IEAI

White Paper

On a Risk-Based Assessment Approach to AI Ethics Governance

Christoph Lütge
Ellen Hohma
Auxane Boch
Franziska Poszler
Caitlin Corrigan

June 2022

In April 2021, the European Commission published a proposal for regulation to harmonize rules on Artificial Intelligence, with a goal of promoting both excellence in AI and trustworthy AI (Regulation 2021/0106). Key to this proposal is a "risk-based approach" to AI governance. Under this approach, AI-enabled systems can be considered minimal or no risk, limited risk, high-risk or prohibited (such as real-time biometric identification). The classification in these categories has real implications in terms of the scrutiny, transparency and regulation that products (and their suppliers) will face.

Nevertheless, concrete instructions for how to independently conduct corresponding risk classifications in the first place still have to be developed. Uncertainty in how to accurately assess risk may leave suppliers in doubt about the true level of risk that their AI poses to society. Consequently, suppliers may either over-evaluate their technology's riskiness, hindering effective market roll-out, or under-evaluate their technology's riskiness, leading to the (under-regulated) introduction of high-risk AI into society. The EU AI Act proposes, for example, that the execution of adequate risk assessment is necessary for high-risk systems, but leaves open the question of what such a methodology concretely entails. Therefore, key concepts raised in the EU AI Act deserve clarification and extension, namely:

- What do we mean by risk?
- What are the different dimensions of risk with regard to AI? and
- How can we properly assess risk in order to classify and regulate AI-systems in an effective and efficient way?

In order to turn comprehensive policy into effective practice, developers, regulators and third party certifiers need applied assessment mechanisms to determine ethical and societal risks for specific AI applications. All these stakeholders (governments, monitors, companies and users) have limited resources and capacity to carry out (as well as comprehend) risk assessments or the regulatory oversight that follows when a system is classified as "high-risk". Therefore, in order to promote the most efficient and effective use of resources for all, we need better tools to define what actually needs attention, oversight and effective scrutiny.

With this in mind and as it one of IEAI's main ambitions is to turn ethics into practice (e.g., Lütge, 2017; Lütge et al., 2021; Lütge & Uhl, 2021), our research team has outlined a first step towards operationalizing an applied "risk-based assessment approach" to AI ethics governance. Using this approach, individuals can conceptualize, assess, identify and visualize the many dimensions, the loci and extents of ethical and societal risks posed by a particular AI application. The blueprint we outline in the following sections follows a three step approach that builds off existing practices, including (1) Granular Risks Assessment, (2) Granular Risks Visualization and (3) Holistic Risk Categorization.

Applying Risk-Based Assessment to AI ethics

Risks are defined as a general probability of negative consequences to actions (Cambridge Dictionary, 2022). In the context of AI applications, those risks need to be considered within the scope of a society as possible threats for groups or individuals, whether legal or ethical. Additionally, the perception of such risks depends on the value attached to them, and thus risks are, to a certain extent, culture dependent (Dietz & Schwom, 2017).

Nevertheless, unified steps have been identified to produce a comprehensive risk assessment for AI applications (Australian and New Zealand Standards, 2004; Renn, 2008; Nagböl et al., 2021):

1. Identify and estimate the likelihood of occurrence of a risk,
2. Analyze the nature, intensity and level of impact a risk presents, and
3. Evaluate and prioritize according to the significance of a risk building off standards and terms of reference

According to the European AI Act proposal (Regulation 2021/0106), this assessment needs to include quantitative or qualitative measures prior, during and after the implementation of the technology in society,

and should be done on the AI within a system and on the purpose of the overall system (Ezeani et al., 2022). This assessment can then allow for adapted risk management actions to reduce the impact of an identified issue to an acceptable standard. For this endeavor, we will next review existing risk assessment approaches proposed by policymakers, academics and industry and investigate their common attributes and shortcomings.

Risk assessment approaches by policymakers

Following the recent exponential growth of the use of AI in society, governments and international organizations are developing and proposing more and more risk assessment guidelines and tools to prevent foreseeable harm to populations. Here we outline several of these guidelines.

The AI Act: In April 2021, the European Commission was the first to propose a legal regulatory framework for AI-systems and the risks they might bring (Regulation 2021/0106- The AI Act). The main aim of this proposal is to ensure the respect of fundamental rights for humans and businesses, as stated in the EU charter of Fundamental Rights (Charter 2012/C 326/02), and the proportional limitations to the freedom to conduct business, art and science.

The AI Act proposes a classification in four levels of risks for AI-systems based on their threat to health, safety and fundamental rights, namely (1) unacceptable, (2) high, (3) limited and (4) minimal. Proportionate requirements and obligations for regulation to follow accordingly based on the level of risk. The proposition for harmonized standards proposes tools for risk assessment based on specific domains and intention of use. For instance, AI used in education and vocational training, creditworthiness, and law enforcement are classified as high-risk automatically because of the domains they are employed in. However, it is important to note that this proposition is still under review and being modified at the time of the publication of this report.

The OECD Framework for the Classification of AI systems: The Organization for Economic and Co-operation and Development proposed a first look into such a framework for policy makers in 2022 (OECD, 2022) following a four pronged evaluation of risks: (1) the context for deployment, intended use, and sector, (2) data governance impacting the system outputs, (3) the type of algorithm used and its transparency, autonomy, and privacy abilities and (4) the performance and outputs of the system. This proposition is accompanied by values and principles (OECD, 2021) aligning with the European commission AI High Level Expert Group's 7 key principles for trustworthy AI (AI HLEG, 2019), and the United Nation Sustainable Development Goals (UN SDGs; United Nations, 2015).

The UNESCO Recommendations: Similar principles supporting risk identification for policy makers were proposed by the United Nations Educational, Scientific and Cultural Organisation in November 2021 in their *Recommendations on the Ethics of Artificial Intelligence* (UNESCO, 2021). In addition to the already mentioned frameworks, those recommendations emphasize the risks for the environment and ecosystems and the need for education, awareness and literacy of the general population to reduce involuntary misuse and misconception on AI systems. Therefore, risks are to be evaluated for society as a whole through SDGs, fundamental freedom, human dignity and human rights respect and protection assessment throughout the life-cycle of the system.

National Level Initiatives: On the national level, countries such as Germany, the UK, and the USA have also proposed interesting approaches to risk assessment of AI systems. The Artificial Intelligence Strategy (AIS) group of the German Federal Government proposed in 2020 a strategy for AI built on a risk-based approach in five levels (German Federal Government, 2020). In their proposition, risks lie for example with data privacy and fairness issues, sustainability and education for the general populations. In the UK, the Centre for Data Ethics and Innovation (2020) proposed in their *AI Barometer Report* an analysis of risks to be evaluated in AI systems. The document highlights the need for an assessment of algorithmic bias and discrimination, explainability, and data privacy, while emphasizing the need for state regulatory capacity

and the general risk of losing public trust in institution and AI if not done adequately. Finally, in the U.S., the Department for Homeland Security proposes a risk assessment test for AI systems looking at factors such as "debugability" and "evolvability" of the tool (Public-Private Analytic Exchange Program Partners, 2018). In Parallel, the National Institute of Standards and Technology is working at developing a Risk management Framework for AI (NIST, 2021), with the inclusion in their approach of unintentional and unanticipated outcomes.

Ultimately, AI risk assessment is a widespread approach across countries and public organizations, with strong similar values at its heart, to avoid foreseeable and unintended adverse consequences on society and individuals. Only few approaches propose practical and clear tools and sandboxes to be used to evaluate AI systems at this point, and none is yet recognized as meeting the needs of practitioners.

Risk assessment approaches by academia

Next to legislative drafts from policymakers, scholars have added to this discussion by clarifying the concept of risk in the field of AI ethics in order to propose roadmaps on how to assess, design and develop 'ethical' technology. For example, Wirtz et al. (2022) provide a structured literature review on AI governance approaches. Based on this, they highlight the variety of AI risk types (such as economic, social, ethical and legal) and propose "an *integrative AI governance framework matching AI risks with guidelines*". More practitioner-oriented, Floridi et al. (2022) published a procedure called *capAI* for assessing an AI systems' conformity with the EU AI Act. Their work aims to serve companies as a governance tool to assess technologies in terms of legal compliance, ethical soundness and technical robustness. These frameworks and assessments highlight the importance of not only paying attention to hard impacts, i.e., 'quantifiable risks' but ever more so to consider soft impacts, i.e., ethical implications of those technologies (Kiran et al., 2015). However, scholars have already illustrated that quantification can also be achieved for 'soft impacts' and provided corresponding baselines for measurement. For example, Wernaart (2021) adopts six dimensions for determining the moral intensity of a situation (that follows from the introduction / adoption of a technology). These dimensions are: (1) *magnitude of consequence* (i.e., the sum of harm and benefits), (2) *social consensus* (around a particular action / issue), (3) *probability of effect* (i.e., probability that the effect will indeed take place), (4) *temporal immediacy* (i.e., the timespan between action and consequence), (5) *proximity* (i.e., sense of nearness towards the affected stakeholders) and (6) *concentration of effect* (i.e., focused on a single individual / group of individuals vs. spread over society). According to Wernaart (2021), with the help of these dimensions, it can be determined whether the situation is 'low stake' or 'high stake', which then has practical implications for the programming of a technology (e.g., who the moral authority is allowed to be).

Overall, it can be stated that "previous governance approaches lack a link between a risk analysis and the resulting guidelines" (Wirtz et al., 2022). Scholars even remark that AI ethics researchers have previously placed too much focus on the 'what' instead of the 'how' (Morley et al., 2021). Nevertheless, academics do provide some valuable insights, concepts and measures that can guide the development of a more practical AI risk assessment approach that is in sync with ethical guidelines / requirements.

Risk assessment approaches by the industry

There is a need to translate the existing conceptual frameworks to practical methodologies and tools to properly address the determined risks of AI systems. Similarly, both research and industry frequently call for a high level of interdisciplinarity in tackling AI risks (Bartneck et al., 2021; Kriebitz et al., 2022). This calls for the additional involvement of industry to provide practical solutions. In fact, technical tools to address risks linked to AI ethics have been proposed by many AI developers. Most of them target specific AI ethics principles, among them often fairness and transparency (Ayling & Chapman, 2021). To tackle fairness problems, toolkits, such as IBM's *AI Fairness 360 suite*, have been developed aiming at exhibiting or even removing biases in the used datasets or AI models. Techniques leveraging transparency highly

complement fairness improving methods, as sources of bias can be more easily spotted. Further transparency enhancing systems often equipped with XAI techniques can help reveal hidden risks within the system. Also, non-technical methodologies targeting the assessment of specific AI ethics principles have been proposed. Especially participatory design processes raising awareness for ethical issues early in the development have been suggested – for example in the form of workshops – to increase diversity and include various stakeholder interests, such as those of civil society (Ayling & Chapman, 2021). Furthermore, methodologies that aim at providing a general risk check (e.g., relying on the 7 key principles for trustworthy AI) or categorization have been developed mostly in a non-technical format. Systematic risk management processes targeting specific shortcomings, such as stakeholder engagement (e.g., Clarke, 2019) or adaptation to organizational viewpoints (e.g., Felländer et al., 2021) can complement this approach.

Examples from practice indicate that many of these initiatives already have had an effect. Many companies (e.g., BMW[1], Novartis[2]) recognize the risks linked to AI and have created codes of conducts for responsible use. For example, Ezeani et al. (2021) gathered examples for how industry stakeholders can begin to discuss and adapt AI risk assessment approaches. The Federation of European Risk Management Associations (FERMA), a consortium of 21 risk management associations, for instance, presents an "AI Risk Management Roadmap" investigating risks along multiple dimensions (Ezeani et al., 2021). Further, several partnerships and initiatives, such as Partnership on AI or The Software Alliance (BSA), have been founded to jointly discuss AI risks and solutions to them among practitioners (Ezeani et al., 2021). However, those examples also show that a standardized and uniform risk assessment procedure has not yet been reached, which hinders broader adoption of fitting governance.

Shortcomings of existing risk assessment approaches

Although significant research has been dedicated to solving the issue of assessing AI risks, and a variety of tools have already been developed for this purpose, there are still shortcomings of what has been proposed so far. A general challenge to creating comprehensive AI risk assessment approaches is linked to the subjective and contextual nature of risk, as well as its constant evolution and change (Corvellec, 2010). This results in diverse methodologies that are specifically designed to address particular types of risks or certain industries (Corvellec, 2010).

While on the one hand it would be handy to have a standardized and concrete procedure, on the other hand, it is questionable whether it would actually be practicable or even feasible to do so (Ezeani et al., 2021). This trade-off between generalizability and practicability and the resulting shortage of a unified framework creates another challenge for developing tools for AI risk assessment which is the current lack of regulatory obligation and requirements for any utilization of impact assessment methodologies (Ayling & Chapman, 2021). Certain official standards for risk assessment or management exist (e.g., ISO 31000, IEEE 7010-2020), however, none of them are mandatory. The yet to be finalized and ratified EU AI Act will surely provide some guidance in this regard. Still, critiques have already been voiced as to whether the Act can satisfy demands for clarity, concretization and extensive applicability (e.g., Ebers, 2020; Hacker & Passoth, 202).

Risk assessment/management methodologies and tools can help solve the challenge of concrete applicability. However, major shortcomings have been identified with many of them. First, some methodologies fail to communicate identified risks to the responsible actors. Often, outputs of risk assessment methods are used for management purposes, instead of feeding back to developers to improve

[1]https://www.bmwgroup.com/content/dam/grpw/websites/bmwgroup_com/downloads/ENG_PR_CodeOf EthicsForAI_Short.pdf

[2]https://www.novartis.com/about/strategy/data-and-digital/artificial-intelligence/our-commitment-ethical-and-responsible-use-ai

systems and practices (Ayling & Chapman, 2021). Most tools are designed for internal self-assessment (Ayling & Chapman, 2021). Therefore, enactment upon identified system shortcomings and communication to, or supervision by, dedicated authorities is not ensured. Second, not all 'voices' are equally represented and heard during risk assessment and management. Especially user or society opinions are often underestimated and left out of the assessment process (Ayling & Chapman, 2021). Most AI risk assessment tools target the product development phase and, thus, focus on development, delivery and quality assurance roles (Ayling & Chapman, 2021). This may cause problems, in particular for societal challenges such as AI ethics, that can only reflect realistic societal demands if debated broadly. Finally, the current design of many risk assessment methods, and especially the technical approaches to them, are criticized for not making use of all capabilities available in an organization. A tension has been identified between suggested template-based analyses and expert intuition (Nagbøl et al., 2021). Further, the full potential of many risk management techniques, such as participation process, base-line study, life-cycle assessment, change measurement or expert committees, has not been entirely harnessed yet (Ayling & Chapman, 2021).[3]

Given the above mentioned shortcomings of current approaches, we have identified several major limitations of current risk assessment methodologies for AI applications. This includes:

- The lack of standardized and uniform risk assessment procedure for all AI systems.
- The lack of practical and clear tools and sandboxes to be used by all practitioners.
- The lack of obligation for AI providers to produce (ongoing) risk assessment for AI systems throughout their lifecycle.
- Underdeveloped inclusion mechanisms for users and society in risk assessment tools that mainly target the product development phase.
- Unclear assessment feedback mechanisms that often are not connected to the responsible practitioner.
- Undervaluing experts' intuition in assessment tools.

With the goal to overcome the prevailing challenges to creating effective tools mentioned above, in the following section we provide a more unified and practical approach to AI ethics risks governance by suggesting the underlying methodology for an AI ethics risk assessment tool.

[3] Since the focus of this white paper is risk assessment, we refrain from elaborating risk management techniques here. Further details about specific risk management techniques can be found, for example, in Ayling and Chapman (2021) or see Corrigan (2022) for an overview of the use of multiple and co-governance mechanisms for AI governance.

Creating a unified and practical approach to AI ethics risks governance

Our proposed approach builds on, integrates and enhances past efforts and insights generated by governments, industry and academics with regard to risk assessment. Specifically, from governmental' efforts, we adopt the proposition of classifying AI risks into levels (such as high, medium and low risk), using key requirements for trustworthy AI (e.g., human agency and oversight), as well as drawing on developed checklist drafts that can guide assessments. From the industry efforts, we prioritize the need to develop concrete procedures and practical tools that can be deployed in companies for risk assessment. From academia, we apply a selection of the moral intensity dimensions (e.g., magnitude of consequences) in our framework. These dimensions not only provide a more nuanced measure of the extent of ethical implications, but also coincides well with the conception of risk which, by definition, considers the magnitude of consequences, as well as its probability.

Origin of approaches	Aspects we draw upon
Policymakers	• AI risk classification • Key trustworthy AI requirements checklists
Industry	• Practical toolkits
Academia	• Moral intensity dimensions

Table 1. Comparison of reused approaches from the different fields.

Key Features

Our risk-based approach establishes easy and practical steps to be followed by users in the endeavor to inform them about risk potentials and consequently, needed interventions in regard to any of their particular technological applications. To do so, the following features are integrated in our approach:

- **Granular risks self-assessment & quantification:** In the form of a questionnaire, users are guided through the process of risk assessment. The risk assessment is organized according to particular risks (e.g., human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination; societal and environmental wellbeing; accountability). For each of these risks, the risk intensity (composed of its prevalence, magnitude and probability) as well as reaction probability (composed of its proximity, corresponding social discourse and temporal immediacy) are assessed. Based on the risk self-assessment, our approach offers a concrete risk quantification in the form of a Reaction Urgency Index.
- **Granular risks visualization:** The information retrieved from the Reaction Urgency Index is summarized and pictured in a risk Reaction Urgency Matrix. This visualization allows users to review the results of the risk-assessment for each individual risk of a particular technological application.
- **Holistic risk categorization:** Furthermore, the value of the two dimensions of the Reaction Urgency Index (i.e., risk intensity and reaction demand) determine the categorization of the overall risk of the technology into low, medium or high stakes. From this, in one glance, a general feel / knowledge of the technology's overall riskiness can be obtained.

Figure 1 outlines how our proposed approach integrates the key features described above. The concrete steps to be taken by the user of this tool and the functionality of key features will be elaborated in more detail in the following section.

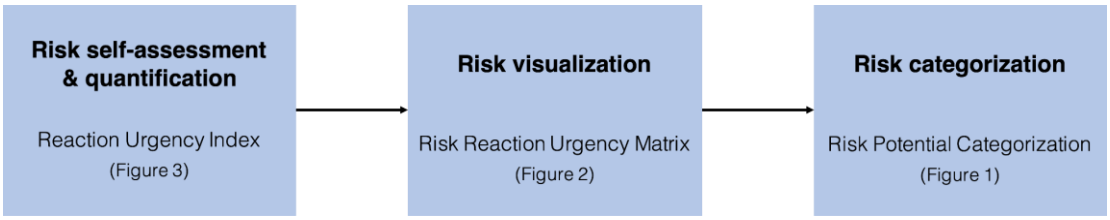

Figure 1: Proposed risk assessment process.

Process Description

Our approach addresses and unifies the identified key features through a *Risk Reaction Urgency Matrix*, based on a *Reaction Urgency Index*. Incorporating risk intensity along with reaction demand, the Reaction Urgency Index displays how severe a certain risk and its impacts are, as well as indicates the necessity for an AI provider to initiate countermeasures. Based on this information, the Risk Urgency Matrix (depicted in Figure 2) can be derived. Figure 2 outlines the matrix that is built off the seven key requirements for trustworthy AI and their sub-categories (AI HLEG, 2019) and structures identified risks according to these principles. Scores are calculated per sub-category for each key principle to break down the quantification, giving more detailed insights and increasing the assessments practicability. Results are presented with a heatmap to help direct the AI provider's reaction per risk subcategory (darker shades of blue reflect higher risk reaction urgency scores) and illustrate where risks are stronger and increasing mitigating reaction capacities is required.

Risks linked to...						
Human agency and oversight	**Technical robustness and safety**	**Privacy and data governance**	**Transparency**	**Diversity, non-discrimination and fairness**	**Societal and environmental well-being**	**Accountability**
Fundamental Rights	Resilience to attack and security	Privacy and data protection	Traceability	Avoidance of unfair bias	Sustainable and environmentally friendly AI	Auditability
Human agency	Fallback plan and general safety	Quality and integrity of data	Explainability	Accessibility and universal design	Social impact	Minimization and reporting of negative impacts
Human oversight	Accuracy Reliability and Reproducibility	Access to data	Communication	Stakeholder Participation	Society and Democracy	Trade-offs Redress

Figure 2: Risk Reaction Urgency Matrix. Darker shades of blue reflect higher risk reaction urgency scores.

Risk Reaction Urgency: To obtain the risk reaction urgency scores that determine the shade of blue used in the heatmap-matrix, we propose a 3-step process including (0) a *preliminary qualitative overview*, assessing detailed facets of the particular risk categories, (1) the calculation for *Risk Intensity*, determining how strong the risks are and (2) the *Reaction Demand* assessment, defining how important a reaction upon determined risks will be.

Previous literature has already extensively focused on developing checklists for qualitative AI risk assessment. We therefore propose to reuse existing approaches for step 0, the preliminary qualitative overview, focusing on an exploratory evaluation of the AI technology to be investigated. The Assessment List for Trustworthy Artificial Intelligence (ALTAI) can serve as a first point of reference and complements our Reaction Urgency Matrix, as both build off the seven key principles for trustworthy AI. However, other risk assessment checklists than the ALTAI can also be considered as baseline.

For calculating the suggested Risk Reaction Urgency Index, we reuse the moral intensity score by Wernaart (2021) and adapt it to the context of AI risk determination. Figure 3 outlines the proposed calculation process including steps 1 and 2 of our proposed process, the Risk Intensity estimation and the Reaction Probability evaluation.

Three concepts are considered to assess Risk Intensity per defined risk (sub-)category. First, a risk's *prevalence*, determining the ubiquity of consequences resulting from the defined risk categories. Second, a risk's *magnitude*, assessing how strong consequences affect those that are impacted. And third, a risk's *probability*, estimating the likelihood of negative consequences resulting from defined risks.

An AI provider's reaction demand constitutes the necessity to take action and is measured using three concepts. *Proximity* indicates the strength with which the AI provider perceives negative consequences resulting from risks of certain (sub-)categories. *Social discourse* around consequences and related mitigation strategies reflects how important a reaction to a certain risk is for society. Finally, through *temporal immediacy*, the timeline and potential urgency of a risk reaction is determined.

The concrete items to measure our suggested concepts can be presented in the form of a practical checklist and all items can be rated on a scale, e.g., from low to high. Like this, an evaluation can be easily determined by summing all concept scores for risk intensity and reaction demand individually and jointly to obtain the final Reaction Urgency Index per (sub-)category, which will later inform the shading in the Risk Reaction Urgency Matrix.

A Use Case: To outline an example assessment, according to an expert workshop (IEAI, 2022), for technological applications in the mobility sector, such as autonomous vehicles, the assessment of particular risks could transpire as follows: the risk intensity for the principle "technical robustness and safety" would be rated high due to the complexity and wide variety of system properties that could open up technical issues leading to road accidents, potentially involving great magnitudes of injuries. Similarly, the reaction probability would be rated high, since the corresponding social backlash (i.e., high rating for social discourse) can be expected to be high as well. Therefore, the Reaction Urgency Index and subsequently the categorization for the principle "technical robustness and safety" would be considered high (i.e. shaded dark blue in the matrix).

		Risks linked to...								
		Human agency and oversight			Technical robustness and safety	Privacy and data governance	Transparency	Diversity, non-discrimination and fairness	Societal and environmental well-being	Accountability
		Fundamental rights	Human agency	Human oversight	...					
Risk Intensity	**Prevalence**	How many people would be affected by risks related to this (sub-)principle?								
	Magnitude	How strong would those that are impacted be affected by risks related to this (sub-)principle?								**Σ**
	Probability	How likely are (without further countermeasures) negative consequences of risks related to this (sub-)principle?								
Reaction Demand	**Proximity**	How strong are you impacted by risks related to this (sub-)principle?								
	Social Discourse	How impactful is the societal reaction answering risks related to this (sub-)principle?								**Σ**
	Temporal Immediacy	How urgent is the reaction to risks related to this (sub-)principle?								
	Σ	*Reaction Urgency Index*								

Figure 3: Reaction Urgency Index calculation.

Risk Reaction Urgency Matrix: The resulting Risk Reaction Urgency Matrix can help organizations assess in detail where risks of AI technologies lie and where mitigating reactions can and should be determined. Both use purposes, internal self-assessment or external justification, are conceivable. Further, a Risk Reaction Urgency Index can help classify an AI system's overall categorisation of risk potential and, for instance, needed actions if certain thresholds are exceeded in total or per subcategory. A proposition of risk potential categorization can be found in Figure 4. Depending on the derived risk stakes, mandatory or non-mandatory response measures may be established. For low risk stakes, for example, the AI provider could be responsible for appropriate regular reassessment and reporting, while high risk stakes could require monitoring by authorities or joint definition of countermeasures. For medium risk stakes, a participatory design process or regular monitoring through authorities may be envisaged.

	Low risk intensity	**High risk intensity**
Low reaction demand	Low stake	Medium stake
High reaction demand	Medium stake	High stake

Figure 4: Risk potential categorization of the overall AI application based on our proposed index calculation.

Discussion

The approach proposed here builds on existing work but supports a more holistic and risk-oriented approach. By re-using recognised and appreciated work proposed by the High Level Expert Group (AI HLEG, 2019), and aligning with the risk based approach of the AI Act (Regulation 2021/0106), we hope to clarify a path for risk assessment of low, medium and high risk AI systems alike. As regulation calls for risk management, our tool participates in paving the way towards the mitigation of dangers arising from the use of any AI-system.

A priority for the future of applied AI ethics, and thus risk assessment and management, is transparency. Through a practical and understandable tool, our approach bridges the gap between outside stakeholders and producers / practitioners to better understand risks related to AI.

Through the creation of a Reaction Urgency Index, we want to support companies in pointing out specific risks arising from their product, while at the same time supporting creators and practitioners in prioritizing their work in a realistic timeline to keep the quality of the product up and its risks low. It is however important to note that our tool does not aim to standardize risk assessment to the point of losing sight of expert opinion. The subjectivity of the two first steps of our approach keep at the center of the process the need for expertise and specific knowledge, while allowing for temporization with numbers and figures from the market.

As an overview, Table 1 highlights why our solution addresses the challenges identified in the previous section.

Challenges	**Our Approach**
• A standardized and uniform risk assessment procedure for all AI-systems has not yet been reached.	Our approach is **sector and system holistic**. All AI systems can be considered through its scope, regardless of its high or low complexity as it builds off **moral and practical graspable concepts**.
• Current approaches do not offer practical and clear tools and sandboxes to be used by all practitioners. • There is a lack of obligation to produce (ongoing) risk assessment for AI tools throughout their lifecycle. • The risk assessment tools mainly target the product development phase which can lead to a lack of consideration of risks for users and society.	Each item of our assessment tool is **clearly identifiable and understandable** by AI practitioners, while being **accessible** for any non-technical person. Our tool allows for **active and reactive self-assessment** throughout the life-cycle of the AI, pinpointing possible risk on the long run arising from the use of the system, with a possible **identification of the most underserved groups**.
• The assessment feedback is sometimes not given to the responsible practitioner. • Experts' intuition can sometimes be removed from the equation due to strict tools, entailing the loss of expert knowledge and intuition.	Experts practitioners will be using this tool on their own, allowing for **subjective expertise assessment**, and **justifiable scoring** in general, while pinpointing specific areas creating specific risks and thus **facilitating identification of responsible actors** to inform.

Table 2. Comparison of identified challenges and offered solutions through our approach.

Conclusion

A current challenge of assessment tools for AI applications at this stage is that they are not fully agreed upon by different actors of the AI ecosystem. Governmental and international organizations' principles are too broad, while industry's propositions are often only tailored to specific sectors or systems. Additionally, each tool usually does not allow for discussions among different disciplines and stakeholders, as the tool might be too technical, or too broad and thus not applicable. The blueprint solution we offer follows a three step approach building off existing practices e.g. (1) Granular Risks Assessment, (2) Granular Risks Visualization and (3) Holistic Risk Categorization. The prototype proposed also refers to international organizations principles, and is transparent regardless of the level of expertise of the reader, while keeping its technical accuracy and expertise flexibility. Thus, the Risk Urgency Index is a practical, applicable, understandable and monitorable tool. Companies could use it throughout the lifecycle of the AI to ensure compliance with ethical principles, and react at adequate time if need be.

Our approach builds off existing work.Therefore, many of the components have already been studied, reviewed and validated, or even used in practice. Nevertheless, our tool is not yet to be understood as a ready-to-use concept, but rather a suggested first step or blueprint for the underlying methodology of a risk assessment tool that still needs to be enriched with missing information in the future.

Specifically, in future studies, we propose to develop concrete and justified items to measure the proposed concepts for risk intensity and reaction demand as well as validating the resulting questionnaires in field experiments. Further, thresholds and weights of the two dimensions risk intensity and reaction demand need to be determined. Questions about whether one dimension can offset another or whether some concepts used for measuring the two dimensions should receive higher weights than others still need to be further clarified. In particular, for a derivation of an overall risk potential of a certain technology, concrete distinctions on when an AI system should be classified as high, medium or low risk need to be determined. The resulting imposed actions from the risk classification and whether or not those actions should be obligatory can then be aligned. Finally, a future step for our proposed approach is to test and validate it in practice in order to demonstrate a proof of concept and further adapt the process to organizational demands.

References

Australian/New Zealand Standards (2004), Risk Management: AS/NZS 4360, Standards Australia International and Standards, New Zealand.

Ayling, J., & Chapman, A. (2021). Putting AI ethics to work: are the tools fit for purpose? *AI and Ethics*, *1*-25.

Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021*). An introduction to ethics in robotics and AI* (p.117). Springer Nature.

Charter 2012/C 326/02. Charter of Fundamental Rights of the European Union. European Union. https://www.refworld.org/docid/3ae6b3b70.html

Centre for Data Ethics and Innovation. (2020). *AI Barometer 2020*. GOV.UK. https://www.gov.uk/government/publications/cdei-ai-barometer

Clarke, R. (2019). Principles and business processes for responsible AI. *Computer Law & Security Review*, *35*(4), 410-422.

Corrigan, C. (2022). Lessons learned from co-governance approaches – Developing effective AI policy in Europe. *Digital Ethics Lab Yearbook 2021*, Springer. (Forthcoming)

Corvellec, H. (2010). Organizational risk as it derives from what managers value: A practice-based approach to risk assessment. *Journal of Contingencies and Crisis Management*, 18(3), 145-154.

Ebers, M. (2020). Regulating Explainable AI in the European Union. An Overview of the Current Legal Framework(s). Liane Colonna/Stanley Greenstein (eds.), *Nordic Yearbook of Law and Informatics 2020: Law in the Era of Artificial Intelligence*.

Ezeani, G., Koene, A., Kumar, R., Santiago, N., & Wright, D. (2021). A survey of artificial intelligence risk assessment methodologies: The global state of play and leading practices identified. *Ernst & Young LLP*.

Felländer, A., Rebane, J., Larsson, S., Wiggberg, M., & Heintz, F. (2021). Achieving a Data-driven Risk Assessment Methodology for Ethical AI. *arXiv preprint arXiv:2112.01282*.

Floridi, L., Holweg, M., Taddeo, M., Amaya Silva, J., Mökander, J., & Wen, Y. (2022). capAI-A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act. *Available at SSRN 4064091*.

German Federal Government. (2020). *Artificial Intelligence Strategy of the German Federal Government.* https://www.ki-strategie-deutschland.de/files/downloads/Fortschreibung_KI-Strategie_engl.pdf

Hacker, P. & Passoth, J., (2022). Varieties of AI Explanations under the Law. From the GDPR to the AIA, and Beyond. In: Holzinger, Goebel, Fong, Moon, Müller and Samek (eds.), *Lecture Note on Artificial Intelligence 13200: xxAI - beyond explainable AI*, Springer. http://dx.doi.org/10.2139/ssrn.3911324

High-Level Expert Group on Artificial Intelligence (AI HLEG). (2019). *Ethics Guidelines for Trustworthy AI*. European Commission. https://ec.europa.eu/futurium/en/ai-alliance-Consultation.1.html

IEEE (2020). *IEEE 7010-2020 IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being*.

ISO (2018). ISO 31000 Risk Management.

Kiran, A. H., Oudshoorn, N., & Verbeek, P. P. (2015). Beyond checklists: toward an ethical-constructive technology assessment. *Journal of responsible innovation*, *2*(1), 5-19.

Kriebitz, A., Max, R., & Lütge, C. (2022). The German Act on Autonomous Driving: why ethics still matters. *Philosophy & Technology*, 35(2), 1-13.

Lütge, C. (2017). The German ethics code for automated and connected driving. Philosophy & Technology, 30(4), 547-558.

Lütge, C., Poszler, F., Acosta, A. J., Danks, D., Gottehrer, G., Mihet-Popa, L., & Naseer, A. (2021). AI4People: Ethical Guidelines for the Automotive Sector–Fundamental Requirements and Practical Recommendations. *International Journal of Technoethics (IJT)*, *12*(1), 101-125.

Lütge, C., & Uhl, M. (2021). *Business Ethics: An Economically Informed Perspective*. Oxford University Press, USA.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2021). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Ethics, Governance, and Policies in Artificial Intelligence*, 153-183.

Nagbøl, P. R., Müller, O., & Krancher, O. (2021). Designing a Risk Assessment Tool for Artificial Intelligence Systems. In *International Conference on Design Science Research in Information Systems and Technology* (pp. 328-339). Springer, Cham. https://pure.itu.dk/portal/files/86457234/_PREPRINT_Designing_a_Risk_Assessment_Tool_for_Artificial.pdf

National Institute of Standards and Technology (NIST). (2021). Artificial Intelligence Risk Management Framework. Federal Register, *The Daily Journal of the United States Government*. https://www.federalregister.gov/documents/2021/07/29/2021-16176/artificial-intelligence-risk-management-framework#citation-1-p40811

OECD. (2021). Tools for trustworthy AI : A framework to compare implementation tools for trustworthy AI systems. *Documents de travail de l'OCDE sur l'économie numérique, n° 312*, Éditions OCDE, Paris, https://doi.org/10.1787/008232ec-en

OECD. (2022). OECD Framework for the Classification of AI Systems. *OECD Digital Economy Paper*. https://doi.org/10.1787/cb6d9eca-en

Public-Private Analytics Exchange Program. (2018). *Artificial Intelligence: Using Standards to Mitigate Risks*. https://www.dhs.gov/sites/default/files/publications/ 2018_AEP_Artificial_Intelligence.pdf

Regulation 2021/0106. *Regulation of the European Parliament and of the council laying down harmonized rules on Artificial Intelligence*. European Parliament, Council of the European Union. https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=CELEX%3A52021PC0206

Renn, O. (2008), *Risk Governance: Coping with Uncertainty in a Complex World*, Earthscan, London.

United Nations. (2015). *THE 17 GOALS | Sustainable Development*. Department of Economic and Social Affairs Sustainable Development. https://sdgs.un.org/goals

UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*. https://unesdoc.unesco.org/ark:/48223/pf0000381137

Wernaart, B. (2021). Developing a roadmap for the moral programming of smart technology. *Technology in Society*, 64, 101466.

Wirtz, B. W., Weyerer, J. C., & Kehl, I. (2022). Governance of artificial intelligence: A risk and guideline-based integrative framework. *Government Information Quarterly*, 101685.