on the characteristics, capabilities and limitations of performance of the HRAIS[105], and on the intended and precluded uses of the AI system. In particular, the instructions must inform deployers of:

- the technical capabilities and characteristics of the HRAIS to provide information that is relevant to explain its output[106];
- any known or foreseeable circumstance, related to the use of the HRAIS which may lead to risks to health and safety or fundamental rights
- specifications for the input data, or any other relevant information in terms of the training, validation and testing data sets used[107].

A systematic reading of Articles 13 and 14, together with the provisions on documentation obligations of providers, shows that the common goal of the strongly interconnected provisions on TIO requirements is that of ensuring an operational knowledge of HRAISs and their output, i.e., information that is available to - and understandable by - different relevant operators in a way that allows informed decision-making and the taking of necessary corrective actions. These actions range from intervening on the system, to not using the system or overruling its output.

AI Metrics can play a particularly important role in ensuring that these requirements are met and that this form of operational knowledge is achieved. First, part of the information that providers are required to make available to deployers is **information *about* the metrics** used to measure the compliance of the system with other relevant AI Act requirements. In this respect, art. 13 requires providers to include information on the following in the instructions for use:

- "the level of accuracy, including its metrics, robustness and cybersecurity … against which the HRAIS has been tested and validated and which can be expected"[108];
- "any known and foreseeable circumstances that may have an impact on that expected level of accuracy, robustness and cybersecurity[109];
- "the performance of the HRAIS regarding specific persons or groups of persons on which the system is intended to be used"[110].

Information about the metrics used by providers is particularly relevant to ensure human oversight, as metrics are key for deployers to duly monitor the operation of the HRAIS and to detect and address anomalies, dysfunctions and unexpected performance[111].

Secondly, AI metrics are clearly relevant to ***measure*** the level of **transparency, human oversight and interpretability** of HRAIS: for providers and public authorities, metrics provide an instrument to monitor the extent to which a HRAIS complies with TIO requirements established by the AI Act; for deployers, metrics

---

[103] Art. 14, § 3, AI Act; R(73) AI Act

[104] Art. 13, § 2; R(72) AI Act

[105] Art. 13, § 3(b) AI Act

[106] Art. 13, § 3, b(iv), AI Act

[107] Art. 13, § 3, b(vi), AI Act

[108] Art. 13, § 3, b(ii), AI Act

[109] Ivi

[110] Art. 13, § 3, b(v), AI Act

[111] Art. 14, § 4(a) AI Act

on TIO can be an important instrument to decide whether or how to deploy a certain HRAISs in a specific scenario.

In recent years, a vast literature has investigated the topic of AI explainability (XAI). It is out of scope for the present report to engage in a thorough review of this literature[112] or to engage in a discussion of which of the multiple metrics and methodologies developed are more in line with the requirements established by the AI Act. Nonetheless it is important to point out that the understanding of transparency, interpretability and oversight that the AI Act institutes and makes legally binding does not *necessarily* match the understanding of "explainability" that has been advanced in the literature. In this respect, it is opportune to further stress some of the legal requirements that constrain the choice of any methodologies and metrics - or combination thereof – intended to operationalise and measure the TIO requirements set out in the AI Act. As said, the transparency and interpretability of, and oversight over, a HRAIS depends on the intended purpose of the system and, in particular, the specific persons or groups of persons targeted by the system. Transparency, interpretability and oversight metrics must be adequate to the needs, legal obligations and foreseeable knowledge of the intended deployers. Any metric employed for the purpose of meeting TIO requirements must be capable of contributing to the assessment of the extent to which the natural persons tasked with human oversight:

- properly understand the relevant **capacities and limitations of the HRAIS**
- remain aware of the possible **automation bias**[113]
- correctly **interpret the HRAIS's output**[114].

Furthermore, correctly interpreting the output entails the overseer's capacity to:

- disregard, override or reverse the output[115];
- decide not to use the HRAIS[116];
- intervene in the operations of the HRAIS or interrupt it safely[117].

Moreover, as seen above, the level of transparency of HRAISs must be such as to enable deployers to comply with their obligations. It is important to consider that such obligations might include, among other[118], the obligation to:

- carry out a Fundamental Rights Impact Assessment (FRIA)[119]
- perform a Data Protection Impact Assessment (DPIA)[120]

---

[112] Sovrano, F.; Sapienza, S.; Palmirani, M.; Vitali, F. Metrics, Explainability and the European AI Act Proposal. J 2022, 5, 126–138. https://doi.org/10.3390/j5010010; Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. ACM Comput. Surv. (CSUR) 2018, 51, 1–42; Zhou, J.; Gandomi, A.H.; Chen, F.; Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics 2021, 10, 593

[113] Art. 14, § 4(b) AI Act

[114] Art. 14, § 4(c) AI Act

[115] Art. 14, § 4(d) AI Act

[116] Art. 14, § 4(d) AI Act

[117] Art. 14, § 4(e) AI Act

[118] Art. 26 AI Act

[119] According to art. 27, § 1, AI Act, this obligation applies to i) deployers that are bodies governed by public law, ii) deployers that are private entities providing public services, and iii) deployers of HRAISs intended to be used to evaluate the creditworthiness of natural persons or establish their credit score (see Annex III, § 5 (b)) and iv) deployers of HRAISs intended to be used for risk assessment and pricing in relation to natural persons in the case of life and health insurance (see Annex III, § 5, c).

[120] Art. 35 GDPR

- provide any affected person subject to a decision taken on the basis of the output from a HRAIS,[121] which produces legal effects or similarly significantly affects that person, in a way that they consider to have an adverse impact on their health, safety or fundamental rights, with a **clear and meaningful explanation of the role of the AI system in the decision-making procedure and the main elements of the decision taken**[122].

3.2.3. Towards operationalisation

The provisions of the AI Act examined in the previous subsections **make the importance of AI metrics along the entire AI value chain clear**, from the design and development of High-risk AI systems and GPAIMs to the deployment and post-market monitoring of High-risk AI systems.

Equally, the overview of the relevant provisions of the AI Act has demonstrated how the legal requirements established by the AI Act are **high-level** and can be considered as "**metrics-agnostic"**, leaving it to providers to determine which technical choices are necessary to operationalise legal requirements. For instance, we have seen that the way in which the requirement of accuracy is formulated in the AI Act demands providers to i) determine what counts as an "appropriate level" of accuracy; ii) choose the metrics that are more appropriate to measure accuracy; iii) document and justify such operationalisation choices, while taking into account the specific intended use of the system, the envisaged deployers and the subjects that will be affected by the deployment of the system.

It must be noted that some provisions of the AI Act require a further operationalisation of the essential requirements, with a specific reference to metrics. For what concerns the requirements established for providers of **GPAIMs**[123], the Commission is assigned to promote the drawing up of **codes of practice** that are expected to be published 9 months from the entry into force of the AI Act[124]. Such codes of practice, which may be given general validity within the EU by the Commission, will contain "commitments or measures, including key performance indicators as appropriate, to ensure the achievement of those objectives"[125]. For what concerns HRAIS, firstly, the Commission is put in charge of promoting the adoption of - or to adopt itself – implementing acts and guidelines on the essential requirements[126]. In particular, the Commission is required to promote the cooperation with relevant stakeholders and organisations such as metrology and benchmarking authorities to develop benchmarks and measurement methodologies to address the technical aspects of how to measure the appropriate levels of *accuracy* and *robustness* and *any other relevant performance metrics*[127]. Secondly, as anticipated in section 3.1., the publication of the

[121] Limited to the HRAISs indicated in Annex III, with the exception of systems listed under point 2 thereof, Art. 86, § 1, AI Act

[122] Art. 86 AI Act. - This obligation applies unless a otherwise provided for under Union law, In most cases, in conjunction with the provisions of art. 22 GDPR, i.e., the prohibition of completely automated decision-making and profiling.

[123] Supra, section 3.2.2

[124] Art. 56 AI Act. This is one of the tasks of the AI Office, art. 3(47): 'AI Office' means the Commission's function of contributing to the implementation, monitoring and supervision of AI systems and general-purpose AI models, and AI governance, provided for in Commission Decision of 24 January 2024; references in this Regulation to the AI Office shall be construed as references to the Commission'.

[125] Art. 56, §§ 4, 6, AI Act

[126] Art. 96, § 1(a), AI Act

[127] Art. 15, § 2, R(74), AI Act

harmonised standards that the Commission has requested of CEN and CENELEC[128] will provide technical specifications for all the essential requirements established in the AI Act.

Together, these various sources of technical specifications will undoubtedly be an important contribution to filling the operationalisation gap left open by the provisions of the AI Act on essential requirements.

However, the conceptual premises explicated in Section 2 combined with the legal analysis carried out in this section demonstrate that technical specifications - whatever their level of detail, can inform, but not "solve" once and for all the legal-technical challenge of ensuring – and "measuring" - compliance with the legal requirements established by the AI Act. The majority of the concrete operationalisation choices necessary to ensure compliance with the AI Act will still be in the sphere of providers.

Standardised technical specifications and methodologies, such as AI metrics, will play a crucial role in laying the groundwork for a market of AI products that ensures the protection of fundamental rights. However, technical specifications and metrics can only go so far with respect to this goal. Technical specifications and metrics are called, as it were, to address a moving target.

First, the goal of the AI Act is to ensure the protection of fundamental rights and compliance with technical specifications is one way to achieve this goal, but is not the goal in and of itself. This also implies that adherence to technical specifications does not relieve providers of their responsibility to develop their systems and models in a way that meets the current state of the art[129]. The state of the art, especially in a field like AI, evolves faster than the process necessary to translate technical expertise into standardised technical specifications. Risks that couldn't be even imagined 6 months ago have quickly arisen, and subsequently become foreseeable and addressable, making it incumbent on providers to take an active part in the identification and implementation of the more adequate measures to ensure their compliance with the law.

Secondly, the rationale for classifying an AI system as High-risk and to subject it to the complex set of requirements applicable to that category is to avoid the use of AI systems in specifically pre-defined areas leads to adverse impact on fundamental rights such as:

- human dignity,
- respect for private and family life,
- protection of personal data,
- freedom of expression and information,
- freedom of assembly and of association,
- and non-discrimination,
- right to education
- consumer protection,
- workers' rights,
- rights of persons with disabilities,
- gender equality,
- intellectual property rights,

[128] Commission Implementing Decision on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on artificial intelligence Brussels, 22.5.2023 C(2023) 3215 final

[129] Philippe Portalier, Myths and realities of the presumption of conformity. Scope and relevance of the presumption of product conformity with Union harmonisation legislation in 10 questions and answers, version 1c of 15/5/2017; Guide to application of the PPE Directive 89/686/EEC, Version of 19 October 2015, page 29: http://ec.europa.eu/DocsRoom/documents/13241/attachments/1/translations/

- right to an effective remedy and to a fair trial,
- right of defence and the presumption of innocence,
- right to good administration[130].

Understanding how the use of an AI system can, in concrete scenarios, result in violations of such rights demands a situated assessment that cannot be fully standardised.

One can imagine important questions arising in this context, like *"What counts as suitable metrics to measure accuracy for a system used to make decisions on requests for visa by third country nationals?"*. Or *"What is the level of transparency and interpretability required for a system used to decide what is the best candidate for a job position?"* or *"What if the use of two different metrics leads to incompatible results?*

Answering these kinds of questions requires both going beyond technical standards and metrics and situating both in a broader context that has as a primary aim to ensure protection of fundamental rights.

Finally, the overview of the legal provisions given in the previous sections has shown that the requirements established in the AI Act are strictly interconnected. Compliance with the AI Act makes it necessary to address the essential requirements through a holistic approach that goes beyond any singular requirement and its accompanying technical specification.

All the considerations put forward above inform the research perspective of the macro-project "Metrics for Ethics", which will be discussed in the next section.

[130] R(48), R(52) AI Act

4. The macro-project "Metrics for Ethics": Lessons learned on how operationalise AI Metrics for Legal Protection by Design

This Section summarises the lessons learned through the participation in the macro-project "Metrics for Ethics", conducted under WP5 of the HumanE AI Net project.

Section 4.1 presents the goals and the research carried out in the macro-project. Subsequently, Section 4.2 connects the findings of the macro-project with the concepts of "Agonistic Machine Learning" and Legal Protection by Design.

4.1. From the case study to the "Ethics Dashboard"

The macro-project "Metrics for Ethics" has been developed through the collective work carried out by a team of researchers from the Barcelona Supercomputing Centre (Spain), the Istituto di Calcolo e Reti ad Alte Prestazioni del Consiglio Nazionale delle Ricerche (ICAR-CNR, Italy), the Institute for Systems and Computer Engineering, Technology and Science (INESC TEC, Portugal), Luleå University of Technology (Sweden), Umea University (Sweden) and the Law Science Technology Society Research Group of the Vrije Universiteit Brussel (LSTS-VUB, Belgium).

The team has decided to address the topic of AI Metrics by conducting a case study based on the so-called "German Credit dataset""[131]. This choice is justified by the consideration that the German Credit dataset is both a well-established dataset within the AI community and a dataset that would have ensure a strong connection with the research conducted in the macro-project and the AI Act. The dataset was made available in 2004 by Professor Hans Hofmann in the context of the European project Statlog[132] and, since then, it has been widely used by AI researchers[133]. The dataset contains a sample of 1000 entries representing applicants for a loan to a bank. The credit risk of each entry is classified as good or bad according to 20 attributes (integer, e.g., age or amount of credit requested; categorical, such as credit history and purpose of the credit request; and binary, such as whether the applicant is a foreign worker or not).

Because of these features, the team deemed this dataset appropriate for conducting a case study on AI systems designed for the purpose of creditworthiness evaluation. Creditworthiness evaluation is an intended purpose of AI systems that is relevant under the rules established under the AI Act for High-risk AI systems[134].

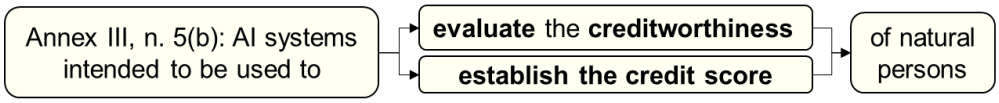
[131] https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data

[132] For an account of the process of collection and curation of this dataset, see Ulrike Grömping, 'South German Credit Data: Correcting a Widely Used Data Set' <http://www1.beuth-hochschule.de/FB_II/reports/Report-2019-004.pdf> and the further literature thereby referenced

[133] For a (partial) list of research conducted on this dataset see https://paperswithcode.com/dataset/german-credit-dataset

[134] Annex III, 5.b, AI Act

Figure 7. AI Act, Annex III, n 5(b)



The goal of the macro-project is to raise awareness about the inherent complexity and context-dependent character of the process of operationalisation of - and compliance with - the requirements established under the AI Act, as well as under relevant AI ethics frameworks. The goal of the case study is therefore not to provide an exhaustive examination of all the legal and ethically relevant aspect of designing, developing and deploying AI systems for creditworthiness evaluation. Rather, the case study seeks to contribute, by providing a hands-on perspective, to the identification of key aspects for research on the relationship between AI metrics on the one hand, and ethical and legal requirements on the other.

Each participant in the macro-project has carried out research that addresses one or more of the requirements that assume relevance with respect to both the compliance with the AI Act and AI ethics frameworks[135].

The team from the Barcelona Supercomputing Centre has engaged in an in-depth review of the literature regarding transparency indicators and metrics. Based on this research, the team has identified a set of indicators that can be used to support the implementation of systems that are **transparent** to the end-user.

The Team from ICAR-CNR has carried out research on the topic of **robustness**, investigating how minimum modifications of attributes in the dataset affect the prediction errors of the model.

The team from Umea University has performed research on **fairness** metrics and **AI trustworthiness**.

The team from the INESC TEC has conducted research on the topic of **explainability**, focusing on different methodologies to explain credit rejection decisions. The team has conducted participatory design research aimed at assessing the clarity, comprehensibility, fairness and bias of different types of local explanations of creditworthiness decisions.

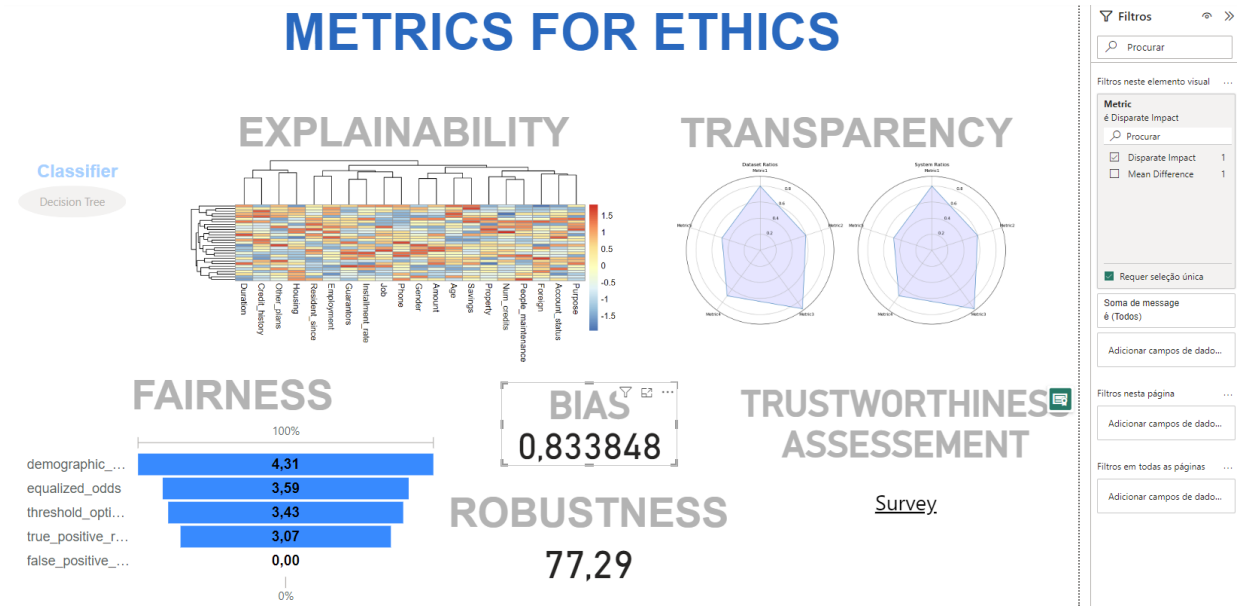
The research carried out by the team from the Luleå University of Technology has focused on the estimation and explanation of bias in the data for training AI models. Relying upon the AI Fairness 360 library, the team has used 7 different metrics to estimate and explain **bias** (or fairness) in the training data and trained a logistic regression model on the data. The metrics are mean difference, disparate impact, consistency, smoothed empirical differential, base rate, number of positives, and number of negatives.

The results of the research carried out by all the research teams have been integrated by the team members of INESC-TEC into the "**Metrics for Ethics Dashboard**". This Dashboard offers to its users an interactive interface to explore the features of the dataset and of model trained on the dataset, with visualisations of

[135] For a more detailed account of the research conducted by each partner, see Sónia Teixeira, Atia Cortez, Dilhan Thilakarathne, Gianmarco Gori, Jack O'Keefe, Marco Minici, Monowar Bhuyan, Nina Khairova, Tosin Adewumi, Carmela Comito, João Gama, Virginia Dignum, 'Integrated tool for evaluating ethics in AI: A Case Study', in Mohamed Chetouani, Andrzej Nowak and Paul Lukowicz (eds), *Handbook of Generative AI for Human-AI Collaboration* (Springer forthcoming)

the effects of the use of different metrics and methodologies that are relevant for compliance with ethical and legal requirements. Users can compare the results of different methods of explanation for algorithmic decision-making and the results of the application of different metrics meant to measure fairness, bias and robustness, as well as engage in a trustworthiness assessment based on a questionnaire.

Figure 8. Screenshot of the Dashboard Metrics for Ethics v. 1.0



The next section will elaborate on how the findings of the macro-project help to answer the question of **"what role AI Metrics can play for Legal Protection by Design in the AI Value Chain**".

4.2. Seeing AI Metrics through the lenses of an Agonistic approach to Machine Learning

The collaborative research conducted in the context of the macro-project has provided support to the finding that, for AI metrics to contribute to Legal Protection by Design, it is key that their development and deployment is informed by the approach that Hildebrandt calls "**Agonistic Machine Learning**". This concept highlights that, to ensure legal protection, it is necessary to incorporate the "adversarial core" of

the Rule of law into AI systems[136], making sure that the design and deployment of AI systems is based on "agonistic debate, built-in falsifiability and a robust constructive distrust"[137].

As an important preliminary step, an agonistic approach challenges providers and deployers to not neglect the so-called "**question 0**". This entails considering whether AI actually is a solution to a problem - or, instead, is a solution *looking for* a problem – or even gives rise to more problems than the ones that it is meant to address. With respect to the use of AI metrics, this implies calling on the relevant actors in the AI value chain to both cultivate a constructive doubt regarding the quantitative methods they adopt to measure compliance with legal requirements as well as to actively identify and consider the **advantages and the limitations of quantification**. To put it differently: "[a]ny metric is just a proxy for what you really care about"[138]. Participating in the macro-project has facilitated a concrete appreciation for the fact that the results of any metric can be informative, but should not be decisive. When it comes to the assessment of whether, or to what extent, a certain legal requirement is complied with, the use of metrics can be an important step, but they are never the be all and end all of the compliance process. Designing, developing and deploying computational technologies with an agonistic approach entail being aware that **things can be computed, and measured, in different ways**. In this respect, the macro-project has yielded concrete insights into the importance of the use of **multiple metrics** by providers[139]. Not only can using a plurality of metrics give providers a more informative and multi-layered picture of what they aim to measure, using a plurality of metrics can also force providers to reckon with the possibility that different metrics can lead to incompatible results. Facing the lack of univocity between different measurements should counteract any temptation to believe that the use of metrics is backed by the incontestable authority of objective quantification. In this respect, the macro-project has demonstrated how the outcomes of metrics form only one small part of a complex and non linear chain of decisions. Only the full picture of such a chain of decisions allows for the proper understanding of the significance of the result of a measurement. In this way, an agonistic approach forces providers to articulate their assumptions and make explicit and duly substantiate their claims as to the rationale and benefits of their design choices[140].

In section 3, it was discussed how the AI Act has turned an articulated set of documentation requirements into positive law, demanding from providers that they draw up technical documentation and instructions for use. Even though the addressees of these documents are Market Surveillance Authorities (and notified bodies) and deployers, the activity of *documenting* also produces an effect for providers: requiring providers to *give an account* of their design practices, documentation obligations thereby facilitating *accountability.* Notably, documentation forces providers to look back and forward simultaneously, to anticipate and (pre-emptively) address the consequences of their choices. In section 2 we have seen that the Legal Protection by Design approach requires taking into account that compliance with the law and legal protection have an inherently situated character. In this respect, an agonistic approach demands that providers and deployers reconcile technological standardisation with the need to account for the effects that an AI solution has in practice, in the context of a specific concrete scenario, with respect to specific persons. This implies the *ex ante* adoption of the measures necessary to allow for meaningful contestation of design choices *ex post,* giving a meaningful avenue to effectively redressing their potentially negative effects.

This is a crucial point, AI metrics can become an instrument that contributes to legal protection by empowering **contestation**. For instance, AI metrics could be used by claimants in a legal proceeding as

---

[136] Mireille Hildebrandt, 'Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning' (2019) 20 Theoretical Inquiries in Law 83, at 107

[137] Ivi

[138] Thomas and Uminsky (n 7).

[139] In this sense, see also Thomas and Uminsky (n 7); High Level Expert Group on Artificial Intelligence (n 8)

[140] For an example of the application of this mindset in the context of AI-driven legal technologies, see Laurence Diver, Pauline McBride, Masha Medvedeva, Arjun Banerjee, Eva D'hondt, Tatiana Duarte, Desara Dushi, Gianmarco Gori, Emilie van Den Hoven, Paulus Meessen, Mireille Hildebrandt, Typology of Legal Technologies, 4 Nov 2022, https://publications.cohubicol.com/typology/.

evidence of the inadequate character of the choices made by a provider to prevent their AI systems from having discriminatory impacts.

5. Conclusions

This report has investigated the issue of the incorporation of fundamental rights protection into the architecture of AI systems by addressing the specific question of “What role can AI metrics play for Legal Protection by Design in the AI Value Chain?”. To this end, the report has integrated legal research with the experience, knowledge and insights gained in the context of the collaborative macro-project “Metrics for Ethics” carried out under WP5 of the HumanE AI Net project.

Section 2 has shown how AI metrics are situated at the intersection of AI practice, AI ethics and law. It was illustrated how the concept of Legal Protection by Design (LPbD) helps to understand the different effects that the use of metrics can produce in each of such domains. Distinguishing LPbD from other “by design" approaches, the report has highlighted the role that metrics play in the context of the operationalisation of legal requirements.

Section 3 has illustrated how metrics assume relevance under EU law. The examination of the provisions of the AI Act on General-purpose AI models and High-risk AI systems has confirmed the finding of section 2. The analysis of the AI Act has shown that AI metrics can play an important part in ensuring compliance with multiple legal requirements, but cannot, *per se*, determine a legal assessment as to whether a certain legal requirement is complied with.

Section 4 has integrated the results of the legal analysis with the lessons learned from the research conducted in the macro-project “Metrics for Ethics”. Building on the insights gained in the macro-project, the report has illustrated how, through an approach informed by the concept of Agonistic Machine Learning, AI metrics can be integrated in the design, development and deployment of AI systems in a way that contributes to Legal Protection by Design.

References

Diver L, Duarte T, Gori G, van den Hoven E and Hildebrandt M, COHUBICOL Research Study on Text-Driven Law, 2023, https://publications.cohubicol.com/assets/uploads/cohubicol-research-study-on-text-driven-law-final.pdf

Diver L, McBride P, Medvedeva M, Banerjee A, D'hondt E, Duarte T, Dushi D, Gori G, van Den Hoven E, Meessen P, Hildebrandt M, Typology of Legal Technologies**,** 4 Nov 2022, https://publications.cohubicol.com/typology/

Dworkin R, Law's Empire (Belknap Press 1986)

Goodhart C, 'Goodhart's law' in Rochon L and Rossi S, The Encyclopedia of Central Banking, (Edward Elgar Publishing 2015), pp. 227–228

Gori G, 'Legal and Computer Rules: An Overview Inspired by Wittgenstein's Remarks' in Alice Helliwell, Alessandro Rossi and Brian Ball (eds), Wittgenstein and Artificial Intelligence, vol II (Anthem Press 2024)

Gori G, 'Rule of law and Positive law' in Diver L, Duarte T, Gori G, van den Hoven E and Hildebrandt M, COHUBICOL Research Study on Text-Driven Law, 2023, https://publications.cohubicol.com/assets/uploads/cohubicol-research-study-on-text-driven-law-final.pdf, pp 39-56.

Grömping U, 'South German Credit Data: Correcting a Widely Used Data Set', http://www1.beuth-hochschule.de/FB_II/reports/Report-2019-004.pdf

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D, A survey of methods for explaining black box models. ACM Comput. Surv. (CSUR) 2018, 51, 1–42. https://doi.org/10.1145/3236009

Hart H, *The Concept of Law* (Oxford University Press 1992)

Hildebrandt M, 'A Vision of Ambient Law' in Roger Brownsword and Karen Yeung (eds), Regulating Technologies (Hart 2008)

——, Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology (Edward Elgar Publishing 2015)

——, 'Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning' (2019) 20 Theoretical Inquiries in Law 83

——, Law for Computer Scientists and Other Folk (Oxford University Press 2020)

——, 'Three framing concepts' in Diver L, Duarte T, Gori G, van den Hoven E and Hildebrandt M, COHUBICOL Research Study on Text-Driven Law, 2023, https://publications.cohubicol.com/assets/uploads/cohubicol-research-study-on-text-driven-law-final.pdf, pp 12-27

——, 'Boundary Work between Computational 'Law' and 'Law-as-We-Know-it'', in Deirdre Curtin, and Mariavittoria Catanzariti (eds), *Data at the Boundaries of European Law*, Collected Courses of the Academy of European Law (Oxford, 2023) https://doi.org/10.1093/oso/9780198874195.003.0002,

Mitchell TM, Machine Learning (McGraw-Hill 1997)

Palumbo G, Carneiro D and Alves V, 'Objective Metrics for Ethical AI: A Systematic Literature Review' 2024 International Journal of Data Science and Analytics, https://doi.org/10.1007/s41060-024-00541-w

Portalier P, Myths and realities of the presumption of conformity. Scope and relevance of the presumption of product conformity with Union harmonisation legislation in 10 questions and answers, version 1c of 15/5/2017

Sovrano, F.; Sapienza, S.; Palmirani, M.; Vitali, F. Metrics, Explainability and the European AI Act Proposal. J 2022, 5, 126–138. https://doi.org/10.3390/j5010010

Teixeira S, Cortez A, Thilakarathne D, Gori G, O'Keefe J, Minici M, Bhuyan M, Khairova N, Adewumi T, Comito C, Gama J, Dignum V, 'Integrated tool for evaluating ethics in AI: A Case Study' in Mohamed Chetouani, Andrzej Nowak and Paul Lukowicz (eds), Handbook of Generative AI for Human-AI Collaboration (Springer forthcoming)

Thomas RL and Uminsky D, 'Reliance on Metrics Is a Fundamental Challenge for AI' (2022) 3 Patterns 100476

Zhou, J.; Gandomi, A.H.; Chen, F.; Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics 2021, 10, 593

Legal sources

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), http://data.europa.eu/eli/reg/2024/1689/oj

Commission Implementing Decision on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on artificial intelligence Brussels, 22.5.2023 C(2023) 3215 final

Regulation (EU) 2023/988 of the European Parliament and of the Council of 10 May 2023 on general product safety, amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council and Directive (EU) 2020/1828 of the European Parliament and the Council, and repealing Directive 2001/95/EC of the European Parliament and of the Council and Council Directive 87/357/EEC, http://data.europa.eu/eli/reg/2023/988/oj

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), http://data.europa.eu/eli/reg/2022/2065/oj

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), http://data.europa.eu/eli/reg/2016/679/oj

Regulation (EU) No 1025/2012 of the European Parliament and of the Council of 25 October 2012 on European standardisation, amending Council Directives 89/686/EEC and 93/15/EEC and Directives 94/9/EC, 94/25/EC, 95/16/EC, 97/23/EC, 98/34/EC, 2004/22/EC, 2007/23/EC, 2009/23/EC and 2009/105/EC of the European Parliament and of the Council and repealing Council Decision 87/95/EEC and Decision No 1673/2006/EC of the European Parliament and of the Council Text with EEA relevance, http://data.europa.eu/eli/reg/2012/1025/oj

Council Resolution of 7 May 1985 on a new approach to technical harmonization and standards, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31985Y0604(01).

Other sources

ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT), https://facctconference.org/

AI Fairness 360 (AIF360), https://github.com/Trusted-AI/AIF360

High Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI', 2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

IBM, Everyday ethics for AI https://www.ibm.com/design/ai/ethics/everyday-ethics/

Mireille Hildebrandt, Tutorial on the proposal for an AI Act, HumanE AI Net, https://www.humane-ai.eu/event/tutorial-on-the-proposal-for-an-ai-act/

Mireille Hildebrandt and Gianmarco Gori, Tutorial on the final text of the AI Act, HumanE AI Net, https://www.humane-ai.eu/event/second-tutorial-on-the-ai-act/

OECD, Catalogue of Tools & Metrics for Trustworthy AI https://oecd.ai/en/catalogue/metrics

UCI Machine Learning Repository, https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data