

A survey of artificial intelligence risk assessment methodologies

The global state of play and leading practices identified



Building a better working world

TRILATERAL RESEARCH



Contents

0	Executive summary	Page 5
1	Introduction	Page 8
2	Legal and regulatory approaches	Page 10
	2.1 Risk-based approaches to AI regulation	
	2.1.0 Proposed EU AI Regulation	
	2.1.1 Recommendation of German Data Ethics Committee	
	2.2 Regulatory policies including elements of AI risk assessment	
	2.2.1 European Commission	
	2.2.2 United States	
	2.2.3 Canada	
	2.2.4 China	
	2.2.5 Singapore	
	2.2.6 India	
	2.2.7 National AI strategies	
3	International organizations	Page 19
	3.1 Council of Europe (CoE)	
	3.2 Global Partnership on AI (GPAI)	
	3.3 Organization for Economic Co-operation and Development (OECD)	
	3.4 United Nations Educational, Scientific and Cultural Organization (UNESCO)	
	3.5 World Economic Forum (WEF)	

4

Standards

Page 23

4.1 ISO/IEC

4.2 IEEE SA

4.3 NIST

4.4 CEN/CENELEC

5

Industry approaches

Page 27

5.1 Wharton School on AI in financial services

5.2 Federation of European Risk Management Associations (FERMA)

5.3 Open Loop

5.4 Partnership on AI

5.5 BSA | The Software Alliance

5.6 Ethics & Compliance Initiative (ECI)

5.7 Healthcare Insurance Reciprocal of Canada (HIROC)

5.8 EY and The Future Society

6

Approaches proposed in civil-society and academic literature

Page 32

6.1 Overview

6.2 Key aspects

7

Conclusions

Page 36

7.1 Key findings

7.2 Leading practices

7.3 Abbreviations used in this survey

8

About this report

Page 46

Foreword

Artificial Intelligence (AI) systems are increasingly being deployed in a wide range of applications. For many of these, a failure of the system would be embarrassing and inconvenient, but otherwise harmless. Failures in other applications, such as credit assessments, recruitment, medical diagnostics, (semi-)autonomous vehicles and power grid load distribution systems, however, would have significant impacts on the lives of individuals, organizations or large parts of society.

Building on the work of high-level expert groups, and studies by academics, industry associations, professional bodies and civil-society, a growing consensus is emerging in favour of risk-based approaches to regulating the use of AI. By adopting a proportionate approach where the complexity of regulatory compliance depends on the risk that the AI system poses, policy makers can fulfil their duty to safeguard without unduly impeding the benefits that AI can bring to society. To successfully use a risk-based approach to AI regulation, however, it is first necessary to have a consistent methodology for assessing the specific risks associated with different AI systems.

This report was commissioned to (1) inform policy makers and regulatory stakeholders about noteworthy approaches to AI risk assessment, including leading practices, and (2) to inform rulemaking on AI risk assessment. The survey covers: legal and regulatory approaches, current work at international bodies, work by standards development organizations, industry approaches and prominent approaches proposed in civil-society and academic literature.


The leading practices identified within this report are categorized within the following areas: risk, risk management, requirements for trustworthiness, and the relevant stakeholders who should be involved for identifying and mitigating AI risk. The survey does not claim to be comprehensive but provides a snapshot of the AI risk assessment landscape in 2021.

It is our hope that this report will prove a valuable resource for informed policy making, and the continuing dialogue on AI regulation.

Dr. Ansgar Koene,

Global AI Ethics and Regulatory Leader





Executive summary

0

Scope of this report

Policymakers, academics, standards bodies, industry, researchers, civil society organizations and many other stakeholders have called for governance of AI systems to include AI risk assessment (AIRA) and have proposed different approaches and methodologies for such assessment frameworks.

These assessment frameworks aim at identifying and assessing risks of AI with the view to mitigating them. This report surveys and assesses the ecosystem of AIRA methodologies. It does not claim to be comprehensive but provides a snapshot of the landscape at a certain point in time. Ultimately, it aims to inform policymakers about the AIRA assessment landscape, including emerging policy trends and leading practices.

Types of risk assessment

There is a diversity of approaches from different actors in the AI ecosystem: policymakers and regulators, international organizations, standards bodies, industry and researchers from academia and associations. Each actor targets different stakeholders, and their AI risk assessment initiatives are developed with different interests and objectives in mind.

At a high level, we distinguish between: (1) **assessment of risks arising from the use of AI:** these might include bias, lack of transparency, discrimination, invasion of privacy, misuse of personal data and damaging trust and (2) **classification of AI systems or applications by risk:** the assessor looks at the risks arising from the use of AI in order to classify the system in a category of risk (e.g., high risk or low risk). This second type of AIRA is specifically relevant in the context of law when the level of AI risk determines the applicable legal obligations.

Related concepts: impact assessment and risk management

AI risk assessment (AIRA), impact assessment (AIIA) and risk management (AIRM) are closely related concepts in the context of AI governance. While risk assessment is about identification, analysis and evaluation of threats and vulnerabilities, impact assessment goes further by considering implications, both positive and negative, for people and their environment. AIRA is also part of AI risk management, which is a broader process. As defined by ISO 31000, risk management is the identification, assessment and prioritization of risks, and the subsequent coordinated and economical application of resources to minimize, monitor and control the likelihood of unintended events.

Key findings and policy trends for the governance of AI governance of AI.

As of August 2021, there is no universally accepted model or common regulatory framework for AI risk assessment, as governments are just beginning to develop policies and laws specific to AI. Nevertheless, we have identified the following trends:

AI principles form benchmarks for AIRA.

- ▶ Many countries and organizations have defined AI principles that largely align with the values-based principles published by the OECD¹ in May 2019, and subsequently supported by the G20 at their 2019 summit.
- ▶ These principles can directly inform AIRA by identifying the risk factors that need to be assessed.
- ▶ There is a widespread trend towards considering risks arising from violation of AI principles (set, for example, at national, European, international levels).
- ▶ An AIRA should be an ongoing process throughout the design, development and deployment of AI technologies.
- ▶ There is an opportunity for high levels of international regulatory alignments on risk assessment since there is broad agreement on AI principles.



AI standards are under development.

The technical community is making good progress on developing standards and guidance for technical implementation of AI risk assessment. Of particular relevance to AIRA are the following draft standards:

- ▶ ISO/IEC 23894 on Artificial Intelligence and Risk Management
- ▶ ISO/IEC 42001 on Artificial Intelligence – Management System
- ▶ ISO/IEC 38507 on Governance implications of the use of artificial intelligence by organizations
- ▶ IEEE P2863 on Recommended Practice for Organizational Governance of Artificial Intelligence
- ▶ IEEE 7000-2021 on Model Process for Addressing Ethical Concerns During System Design
- ▶ IEEE 7010-2020 on Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being Draft NISTIR 8332 on Trust and Artificial Intelligence
- ▶ NIST Special Publication 1270 on A proposal for Identifying and Managing Bias in Artificial Intelligence

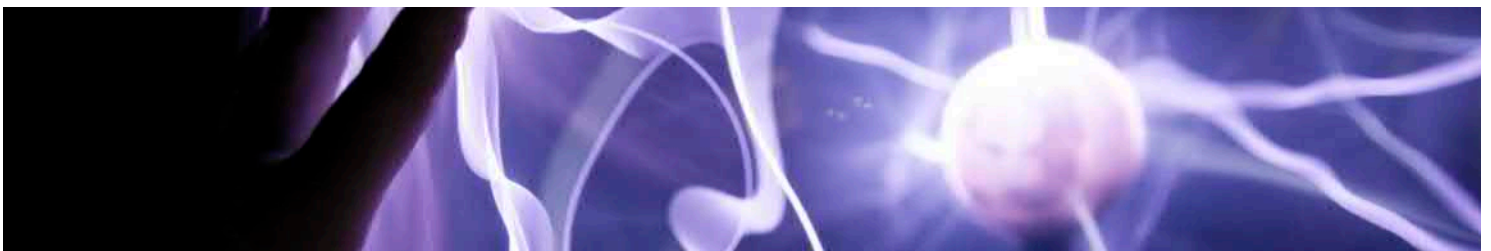
Industry is forming mixed membership associations to promulgate policy guidance.

- ▶ Industry has been pursuing various approaches to artificial intelligence risk assessment and management. Of particular note are associations with a mixed membership of companies, academics, policymakers and/or civil society organizations
- ▶ There is recognition that AIRA will need to take into account the particularities of specific sectors.

Academics and other researchers are playing a role in developing AIRA methodologies.

- ▶ Research papers on proposed AI risk assessment frameworks take inspiration from environmental-, data protection-, privacy-, human rights-, ethics-, socio-economic-impact assessment methods.
- ▶ Researchers recognized early on that assessment frameworks are a useful component for the governance of AI.
- ▶ They have helped to clarify key concepts and proposed structured processes for the assessment of risks emerging from AI. For example, the EC-appointed High Level Expert Group on Artificial Intelligence (AI HLEG), comprising 52 experts, has developed a set of seven principles for trustworthy AI. It has also developed an Assessment List with questions keyed to each of the seven principles.
- ▶ Other academics have noted the distinction between AI audits and AI impact assessments, and have espoused the applicability of human rights impact assessment to AI. They have also called for accountability, explainability, auditability and fairness in AI technology development.
- ▶ There are, however, open questions remaining, e.g., when to conduct an AIRA, the independence of the assessor, whether to publish the results of the AIRA.

While the AIRA ecosystem is still in flux, it is nevertheless possible to identify leading practices, as we have done with this survey (see page 37), which could help inform AI policymaking as well as the promulgation of AIRA.



Introduction

1

While the use of artificial intelligence (AI) can have many benefits, certain aspects of the technology, the way it is developed or used can also exhibit important shortcomings. Stories abound in the press about how AI-driven applications have resulted in discrimination against people. Moreover, AI applications have led to greater intrusions upon personal privacy, more surveillance and social manipulation. In the last few years, politicians, policymakers, regulators, international organizations, civil society organizations, the media and academics in many countries have discussed the need for more oversight over the deployment and use of artificial intelligence in our societies and economies. Artificial intelligence risk assessments (AIRA) are an important component of that oversight. AIRAs are one of the various tools and initiatives developed for the governance of AI, among other tools and initiatives, such as ethics guidelines, regulatory frameworks, audits.

AI Risk Assessments are an important component in the AI governance toolkit.

The survey shows the diversity of approaches to AIRA. We distinguish between AIRA and AI impact assessment (AIIA), although there is significant overlap between the two. Risks typically involve threats and vulnerabilities, while impacts can be positive and/or negative. An AIRA is subsumed within an AIIA as well as within an AI risk management (AIRM) process.

We identify two types of AI risk assessment. The first makes an assessment of the use of AI. The second type classifies AI systems or applications by risk.

Two types of AI risk assessment:

- 1** Identification of specific risks, e.g., bias;
- 2** Classification of an overall level of risk, e.g., high vs. low risk systems.

In the first type, the assessor is looking for risks arising from the use of AI, which might include bias, lack of transparency, discrimination, increasing power asymmetries, invasion of privacy, misuse of personal data, surveillance and damaging trust. The second type, relevant in the context of law and manifested in the European Commission's proposed AI Regulation, classifies AI systems or applications based on risk in order to determine the applicable legal obligations. Under this risk-based approach to regulation, a high-risk AI system may, for example, be subject to stricter rules than a low-risk system.

In the following pages, we sketch the current ecosystem for AI risk assessments, specifically:

- ▶ Section 1 introduces the survey, and the subsequent sections present current initiatives related to AIRA from the perspective of different actors in the AI ecosystem.
- ▶ Section 2-6 provides an inventory of AI risk assessment methodologies/frameworks categorized by stakeholder groups
 - ▶ Section 2: Legal and regulatory approaches (European Commission, Council of Europe, Germany, Canada, the US, China, Singapore and India).
 - ▶ 2.1: Risk-based approaches to AI regulation
 - ▶ 2.2: Regulatory policies including elements of AI risk assessment
 - ▶ Section 3 covers how international organizations are approaching AI risk assessments (e.g., OECD and United Nations)
 - ▶ Section 4 focuses on initiatives being taken by standards bodies
 - ▶ Section 5 reviews industry approaches to AI risk assessments
 - ▶ Section 6 focuses on methods proposed by academic and civil society organizations.
- ▶ Section 7 presents key findings relevant to policymakers, including leading practices.
- ▶ Section 8 contains a list of references underpinning the previous sections as well as a list of abbreviations.

Some of the documents in this survey are not AI risk assessment methodologies per se, but we have included reference to them because they offer useful guidance about what an AI risk assessment could cover as a matter of good practice.

This survey does not purport to be comprehensive but provides a snapshot of the AIRA landscape at a certain point in time (May-Aug 2021) and may serve as a useful starting point for policymakers and those who would like to know more about current frameworks in relation to AI risk.





Legal and regulatory approaches

2

This section is in two parts, the first of which focuses on risk-based approaches to AI regulation, while the second focuses on regulatory policies that include elements of AI risk assessment.

2.1

Risk-based approaches to AI regulation

A 'risk-based approach to regulation' involves the use of a systematized framework of risk classification to categorize type and degree of risk posed by the object or activity being regulated. In the context of AI, it would mean that not all AI systems are treated equally under the law; legal requirements would vary depending on the AI risk classification. Those in favor argue that it is a way of making legal rules fair and proportionate and that it facilitates efficient and effective use of limited regulatory resources. A regulation with a risk classification approach requires an assessor to identify, assess and classify risk in order to determine the applicable category of risk and corresponding legal requirements within the legal framework.

Risk-based approaches to AI regulation pursue a burden of compliance that is proportionate to the risks of a specific AI application.

As AI regulation is at a very early stage – and regional and national governments are just beginning to develop their AI regulatory frameworks – some are considering whether to adopt a risk-based approach. The European Commission’s proposal is the most well-known, but an earlier example comes from the German Data Ethics Commission, both presented below. While the trend may move towards wider adoption of risk-based regulation for AI, it is too early to say with certainty whether other governments will adopt a similar approach. Some have argued against risk-based AI regulation, calling instead for a rights-based approach.² In the latter case, advocates argue that human rights must be respected regardless of a risk level associated with external factors, i.e., rights are non-negotiable. However, the two approaches are not necessarily mutually exclusive. A risk-based approach can centre on the risk of violating rights and how to avoid such violations. This is reflected in article 7.2 of the proposed EU AI Act which includes ‘risk of harm to fundamental rights’ among the criteria to establish is an AI system’s intended use should be classified as ‘high-risk ’

Risk-based and Rights-based approaches to AI regulation can be complementary.

2.1.0 Proposed EU AI Regulation

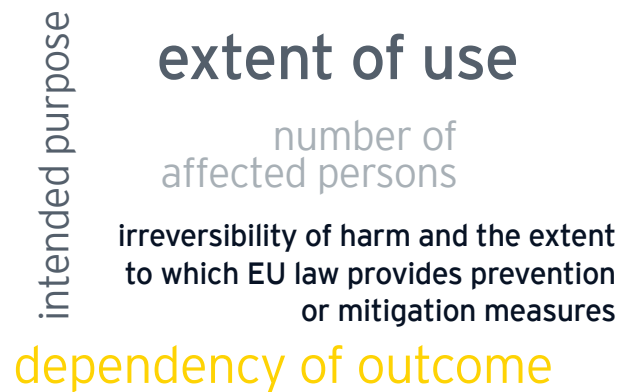
In April 2021, the EC published a proposed Regulation on AI as part of an AI regulatory package for Europe.³ The proposed regulation lays out a framework for classifying AI systems based on risk, with specific legal obligations corresponding to each category of risk. The Commission states that this risk-based approach is intended to ensure the regulatory framework is proportionate to meet the regulatory objectives.

There are four categories of risk within the proposed framework: unacceptable, high, limited/low and minimal. For the purpose of classification, the relevant understanding of risk is “significant risk to the health and safety or fundamental rights of persons”.

The proposal itself lays out a methodology for assessing risk involving assessment of both (1) function performed by the AI within the overall system and (2) specific purpose of the overall system.

- Proposed EU AI Act: Risk Assessment based on:**
- 1** Function performed by the AI within the system
 - 2** Specific purpose for which the system is used

Other criteria for assessing risk are:



Additionally, Annex III of the proposed Regulation identifies high-risk domains: biometric identification and categorization of natural persons; management and operation of critical infrastructure; education and vocational training; employment, workers management and access to self-employment; essential public and private services and creditworthiness; law enforcement; migration and border control; and administration of justice and democratic processes.

High-risk systems are subject to mandatory obligations, including conformity assessments and practices related to risk assessment and management (discussed below). Limited or low-risk systems are subject to transparency requirements. Minimal risk systems are not subject to any regulatory requirements but encouraged to adopt voluntary codes of conduct.

2.1.1 Recommendation of German Data Ethics Committee

In 2019, the German Data Ethics Commission released a set of policy recommendations to the German Federal Government on data and AI systems.⁴ A key recommendation is the development of a risk-based approach to AI regulation, using a five-tiered scale. According to the Data Ethics Commission's recommendation, this would be subject to a partial or complete ban. In determining risk, the Data Ethics Commission recommends assessing "the sociotechnical system as a whole" and calls on the government to develop a criteria-based assessment scheme.

In December 2020, the Federal Government published an updated report on its 2018 AI Strategy⁵ and one of the priority areas is regulatory frameworks. It remains to be seen if the Data Ethics Committee's recommendations will be reflected in this regulatory framework.

The German Data Ethics Committee recommend that a risk based approach on AI regulation should be a five-tiered scale:

- ▶ **Level 1** risk zero or negligible potential harm
- ▶ **Level 2** some potential for harm,
- ▶ **Level 3** regular or significant potential for harm
- ▶ **Level 4** serious potential for harm
- ▶ **Level 5** untenable potential for harm

2.2 Regulatory policies including elements of AI risk assessment

As of July 2021, there are approximately 60 countries worldwide that have developed or are developing AI-related strategies and policies.⁶ Some of those national governments, and the European Union, have contemplated or are contemplating incorporating AI assessments in their AI strategy and regulatory frameworks. The type of required assessment (e.g., risk, impact, human rights, ethics) and degree of specificity for the assessment process differ. The following section presents examples of regulatory initiatives that include elements of AI risk assessments in the European Union, the United States, Canada, China, Singapore and India as well as a reference to national AI strategies.

2.2.1 European Commission

Proposed EU AI Regulation

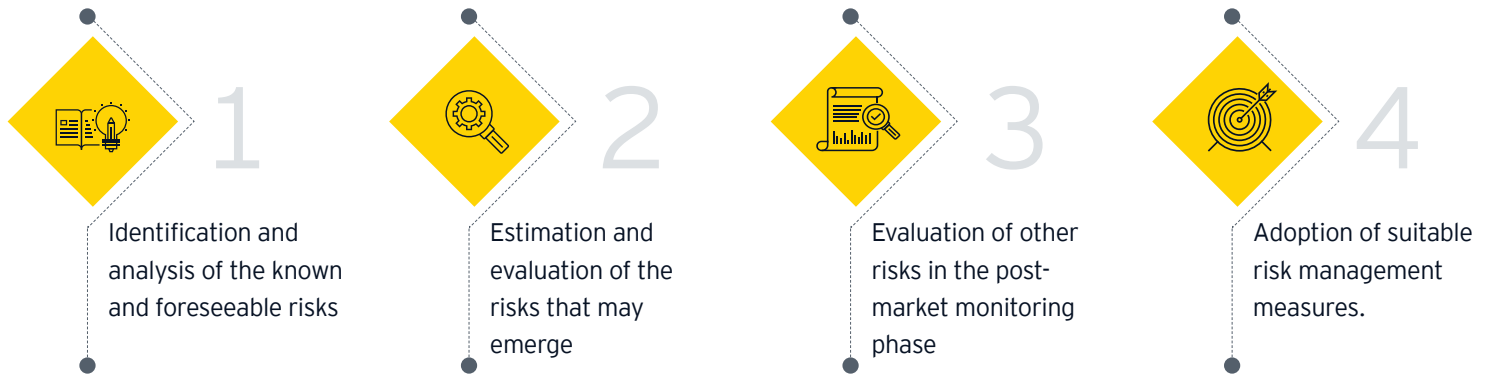
There are no provisions related to methods of risk (or impact) assessment per se in the proposed AI Regulation, beyond the classification outlined in 2.1.0 above. The only assessment mentioned is 'conformity assessment'. However, those designing, developing, testing or putting AI systems on the market in the EU will need to comply with its provisions related to risk management should the Regulation become law.

The proposed Regulation sets mandatory requirements for high-risk AI systems to be regarded as trustworthy and expects providers to conduct conformity assessment procedures before AI systems can be placed on the Union market, including AI systems embedded in other products or services. Other requirements for high-risk AI systems include high-quality data, documentation and traceability, transparency, human oversight, accuracy and robustness to mitigate the risks to fundamental rights and safety posed by AI and that are not covered by other existing legal frameworks.⁷

The proposed Regulation says that an assessment through internal checks for "stand-alone" high-risk AI systems would require a full, effective, properly documented, ex ante compliance with all requirements of the Regulation and compliance with robust quality and risk management systems and post-market monitoring.

As part of good risk management practice, the proposed Regulation indicates that all providers should have a post-market monitoring system to ensure that the possible risks emerging from AI systems after being placed on the market can be efficiently addressed in a timely manner. The EC also expects providers to report to relevant authorities any serious incidents or any breaches from use of their AI systems.⁸

The Regulation describes **four main steps** in a risk management system for AI as follows:



Where risks can't be entirely removed, it says developers and providers should put in place adequate mitigation and control measures and provide adequate information (transparency) regarding potential risks and, where appropriate, training to users. High-risk AI systems also need to undergo conformity assessment procedures before the systems can be placed on the market, which becomes part of the AIRM process in the EU (and elsewhere).

Providers of high-risk AI systems need an AI risk management plan and practice that goes beyond the design, development and testing of an AI system. Art. 61 of the proposed Regulation says they should also monitor the use of the AI system after it is put on the market, especially the performance of the AI system as well as its ongoing compliance with the Regulation. The Commission plans to lay down "detailed provisions establishing a template for the post-market monitoring plan and the list of elements to be included in the plan".

The draft Regulation says regulators will need to have information on how high-risk AI systems have been developed and how they perform throughout their life cycle to verify compliance with the Regulation.⁹

Additionally, the draft Regulation encourages providers of non-high-risk AI systems to adopt codes of conduct to foster voluntary application of the mandatory requirements applicable to high-risk AI systems.¹⁰ The Regulation encourages the drawing up of codes of conduct applied to AI systems that make provisions, for example, to environmental sustainability, accessibility for persons with a disability, stakeholder participation in the design and development of the AI systems and diversity of development teams and performance indicators. Any such codes of conduct could include provisions for a practical AI risk assessment. These provisions suggest that a risk assessment of an artificial intelligence technology should take into account its impact on the environment. But the link between an AIRA and an environmental impact assessment (EIA) goes much deeper, at least as far as the United Nations Economic Commission for Europe (UNECE) Convention on Access to Information, Public Participation in Decision-Making and Access to Justice in Environmental Matters, which was adopted in 1998 in the Danish city of Aarhus, hence it is better known as the Aarhus Convention. The EIA was the antecedent of privacy impact assessment and ethical impact assessment, which have informed AIRA, especially in regard to assessing environmental risks, informing citizens and engaging them in the decision-making process.





High-Level Expert Group on Artificial Intelligence (AI HLEG)

In June 2018, the European Commission appointed a group of 52 experts to provide advice on its artificial intelligence strategy.¹¹ The AI HLEG produced seven requirements for trustworthy AI, which are the guiding principles for the draft EU AI Regulation:

- Human agency and oversight;
- Technical robustness and safety;
- Privacy and data governance;
- Transparency;
- Diversity, non-discrimination and fairness;
- Societal and environmental well-being;
- Accountability.

The AI HLEG says trustworthy AI has three components: (1) it should be lawful, complying with all applicable laws and regulations; (2) it should be ethical, ensuring adherence to ethical principles and values and (3) it should be robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm.

The AI HLEG believes that Trustworthy AI has three components:

- ▶ It should be **lawful**
- ▶ It should be **ethical**
- ▶ It should be **robust**

To support organizations implementing their recommendations, the AI HLEG has developed an Assessment List on Trustworthy Artificial Intelligence (ALTAI).¹² While the ALTAI is not an AI risk assessment per se, its questions, adapted as necessary, can be considered for an AI risk assessment. ALTAI aims to provide an evaluation process for trustworthy AI self-evaluation. Organizations can draw elements relevant to the particular AI system from ALTAI or add elements to it as they see fit, taking into consideration the sector in which they operate. It helps organizations understand what trustworthy AI is, in particular, what risks an AI system might generate. The Expert Group says its Assessment List is firmly grounded in the protection of people's fundamental rights and it encourages the involvement of all relevant stakeholders, including, for example, AI designers and AI developers of the AI system; data scientists; procurement officers or specialists; front-end staff that will use or work with the AI system; legal/compliance officers; management. It says that, prior to self-assessing an AI system with this Assessment List, a fundamental rights impact assessment (FRIA) should be performed.

The ALTAI is divided into seven main sections with questions keyed to each of the guiding principles cited above. Examples of its questions are the following relating to resilience to attack and security:

- ▶ Could the AI system have adversarial, critical or damaging effects (e.g., to human or societal safety) in case of risks or threats such as design or technical faults, defects, outages, attacks, misuse, inappropriate or malicious use?

- ▶ Is the AI system certified for cybersecurity (e.g., the certification scheme created by the Cybersecurity Act in Europe) or is it compliant with specific security standards?
- ▶ How exposed is the AI system to cyber-attacks?
 - ▶ Did you assess potential forms of attacks to which the AI system could be vulnerable?
 - ▶ Did you consider different types of vulnerabilities and potential entry points for attacks such as:
 - ▶ Data poisoning (i.e., manipulation of training data);
 - ▶ Model evasion (i.e., classifying the data according to the attacker's will);
 - ▶ Model inversion (i.e., infer the model parameters)

2.2.2 United States

The US government has defined trustworthy AI as “designed, developed, and used in a manner that is lawful, fair, unbiased, accurate, reliable, effective, safe, secure, resilient, understandable, and with processes in place to regularly monitor and evaluate the AI system’s performance and outcomes”.¹³ These benchmarks are similar to, but not as detailed as the AI HLEG’s Assessment List for Trustworthy Artificial Intelligence. Nevertheless, they are useful benchmarks to be included as a good practice in an AIRA. In its consideration of risks, an AIRA would need to question each step in the design, development and deployment of AI and whether the algorithm or AI-enabled process was lawful, fair, unbiased, etc. If the AI was assessed as not being lawful, of being unfair, biased, etc., then it presents a risk to the developers and those putting the AI on the market.

The following initiatives in the US are particularly germane for those developing AI risk assessment guidance.

Proposal for an Algorithmic Accountability Act

In 2019, Congress introduced the Algorithmic Accountability Act, the first US national attempt to regulate algorithms.¹⁴ The bill would have authorized the US Federal Trade Commission (FTC) to issue regulations mandating specific commercial entities to conduct impact assessments of high-risk automated decision systems (ADS). The assessment would have been required to describe the system in detail, assess the relative costs and benefits, determine the risks to the privacy and security of

personal information, and explain the steps taken to minimize those risks. If possible, the assessment would have been conducted in consultation with external third parties, but there would have been no requirement to publicly disclose the findings of the assessment. The bill did not advance, but it indicates a possible direction for federal regulation of AI systems in the US in the future.

US Department of Homeland Security (DHS)

A DHS study identifies factors to consider when formulating standards to manage the national security risks arising from use of AI for decision support systems.¹⁵ It states:

To build trust, fairness, transparency, and accountability of AI to curtail error and misuse while ensuring functionality and securing against attacks, algorithms should undergo the “illities” test. The test looks at reliability, accountability maintainability, functionality, debug-ability, evolve-ability, fragility, and vulnerability. Algorithms should also incorporate ethical, legal, privacy, transparency, and bias concerns. The main challenge is the complexity and difficulty in getting the public and private sector that have different incentive structures to agree on specific standards.

US Department of Homeland Security (DHS) recommends that to build fairness, transparency and accountability algorithms should undergo the “illities” tests:

- | | |
|-------------------|------------------|
| ▶ Reliability | ▶ Debug-ability |
| ▶ Accountability | ▶ Evolve-ability |
| ▶ Maintainability | ▶ Fragility |
| ▶ Functionality | ▶ Vulnerability |

The study postulates an AI risk model that is conceived as a cycle of generalised risk and then goes on to say:

Starting with assets that are compromised by threats, which exploit vulnerabilities that are exposed to risks, which are mitigated by controls to protect assets. The inner ring represents the same cycle for an example AI-related system. Start with training data compromised by actors who feed malicious data to AI systems, which are designed with narrow purposes and algorithms that lack pressure testing. Decisions derived from algorithms that are not scrutable require standardized architecture to protect the compromised training data.

The model offers a straight-forward conceptualisation of factors that should be taken into account in an AI risk assessment.

2.2.3 Canada

Canada's federal government has developed an Algorithmic Impact Assessment (AIA) tool to support its Directive on Automated Decision-Making.¹⁶ The tool is a questionnaire that determines the impact level of an automated decision-system. It is composed of 48 risk and 33 mitigation questions. Assessment scores are based on many factors, including systems design, algorithm, decision type, impact and data. The AIA was based on best practices in algorithmic impact assessment, according to the government, in consultation with both internal and external stakeholders, including academia, civil society and other public institutions. While the AIA was designed to help departments and agencies better understand and manage the risks associated with automated decision systems, it is available to the public for sharing and re-use under an open licence.¹⁷ Its questions can be a source of good practice for inclusion in an AI risk assessment process.

Canada's federal government Algorithmic Impact Assessment tool was co-designed by government, academia, civil-society and public institutions. It consists of 48 risk and 33 mitigation questions including system design, algorithm, decision type, impact and data.

2.2.4 China

In 2017, China's State Council launched the Next Generation Artificial Intelligence Development Plan (AIDP) as part of China's AI development strategy. The AIDP provides a longitudinal perspective on China's strategic situation in AI, including its comparative capabilities, opportunities and potential risks. An AI risk assessment model has not been developed in the AIDP to date. More recently, the Beijing Artificial Intelligence Principles,¹⁸ published by the Beijing Academy, is a key development in the

field of AI in China as it proposes the ethical governance of AI. Many of the principles will resonate with stakeholders in other countries who seek to translate principles into policy and procedure. The Beijing AI principles are subsumed within four major areas: Research and Development, Use, Governance, and Release and Endorsement. For example, under R&D, the principles include the following:

- For Humanity:** The R&D of AI should serve humanity and conform to human values as well as the overall interests of mankind. Human privacy, dignity, freedom, autonomy, and rights should be sufficiently respected. AI should not be used against, or utilized to harm, human beings.
- Be Responsible:** Researchers and developers of AI should have sufficient considerations for the potential ethical, legal, and social impacts and risks brought in by their products and take concrete actions to reduce and avoid them.
- Control Risks:** Continuous efforts should be made to improve the maturity, robustness, reliability, and controllability of AI systems, so as to ensure the security for the data, the safety and security for the AI system itself, and the safety for the external environment where the AI system deploys.
- Be Ethical:** AI R&D should take ethical design approaches to make the system trustworthy. This may include, but not limited to: making the system as fair as possible, reducing possible discrimination and biases, improving its transparency, explainability, and predictability, and making the system more traceable, auditable and accountable.
- Be Diverse and Inclusive:** The development of AI should reflect diversity and inclusiveness, and be designed to benefit as many people as possible, especially those who would otherwise be easily neglected or underrepresented in AI applications.

Hence, an AIRA in China would need to take into account the above principles, among others, and avoid or minimize the risks that might arise from violating these principles.



2.2.5 Singapore

In 2019, Singapore launched its national AI strategy and has a national AI programme known as “AI Singapore”.¹⁹ A key element of the strategy is a model framework for AI governance that has been developed to address growing concerns about the risk of harm associated with the use of AI technologies. The government says it will publish assessment guides for organizations to assess the alignment of their AI governance processes with the Model AI Governance Framework. The first edition of the model framework, which is industry-, technology – and algorithm-agnostic, translates relevant ethical principles into practices that can be applied in an AI deployment process so that organizations can operationalise these principles. A second edition was published a year later, in January 2020. It provides guidance to organizations to adopt a risk-based approach when

implementing measures by identifying features or functionalities with the greatest impact on stakeholders; and considering which measure would be most effective in building trust with stakeholders. Annex A of the National Strategy lists various AI principles, including accountability, accuracy, auditability, explainability, fairness, human centricity and well-being, human rights alignment, inclusivity and progressiveness, responsibility, transparency, robustness, security and sustainability. Thus, an AIRA in Singapore would need to take into account these principles. Singapore, together with the World Economic Forum, has also developed an Implementation and Self-Assessment Guide for Organizations (ISAGO) to help organizations to assess the alignment of their AI governance practices with the Model Framework.²⁰ The ISAGO includes various questions (and answers) that are pertinent to AIRA. For example:

Singapore proposes that AIRA should take the following into account:

- ▶ Features or functionalities with greatest impact on stakeholders
- ▶ Measures that would be most effective in building trust

AI Principles of: accountability, accuracy, auditability, explainability, fairness, human centricity/well-being, human rights, inclusivity, progressiveness, responsibility, transparency, robustness, security and sustainability.

For example:

Has your organization considered conducting an assessment on whether the expected benefits of implementing the identified AI solution in a responsible manner (as described in the Model Framework) outweigh the expected costs?

Did your organization consider whether the decision to use AI for a specific application/use case is consistent with its core values and/or societal expectations?

Does your organization have an existing risk management system that can be expanded to include AI-related risks?

Did your organization implement a risk management system to address risks involved in deploying the identified AI solution (e.g., personnel risk or changes to commercial objectives)?

2.2.6 India

India does not have an AI risk assessment template, but the national policy think tank, NITI Aayog (National Institution for Transforming India), released its National Strategy for Artificial Intelligence (NSAI) in 2018.²¹ The strategy, called ‘#AI for All’, aims to position India as the ‘AI Garage of the World,’ and the strategy identifies frameworks for regulation using AI as well as mitigation measures for possible risks. The corresponding regulations framing the use of AI are information technology law, discrimination law, competition law, consumer protection law and sectoral regulation. The National Strategy encourages self-regulation: “Data Privacy Impact Assessment Tools can be used by AI developers and enterprises adopting AI solutions to manage privacy risks” [p. 87]. It gives implicit support to risk assessments of AI: “Any strategy document on promoting AI necessarily needs to be conscious of the probable factors of the AI ecosystem that may undermine ethical conduct, impinge on one’s privacy and undermine the security protocol. Appropriate steps to mitigate these risks need to be an integral part of any such strategy [p. 85]”. The strategy does not explicitly mention an AIRA, but it does identify many AI issues requiring regulatory attention and many such issues would need to be taken into account in an AIRA.

2.2.7 National AI strategies

The EC and OECD have compiled a database of national AI policies and strategies, which contains more than 650 AI policies and strategies from more than 60 countries.²² National strategies differ in both the scope and approaches to the regulation of AI, ranging from high-level strategies with different policy initiatives to concrete action plans with specific milestones and time frames (e.g., Bulgaria, Estonia and Hungary). AI policies can also be incorporated in a wider strategy of digital transformation, as for Slovakia. Recently published or updated strategies (e.g., Germany and Spain) include policy initiatives in response to the COVID-19 pandemic and sustainability issues, such as environmental and climate change.

National actions to address ethical concerns differ across countries in terms of strategic approach and level of focus. Many governments have established AI ethics committees and councils.

Many governments also implement monitoring and reward systems for compliance with principles for trustworthy AI. Malta has developed an AI certification framework, issued by the Malta

Digital Innovation Authority (MDIA), which recognises that the AI systems of successful applicants have been developed in an ethical, transparent and socially responsible manner. Similar quality seals or labels – acting as hallmarks for a responsible approach in AI – have been adopted in other countries such as Denmark and Germany. The Czech Republic, Italy, Lithuania and Spain are considering developing them as well. Similarly, the AI registers set up by the cities of Amsterdam and Helsinki aim to ensure a secure, responsible and transparent use of AI algorithms.



International organizations

3

3.1

Council of Europe (CoE)

The Council of Europe (CoE) is an international organization founded to uphold human rights, democracy and the rule of law in Europe. Due to this particular focus, the work of the CoE on AI has focused on human rights-related risks and impacts.

The CoE Council of Ministers and the CoE Commissioner for Human Rights have made recommendations for human rights impact assessments (HRIA) of AI systems. In its “Recommendation on the human rights impacts of algorithmic systems”, the Council of Ministers calls on Member States to require mandatory HRIA of high-risk algorithmic systems used by governments and encourage HRIA for the private sector. In “Unboxing AI: 10 steps to protect human rights”, the Commission for Human Rights calls for Member States to establish a legal framework setting out a HRIA procedure and lays out some elements of a HRIA.²³ One of those elements is that the HRIA should be conducted by an independent third party to help discover, measure and/or map human rights impacts and risks over time. Furthermore, it says, self-assessments and external reviews should not be limited to an evaluation of the models or algorithms behind the AI system, but should include an evaluation of how decision-makers might collect or influence the inputs and interpret the outputs of such a system. It should also include an assessment of whether an AI system remains under meaningful human control throughout the AI system’s life cycle. Another element is that the HRIA must set out the measures, safeguards and mechanisms envisaged for preventing or mitigating risks.

The Council of Europe recommends that the Human Rights Impact Assessment of AI should include an evaluation of how decision-makers might collect or influence the input and interpret the outputs of such a system.

3.2

Global Partnership on AI (GPAI)

The Global Partnership on AI (GPAI) is an international, multi-stakeholder initiative to promote responsible AI. There are 13 founding member countries, with the secretariat hosted at the OECD. The GPAI has not done any work yet on risk assessment

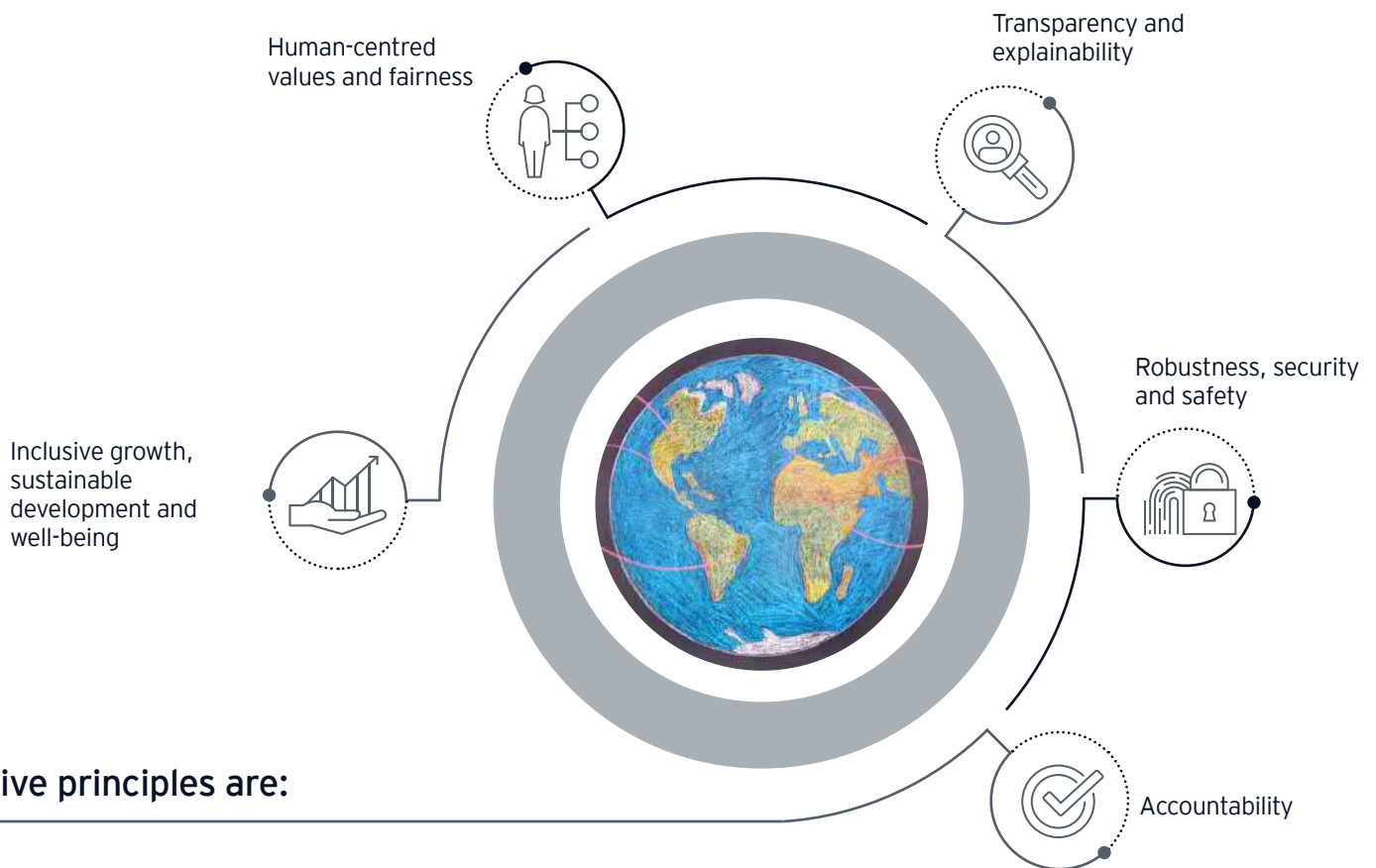
but its Working Group on AI has proposed creating an internal committee on Issues and Means of Governance, which it has indicated may be tasked with working on AI assessment.²⁴

3.3

Organizations for Economic Co-operation and Development (OECD)

Multiple initiatives at the OECD relate to AI risk assessment. In May 2019, OECD member countries adopted the OECD Principles on AI to promote the “responsible stewardship of trustworthy AI”.²⁵ The principles are the first intergovernmental standard

on AI; while they are not legally binding, member countries are expected to “do their utmost to fully implement” them. The Principles do not refer to AI risk assessment, but they could be the basis for a risk or impact assessment framework.



The five principles are:

Related to AI risk classification, in February 2021, the OECD AI Network of Experts proposed a Framework for classifying AI systems to help policymakers who favor a risk-based approach to regulation determine how to classify AI systems by risk.²⁶ The proposed framework, which includes an ‘Illustrative ethical and societal risk assessment’, classifies AI systems according to: (1) context, (2) data and input, (3) AI model, and (4) task and output. The proposed Framework was open for public consultation through June 2021 and is now under revision.

Related to AI risk assessment, a working group on implementing trustworthy AI is also developing a public database of tools for trustworthy AI. One of the types of procedural tools identified for inclusion in the database is risk management tools, which will likely include AI risk and impact assessments. The framework for the database was published in June 2021, but the database is not yet live.

3.4 United Nations Educational, Scientific and Cultural Organization (UNESCO)

The United Nations Educational, Scientific and Cultural Organization (UNESCO) is currently developing a Recommendation on the ethics of artificial intelligence. The Recommendation, directed at UNESCO Member States, outlines specific recommendations for policy action. While the most recent draft (from September 2020)²⁷ does not refer to AI risk assessment frameworks specifically, there are recommendations for ethical impact assessments that include AI risk assessment, prevention, mitigation and monitoring. For example, it states that “In the event of possible occurrence of any harm to human beings or the environment and ecosystems, the implementation of procedures for risk assessment and the adoption of measures in order to preclude the occurrence of such harm should be ensured.” However, it does not detail those procedures, but it does advocate “engagement with a broad range of appropriate stakeholders guided by international human rights law, standards and principles, making use of social dialogue, as well as ethical deliberation, due diligence, and impact assessment”. It also goes on to say: “Participation of different stakeholders throughout the AI system life cycle is necessary for inclusive AI governance, sharing of benefits of AI, and fair technological advancement and its contribution to development goals.”

The draft Recommendation says that Member States should adopt a regulatory framework that sets out a procedure, particularly for public authorities, to carry out ethical impact assessments on AI systems to predict consequences, mitigate risks, avoid harmful consequences, facilitate citizen participation and address societal challenges. The assessment should also establish oversight mechanisms, including auditability, traceability and explainability that enable the assessment of algorithms, data and design processes.

Once adopted (expected in November 2021), the Recommendation will not be legally binding, but UNESCO Member States will be required to report on the measures taken in relation to the Recommendation.

UNESCO recommends that ethical impact assessments of AI should include:

- ▶ Risk assessment
- ▶ Prevention
- ▶ Mitigation
- ▶ Monitoring

3.5 World Economic Forum (WEF)

The World Economic Forum (WEF) is an independent international organization for public-private partnerships, providing a platform for collaboration on projects. Multiple projects involving the WEF concern assessment of the risks that can arise from the use of AI for both the public and private sectors.

For governments, the WEF has created an ‘AI Procurement in a Box: Toolkit’ that includes an AI risk assessment tool. The assessment guidelines in the toolkit identify key variables for consideration related to data, field of use, socio-economic impact, financial consequences for agency and individuals, and impact of the AI system on processes, employees and core business. Although we don’t know which governments are using the toolkit, the WEF says its guidelines were developed by the WEF in

consultation with a multistakeholder community. Project fellows from the UK Government’s Office for AI, Deloitte and Salesforce worked with Forum staff, and in partnership with representatives from government, academia, civil society and the private sector.²⁸

WEF created a ‘AI Procurement in a Box: Toolkit’ which includes an AI risk assessment tool identifying key variables such as:

- ▶ Data
- ▶ Field of use
- ▶ Socio-economic impact
- ▶ Financial consequences for agency and individuals
- ▶ Impact processes
- ▶ Employees and core business



A pilot project with the Government of New Zealand, 'Reimagining Regulation for the Age of AI', was a WEF effort to co-design a regulatory framework for AI that included general discussion of the need for AI risk assessment and set of assessment guidelines for procurement.²⁹ Guidelines included, for example, applying risk/benefit assessment frameworks across the life cycle of AI, laying out risk prioritization schemes and defining performance metrics to ensure AI is fit for purpose. Specific to regulation on AI and facial recognition, the WEF has also proposed a Framework for Responsible Limits on Facial Recognition that includes a set of questions for AI risk assessment.³⁰

For the private sector, the WEF partnered with the Government of Singapore to develop the 'Implementation and Self-Assessment Guide for organizations (ISAGO)' deploying AI at scale, a complement to Singapore's 'Model AI Governance Framework and Assessment Guide for AI'. The WEF has also created an 'Empowering AI Leadership: Toolkit for Boards of Directors' with tailored tools and resources on general AI risk identification and assessment. The WEF is currently developing a similarly tailored 'Toolkit for C-Suite Executives'. While not an AIRA framework, these resources illustrate how the WEF views AI risk identification as something that the private sector needs to take seriously, and which should be done as part of robust risk management.

'Reimagining Regulation for the Age of AI' includes guidelines for:

- ▶ Applying risk/benefit assessment across the AI life cycle
- ▶ Laying out risk prioritization schemes
- ▶ Defining performance metrics

Standards

4

This section provides an overview of the standards relating to risk management and governance being developed by standard-setting organizations at the international and regional level. In this case, AI risk assessment is part of the risk management and governance process.

At the international level, the International Organization for Standardization (ISO), the International Electrotechnical Commission (IEC) and the Institute of Electrical and Electronics Engineers Standards Association (IEEE SA) have initiated the development of standards for risk management and governance. At the national level, the US National Institute of Standards and Technology (NIST) is developing a series of documents and workshops to establish a risk management framework and corresponding standards for trustworthy AI. Similarly, the European Committee for Standardization (CEN) and European Committee for Electrotechnical Standardization (CENELEC) are working together to develop standards suitable for the European market and underpinned by European legislation.

4.1

ISO/IEC

ISO/IEC 23894 – Information Technology – Artificial Intelligence – Risk Management
Stage: 30.60, Committee Draft (CD)

ISO 23894 provides guidelines on managing risks during the development and application of AI techniques and systems and assists organizations to integrate risk management into their AI-related activities and functions.³¹ It has three main substantive sections on the principles of AI risk management, framework and processes and several annexes on objectives, risk sources, controls and their relationship, and a typical mapping between risk management processes and an AI system life cycle.

The document is based on ISO 31000 (Risk management – guidelines, 2018) and follows the same approach, but also notes some things that are special to AI and that should be considered in the risk management process. For example, organizations are encouraged to dialog with a diverse group of internal and external stakeholders regarding the benefits and consequences of AI systems and to take account of their knowledge and views to improve awareness and informed risk management. AI systems also raise specific issues that need to be accounted for in the risk management process, such as transparency, explainability, human oversight, equitable outcomes, privacy, freedom of expression, fairness, safety, security, employment, environmental and human rights. In addition, organizations need to be alert to the prospect of previously unknown risks.

ISO/IEC AWI 42001 Information Technology – Artificial intelligence – Management system
Stage: 20.00, Preparatory

Taking a risk-based approach, ISO 42001 aims to provide requirements for establishing, implementing, maintaining and continually improving an artificial intelligence (AI) management system.³² The standard sets out guidelines for the deployment of applicable controls to measure the effectiveness and efficiency of such a process. It is also intended to aid organizations to develop

or use AI responsibly in pursuit of their objectives and to meet applicable regulatory requirements, stakeholder obligations and expectations. ISO 42001 is being designed to serve as an auditable and certifiable standard that fits into an organization's governance, risk and compliance (GRC) ecosystem.

ISO/IEC DIS 38507 – Information technology – Governance of IT – Governance implications of the use of artificial intelligence by organizations
Stage: 40.20, Enquiry

To ensure effective, efficient and acceptable use of AI within organizations, ISO 38507 aims to provide guidance for members of an organization's governing body.³³ Regarding the governance oversight of AI, the standard requires governing bodies to ensure that policies are in place to cover the appropriate and responsible use of AI by establishing a chain of responsibility, accountability, authority and ensuring that the potential delegation of authority is clearly defined and agreed both within the organization and, where appropriate, between different parties in any value chain. At the level of compliance management, care should be taken to extend compliance processes to accommodate the speed, scope and sophistication of the AI system. The standard document is cross-organizational and can be used by multiple stakeholders for current and future AI governance.

4.2

IEEE SA

IEEE P2863 – Recommended Practice for Organizational Governance of Artificial Intelligence
Stage: Under Development

This recommended practice details governance criteria such as safety, transparency, accountability, responsibility and bias reduction, as well as process steps for effective implementation, performance, auditing, training and compliance in the development or use of artificial intelligence within organizations.³⁴

IEEE 7000-2021 – IEEE Model Process for Addressing Ethical Concerns During System Design
Stage: Published

The standard establishes a set of processes by which organizations can include consideration of ethical values throughout the stages of concept exploration and development. This standard supports management and engineering in transparent communication with selected stakeholders for ethical values elicitation and prioritization. It involves traceability of ethical values through an operational concept, value propositions, and value dispositions in

the system design. The standard describes processes that provide for traceability of ethical values in the concept of operations, ethical requirements, and ethical risk-based design. It is applicable for all sizes and types of organizations using their own life cycle models.

The Ethical Risk-based Design process translates value-based requirements into design characteristics and determines controls that can mitigate risks to values. Controls are system requirements or organizational policies and procedures. As ethical value requirements (EVR) are instantiated in the system design, the value dispositions are validated for incorporation of the specified values.³⁵

IEEE 7010-2020 – IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being

Stage: Published

This recommended practice provides contextual measures of well-being that aim to facilitate the use of a well-being impact assessment (WIA) process to proactively increase and contribute to safeguarding human well-being throughout the life cycle of

autonomous and intelligent systems (A/IS)³⁵. The document considers the whole life cycle of autonomous and intelligent systems and provides information on the impacts that should be considered at each stage by the developers of these systems across varied sectors. It is intended to help developers to establish a concept of human well-being in relation to A/IS, identify areas of improvement, inform risk mitigation strategies, assess performance, identify intended and unintended users, uses and impacts on human well-being of A/IS.³⁶

The IEEE Applied Artificial Intelligence Systems (AIS) Risk and Impact Framework Initiative

Stage: Under Development

This initiative aims to understand and propose an applied risk framework or assessment. The initiative will review and use existing risk approaches in the areas of finance and cybersecurity, identify gaps introduced by AI and determine an approach for developing an applied AI risk assessment that determines the level of risk and identifies approaches to mitigate that risk. Proposed outcomes include the creation of a general framework for AI risk assessment, white papers and the subsequent development of recommended practices and guides.

4.3

NIST

Draft NISTIR 8332: Trust and Artificial Intelligence

Stage: Draft Published

Recognising that the dynamic between the AI user and the AI system is both a partnership and a relationship where user trust is an essential element, this report provides an overview of the trust challenges associated with AI and establishes the importance of user trust for the development of AI systems.³⁷ Building on Bill Gates' conceptualisation of a system's trustworthiness attributes, IEEE and ISO define the trustworthiness of a computer system as (1) reliance that can be justifiably placed on the service it delivers of an item and (2) the ability to function as and when required. In the second definition, the report notes that it encourages

the creation of characteristics that an AI must have to be trustworthy. Therefore, the development of these characteristics, how they should be measured and what those measurements should be, based on a given AI use case, are essential to the development of an AI system. User trust in AI is then based on the perception of its trustworthiness, with trust being seen as a function of users' perception of the technical characteristics of trustworthiness. The report notes that trust in AI will depend on how the human user perceives the system. If the AI system has a high level of technical trustworthiness and the values of the trustworthiness characteristics are perceived to be good enough for the context of use, and in particular the risk inherent in that context, then the likelihood that AI users will trust it increases. This report is intended to complement the work being done on the trustworthiness of AI systems.

Draft NIST Special Publication 1270: A proposal for Identifying and Managing Bias in Artificial Intelligence

Stage: Draft Published

This report proposes a strategy for managing AI bias and outlines prominent biases that are prevalent in AI and can contribute to societal harms.³⁸ It is one of a series of NIST specific documents and workshops in pursuit of a risk management framework and accompanying standards for trustworthy and responsible AI. The report notes that current approaches tend to categorize bias by type, use case or industry sector, but are not always able to provide the broad perspective needed to effectively manage bias as a context-specific phenomenon. The report proposes an approach, modeled on the AI life cycle, that includes three distinct stages: pre-design, design and development, and deployment. At the pre-design stage, planning, problem specification,

background research and data identification and quantification are undertaken. Biases may arise in the problem formulation and decision-making phase, as many downstream processes depend on the decisions made at this stage. The crucial aspect of these decisions is who – individual or group – makes them and the power dynamics within the team. At the design and development stage, modelling, engineering and validation are undertaken. Relevant stakeholders, including software designers, engineers and data scientists, implement risk management techniques in the form of algorithmic auditing and improved metrics for validation and evaluation. During the deployment stage, users begin to interact with the developed technology and use it in ways not intended by its designers. Biases at this stage can be demonstrated by the discriminatory impact of technologies, their use in a real versus simulated context and contextual gaps that result in performance gaps.

4.4

CEN/CENELEC

A new CEN-CENELEC Joint Technical Committee (JTC) 21 “Artificial Intelligence” was established in 2021. It will be responsible for the development and adoption of standards for AI and related data. Among the initial set of proposed work items

is an ad hoc committee to develop a report on AI Conformity Assessment that is intended to provide guidance on assessing compliance with the requirements for high-risk systems under the EC’s proposed AI Regulation (see section 2.1.0).



Industry approaches

5

In this section, we briefly review various industry approaches to artificial intelligence risk assessment and management. Of particular note are associations with a mixed membership of companies, academics, policymakers and/or civil society organizations.

5.1

Wharton School on AI in financial services

A group of about 40 academics and executives from the financial services and technology industries, formed in 2019 under the aegis of the Wharton School at the University of Pennsylvania, have produced a white paper on AI in the financial services industry.³⁹ The white paper details the opportunities and challenges of implementing AI strategies by financial firms and how they could identify, categorize and mitigate potential risks by designing appropriate governance frameworks. The white paper is intended for discussion purposes only and not to serve as a prescriptive roadmap for implementing AI/ML tools or as a comprehensive inventory of risks associated with the use of AI.

The paper recognizes that AI could lead to privacy issues and/or potentially discriminatory or unfair outcomes, if not implemented with appropriate care. It says institutions could adopt measures to mitigate AI risks such as oversight and monitoring, enhancing explainability and interpretability, differential privacy and watermarking. Each institution should assess its own AI uses, risk assumptions and design governance frameworks that fit their particular circumstances.

Wharton School recommends that measures to mitigate AI risks such as:

- ▶ **Oversight and monitoring,**
- ▶ **Enhancing explainability and interpretability,**
- ▶ **Differential privacy**
- ▶ **Watermarking**

It categorizes AI risks in four major groups:

- Data-related risks
- AI/ML attacks
- Testing and trust
- Compliance

It highlights the potential for AI governance frameworks to help organizations learn, govern, monitor and mature AI adoption.

An AI governance framework might begin with an organization's identifying key stakeholders and creating a working group or 'coalition' to review AI-enabled initiatives. The 'coalition' should take into account data ethics, privacy rights, applicable regulatory considerations, whether the data on which the AI system is being trained is suitable (i.e., was it provided for this purpose or is it being repurposed from some other use?), whether notice of such use may be required to third parties, whether the data set is appropriately safeguarded (e.g., via access controls and encryption protocols) and the manner of supervisory oversight.

5.2 Federation of European Risk Management Associations (FERMA)

FERMA brings together 21 risk management associations in 20 European countries, representing some 5,000 practitioners. It produced a paper titled "AI applied to risk management", which assesses the impacts of AI on the use of its Enterprise Risk Management (ERM) framework, the benefits and opportunities for risk managers in organizations that apply AI and includes an appendix with a template "Your AI Risk Management Roadmap".⁴⁰

Your AI Risk Management Roadmap" assesses AI risk along five dimensions:

- ▶ Governance & culture
- ▶ Strategy & objective-setting
- ▶ Performance
- ▶ Review & revision
- ▶ Information communication & reporting

FERMA introduces the paper with the observation that, as with every new technology, there are emerging risks and challenges related to AI. Questions on ethics, equality, biases, trustworthiness and reliability of decisions made or suggested by

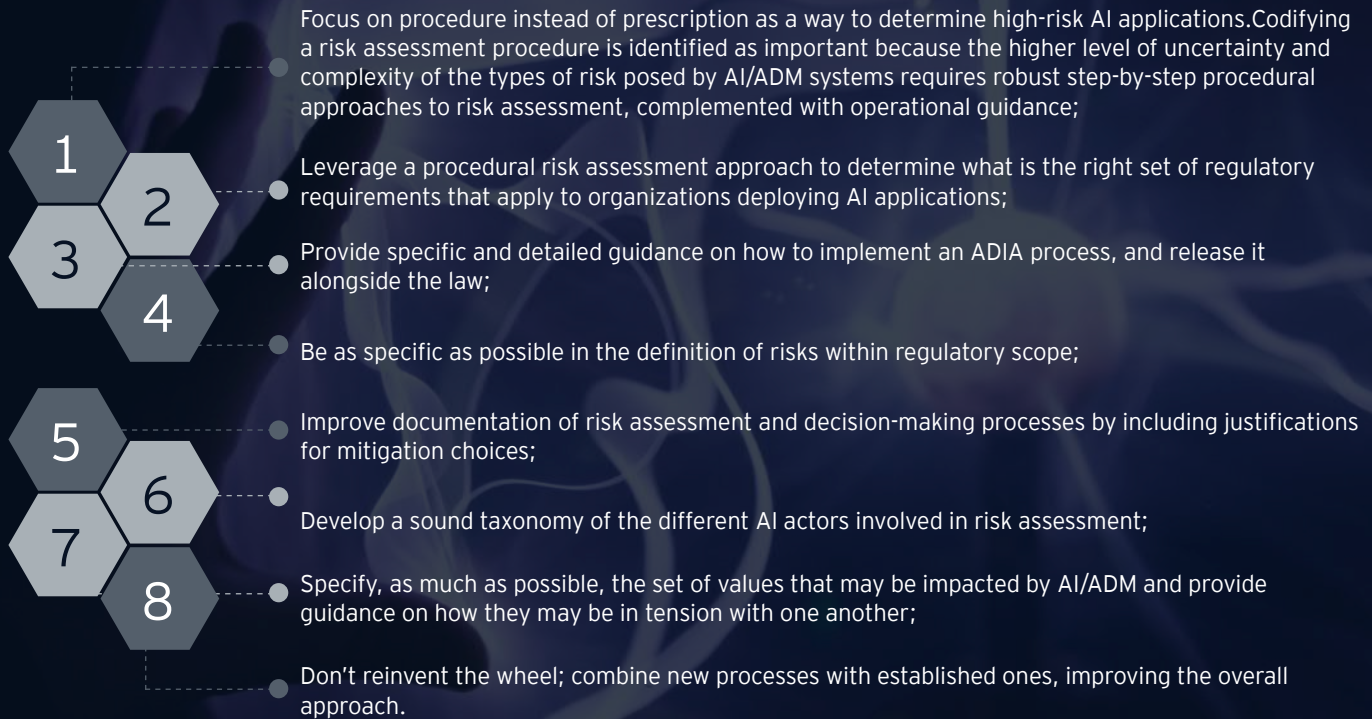
AI are just some of the topics that need attention. Transparency of algorithms and increasing cyber risks are also issues the risk manager will need to address. The paper aims to guide risk managers on applying AI. It goes from basic learning to identifying new risks and developing their own strategy regarding implementation of AI. To manage AI risks in a secure, vigilant and resilient manner, organizations need to analyse their risk profile through the components of their risk management framework. The 2017 ERM Framework illustrates how to use a risk management framework to capture and follow the new risks created by AI. It has five main components: Governance & culture; Strategy & objective-setting; Performance; Review & revision; Information communication & reporting. Under each heading, FERMA makes recommendations. In major risk categories, FERMA identifies 14 themes and processes as well as 26 key, AI-related risk issues, including, for example, environmental impacts. It has a table outlining the risk management process. It promulgates an AI roadmap as a dynamic process. FERMA describes each of the steps in the text and offers a template for the user's AI risk management roadmap.

5.3 Open Loop

Open Loop brings together policymakers and technology companies to help develop policies around AI and other emerging technologies. Initiated and supported by Facebook, Open Loop (www.openloop.org) builds on the collaboration of regulators, governments, tech businesses, academics and civil society representatives. Open Loop members co-create policy prototypes and test new and different approaches to laws and

regulations before they are enacted. Open Loop has produced a report⁴¹ presenting the findings and recommendations of the Open Loop policy prototyping program on Automated Decision Impact Assessment (ADIA). The policy prototype has two parts: a prototype law (drafted as legal text) and prototype guidance (drafted as a playbook).

Based on the results of the prototyping exercise and the feedback on the prototype law and playbook, the report advises lawmakers formulating requirements for AI risk assessments to consider the following recommendations:



5.4 Partnership on AI

The Partnership on AI conducts research on AI technologies including machine perception, learning and automated reasoning.⁴² It has an open platform for discussion and engagement about AI and its influences on people and society. AI researchers from Apple, Amazon, DeepMind and Google, Facebook, IBM and Microsoft created the Partnership in late 2016 and were joined in 2017 by six not-for-profit Board members. It now has 100+ member organizations.

The Partnership has published various blogs, papers and reports on topics such as Framework for Promoting Workforce

Well-being in the AI-Integrated Workplace, AI and Shared Prosperity Initiative, The Role of Demographic Data in Addressing Algorithmic Bias, Publication Norms for Responsible AI, Bringing Facial Recognition Systems to Light, Closing Gaps In Responsible AI, About ML – Annotation and Benchmarking on Understanding and Transparency of Machine learning Life cycles. While many of these documents are relevant to AI risk assessment and AI risk management, none of them is an AIRA methodology per se. Nevertheless, they offer useful guidance for those undertaking AIRAs.

The Software Alliance, also known as BSA, is a trade group created by Microsoft in 1988 to represent commercial software makers. The Software Alliance says an AI risk management process should include two key elements:

The BSA recommends that AI risk management processes should include two key elements:

- 1** A governance framework to support the organization's risk management functions.
- 2** A scalable process for performing an impact assessment to identify and mitigate risks.

In support of the process, it has produced a document "Confronting Bias: BSA's Framework to Build Trust in AI",⁴³ which sets forth an AI Bias Risk Management Framework that organizations can use to perform impact assessments to identify and mitigate risks of bias that may emerge throughout an AI system's life cycle. The Framework is organized around the design, development and deployment phases involved in the creation and use of an AI system. It identifies different types of biases that can creep into each phase. In the design phase, comprising project

conception and data acquisition, it spots problem formulation, historical, sampling and labelling biases. In the development phase, comprising data preparation, model definition and validation, it identifies proxy and aggregation biases. In the deployment phase, it cautions against deployment and misuse biases.

The BSA document positions AI impact assessments as an important assurance mechanism that promotes accountability and enhances trust that high-risk AI systems have been designed, developed, tested and deployed with sufficient protections in place to mitigate the risk of harm. Furthermore, AI impact assessments are viewed as an important transparency mechanism that enable stakeholders involved in the design, development and deployment of an AI system to communicate about its risks and ensure that responsibilities for mitigating those risks are clearly understood.

Although BSA highlights the importance of AIIA, it also cautions that it believes that it is impossible to eliminate bias from AI systems because there is no universally agreed upon method for evaluating whether a system is operating in a manner that is "fair".

The US-based Ethics & Compliance Initiative (ECI) was established as a best practice community of organizations committed to creating and sustaining high quality ethics and compliance programs. ECI has a mixed membership comprising companies, universities and non-governmental organizations.

While there are lots of ethics-related initiatives, ECI is of interest for AI risk assessment, in particular, for its paper with a section on the Ethical Implications of AI.⁴⁴ The paper has a table with five principles that should be applied when considering the ethical implications of a new AI development, which are Transparency, Business Strategy, Trust, Privacy and Security, and Values and Social Impact. For each of the principles, it lists considerations,

risks, impacts and enablers. The subsequent pages of its paper elaborate each of these, in the course of which it makes two important points:

- ▶ Application of AI can produce unintended real or perceived consequences.
- ▶ There may be industry-specific or application-specific considerations. For example, AI application to personnel systems, law enforcement systems or commercial product manufacturing may require different transparency practices. Such practices should be addressed as principles, policies and procedures, training and communications are developed.

It advises reaching out to internal stakeholders. Product developers, programmers, engineers and data scientists are advised to get to know their compliance, ethics, human resources, IT infrastructure/security, operations, sales, marketing and legal teams—and vice versa. It also encourages organizations to provide forums and oversight for impact discussions. For example, there may be a need for public discourse on dislocation of markets. It says that transparency needs differ by stakeholder group and the way AI is being applied. It asks how an AI system might impact employees, customers, industry and society.

AI risk assessments require a multi-disciplinary approach with input across organizational functions involving: product developers, programmers, engineers, data scientists, compliance, ethics, human resources, IT infrastructure/security, operations, sales, marketing and legal teams

5.7 Healthcare Insurance Reciprocal of Canada (HIROC)

Toronto-based HIROC is a healthcare safety advisor, offering a range of insurance products and support throughout an individual’s healthcare claim. It has produced a document on “Artificial intelligence: Risk management in health care”⁴⁵, which provides guidance in the development and implementation of AI applications in healthcare. The paper focuses on the identification, management and mitigation of risks. It makes the point that “a careful and deliberate assessment of the risks must be taken prior to the start of each new [AI] project”.

HIROC believes that “a careful and deliberate assessment of the risks must be taken prior to the start of each new [AI] project” – the assessment should cover ethical, performance, governance, communication and security risks

It is intended to support healthcare organizations in establishing their own processes for identifying, initiating, prioritising, overseeing and governing AI-based projects in which they participate.

The paper identifies several broad categories of risk – ethical, performance, governance, communication and security risks. It includes guiding principles to support organizations with the oversight and governance of AI applications. HIROC provides organizations with a framework for developing and managing AI strategies in alignment with their organizational objectives.

The report includes questions and recommendations to help the assessor regarding value proposition, governance, methodology, transparency, data integrity, privacy and security.

Such questions can serve as a model or template in an AIRA.

5.8 EY and The Future Society

EY and The Future Society produced a survey report on trust gaps between policymakers and companies.⁴⁶ The 2019/2020 survey of 71 policymakers and 284 companies across 55 countries

found that companies are “misaligned” with policymakers on key ethical AI risk issues: market or competitive risk, reputational risk, compliance risk and legal risk.



A woman with long dark hair and glasses is looking down at a smartphone she is holding in her hands. The background is dark and slightly blurred.

Approaches proposed in civil-society and academic literature

6

6.1

Overview

Various researchers from academia and civil society organizations have proposed assessment frameworks for AI systems to mitigate the negative impacts of AI on individuals, the society and the environment. Most assessment frameworks proposed by researchers draw from various predecessors, including environmental impact assessment (EIA), data protection impact assessment (DPIA) or privacy impact assessment (PIA), ethical impact assessment (EtIA), human rights impact assessment and/or socio-economic impact assessment (SEIA). More specifically, a number of the assessments reviewed are inspired by DPIA or PIA. However, DPIAs and PIAs are mainly focused on protection of personal data and do not engage with the broader range of impacts of AI on individuals, society and the environment, such as issues of potential biases and discrimination or surveillance. The need to engage with these broader issues and to mitigate these risks is to a large extent behind the development of these assessment frameworks.

We have focused our analysis on 11 frameworks that were identified as most relevant to the assessment of risks and/or impacts that the deployment of AI may bring about.

	Source	Name of the assessment	Year
1	Ada Lovelace Institute and DataKind UK (ALIDUK)	Discusses four types: <ul style="list-style-type: none"> ▸ Bias audit ▸ Regulatory inspection ▸ Algorithmic risk assessment ▸ Algorithmic impact evaluation 	2020
2	Calvo et al.	Human impact assessment for intelligent systems	2020
3	Diakopoulos et al.	Social impact statement for algorithms	Undated
4	Kaminski and Malgieri	Algorithmic impact assessments	2020
5	Mantelero	Human rights, social and ethical impact assessment for AI and big data	2018
6	Mantelero and Esposito	Human rights impact assessment in the development of AI data-intensive systems	2021
7	Platform for the Information Society (ECP)	AI impact assessment	2018
8	Reisman et al.	Algorithmic impact assessment	2018
9	Schiff et al.	Impact assessment for responsible AI	2020
10	Selbst	Algorithmic impact statement	2017
11	Zicari et al.	Assessment of AI systems' trustworthiness (Z-Inspection)	2021

The report by the Ada Lovelace Institute and Data Kind UK (ALIDUK) published in 2020 defines various tools for assessing algorithmic systems and presents the state of art of these tools.⁴⁷ It distinguishes between audits and impact assessments.

The Calco et al. 2020 article promotes the development and use of human impact assessment for AI systems, drawing in particular from previous environmental impact assessments and data protection impact assessments.

The Diakopoulos et al. framework was developed by 12 researchers of the Fairness, Accountability and Transparency in Machine Learning community.⁴⁸ Its main aim is to “help developers and product managers design and implement algorithmic systems in publicly accountable ways”. It proposes a social impact statement for algorithms to ensure compliance with the five principles for accountable algorithms that they developed, i.e., responsibility, explainability, accuracy, auditability and fairness.

The Kaminiki and Malgieri 2020 article proposes an algorithmic impact assessment built on the GDPR’s Data Protection Impact Assessment. It highlights how such impact assessment may serve both as a governance tool and a safeguard of individual rights.⁴⁹

Alessandro Mantelero published two papers specifically on impact assessment related to AI highlighting the central role of human rights for such assessment. The 2018 paper presents “a blueprint for a human rights, social, and ethical impact assessment”, which is a self-assessment model for AI centered on human rights.⁵⁰ The 2021 article by Mantelero and Maria Samantha Esposito proposes a methodology and a model for human rights impact assessment.⁵¹

The Platform for the Information Society (ECP) is a Dutch “independent platform where government, business community and social institutions come together to exchange knowledge and cooperate to exploit the opportunities that the information

society provides, and to mitigate the threats posed by the online world” (<https://all-digital.org/ecp/>). In 2018, the ECP published a report “Artificial Intelligence Impact Assessment” that proposes an assessment process to identify ethical and legal consequences of AI systems.⁵²

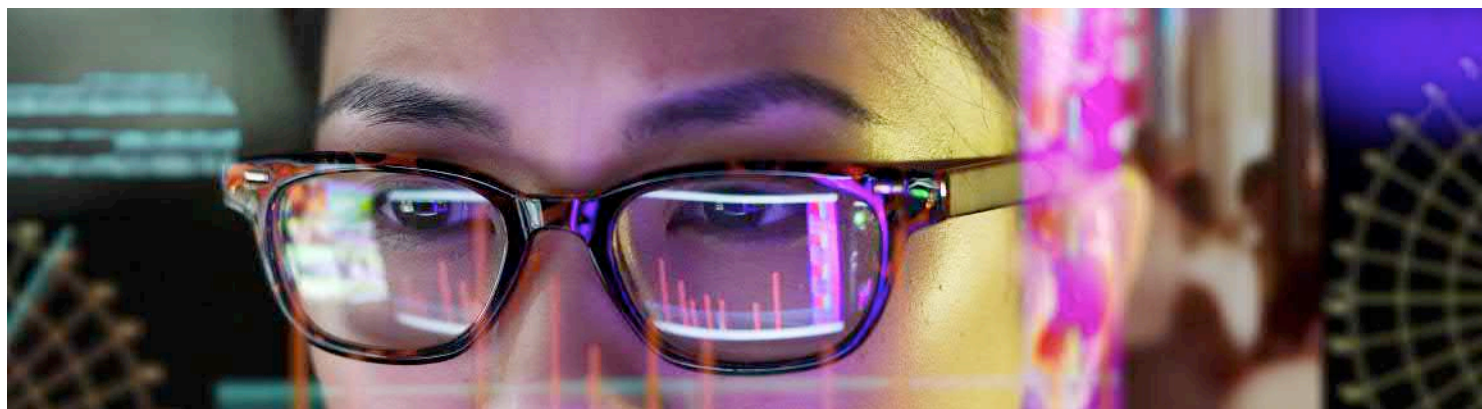
The 2018 report by Reisman et al, Jason Schultz, Kate Crawford and Meredith Whittaker from the AI Now Institute proposes a process for an algorithmic impact assessment to ensure public agency accountability. The report presents both the process from pre-acquisition review to regular review during use and the content of the assessment.⁵³

The 2020 article “Principles to Practices for Responsible AI: Closing the Gap” by Schiff et al highlights the value of conducting impact assessments to ensure implementation of principles for Responsible AI into practice.⁵⁴ Their framework builds on the IEEE 7010 standard.

The 2017 Andrew D. Selbst article, “Disparate impact in big data policing”,⁵⁵ developed the impact assessment proposal as part of a study on discriminatory impacts of the use of big data in policing. It is built on environmental impact statements and, although it is developed specifically for law enforcement agencies, the author notes that this impact assessment process can be used more broadly to assess AI systems used in other sectors.

The Zicari et al 2021 “Z-Inspection” framework proposes a process to assess the trustworthiness of AI. It builds upon the latest developments at the EU-level toward trustworthy AI (see section 2 of the present report).

Most of the assessment frameworks analysed in this section have been proposed, but to the best of our knowledge, not actually used. An exception is the framework by Reisman et al. on which the Canadian government based its Algorithmic Impact Assessment (see section 2 of the present report). The field of impact assessments for AI is therefore still largely in development.



Four key aspects were identified in the frameworks proposed by academic and civil-society researchers:



1

The stage at which the assessment is proposed to take place, pre-deployment (e.g., during design) vs. post-deployment potentially on an ongoing or recurring basis.

2

Who is expected to conduct the risk and/or impact assessment. Should it be conducted by data controllers, by auditors (either internal or external to organization), by the developers of the algorithm themselves, by the agency/organization seeking to procure an AI system.

3

Whether the results of the assessment proposed are to be made public, and if this should be conditional on factors related to the “legitimate interests of the data controller, such as confidentiality of information, security and competition”

4

Whether these frameworks include a “no-go” clause, i.e., a clause according to which, if the results of the assessment indicate that the AI system is too risky or that risks cannot be mitigated, then the AI system should not be developed, procured and/or used.

Conclusions

7

From our survey, we found different approaches to artificial intelligence risk assessment, although many of them share commonalities in processes and principles. We reviewed approaches in Europe, the US, China and other countries as well as the AI risk assessment approaches of international organizations, standards bodies, industry and researchers.

7.1

Key findings

AI risk assessment approaches were identified as fundamentally belonging to two distinct types. The first assesses risks arising from the use of AI. The second type classifies AI systems or applications by risk.

Both approaches build on antecedents – from environmental impact assessments, privacy impact assessments, ethical impact assessments and human rights assessments, among others. For example, in Europe, the Aarhus Convention⁵⁶ obliges organizations to conduct an EIA where there is a risk of environmental damage. It also entitles the public to be informed and consulted in an assessment. PIA methodologies, pioneered in New Zealand, Canada and Australia, also called for independent assessors, consultation with the public and publication of the assessment (either the whole assessment or a summary). US government agencies publish a list of the PIAs on their websites. AI risk assessments may have their closest analogue with ethical impact assessments (EtiAs), which consider a wider range of issues than privacy and data protection – they address issues such as bias, discrimination, surveillance, dignity, power asymmetries, which are all relevant in an AIRA. Like PIAs and human rights impact assessments, EtiAs use questions as a means of assessing risks, a practice that is prevalent in AIRA as well.

Issues that have arisen in these antecedents are germane to AIRA as they make the assessment more or less ambitious and open to scrutiny. These include:

The stage at which the assessment is proposed to take place (at design stage, before procurement and/or during use) – While a risk assessment that begins at the design stage and follows through to development and deployment will be more expensive than one that focuses only on deployment, it will be more rigorous and credible.

The intended user of the assessment (and potential conflicts of interest that need to be avoided) – For the assessment to be credible, the assessor should not be beholden to the user. Even within the same organization, the assessor should have a mandate from senior management to assess the risks and not to whitewash them.

The release of the assessment results to the public – While making the results public enables independent scrutiny, organizations will need to consider commercial and competitive sensitivities. An alternative would be to make a summary of the assessment public or set out the reasons to the regulator for not making the assessment public.

Whether the framework includes a “no go” clause in case risks are assessed to be too high – Senior management will need to decide if a particular artificial intelligence scheme is really worth pursuing if it will result in bad publicity and reputational damage.

7.2 Leading practices

Based on the survey presented in this report we have identified a number of leading practices, supported by two or more of the stakeholders that were cited in this study. These leading practices provide a good starting point for risk based AI regulation, but

any framework will need to be tailored to fit into the legislative tapestry of the jurisdiction where it is applied. In doing so, special consideration needs to be given to the regulatory ecosystem that applies to the AI related digital infrastructure of data, privacy legislation, and information security, among others.

7.2.0 – Categorization of Risk

1 Categorize risk based on the application for which the AI is used, rather than technical features alone, assessing both:

- the purpose of the overall system
- the function performed by the AI component in the system.

2 Risk dimensions include: ethical, performance, governance, communication and security risks

3 Assess risk across multiple aspects, including: impact on fundamental rights of persons (including risk to health and safety), extent of use, intended purpose, number of affected persons, availability of (human) alternatives, irreversibility of harm, and the extent to which law provides prevention and/or mitigation measures.

4 Categorize risk into multiple levels (three to more) to enable proportional regulatory intervention that distinguishes between unacceptable uses, high-risk applications, limited/low and minimal risk applications.

7.2.1 – Risk management

- 1 Risk management systems for AI should include:
 - a. identification and analysis of the known and foreseeable risks,
 - b. estimation and evaluation of the risks that may emerge and
 - c. evaluation of other risks in the post-market monitoring phase adoption of suitable risk management measures.

7.2.2 – Requirements for Trustworthiness of AI

- 1 Trustworthiness requirements include:
 - a. lawfulness, complying with all applicable laws and regulations;
 - b. ethical, ensuring adherence to ethical principles and values;
 - c. robust, both from a technical and social perspective to avoid unintentional harm.

- 2 In all phases of its life cycle (design, development, and use) the AI system must:
 - a. operate in a manner that is lawful, fair, unbiased, accurate, reliable, effective, safe, secure, resilient, understandable,
 - b. have processes in place to regularly monitor and evaluate the AI system's performance and outcomes.

- 3 Properties that should be tested for include reliability, accountability, maintainability, functionality, debug-ability, evolve-ability, fragility, vulnerability, privacy, transparency and bias.

7.2.3 – Relevant stakeholders for identifying and mitigating AI risk

- 1 Given the scale of the challenges associated with AI, the mobilisation of a diverse set of participants, including businesses, consumer organizations, trade unions and other representatives of civil society bodies is essential.

The broad diversity of methods and applications where AI is being deployed makes it unlikely that cross-sectoral AI legislation will be able to address all the nuances that a good regulatory

environment requires. It is therefore anticipated that such legislation will have to be augmented with sector-specific approaches that can address specialized AI risk factors.



Endnotes

7.3

Abbreviations used in this survey

A/IS	Autonomous and intelligent systems
ADIA	Automated Decision Impact Assessment
AGI	Artificial general intelligence
AI	Artificial intelligence
AIDP	New Generation Artificial Intelligence Development Plan
AI HLEG	High-Level Expert Group on Artificial Intelligence
AIRA	Artificial intelligence risk assessment
AIRS	Artificial Intelligence/Machine Learning Risk & Security Working Group
AIIA	Artificial intelligence impact assessment
AIRM	Artificial intelligence risk management
ALTAI	Assessment List for Trustworthy Artificial Intelligence
AIA	Algorithmic Impact Assessment
BSA	Business Software Alliance
CEN	European Committee for Standardization
CENELEC	European Committee for Electrotechnical Standardization
CD	Committee Draft
DHS	Department of Homeland Security
DPIA	Data protection impact assessment
EC	European Commission
ECI	Ethics & Compliance Initiative
ECOACT	Equal Credit Opportunity Act
EIA	Environmental impact assessment

ERM	Enterprise Risk Management
EtIA	Ethics impact assessment
EU	European Union
FCC	Federal Communications Commission
FCRA	Fair Credit Reporting Act
FDA	Food & Drug Administration
FERMA	Federation of European Risk Management Associations
FTC	Federal Trade Commission
GDPR	General Data Protection Regulation
GRC	Governance, risk and compliance
HLEG-AI	High Level Expert Group on Artificial Intelligence
HRIA	Human rights impact assessment
IEC	International Electrotechnical Commission
IEEE SA	Institute of Electrical and Electronics Engineers Standards Association
ISO	International Organization for Standardization
JTC	Joint Technical Committee
ML	Machine Learning
NGO	Non-Governmental Organization
NITI Aayog	National Institution for Transforming India
NSAI	National Strategy for Artificial Intelligence
OECD	Organization for Economic Co-operation and Development
PIA	Privacy Impact Assessment
SEIA	Socio-economic impact assessment
UNESCO	United Nations Educational, Scientific and Cultural Organization
WEF	World Economic Forum
WIA	Well-being impact assessment

Endnotes

- 1 <https://oecd.ai/en/ai-principles>
- 2 Hidvegi, Fanny, Daniel Leufer and Estelle Massé, “The EU should regulate AI on the basis of rights, not risks”, Access Now, 17 Feb 2021. [https:// www.accessnow.org/eu-regulation-ai-risk-based-approach/](https://www.accessnow.org/eu-regulation-ai-risk-based-approach/).
- 3 European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, Brussels, COM(2021) 206 final, 21 Apr 2021. <https://digital-strategy.ec.europa.eu/en/policies/strategy-artificial-intelligence>
- 4 Federal Government of Germany, Artificial Intelligence Strategy, 2018. <https://www.ki-strategie-deutschland.de/home.html>
- 5 Federal Government of Germany, Artificial Intelligence Strategy, 2020 Update. https://www.ki-strategie-deutschland.de/files/downloads/Fortschreibung_KI-Strategie_engl.pdf
- 6 OECD, State of Implementation of the OECD AI Principles: Insights from National AI Policies, OECD Digital Economy Papers, No. 311, June 2021. https://www.oecd-ilibrary.org/science-and-technology/state-of-implementation-of-the-oecd-ai-principles_1cd40c44-en
- 7 European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, Brussels, COM(2021) 206 final, 21 Apr 2021, p. 7.
- 8 Ibid, recital 78.
- 9 Ibid., recital 46.
- 10 Ibid, recital 81.
- 11 Independent High-Level Expert Group on Artificial Intelligence (AI HLEG) set up by the European Commission, Ethics Guidelines for Trustworthy AI, European Commission, 8 Apr 2019. <https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai>
- 12 Independent High-Level Expert Group on Artificial Intelligence (AI HLEG) set up by the European Commission, The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, European Commission, July 2020. <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>
- 13 <https://www.ai.gov/naiio-directors-corner-may-05-2021/>
- 14 US Congress, A Bill to direct the Federal Trade Commission to require entities that use, store, or share personal information to conduct automated decision system impact assessments and data protection impact assessments (Algorithmic Accountability Act of 2019), H.R.2231, introduced in the House 10 April 2019. <https://www.congress.gov/bill/116th-congress/house-bill/2231/text>.
- 15 US Department of Homeland Security and Office of the Director of National Intelligence, AI: Using standards to mitigate risks, 2018.
https://www.dhs.gov/sites/default/files/publications/2018_AEP_Artificial_Intelligence.pdf

- 16 Government of Canada, Algorithmic Impact Assessment (AIA), 2020.
<https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html>. See also Government of Canada, AIA self-assessment questionnaire, 2020. <https://canada-ca.github.io/aia-eia-js/>. Corriveau, Noel, The Government of Canada's Algorithmic Impact Assessment: Towards Safer and More Responsible AI, 2018.
https://aiforsocialgood.github.io/2018/pdfs/track2/83_aisg_neurips2018.pdf
- 17 <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html>
- 18 BAAI, "Beijing AI Principles", 2019. <https://www.baai.ac.cn/news/beijing-ai-principles-en.html>
- 19 Smart National Singapore, "National AI Strategy: The next key frontier of Singapore's Smart Nation Journey", 2021. <https://www.smartnation.gov.sg/why-Smart-Nation/NationalAIStrategy>
- 20 Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations, Jan 2020. <https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGIsago.pdf>
- 21 NITI Aayog, "National Strategy for Artificial Intelligence – #AI for all", Discussion Paper, 2018. <https://www.niti.gov.in/national-strategy-artificial-intelligence>
- 22 The text on national AI strategies has been extracted from Van Roy, V., F. Rossetti, K. Perset and L. Galindo-Romero, AI Watch – National strategies on Artificial Intelligence: A European perspective, Publications Office of the European Union, Luxembourg, 2021. <https://op.europa.eu/en/publication-detail/-/publication/619fd0b5-d3ca-11eb-895a-01aa75ed71a1/language-en/format-PDF/source-search>
- 23 <https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64>
- 24 Global Partnership on AI (GPAI), Responsible AI Working Group Report GPAI – Montreal Summit 2020, Executive Summary, Nov 2020. <https://www.gpai.ai/projects/responsible-ai/gpai-responsible-ai-wg-report-november-2020-summary.pdf>
- 25 Organization for Economic Co-operation and Development (OECD), Recommendation of the Council on Artificial Intelligence, Paris, Adopted 22 May 2019.
<https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>
- 26 Organization for Economic Co-operation and Development (OECD), OECD Framework for the classification of AI systems, May 2021. <https://oecd.ai/classification>
- 27 UNESCO, Recommendations on the ethics of Artificial Intelligence, SHS/BIO/AHEG-AI/2020/4 REV.2, 7 Sept. 2020. <https://unesdoc.unesco.org/ark:/48223/pf0000373434>
- 28 World Economic Forum (WEF), AI Procurement in a Box: AI Government Procurement Guidelines, June 2020, p. 5. <https://www.weforum.org/reports/ai-procurement-in-a-box>.
- 29 World Economic Forum (WEF), Reimagining Regulation for the Age of AI: New Zealand Pilot Project, White Paper, June 2020. <https://www.weforum.org/whitepapers/reimagining-regulation-for-the-age-of-ai-new-zealand-pilot-project>
- 30 World Economic Forum (WEF), Framework for Responsible Limits on Facial Recognition Use Case: Flow management, White Paper, 2020.
http://www3.weforum.org/docs/WEF_Framework_for_action_Facial_recognition_2020.pdf.

- 31 ISO/IEC 23894 – Information Technology – Artificial Intelligence – Risk Management. <https://www.iso.org/standard/77304.html>
- 32 ISO/IEC AWI 42001 Information Technology – Artificial intelligence – Management system. <https://www.iso.org/standard/81230.html>
- 33 ISO/IEC DIS 38507 – Information technology – Governance of IT – Governance implications of the use of artificial intelligence by organizations. <https://www.iso.org/standard/56641.html>
- 34 IEEE P2863 – Recommended Practice for Organizational Governance of Artificial Intelligence. <https://standards.ieee.org/project/2863.html>
- 35 IEEE 7000-2021 – IEEE Model Process for Addressing Ethical Concerns During System Design. <https://standards.ieee.org/standard/7000-2021.html>
- 36 IEEE 7010-2020 – IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being. <https://standards.ieee.org/standard/7010-2020.html>
- 37 <https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8332-draft.pdf>
- 38 <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270-draft.pdf>
- 39 Artificial Intelligence/Machine Learning Risk & Security Working Group (AIRS), How Can Financial Institutions Prepare for AI Risks?, Wharton AI for Business, 13 Apr 2021.
- 40 FERMA, Artificial Intelligence applied to risk management, [2020?].
<https://www.ferma.eu/app/uploads/2019/11/FERMA-AI-applied-to-RM-FINAL.pdf>
- 41 Norberto Nuno Gomes de Andrade and Verena Kontschieder, AI Impact Assessment: A Policy Prototyping Experiment, 2021. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3772500
- 42 <https://www.partnershiponai.org/about/>
- 43 BSA | The Software Alliance, Confronting Bias: BSA's Framework to Build Trust in AI, 2021. <https://ai.bsa.org/wp-content/uploads/2021/06/2021bsaibias.pdf>
- 44 Ethics & Compliance Initiative (ECI) AI Working Group, Navigating Artificial Intelligence to Optimize Risk Management and Minimize Ethical Implications, ECI, 2020. https://www.ethics.org/whitepaper-optimizing-risk-management-using-artificial-intelligence/#lp_ai_form
- 45 HIROC, Artificial intelligence: Risk management in health care, August 2020.
https://www.hiroc.com/system/files/resource/files/2020-10/Artificial%20Intelligence%20-%20Risk%20Management%20in%20Healthcare_2.pdf
- 46 EY & The Future Society, Bridging AI's trust gaps: Aligning policymakers and companies, July 2020. https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/ai/ey-bridging-ais-trust-gaps-report.pdf
- 47 Ada Lovelace Institute, and DataKind UK, "Examining the Black Box. Tools for Assessing Algorithmic Systems", London, UK, 2020. <https://www.adalovelaceinstitute.org/case-study/examining-the-black-box/>
- 48 Diakopoulos, Nicholas, Sorelle Friedler, Marcelo Arenas, et al, "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms", FAT/ML, [2018?]. <https://www.fatml.org/resources/principles-for-accountable-algorithms>

- 49 Kaminski, Margot E., and Gianclaudio Malgieri, "Algorithmic Impact Assessments under the GDPR: Producing Multi-Layered Explanations", *International Data Privacy Law*, Vol. 00, no. 0, 2020, pp. 1-20. <https://academic.oup.com/idpl/advance-article/doi/10.1093/idpl/ipaa020/6024963>
- 50 Mantelero, Alessandro, "AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment", *Computer Law & Security Review*, Vol. 34, Issue 4, 2018, pp. 754-72. <https://www.sciencedirect.com/science/article/pii/S0267364918302012>
- 51 Mantelero, Alessandro, and Maria Samantha Esposito, "An Evidence-Based Methodology for Human Rights Impact Assessment (HRIA) in the Development of AI Data-Intensive Systems", *Computer Law & Security Review*, Vol. 41, July 2021, pp. 1-35. <https://www.sciencedirect.com/science/article/pii/S0267364921000340>
- 52 Platform for the Information Society (ECP), "Artificial Intelligence Impact Assessment", 2018. <https://ecp.nl/wp-content/uploads/2019/01/Artificial-Intelligence-Impact-Assessment-English.pdf>
- 53 Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker, "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability", AI Now Institute, April 2018. <https://ainowinstitute.org/aiareport2018.pdf>
- 54 Schiff, Daniel, Bogdana Rakova, Aladdin Ayes, Anat Fanti and Michael Lennon, "Principles to Practices for Responsible AI: Closing the Gap", ArXiv, 8 June 2020. <https://arxiv.org/abs/2006.04707>
- 55 Selbst, Andrew D., "Disparate Impact in Big Data Policing", *Georgia Law Review*, Vol. 52, no. 1, 2017, pp. 109-95. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2819182
- 56 <https://ec.europa.eu/environment/aarhus/>





Acknowledgements

Authors

Gabriella Ezeani

Research Analyst, Trilateral Research

Dr. Ansgar Koene

Associate Director, Ernst & Young LLP (UK), and EY Global AI Ethics and Regulatory Leader

Richa Kumar

Research Analyst, Trilateral Research

Nicole Santiago

Senior Research Analyst, Trilateral Research

David Wright

Director, Trilateral Research

Contributors

Douglas Bell

Director, Ernst & Young LLP (US), and EY Global Public Policy

Cathy Cobey

Partner, Ernst & Young (Canada), and EY Global Trusted AI Leader

Nigel Duffy

Partner, Ernst & Young LLP (US), and EY Global Artificial Intelligence Leader

Nina Emordi

Associate Director, Ernst & Young LLP (UK), and EY Global Public Policy

Eugène Goyne

Associate Partner, Ernst & Young LLP (Hong Kong), and EY Asia Pacific Financial Services Regulatory Lead

Josh B Konvisser

Partner, Ernst & Young LLP (US) and EY Americas General Counsel Office

Katie Kummer

Partner, Ernst & Young LLP (US), and EY Global Deputy Vice Chair, Public Policy

Ivan Mannino

Senior Associate, Ernst & Young (Belgium), and EMEIA Public Policy

Anaïs Resseguier

Senior Research Analyst, Trilateral Research

Fatima Hassan-Szlamka

Associate Director, Ernst & Young LLP (US), and EY Global Public Policy

For further information, please contact:

Dr. Ansgar Koene, EY, Global AI Ethics and Regulatory Leader

ansgar.koene@uk.ey.com

About this report

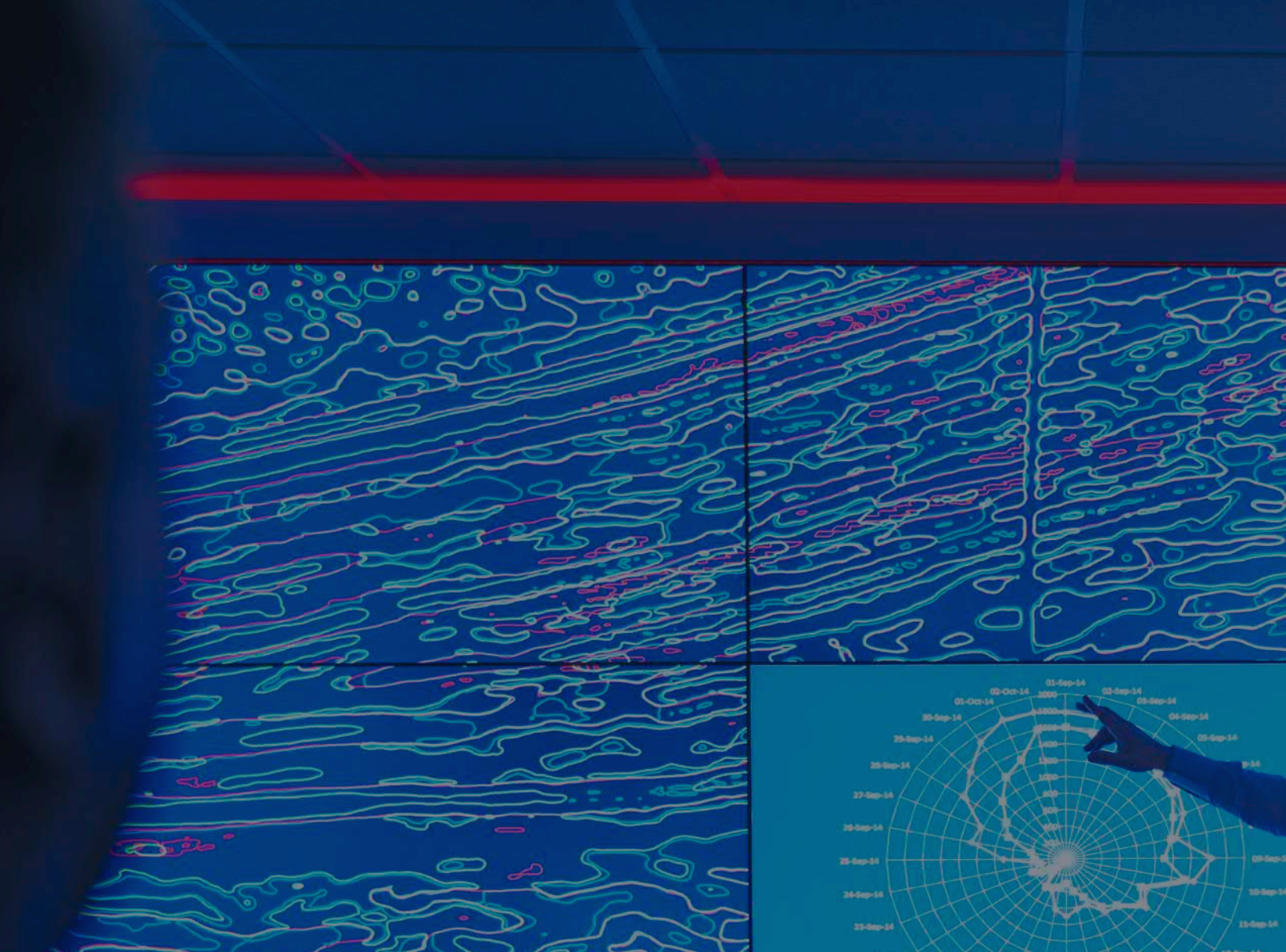
8

Trilateral Research worked in close co-operation with EY teams to produce this survey of artificial intelligence risk assessments to enable organizations across many sectors to mitigate the risks and respond to the challenges raised by using AI today. The Trilateral team drew on its pioneering work in privacy, data governance and impact assessments, notably the first book on privacy impact assessment (Wright, 2008); the first article on ethical impact assessment (Wright, 2010) and the first article on surveillance impact assessment (Wright & Raab, 2012). For the present survey, Trilateral benefited from EY's Ansgar Koene's advice and suggestions.

The authors acknowledge with thanks the inputs and comments from Trilateral colleague Anaïs Resseguier as well as our collaborators from Ernst & Young.

Trilateral, founded in 2004, is a research, consulting and AI-technology development company. Trilateral's teams include social scientists, data scientists, ethical, legal and human rights experts who work across the technology-social disciplinary divide. They apply rigorous, cutting-edge research when developing and assessing new technologies to achieve sustainable innovation and measurable impact. Trilateral's work spans a wide range of clients, from the European Commission to the UK and Irish governments, law enforcement agencies, universities, hospitals and many others, focusing on enhancing societal well-being. Trilateral has partnered in or co-ordinated more than 70 EU-funded projects. It has offices in London and Waterford, Ireland, with staff based in 10 countries. Trilateral lives and breathes its motto of turning research into innovation.

For more information, visit the Trilateral website at <https://www.trilateralresearch.com/>



EY | Building a better working world

EY exists to build a better working world, helping to create long-term value for clients, people and society and build trust in the capital markets.

Enabled by data and technology, diverse EY teams in over 150 countries provide trust through assurance and help clients grow, transform and operate.

Working across assurance, consulting, law, strategy, tax and transactions, EY teams ask better questions to find new answers for the complex issues facing our world today.

EY refers to the global organization, and may refer to one or more, of the member firms of Ernst & Young Global Limited, each of which is a separate legal entity. Ernst & Young Global Limited, a UK company limited by guarantee, does not provide services to clients. Information about how EY collects and uses personal data and a description of the rights individuals have under data protection legislation are available via ey.com/privacy. EY member firms do not practice law where prohibited by local laws. For more information about our organization, please visit ey.com.

Ernst & Young LLP is a client-serving member firm of Ernst & Young Global Limited operating in the US.

© 2021 Ernst & Young LLP.
All Rights Reserved.

SCORE number. EYG no. 009797-21Gbl

2109-3873062
ED None

This material has been prepared for general informational purposes only and is not intended to be relied upon as accounting, tax, legal or other professional advice. Please refer to your advisors for specific advice.

ey.com