The EU AI Act and the Wager on Trustworthy AI

by

Alejandro Bellogín, Oliver Grau, Stefan Larsson, Gerhard Schimpf, Biswa Sengupta, Gürkan Solmaz

1 INTRODUCTION

Artificial intelligence (AI) systems are increasingly supplementing or taking over tasks previously performed by humans. On the one hand, this relates to low-risk tasks such as recommending books or movies, or making purchase recommendations based on previous buying behavior. But it also includes crucial decision-making by highly autonomous systems. Many current systems are opaque in the sense that their internal principles of operation are unknown, leading to severe problems for safety and regulation. Once trained, deep learning systems perform well, but they are subject to surprising vulnerabilities when confronted with adversarial images [28].The decisions may be explicated after the fact, but these systems carry the risk of wrong decisions affecting the well-being of people. They may be discriminated against, disadvantaged, or seriously injured. Examples are suggestions on how to select a job applicant, the proper medical treatment for a patient, or how to navigate autonomous cars through heavy traffic. In such situations, several ethical, legal, and general societal challenges arise. At the forefront is the question of who is responsible for a decision made by an AI system. Do we leave the decision to the AI system, or does a human decide in partnership with an AI system? Are there reliable, trustworthy, and understandable explanations for the decisions in each case? Yet the inner workings of many AI systems remain hidden – even from experts. Given the critical role, AI systems play in modern society; this seems in many cases unacceptable. But how can we make complex, self-learning systems explainable? And to what extent is this lack of explanation or broader transparency contributing to a watchful and responsible introduction of AI-systems that have evidenced benefits?

A deeper look at the technical details of AI and technical innovation on their way, like autonomous systems shows an obvious need for technical expertise in the practical technical and societal aspects of AI in the decision-making process. On the other hand, a purely technological perspective may result in regulations that cause more significant societal problems. This paper highlights accurate and realistic technology descriptions that take into account the risk factors as required, for example, by the risk pyramid of the EU AI Act that entered into force in June 2024. To strike such a balance for the public interest, policymakers should prioritize societal and environmental well-being and seek advice from interdisciplinary groups, as the impact of AI and autonomous systems is very difficult to assess by a single group. This more holistic system view is complementary to previous statements focusing on ethical aspects, responsibility and transparency in the development of algorithms [2], specifically on algorithmic systems involving AI and machine learning[1],[21],[29].

Many members of the public, particularly in Europe, exhibit skepticism towards AI and autonomous systems, which often translates into a lack of confidence or a cautious ”wait-and-see” approach [24]. For this technology to develop its beneficial potential, we need a framework of rules within which all players can operate responsibly. For the future of AI systems, specifically in the public spheres, where people express their personal expectations and worries about the potential consequences of AI being used without proper oversight, certain aspects must be taken into account. The following points are crucial for guiding the formulation of policies and regulations related to AI and are essential for the research and development community:

1.1 Supporting research and development in AI and autonomous systems

We recommend advanced research on the governance of implemented AI and automated systems, e.g., transportation. Special care must be taken at an early stage to contribute and adhere to transparent standards for hardware and software that provide insight to carry out the independent safety certifications that are legally required.

1.2 Creating and supporting sustainable solutions

In the light of the UN sustainability development goals, we recommend advancing multidisciplinary research methodologies that integrate social sciences and humanities alongside engineering sciences. Social sciences, such as sociology and anthropology, can provide crucial insights into how people understand, interact with, and trust AI systems. This understanding is vital for designing technologies that are socially acceptable, beneficial, and promote sustainable development. Humanities disciplines, like philosophy, can offer valuable perspectives on ethics, fairness, and the potential impact of AI on human values. This combined approach can lead to developing sustainable and energy-efficient autonomous systems that align with societal well-being.

1.3 Prioritising societal well-being and equal opportunities

We recommend that the legislative processes, especially in adapting existing laws and the new design of liabilities, take an interdisciplinary approach and consult the scientific and technical expertise in trusted AI. This should ideally lead to equal opportunities and fairness in the new business development considering new autonomous systems, preventing monopolies.

1.4 Promoting education on science, technology, social impact, and ethics

To foster responsible and beneficial use of AI, we propose enhancing educational curricula in secondary schools, universities, and technical fields to include fundamental knowledge about AI ethics and its impact on society. Incorporating ethical and social scientific aspects into computer science (CS) curricula, as exemplified by Stanford University's approach, will encourage students to consider "embedded" ethical, legal, or social implications while solving problems. Similarly, in Europe, some institutions teach CS students to relate the ACM Code of Ethics for Professional Conduct [2] to their tasks, fostering a sense of responsibility in their future AI-related endeavors.

The overall level of expertise in all levels of our society about how AI works and operates represents a critical success factor that will ultimately lead to confidence and acceptance of beneficial uses of these technologies in our daily lives. Policymakers, developers, and adopting users of AI systems need to be literate about these technologies and find answers at the intersection of technology, society, and policymaking. Furthermore, we ought to weigh the risks of autonomous systems against the benefits to allay public fears.

The points mentioned here highlight the need for an interdisciplinary and holistic approach to beneficial usage of AI. They set the foundation for a broader involvement of the public on one hand, and the subsequent development of the EU AI Act. To be informed about the endeavors of a supranational governmental organization such as the EU, striving to establish consensus across 27 member states regarding the legal regulation of Artificial Intelligence, is likely to capture the attention of a diverse international readership. This diverse audience includes academics in the field of AI ethics, explainable AI, and risk management as well as professionals who may be called upon to provide technical expertise to lawmakers in other parts of the world.

2 BACKGROUND: THE EU POLICIES ON AI AND ETHICS GUIDELINES

Considered one of the 'lighthouse' projects, public trust in autonomous systems is a crucial issue, well in line with recent awareness in the governance over artificial intelligence [22], [27], [18] expressed in the joint agreement of the EU Commission and EU council's proposal for a new European AI Act [14], as well as the High-Level Expert Group called in by the EU Commission in 2019 [10]. The High-level Expert Group's Ethics Guidelines echo several critical issues on human-centred and transparent approaches pointed to several principled documents [19].

The EU Commission takes a three-step approach: setting out the essential requirements for trustworthy Artificial Intelligence, launching a large-scale pilot phase for feedback from stakeholders, and working on international consensus building for human-centric Artificial Intelligence [12]. Among others, the ACM Europe Technology Policy Council (TPC) [5] collaborates with the EU Commission as a stakeholder and representative of the European computer science community, providing technical input on relevant initiatives. While the Commission looks broadly at an assessment of AI from a general point of view to preserve the values of the European member states, a more comprehensive judgment will result if all

the actors, i.e., owners, designers, developers, and researchers' predictive assessments are taken into account [1], [2], [29], [6]. This process led to the proposal for an AI Act, first published by the European Commission in April 2021, and the final version in force from June 2024, which we will return to below.

3 ESSENTIALS FOR ACHIEVING TRUSTWORTHY AI SYSTEMS

Implementers of AI and autonomous systems must be aware of what we, as responsible citizens, can accept and what's ethical and put laws and regulations in place to safeguard against future tragedies. Trustworthy AI should, e.g., according to the European Commission's High-Level Expert Group on AI, therefore, respect all applicable laws and regulations and a series of requirements for the particular sector. Specific assessment lists aim to help verify the application of each essential requirement. The following list of essentials is taken from the EU document Building Trust in Human-Centric Artificial Intelligence. It results from the work of a European High-Level Expert Group on ethics [11]. Additional perspectives are covered in a report by the Alan Turing Institute [23].

3.1 Developing trust in Autonomous Systems in the Public Sphere

Human Agency and Oversight

The essentials described above in the direction of an explainable and trustworthy AI may be suitable to convince professionals who knowingly interact with AI systems [8], [9]. It would be similarly essential to ensure trust in these systems among the public. However, it is important to note that explainability in AI, particularly in deep neural networks, remains a significant scientific challenge. Some scientists argue that the inherent complexity and the high-dimensional nature of these models make it difficult, if not impossible, to fully explain their outcomes. This skepticism raises critical questions about the feasibility of achieving truly transparent AI systems.

Therefore, ways to establish an individual trust in AI must be sought. However, more than detailed explanations of individual outcomes will be required for the public. In [20], the authors call for building a public regulatory ecosystem based on traceable documentation and auditable AI, with a slightly different emphasis than the one on individual transparency and information for all.

Robustness and verification

Given the complexity, more work needs to be done by interdisciplinary teams bringing together the social sciences and humanities expertise with the computer scientists, software engineers, legal scholars, and political scientists in investigating what meaningful control and verification procedures for AI systems might look like in the future.

Safety, risk issues, and ethical decisions

In the domain of autonomous vehicles, looking at state-of-the-art to avoid collisions, autonomous cars have been trained not only to respect traffic rules rigorously but also to 'drive cautiously', i.e., negotiate and not enforce the right of way. Even in case of unavoidable and dilemmatic situations, legislation is underway to respect the ethical dilemma aka Trolley Dilemma, investigated in [7] and [26].

In the context of public expectations, it's important to understand that there isn't a universally "right" answer when it comes to making decisions in dilemma situations. Primarily, an algorithm should not be constrained to making pre-defined decisions. Nevertheless, ongoing discussions about this topic persist in society. Furthermore, the lack of acceptance for autonomous driving can be attributed to the fact that humans are allowed to make mistakes, whereas there seems to be zero tolerance for any mistakes made by AI.

Cybersecurity

In the cybersecurity domain, other than attacks through the internet, there are also AI-specific attacks, such as adversarial learning, which researchers have successfully demonstrated from the Tencent Keen Security Lab [28]. AI systems such as autonomous vehicles must demonstrably be able to defend themselves and go into a safe mode in case of doubt.

Physical Security

There might be physical attacks, like throwing a paint bag against the cameras to blind an autonomous system or using a laser pointer against the LIDAR. In cases like these, the error handling must be capable of bringing the system into a safe mode in case of doubt.

Data Privacy

People have the right to determine if they want to be “filmed” and whether they want their location, date, and time to be recorded and shared. To build trust, autonomous systems manufacturers must adhere to the data protection principles in the GDPR to ensure that no privacy rights are being violated.

Trust and Human Factors

Different levels of trust and comfort may arise through explanation, e.g., if an autonomous car explains its maneuvers to its passengers and road users outside the vehicle.

Trust and Legal System

The decisive question is who or what caused the error: The human at the wheel? A flawed system? A defective sensor? Complete digitization makes it possible to answer these questions. To do this, however, extensive data must be stored.

Open legal questions that need to be clarified in this context are who owns these data, who has access to the data, and whether this is compatible with privacy protection.

Public administration

The answers to the above must be found because they represent significant concerns of the citizens. In our capacity as members of the ACM Europe TPC, we contribute to the work by the EU Commission and EU Parliament to establish harmonised rules for the use of AI. Our comments from the perspective of autonomous systems can be found in [4].

4 AI LEGISLATION IN THE EU: THE AI ACT

Currently, AI policy work is being done globally in most industrial countries. The Future of Life Institute has partnered with PricewaterhouseCoopers. On their website [16] they offer a dashboard with a wealth of information and references to documents. According to their analysis, the approach to govern AI varies greatly between soft and hard law efforts depends largely on how the following areas of concerns are rated and prioritized by policymakers:

- Global Governance and International Cooperation
- Maximizing Beneficial AI Research and Development
- Impact on the Workforce
- Accountability, Transparency, and Explainability
- Surveillance, Privacy, and Civil Liberties
- Fairness, Ethics, and Human Rights
- Manipulation
- Implications for Health
- National Security
- Artificial General Intelligence and Superintelligence.

Looking at the major players, we see:

United States.

> The White House has published a ’Blueprint for an AI Bill of Rights’, a set of five principles and associated practices to help guide the design, use, and deployment of automated systems to protect the rights of the American public in the age of artificial intelligence. However, there is currently no federal AI regulation in the US, but some states have taken steps to regulate particular use cases and the use of AI in specific industries. For example, California has passed a law that requires companies to disclose the use of automated decision-making in employment and housing. Overall, the strategy is business oriented.
>
> After the appearance of ChatGPT, the U.S. Senate Committee on the Judiciary's Subcommittee on Privacy, Technology and the Law held several hearings with leading AI academics to evaluate the risks of generative AI. In October 2023, the Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, signed by President Biden, arose from a desire to address both the potential benefits and risks of AI.

China.

> China has been actively investing in AI and has taken steps to regulate its use, including developing national AI standards and guidelines for ethical use. The country has also established a national AI development plan that sets out its goals and objectives for the industry.
>
> China has significantly restricted the utilization of generative artificial intelligence. ChatGPT is blocked within the Chinese network, and access to domestic alternatives is granted solely through individual application requests.

Canada.

> Canada has established the ‘Pan-Canadian Artificial Intelligence Strategy’, which aims to promote the responsible development and use of AI. The strategy includes funding for research, development, and innovation in AI, as well as ethical guidelines for its use.

United Kingdom.

> The UK has established the ”AI Council,” which aims to promote the responsible use of AI and advise the government on AI regulation. The council has published guidelines on ethical use. The approach so far aims to ensure that consumers ‘have confidence in the proper functioning of the system’.

The G7.

> During their summit meeting on May 20th, 2023, in Hiroshima, the G7 issued a statement about what they called the ‘Hiroshima AI Process’.
>
> “We recognize the need to immediately take stock of the opportunities and challenges of generative AI, which is increasingly prominent across countries and sectors, and encourage international organizations to consider analysis on the impact of policy developments and Global Partnership on AI (GPAI) to conduct practical projects. In this respect, we task relevant ministers to establish the Hiroshima AI process, through a G7 working group, in an inclusive manner and in cooperation with the OECD and GPAI, for discussions on generative AI by the end of this year. These discussions could include topics such as governance, safeguard of intellectual property rights including copy rights, promotion of transparency, response to foreign information manipulation, including disinformation, and responsible utilization of these technologies” [17]. In October 2023, this was followed by the publication of AI guidelines for a ‘Hiroshima Process’ for advanced AI systems and a code of conduct for developer organisations.

In the EU, preparations for AI regulation began in April 2021, when the EU Commission presented the Artificial Intelligence Act, which sets out horizontal rules for the development, commodification, and use of AI-driven products, services, and systems within the territory of the EU. It should be noted that the EU AI legislation does not regulate AI technology per se, but rather the effect of AI products on the lives of EU citizens. There is no intention to intervene in the development of AI products, but there is a claim to help shape their use in the EU. The regulation provides core artificial intelligence rules that apply to all industries.

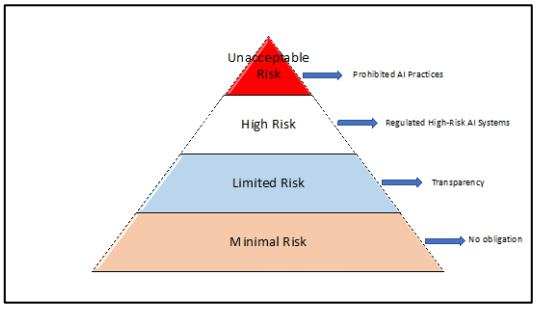

Figure 1The EU Risk Pyramid

The EU AI Act introduces a sophisticated ‘product safety framework’ constructed around four risk categories as evidenced in Figure 1. It imposes requirements for market entrance and certification of high-risk AI Systems through a mandatory CE-marking procedure. To ensure equitable outcomes, this pre-market conformity regime also applies to machine learning training, testing and validation datasets. The Act seeks to codify the high standards of the EU trustworthy AI paradigm, which requires AI to be legally, ethically, and technically robust, while respecting democratic values, and human rights, including privacy and the rule of law.

This is claimed to be the first law worldwide to regulate AI in all areas of life, except the military sector. The legislative process reached a milestone in December 2023, when the the EU Commission, the EU Council and the Parliament managed to reach an agreement in the so-called trilogue. After subsequent approval from votes in the Parliament and the Council, the regulation came into force in June 2024, shifting the attention to Member States to set up supervisory bodies, the standardization bodies to develop harmonized standards for high-risk AI compliance, and for the new AI Office to develop guidelines.

4.1 Who is affected by the new regulation?

Companies that plan to provide or deploy AI systems in the EU (the ”providers and deployers” according to the wording of the Act) are the primary addresses bound by the provisions of the AI Act. They apply regardless of where the systems were developed or are operated from – or when the operation of the systems has an impact on EU citizens. It will take courage and creativity to legislate this convoluted, interdisciplinary issue and will require non-EU, namely U.S. and Chinese companies, to adhere to values-based EU standards before their AI products and services gain access to the European market of 450 million consumers. Consequently, the proposal has an extraterritorial effect.

Given the need for more awareness outside the EU, companies are well advised to start early and learn what is in the EU AI Act and what is needed to meet the compliance criteria.

4.2 The essence of the EU AI Act

The AI act contains the following sections [13], called titles. A collection of all publicly available documents and amendments since the initial proposal to the AI Act as of July 2023 may be found here [15].

Chapter I: General Provisions	Outlines the proposal's scope and how it would affect the market once in place.
Chapter II: Prohibited AI Practices	Defines AI systems that violate fundamental rights and are categorized at an unacceptable level of risk.
Chapter III: High-Risk AI Systems	Covers the specific rules for classifying AI systems as high risk, the connected requirements and obligations for Providers and Deoplyers and other parties.
Chapter IV: Transparency Obligations for Providers and Deployers of Certain AI Systems and GPAI Models:	Lists transparency obligations for systems that interact with humans, detect emotions, or determine social categories based on biometric data, or generate or manipulate content (e.g., ‘deep fakes’).
Chapter V: General Purpose AI Models	Classification Rules, Obligations for Providers of General Purpose AI Models, and GPAI Models with Systemic Risk.
Chapter VI: Measures in Support of Innovation	AI Regulatory Sandboxes, Testing of High Risk AI Systems in Real World Conditions..
Chapter VII: Governance:	Establishing the Act’s governance systems, including the AI Office and the AI Board, and monitoring functions of the European Commission and national authorities.
Chapter VIII: EU Database for High-Risk AI Systems	EU Database for High Risk AI Systems listed in Annex III.
Chapter IX: Post-Market Monitoring, Information Sharing, Market Surveillance	Sharing of Information on Serious Incidents, Supervision, Investigation, Enforcement and Monitoring in Respect of Providers of General Purpose AI Models.
Chapter X: Codes of Conduct and Guidelines	Guidelines from the Commission on the Implementation of this Regulation.
Chapter XI: Delegation of Power and Committee Procedure	Exercise of the Delegation and Committee Procedure.
Chapter XII: Confidentiality and Penalties	Administrative Fines on Union Institutions, Agencies and Bodies. Fines for Providers of General Purpose AI Models.
Chapter XIII: Final Provisions	Amendments to several articles in other legislation

Table 1 Contents of the EU AI Act

4.3 The risk pyramid of the AI Act

The main guiding point of the AI Act is the risk pyramid with a core focus on high-risk applications. The risk levels, as depicted previously in Figure 1, are summarised below.

Unacceptable Risk

This category delineates which uses of AI systems carry an unacceptable level of risk to society and individuals and are thus prohibited under the law. These prohibited use cases include AI systems that entail social scoring, subliminal techniques, biometric identification in public spaces, and exploiting people's vulnerabilities. In these uses, the AI Act describes when and how exceptions may be made, such as in emergencies related to law enforcement and national security.

High-Risk

Requirements related to high-risk systems are at the crux of this proposed regulation, such as compliance with risk mitigation requirements like documentation, data safeguards, transparency, and human oversight. The list of high-risk AI systems that must deploy additional safeguards is lengthy and can be found in Art. 6, Annex III of the Act.

Explainability plays a crucial role in ensuring that AI systems are transparent and trustworthy, particularly in domains where the risk of harmful decisions is high, e.g., in the medical domain, where a false negative may be as harmful as a false positive. The EU AI Act requires that AI systems provide information on their decision-making process so that individuals can understand the basis for the AI system's outputs and that they are not used to manipulate behavior. Additionally, the requirement for human oversight and control over high-risk AI systems is based on the principle that there must be a human in the loop to make decisions that have significant consequences for individuals' rights and safety [25]. The EU AI Act aims to ensure that AI systems are developed and deployed responsibly and transparently, considering the potential impact on individuals' rights and safety.

Limited-Risk

Limited-risk AI systems have much fewer obligations to providers, and users must follow compared to their high-risk counterparts. AI systems of limited risk must follow certain transparency obligations outlined in Title IV of the proposal. Examples of systems that fall into this category include biometric categorization, or 'establishing whether the biometric data of an individual belongs to a group with some predefined characteristic to take a specific action', emotion recognition, and deep fake systems.

Minimal-Risk

The proposal's language describes minimally risky AI systems as all other systems not covered by its safeguards and regulations. There are no requirements for systems in this category. Of course, businesses with multiple kinds of AI systems must ensure compliance with each appropriately.

4.4 Handling of General Purpose AI with or without Systemic Risk

As a result of the increased general capabilities of several new AI models during the spring of 2023, and likely the broad adoption of ChatGPT, there were intense public debates and a delay of the EU Parliament's proposal for the AI Act. The proposal, from June 2023, came to include rules that the earlier proposals did not, on "foundation models" (see definition in Art. 3) and responsibilities linked to providers of generative AI (see, for example [15]). These proved to be part of the most intensely negotiated aspects of the AI Act, which solidified into a set of obligations for all providers of General Purpose AI, GPAI, that also included a second tier with additional obligations for GPAI models that (see Chapter V) are "having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain" (Art. 3(65)).

In brief, all providers of GPAI models must:

- Draw up technical documentation, including training and testing process and evaluation results.
- Draw up information and documentation to supply to downstream providers that intend to integrate the GPAI model into their own AI system in order that the latter understands capabilities and limitations and is enabled to comply.
- Establish a policy to respect the Copyright Directive.
- Publish a suffciently detailed summary about the content used for training the GPAI model.
- Free and open licence GPAI models – whose parameters, including weights, model architecture and model usage are publicly available, allowing for access, usage, modification and distribution of

the model, only have to comply with the latter two obligations above. This exception does not apply to general-purpose AI models with systemic risks.

The GPAI models are considered systemic when the cumulative amount of compute used for its training is greater than 10^25 floating point operations per second (FLOPS)**.** Providers must notify the Commission if their model meets this criterion within 2 weeks. The provider may present arguments that, despite meeting the criteria, their model does not present systemic risks. The Commission may decide on its own, or via a qualified alert from the scientific panel of independent experts, that a model has high impact capabilities, rendering it systemic.

We consider the assumption of that more compute in the training of a model should directly equal risks for negative impact on public health, safety, public security etc. to be quite a leap. The Commission is also mandated to change how "systemic risk" is allocated (in Art. 51(3)), which may be both meaningful in terms of how AI evolves, but also opens for legal unpredictability.

In addition to the five obligations for GPAI above, providers of GPAI models with systemic risk must also:

- Perform model evaluations, including conducting and documenting adversarial testing to identify and mitigate systemic risk.
- Assess and mitigate possible systemic risks, including their sources.
- Track, document and report serious incidents and possible corrective measures to the AI Office and relevant national competent authorities without undue delay.
- Ensure an adequate level of cybersecurity protection.

In response to the complexities of AI regulation, the EU has established an AI Office to facilitate coordination on cross-border cases. However, the resolution of intra-authority disputes remains the responsibility of the Commission.

5 ASSESSMENT AND HOW TO COPE WITH THE EU AI ACT

For anyone wishing to put an AI system into operation in the EU, the AI act serves as a reminder for developers to always prioritize the well-being of individuals and society as a whole. They must first assess the risk, enter the system into a database and, depending on the risk class, comply with requirements relating to transparency and security. It is expected that it will be particularly challenging for high-risk applications to obtain approval for the EU market. There will be a grace period until the law is converted into national law and comes into force. Nevertheless, developers should analyse the respective compliance requirements at an early stage to adapt the development process accordingly. The strategy includes the following key elements:

- Informing and training employees about the regulations and their obligations under the law. These cannot be understood without addressing the EU's rationale for this law and the expectations or EU citizens regarding trustworthy AI. Researchers and Developers must understand that automated and algorithmic decision making should be based on the principles and values enshrined in the Charter of Fundamental Rights, such as human dignity, equality, justice and equity, non-discrimination, informed consent, private and family life and data protections together with principles and values of Union law, such as non-stigmatization, and individual and social responsibility. Support from an interdisciplinary working group should therefore be planned for.
- During the design of the systems, attention should be paid to transparency [6], the nature and quality of the training data and its documentation because of a later evaluation by external reviewers. This also includes the establishment of a risk management system (see Table 2).
- Continuous investment in research and development, especially in rapidly evolving methods of AI explainability, see [8][9]. Once an AI system is explainable, it will build trust and forms a step toward acceptance and approval.
- Collaborate with other companies and organisations in the industry to share information and best practices for compliance. This can help reduce costs and ensure that all parties are on the same page when it comes to compliance.

I. Define Use Case	- Risk Self-Assessment

II. Evaluation of risk level and compliance requirements	- Limited Risk
	  ○ Accessible disclosure of concrete user information
	- High Risk (Annex III and Annex VIII)
	  ○ Risk management
	  ○ Data and Data Governance
	  ○ Human oversight
	  ○ Technical Documentation
	  ○ Transparency and provision of information to users
	  ○ Accuracy, robustness, and cybersecurity.

III. Compliance Assessment	- Internal control (see AI act Annex VI) - External control / Quality management (see AI act Annex VII)

Table 2 Management System

ACKNOWLEDGEMENTS

This work was done while working for the ACM Europe Technology Policy Committee (TPC) on autonomous systems. We are grateful for support and discussions with Chris Hankin, chair of the TPC. Further information may be found on the TPC website [5] and in prior publications [1], [3].

REFERENCES

[1]. ACM (2017) ACM Principles for Algorithmic Transparency and Accountability https://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf Last Access in February 2023.

[2]. ACM (2018) The ACM Code of Ethics and Professional Conduct https://www.acm.org/binaries/content/assets/about/acm-code-of-ethics-booklet.pdf Last Access on February 2023.

[3]. ACM (2018) When Computers Decide: European Recommendations on Machine-Learned Automated Decision Making https://www.acm.org/binaries/content/assets/public-policy/ie-euacm-adm-report-2018.pdf Last Access in February 2023.

[4]. ACM (2021) ACM Europe TPC Comments on Proposed AI Regulations https://europe.acm.org/binaries/content/assets/public-policy/europe-tpc-comments-ai-consultation.pdf Last Access in February 2023.

[5]. ACM Europe Technology Policy Committee, https://www.acm.org/public-policy/europe-tpc, Last Access in February 2023.

[6]. ACM(2022) ACM Technology Policy Council Statement on Principles for Responsible Algorithmic Systems, https://www.acm.org/binaries/content/assets/public-policy/final-joint-ai-statement-update.pdf Last Access in December 2022.

[7]. Awad, E., Dsouza, S., Kim, R. et al. The Moral Machine Experiment. Nature 563, 59–64 (2018). https://doi.org/10.1038/s41586-018-0637-6 Last Access in December 2022.

[8]. Balasubramanian, V., "Towards Explainable Deep Learning" Communications of the ACM Vol. 65 Issue 11 (November 2022) pp 68-69 https://dl.acm.org/doi/pdf/10.1145/3550491 Last Access in February 2023.

[9]. Barredo Arrieta, A., et.al. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI" https://www.sciencedirect.com/science/article/abs/pii/S1566253519308103 Last Access in December 2022

[10]. EU (2019) Building Trust in Human-Centric Artificial Intelligence https://digital-strategy.ec.europa.eu/en/library/communication-building-trust-human-centric-artificial-intelligence Last Access in December 2022

[11]. EU (2019) https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai Last Access in August 2023

[12]. EU (2021) https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment Last Access in August 2023

[13]. https://artificialintelligenceact.eu/the-act/ Last Access in May 2024

[14]. EU (2022) https://www.consilium.europa.eu/en/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights/ Last Access in January 2023

[15]. EU AI Act, Collection of publicly available documents as of May 2024 https://www.kaizenner.eu/post/aiact-part3 Last Access in May 2024

[16]. Future of Life https://futureoflife.org/resource/ai-policy/ Last Access in January 2023

[17]. G7 Meeting Hiroshima May 2023 https://www.g7hiroshima.go.jp/en/documents/ Last Access in August 2023

[18]. IEEE "Ethically Aligned Design", First Edition, Last Access in October 2021

[19]. Jobin, Anna, Marcello Ienca, and Effy Vayena. "The global landscape of AI ethics guidelines." Nature Machine Intelligence 1 (2019): 389-399 Last Access in August 2023

[20]. Knowles, Bran, Richards John: The Sanction of Authority: Promoting Public Trust in AI FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency March 2021 Pages 262–271

[21]. Larsson, S. & Heintz, F. (2020). Transparency in artificial intelligence. Internet Policy Review, 9(2). https://policyreview.info/concepts/transparency-artificial-intelligence Last Access in October 2021

[22]. Larsson, Stefan. "On the governance of artificial intelligence through ethics guidelines." Asian Journal of Law and Society7.3 (2020): 437-451 Last Access in October 2021

[23]. Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. https://zenodo.org/record/3240529 Last Access in October 2021

[24]. Marketplace (2021) Self-Driving Cars Might Never Be Able to Drive Themselves, ACM Opinion. https://cacm.acm.org/opinion/interviews/252899-self-driving-cars-might-never-be-able-to-drive-themselves/fulltext https://www.marketplace.org/shows/marketplace-tech/self-driving-cars-might-never-drive-themselves/ Last Access in January 2023

[25]. Middelton, S., Letouzé, E., et.al. "Trust, Regulation, and Human-in-the-Loop AI within the European Region" Communications of the ACM Vol. 65 Issue 6 (April 2022) pp 64-68 Last Access in January 2023

[26]. Noah J. Goodall: Machine Ethics and Automated Vehicles Pre-print version. Published in G. Meyer and S. Beiker (eds.), Road Vehicle Automation, Springer, 2014, pp. 93-102. Available at https://www.researchgate.net/publication/300567119_Machine_Ethics_and_Automated_Vehicles Last Access in January 2023

[27]. Shneiderman, Ben "Responsible AI: bridging from ethics to practice" Communications of the ACM Vol. 64 Issue 8 (August 2021) pp 32–35 Last Access in January 2023

[28]. Small stickers on the ground trick Tesla autopilot into steering into opposing traffic lane https://boingboing.net/2019/03/31/mote-in-cars-eye.html Last Access in December 2022

[29]. Villani, C. (2018) For a Meaningful Artificial Intelligence https://comite-etica.upc.edu/ca/actualitat/media/missionvillani_report_eng-vf.pdf/view Last Access in May 2024