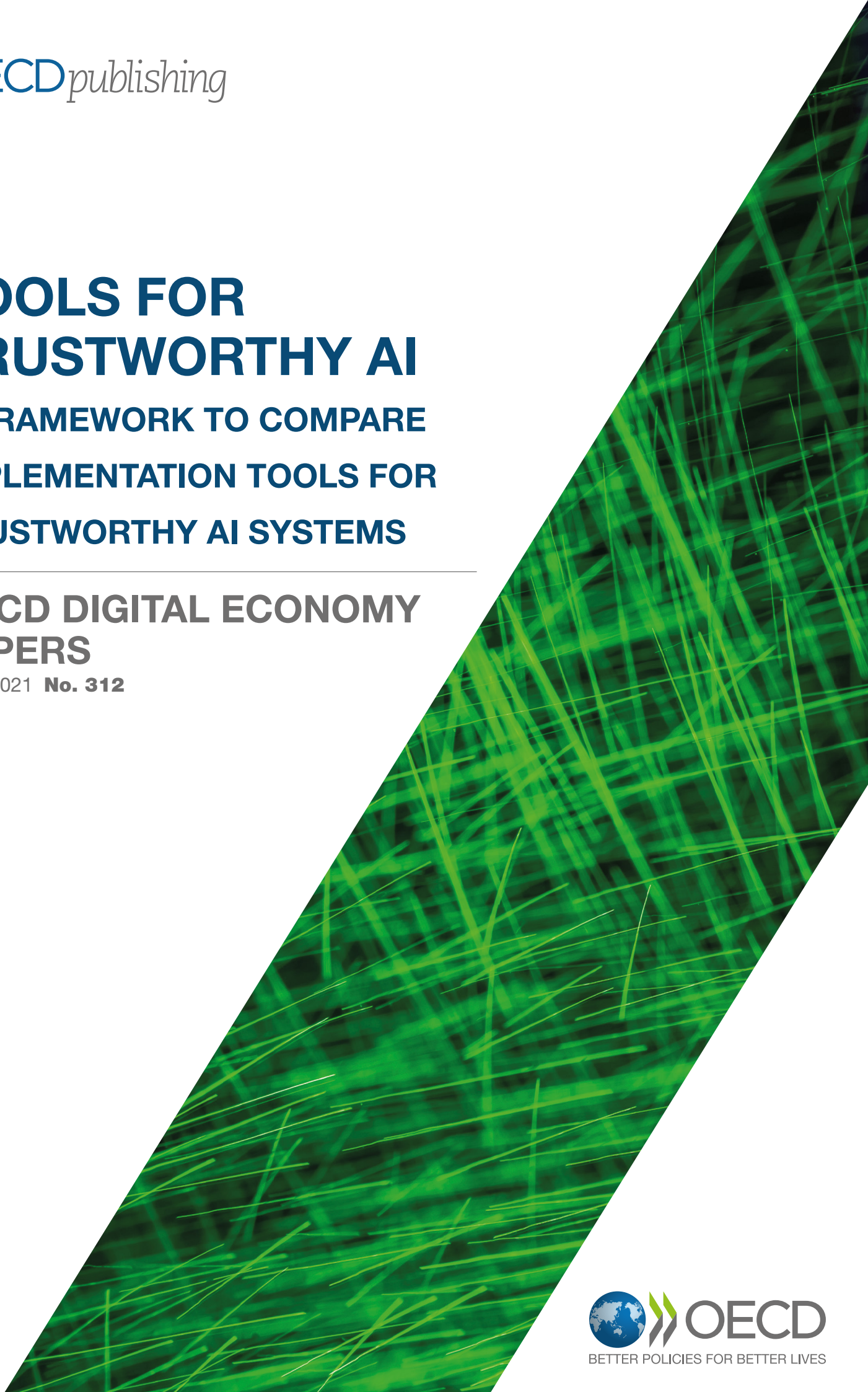
OECD publishing

TOOLS FOR TRUSTWORTHY AI

A FRAMEWORK TO COMPARE IMPLEMENTATION TOOLS FOR TRUSTWORTHY AI SYSTEMS

OECD DIGITAL ECONOMY PAPERS

June 2021 **No. 312**



OECD
BETTER POLICIES FOR BETTER LIVES

Foreword

This document presents the work conducted by the OECD Network of Experts on AI (ONE AI) working group on implementing Trustworthy AI to develop a framework for comparing tools and practices to implement trustworthy AI systems, as requested by the Committee on Digital Economy Policy (CDEP). The work was developed over ten virtual working group meetings. The group agreed on this draft at its tenth meeting on 5 May 2021.

This paper was approved and declassified by the CDEP on 15 April 2021 and prepared for publication by the OECD Secretariat. For more information, please visit www.oecd.ai.

Note to Delegations:

This document is also available on O.N.E under the reference code:

DSTI/CDEP(2020)14/FINAL

Table of contents

Tables

Figures

Boxes

Executive summary

AI policy discussions have moved from principles to implementation. As artificial intelligence (AI) advances across economies and societies, the technical, business, academic and policy stakeholder communities are actively exploring the best ways to encourage the design, development, deployment and use of AI that is human-centred and trustworthy, to maximise its benefits while minimising risks. The challenge is to ensure that the outcomes of AI systems promote shared wellbeing and prosperity while protecting individual rights and democratic values.

Efforts to implement trustworthy AI exist but are scattered. Many tools, instruments and structured methods to facilitate the implementation of the OECD AI Principles exist and are being developed to help AI actors – anyone playing an active role in the AI system lifecycle, including organisations and individuals that deploy or operate AI – navigate the challenges involved in building and deploying trustworthy AI. However, information about these tools is often sparse, hard to find and often detached from broader international policy discussions.

AI actors need a common framework to compare tools for trustworthy AI. To contribute to the wider adoption of the OECD AI Principles, it is essential to share the experiences and lessons learned of those who have already implemented them. This includes collecting and disseminating concrete tools, practices and approaches under a common framework that is accessible and allows for comparability.

The OECD framework of tools for trustworthy AI identifies relevant tools for developing, using and deploying trustworthy AI systems. Tools are classified according to systems' specific needs and contexts. While the framework is not designed to assess the quality or completeness of an individual tool, it does provide the means for comparing and analysing tools in different use contexts.

Based on the framework, a regularly updated database of tools for trustworthy AI will be built and made accessible to all via OECD.AI. The interactive database will provide AI actors and policy makers with information on the latest tools to help ensure that AI systems in different contexts abide by the principles of human rights and fairness; transparency and explainability; robustness, security, and safety and accountability.

Synthèse

Les discussions politiques autour de l'intelligence artificielle (IA) sont passées de l'énonciation de principes à leur mise en œuvre concrète. À mesure que l'IA progresse au sein des économies et des sociétés, les diverses parties prenantes techniques, commerciales, universitaires et politiques cherchent activement des moyens d'inciter la conception, le développement, le déploiement et l'utilisation d'une IA centrée sur l'humain et digne de confiance, afin d'en décupler les bénéfices tout en en minimisant les risques. Dans cette perspective, le défi principal consiste à faire en sorte que les résultats produits par les systèmes d'IA favorisent le bien-être sociétal et la prospérité générale tout en protégeant les droits individuels et les valeurs démocratiques.

Si un certain nombre d'initiatives ont déjà été mises en œuvre pour déployer une IA digne de confiance, celles-ci restent dispersées. En effet, de nombreux outils, instruments et méthodes de travail aidant à la mise en œuvre des Principes de l'OCDE sur l'IA sont disponibles ou en cours d'élaboration pour accompagner les acteurs de l'IA – soit tout agent jouant un rôle actif dans le cycle de vie du système d'IA dont les organisations et les individus qui déploient ou exploitent l'IA – dans les processus d'élaboration et de déploiement d'une IA digne de confiance. Cependant, l'information disponible sur ces instruments s'avère souvent rare, difficile d'accès et décorrélée des discussions politiques internationales plus générales qui s'y rapportent.

Les acteurs de l'IA ont donc besoin d'un cadre commun pour comparer et contextualiser les instruments qui existent pour promouvoir une IA digne de confiance. Pour contribuer à une adoption plus large des Principes de l'OCDE sur l'IA, il est en effet essentiel de partager les expériences et enseignements tirés par ceux qui ont déjà commencé à les appliquer. Il s'agit notamment de faire l'inventaire et de diffuser les instruments, pratiques et approches concrètes qui existent, tout en les analysant dans un cadre commun qui soit accessible et permette de les comparer.

Le cadre de l'OCDE pour les instruments au service d'une IA digne de confiance identifie des instruments pertinents pour le développement, l'utilisation et le déploiement de systèmes d'IA digne de confiance. Ces instruments sont répertoriés suivant les besoins et contextes spécifiques des systèmes d'IA considérés. Si ce cadre n'a pas été conçu pour évaluer la qualité ou le caractère exhaustif des instruments pris individuellement, il permet néanmoins de les comparer et de les analyser dans différents contextes d'utilisation.

En outre, ce cadre sert de structure à une base de données sur les instruments mis en œuvre pour une IA digne de confiance, qui sera développée et mise en accès public sur OECD.AI prochainement. Cette base de données interactive fournira aux acteurs de l'IA et aux décideurs politiques des informations sur les instruments les plus récents. Cette initiative vise notamment à contribuer à garantir que les systèmes d'IA respectent les droits de l'homme et les principes d'équité, de transparence et d'explicabilité, de robustesse, de sécurité, de sûreté et de responsabilité et ce dans différents contextes.

Introduction

This report provides a framework for comparing tools and practices to implement trustworthy AI systems as set out in the OECD AI Principles, *i.e.* AI systems that benefit people and planet; uphold human rights, democratic values and fairness; are transparent and explainable; robust, secure and safe; and whose operators are accountable (OECD, 2019a).

Many AI actors have been developing tools to help address the challenges of building AI systems that are trustworthy (Box 1). This framework collects, structures and shares the latest information and insights on tools and methods for implementing trustworthy AI. The framework addresses needs as they arise throughout all phases of the AI system lifecycle, for various systems and contexts. This means that the framework can also be used to compare tools in different use contexts.

Experts from all stakeholder groups participating in the OECD Network of Experts on AI (ONE AI) worked together to develop the framework from February 2020 to March 2021 (Annex B).

The framework will serve as the basis for developing an interactive, publicly available database on the OECD.AI Policy Observatory. The database will enable policy makers and practitioners to quickly identify practical and actionable information on tools that match their contexts and requirements. This includes the type of tool, its scope, the problem to address and resource requirements for implementation. To date, the framework does not assess the quality or completeness of an individual tool.

Box 1. What is trustworthy AI?

Trustworthy AI refers to AI systems that embody the OECD AI Principles; that is, AI systems that respect human rights and privacy; are fair, transparent, explainable, robust, secure and safe; and the actors involved in their development and use remain accountable. The Principles constitute the first AI standard at the intergovernmental level. They focus on how governments and other actors can shape a human-centric approach to trustworthy AI. The Principles were adopted in May 2019 by the 37 OECD member countries and five non-member countries, and endorsed by the G20 in June 2019.

The OECD AI Principles provide five values-based principles for the responsible stewardship of trustworthy AI:

- *Inclusive growth, sustainable development and wellbeing:* Stakeholders should engage in creating credible AI that can contribute to inducing outcomes that are beneficial for people, as well as for the planet.
- *Human-centred values and fairness:* The values of human rights, democracy, and rule of law should be incorporated throughout the AI system's lifecycle, while allowing human intervention through safeguard mechanisms.
- *Transparency and explainability:* AI actors that develop or operate AI systems should provide information to foster an overall understanding of the systems among stakeholders, in which

people affected by AI systems could comprehend the outcome and challenge the decision when needed.

- *Robustness, security and safety:* AI systems need to function appropriately while ensuring traceability, while AI actors need to apply systematic risk management approaches to mitigate safety risks.
- *Accountability:* AI actors should respect the principles and should be accountable for the proper operation of AI systems.

The OECD AI Principles also contain five recommendations for national policies and international co-operation. The recommendations include: 1) investing in AI research and development; 2) fostering a digital ecosystem for AI; 3) shaping an enabling policy environment for AI; 4) building human capacity and preparing for labour market transformation; and 5) international co-operation for trustworthy AI (OECD, 2019a; OECD, 2019b)

From Principles to practice

The fast-paced and far-reaching changes from AI technologies offer dynamic opportunities for economic and social sectors. To maximise the benefits of AI while mitigating its risks, technical, business, academic, and policy stakeholder communities are actively exploring how best to encourage the design, development, deployment and use of AI that is human-centred and trustworthy. The outcomes of AI systems should foster shared wellbeing and prosperity while safeguarding human rights and democratic values.

Since the adoption of the OECD AI Principles in May 2019, the OECD has focused on helping policy makers and other stakeholders implement these Principles in practice. In early 2020, the OECD launched the AI Policy Observatory (OECD.AI) and the OECD Network of Experts on AI. The Network of Experts formed a working group on Implementing Trustworthy AI, comprised of representatives from government, business, labour unions, academia and the technical community (Box 2).

Box 2. The OECD Network of Experts on AI and its working group on implementing trustworthy AI

The OECD Network of Experts on AI (ONE AI) provides policy, technical and business expert input to inform OECD analysis and recommendations. It is a multi-disciplinary and multi-stakeholder group. ONE AI also provides the OECD with an outward perspective on AI, serving as a platform for the OECD to share information with other international initiatives and organisations. The Network raises awareness about trustworthy AI and other policy initiatives, particularly in instances where international co-operation is useful.

The mission of the ONE AI working group on implementing Trustworthy AI (the “working group”) is to identify practical guidance and standard procedural approaches that lead to trustworthy AI. To do so, the working group has developed a practical framework to collect concrete examples of tools and approaches to help implement each of the five values-based OECD AI Principles in different sectors and operational contexts. These tools will serve AI actors and decision-makers as they seek to implement effective, efficient and fair AI-related policies.

The working group is co-chaired by Adam Murray, ONE AI Chair and US delegate to the OECD Committee on Digital Economy Policy (CDEP); Carolyn Nguyen, Director of Technology Policy, Microsoft; and Barry O’Brien, Government and Regulatory Affairs Executive, IBM. The group has been meeting virtually every 3 to 4 weeks since May 2020.

A framework of tools for trustworthy AI

In June 2020, the working group on Trustworthy AI designed and ran a survey to take stock of initiatives to implement trustworthy AI in diverse operational contexts, collecting and reviewing submissions from a wide range of stakeholder types to develop the framework (Annex A).

Zooming in on tools

The working group agreed to focus its analysis on tools, understood as instruments and structured methods that can be leveraged by others to facilitate their implementation of the AI Principles (e.g. toolkits to check for biases or robustness in an AI system, risk management guidelines, educational material)[1]. There are three types of tools that can be classified as technical, procedural, or educational (Table 1).

Table 1. Technical, procedural and educational tools

Approach	Type of tool
Technical	Toolkits / toolboxes / software tools
	Technical documentation
	Technical certification
	Technical standards
	Product development / lifecycle tools
	Technical validation tools
Procedural	Guidelines
	Governance frameworks
	Product development / lifecycle tools
	Risk management tools
	Sector-specific codes of conduct
	Collective agreements
	Certification
	Process-related documentation
	Process standards
Educational	Change management processes
	Capacity / awareness building
	Inclusive design guidance
	Educational materials / training programmes

Technical approaches

Technical tools for trustworthy AI aim to address specific AI-related issues from a technical angle, including bias detection, transparency and explainability of AI systems, performance, robustness, safety and security against adversarial attacks. They include toolkits, software tools, technical documentation,

certification and standards, product development or lifecycle tools, and technical validation tools (Table 2).

A sizeable proportion of the technical tools submitted originate from large private sector companies, such as IBM, Google and Microsoft. Many of these technical tools to develop and use trustworthy AI exist as open-source resources, which facilitates their adoption and allows for crowdsourcing solutions to software bugs. Many of these tools allow developers and others to check AI systems for reliability and fairness.

Table 2. Selection of technical tools to implement trustworthy AI

Objective	Tool	Description
Fairness	AT&T software System to Integrate Fairness Transparently (SIFT)	Software system to integrate mechanised and human-in-the-loop components in bias detection, mitigation, and documentation of projects at various stages of the machine learning lifecycle.
	Microsoft Fairlearn	Open-source toolkit to assess and improve the fairness of machine learning models. Contains an interactive visualisation dashboard and bias mitigation algorithms to help navigate trade-offs between fairness and model performance.
	LinkedIn Fairness Toolkit (LiFT)	Open-source toolkit to enable measurement of fairness according to a multitude of fairness definitions in large-scale machine learning workflows.
	Google What-If Tool	Open-source software tool to visually inspect and explore machine learning model performance and data across multiple hypothetical situations, with minimal coding required.
	IBM AI Fairness 360	Open-source toolkit to help detect and mitigate unwanted bias in machine learning models and datasets. Provides approximately 70 metrics to test for biases, and 10 algorithms to mitigate bias in datasets and models.
Transparency	IEEE Standard for Transparency of Autonomous Systems	Technical standard to describe measurable and testable levels of transparency, so that autonomous systems can be assessed and levels of compliance determined.
	Google Model Card Toolkit	Documentation framework for sharing the essential facts of a machine learning model in a structured, accessible way, providing an overview of what the model is intended to do, how it was architected, trained, and its limitations.
Explainability	Google Cloud Explainable AI service	Software to help developers get explanations on the outcomes of their models. Can be applied to the AI models trained on tabular, image, and text data. Not open source.
	IBM AI Explainability 360 Toolkit	Open-source toolkit of algorithms, code, guides, tutorials, and demos to support the interpretability and explainability of machine learning models.
	Microsoft InterpretML	Open-source toolkit containing machine learning interpretability algorithms to help understand model predictions.
Robustness	IBM Adversarial Robustness 360 Toolkit	Open-source toolkit for machine learning security. It provides tools to evaluate, defend, certify and verify machine learning models and applications against the adversarial threats of evasion, poisoning, extraction, and inference.

Note: illustrative, non-exhaustive examples.
Source: June 2020 survey of the OECD Network of Experts on AI, working group on implementing Trustworthy AI.

Procedural approaches

Procedural tools for trustworthy AI provide operational or process-related implementation guidance. They encompass guidelines; governance frameworks; product development, lifecycle, and risk management tools; sector-specific codes of conduct and collective agreements; and process certifications and standards (Table 3).

Compared to technical tools – where there is high private sector participation – procedural tools to implement AI systems ethically and inclusively are produced by a wider variety of stakeholders,

including governments and trade unions. Some procedural tools for transparency and explainability emphasise the importance of documenting the development and deployment of AI systems and propose governance frameworks for their implementation.

Table 3. Selected procedural tools to implement trustworthy AI

Objective	Tool	Description
Inclusive implementation	German Trade Union Confederation's Good Work by Design	Guidelines for trustworthy implementation of AI systems in the workplace, with the goal of gaining acceptance among the workforce.
	Negotia AI Governance System	Framework for the governance of AI systems in the workplace. It aims to equip unions, workers, and managers with a set of tools to manage and govern algorithmic systems responsibly.
	Google People + AI Guidebook	Guidelines to help user experience professionals and product managers follow a human-centred approach to AI. They are structured based on the product development cycle and contain worksheets to help turn guidance into action.
	IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being	Standard to enable programmers, engineers, and technologists to better consider how the products and services they create can increase human wellbeing based on a wider spectrum of measures than growth and productivity alone.
Ethical implementation	IBM Everyday Ethics for AI	Practical guidelines for designers and developers of AI solutions, including key questions for team members to consider, and some topical examples on ethical issues.
	IEEE Ethics Certification Program for Autonomous and Intelligent Systems	Certification programme to create specifications for AI processes that advance transparency, accountability and reduction in algorithmic bias.
	IEEE Trusted Data & AI Systems Playbook for Finance Initiative	Sector-specific guidelines to curate, summarise, and contextualise trusted data and AI best practices for the financial sector around design principles, standards, and certifications.
	Denmark Algorithm Test	Practical guide for companies to use AI systems responsibly and ethically. It consists of two tests with six questions each on bias and transparency, leading to recommendations for companies and developers on how they can improve their algorithms.
Transparent and explainable implementation	Microsoft Datasheets for Datasets	Tool for documenting the datasets used for training and evaluating machine learning models to increase dataset transparency and facilitate better communication between dataset creators and dataset consumers.
	IBM AI Factsheets 360	Governance approach to the AI lifecycle and methodology for assembling information about an AI model's important features (including a collection of templates, worked examples, research papers and other resources).
	UK Information Commissioner's Office "Explaining decisions made with AI"	Guidance for organisations on how to implement explainable AI solutions in compliance with a range of legislation, including data protection legislation. It advises on how to build and operate systems that allow explanations to be provided to individuals that are affected by the decisions made by the system.

Note: illustrative, non-exhaustive examples.
Source: June 2020 survey of the OECD Network of Experts on AI, working group on implementing Trustworthy AI.

Educational approaches

Educational tools for trustworthy AI encompass mechanisms to build awareness, inform, prepare or upskill stakeholders involved in or affected by the implementation of an AI system. They include change management processes; capacity and awareness building tools; guidance for inclusive AI system design; and training programmes and educational materials (Table 4).

Depending on the implementation context, educational tools are designed to serve different audiences. They can be wide-reaching and open to the public at large or focus on a specific group affected by the implementation of an AI system, such as SMEs or workers.

Table 4. Selection of educational tools to implement trustworthy AI

Target audience	Tool	Description
Businesses	Denmark Data Ethical Dilemma Game	Educational game to create business awareness around the challenge of developing responsible and ethical AI solutions. The game seeks to stimulate reflections and perspectives on the work with data through common ethical dilemmas. It targets SMEs.
Workplace actors	Negotia AI Governance System	Framework for the governance of AI systems in the workplace. It aims to equip unions, workers, and managers with a set of tools to manage and govern algorithmic systems responsibly, including by raising workers' understanding of such systems.
General public	Finland AI Course "Elements of AI"	Free online courses combining theory with practical exercises to encourage people to learn the basics of AI, its impacts and how it is created.
	VIRT-EU Service Package	Capacity-building material and practical resources to help develop ethically-informed AI systems, based on three different ethical frameworks: virtue ethics, care ethics and capabilities approach.

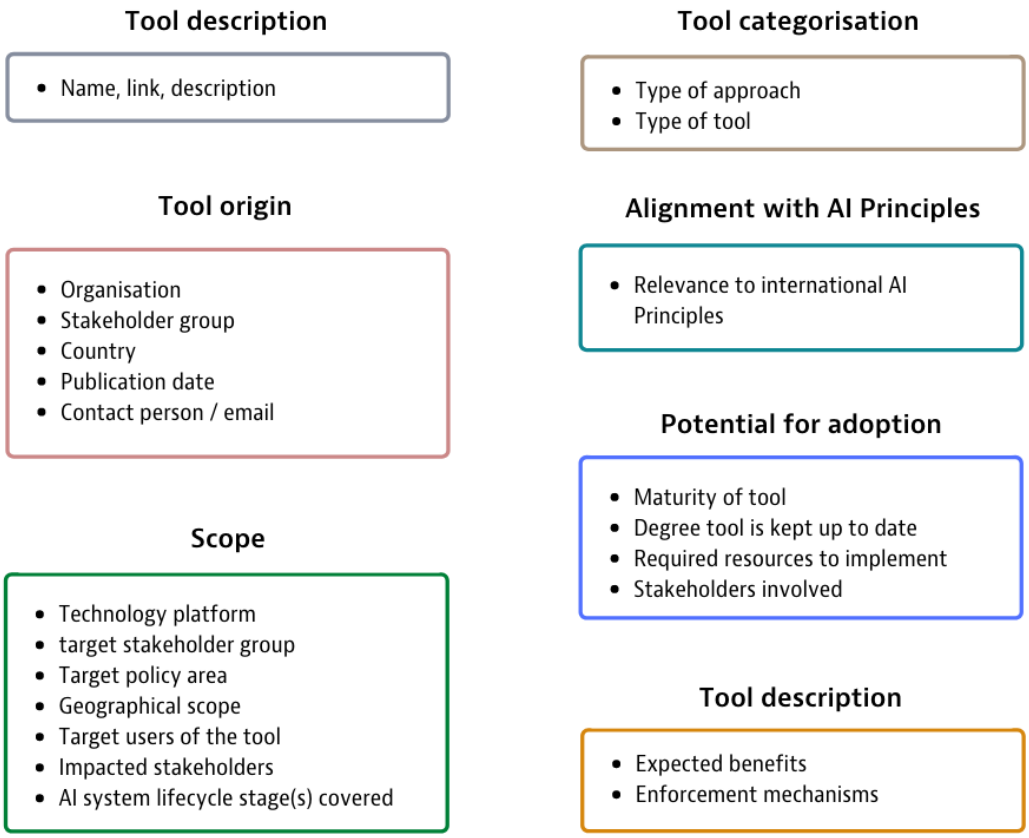
Note: illustrative, non-exhaustive examples.
Source: June 2020 survey of the OECD Network of Experts on AI, working group on implementing Trustworthy AI.

Building a framework

The goal of the present framework is to structure information and facilitate comparison between different tools used for different purposes and in different contexts. It does not aim to provide any qualitative assessment. The seven key dimensions of the framework are (Figure 1):

- **The tool description**, including the tool's name, background information and hyperlinks to additional information;
- **The tool origin**, including the organisation, stakeholder group and geographical region from which the tool originates; the date of publication; and contact details of the person submitting the tool;
- **The tool categorisation**, including the type of approach – technical, procedural or educational – and the type of tool (e.g. toolkits, standards, guidelines, governance frameworks, certifications and educational materials, etc.);
- **The scope**, including platform specificity; target users, policy areas, stakeholder groups; geographical scope; impacted stakeholders; and AI system lifecycle stage(s) covered by the tool;
- **The alignment with international AI Principles**, including the tool's relevance to the OECD AI Principles and the European Commission's key requirements for trustworthy AI[2];
- **The adoption potential**, including the maturity of the tool and the degree to which it is kept up to date; the required resources and legal conditions to use the tool; and the stakeholders who need to be involved in using the tool;
- **Implementation incentives**, including the expected benefits from using the tool and the possible enforcement mechanisms that may facilitate the use of the tool.

Figure 1. High-level structure of the framework of tools for trustworthy AI



The framework underwent several iterations of expert testing and validation to assess robustness and comprehensiveness so it can serve as a reference point for actors seeking to implement trustworthy AI. It was complemented with relevant research from the Global Partnership on AI (The Future Society, 2020), the Open Community for Ethics in Autonomous and Intelligent Systems (Institute of Electrical and Electronics Engineers, 2021) and the Arizona State University (Gutierrez, Marchant, Carden, Hoffner, & Kearl, 2020). Figure 2 provides a detailed overview of the framework.

Applying the framework

Figure 3 illustrates the use of the framework by detailing seven specific tools covering different characteristics and objectives, stakeholder groups, and use contexts; three of which are technical (LinkedIn Fairness Toolkit, Google Model Cards Toolkit and IBM Adversarial Robustness 360 Toolbox); two procedural (IEEE Ethics Certification Program for Autonomous and Intelligent Systems and Microsoft Datasheets for Datasets); and two educational (Danish Data Ethical Dilemma Game and the Negotia AI Governance System).

Of the seven tools detailed in Figure 3, four originated from the private sector, one from the technical community, one from the public sector and one from various organisations (trade union, private sector and public sector). They range from technical toolkits and documentation to guidelines, sector-specific codes of conduct and governance frameworks.

Additionally, the tools have a broad scope, targeting many different types of users and several stages of the AI system lifecycle. Most tools have an international reach, with the exception of two that have a national reach. The prevalence of open-source or free-to-use tools is noteworthy.

The tools included in Figure 3 require medium or high level domain expertise and data to implement and low or medium financial resources.[3] Each tool aims to generate at least two benefits, the most prevalent being 'reduction in the risk of AI system failure', followed by 'increased quality of results of AI system'. The two educational tools with a national scope (Danish Data Ethical Dilemma Game and the Negotia AI Governance System) and the tool originating from the technical community (IEEE Ethics Certification Program for Autonomous and Intelligent Systems) foresee explicit enforcement mechanisms.

Tools to implement trustworthy AI often target issues related to bias in AI systems, for example by integrating human-in-the-loop detection mechanisms – such as interactive dashboards – to allow people to visualise and inspect model performance across different configurations of the variables. The importance of fairness is reflected in Figure 3 under the section "Alignment with international AI Principles": five of the seven tools are highly relevant to the Principles of human-centred values and accountability. Four tools are highly relevant to the Principle of building human capacity and preparing for labour market transformation.

The examples illustrate how the framework can help gather and share tools, practices and approaches to implement trustworthy AI in a comparable and accessible manner.

[1] The other types of initiatives collected were use cases of AI solutions aiming to achieve a specific goal in a specific context (e.g. using AI for fraud prevention, disease detection, etc.) as well as documents and reports providing an overview or flagging an issue related to a specific AI topic (Annex A).

[2] In particular its requirement on human agency and oversight that refers to AI systems that empower human beings, allowing them to make informed decisions and fostering their fundamental rights. It includes proper oversight mechanisms, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches (European Commission, 2019).

[3] Subjective assessments from the organisations submitting the tools.

Figure 2. Framework of tools for trustworthy AI

Type	Field	Definition	Options (if applicable)
Tool description	Name	The name of the tool	
	Link	A link to an up-to-date document	
	Description	A brief summary of the tool and its purpose	
Tool origin	Organisation	The organisation that developed the tool	
	Stakeholder group	The stakeholder group from which the initiative originates	Academia; Trade union/worker representative; Private sector; Civil society; Technical community; Public sector; International governmental organisation; Other
	Country	The country or region where the initiative originated	International; OECD countries; List of regions; List of countries; Other
	Date of publication	Date the tool was published in its first version	
	Contact email	Email of the contact person for the tool (not for public use)	
Tool categorisation	Type of Approach	High-level category of the tool	Process-related approach; Technical approach; Educational approach; Other
	Type of Tool	Category of the tool	Toolkits/toolboxes/software tools; Technical documentation; Technical certification; Technical standards; Product development/lifecycle tools; Technical validation tools; Guidelines; Governance frameworks; Risk management tools; Sector-specific codes of conduct; Collective agreements; Certification; Process-related documentation; Process standards; Change management processes; Capacity/awareness building tools; Inclusive design guidance; Educational materials/training programmes; Other
Scope	Technology platform	The technology platform(s) that the tool can be used for	Platform neutral; Platform specific; Multi-platform; Other
	Target stakeholder group	The stakeholder group where the tool is expected to be implemented	Academia; Trade union/worker representative; Private sector; Civil society; Technical community; Public sector; International governmental organisation; Other
	Primary and secondary policy area	The policy area(s) where the tool is expected to be implemented	Agriculture; Competition; Corporate governance; Development; Digital Economy; Economy; Education; Employment; Environment; Finance and insurance; Health; Industry and entrepreneurship; Innovation; Investment; Public governance; Science and technology; Social and welfare issues; Tax; Trade; Transport; All of the above; Not applicable; Other
	Geographical scope	The country or region that the initiative targets	International; OECD countries; List of regions; List of countries
	Target users of the tool	Users who are expected to use the tool to implement a project	AI system business leader; AI system technical developers; IT specialists; Researchers; AI system operators; Executive management; Government agencies; Data scientists; Project managers; HR managers; All employees; Other
	Impacted stakeholders	Groups of people that will be impacted by the implementation of the tool	Employees; Specific policy communities; Consumers; Regulators; Management; Other
	AI system lifecycle stage(s) covered	The stages of the AI system lifecycle that the tool helps to implement	Planning & design; Data collection & processing; Model building & interpretation; Verification & validation; Deployment; Operation & monitoring; All stages
Alignment with international AI Principles	Relevance to international AI Principles	Grade relevance to international AI Principles	Values-based Principles: Socio-economic and environmental impacts; Human-centred values & fairness; Transparency & explainability; Robustness, security, safety; Accountability; Human agency and oversight. Recommendations for policy makers: Investing in research; Data, compute, technologies; Enabling policy environment; Jobs, skills, transitions; International co-operation
Potential for adoption	Maturity of the tool	Project phase the tool is currently in	Project stage; In development; Running code; Implemented in one project; Implemented in multiple projects; Not relevant anymore; Other
	Degree tool is kept up to date	How the tool is kept up to date with evolving standards, requirements, etc.	No update mechanism planned; Periodic review; Always up to date; Other
	Degree of free use of the tool	Legal conditions for using the tool	Subscription fee; One-time license fee; Free-to-use (creative commons); Open source; Other
	Required resources to implement	The extent to which certain resources are needed to implement/use the tool	IT skills; Domain expertise; Data; IT infrastructure; Operational infrastructure; Financing
	Stakeholders involved	Stakeholders who will be involved in the implementation and operation of the tool	IT employees; Operations employees; All employees; Business unions; Trade unions/worker representatives; Clients; Suppliers; Government agencies; Other
Implementation incentives	Expected benefits	Expected benefits from using the tool	Reduction in risk of AI system failure; Reduction in cost of AI system implementation; Faster implementation of an AI system; Increased quality of AI system results; Improved ability of AI system's implementation to scale; Responsible implementation of AI system; Other
	Enforcement mechanisms	Enforcement mechanisms attached with the usage of this tool	Internal mediation (ombudsman); Ethics board; Certification; Enforcement body; Governmental regulation; Log registrars; Reporting frameworks; Collective agreements; N/A; Other

Figure 3. Illustrating the framework through selected examples of tools

		LinkedIn Fairness Toolkit (LiFT)	Google Model Card Toolkit	IBM Adversarial Robustness 360 (ART) Toolbox	IEEE Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS)
Tool origin	*Organisation*	LinkedIn	Google	IBM	IEEE SA
	Stakeholder group	Private sector	Private sector	Private sector	Technical community
	Country/Region	North America	United States of America (USA)	International	International
	Date of publication	2020	2020	2018	2018
Tool categorisation	*Type of Approach*	Technical	Technical	Technical	Procedural
	Type of Tool	Toolkit	Toolkit / Technical documentation	Toolkit / Software tools	Sector-specific codes of conduct
Scope	*Technology platform*	Platform neutral	Platform specific (TensorFlow Extended)	Platform neutral	Platform neutral
	Target stakeholder group	All	Academia, Private sector, Technical community	Academia, Private sector, Technical community, Public sector	Trade union / worker representative, Private sector, Technical community, Public sector
	Policy area	All	Digital economy / Corporate governance	All	All
	Geographical scope	International	International	International	International
	Target users of the tool	AI system business leader, AI system technical developers, AI system operators, Data scientists	AI system business leader, AI system technical developers, IT specialists, AI system operators, Data scientists	AI system technical developers, Data scientists, IT specialists	AI system business leader, AI system technical developers, IT specialists, AI system operators, Government agencies, Executive management
	Impacted stakeholders	Employees, Consumers	Employees, Consumers, Regulators, Specific policy communities	Employees	Employees, Consumers, Regulators, Specific policy communities
	AI system lifecycle stage(s) covered	All stages	Planning & design, Model building & interpretation, Verification & validation, Deployment, Operation & monitoring	Model building & interpretation, Verification & validation, Deployment, Operation & monitoring	All stages
Alignment with AI Principles	*Relevance to: Socio-economic and environmental impacts*	Medium	Not relevant	Not relevant	Low
	Relevance to: Human-centered values and fairness	High	High	Not relevant	Medium
	Relevance to: Transparency and explainability	Medium	High	Not relevant	High
	Relevance to: Robustness, security and safety	Low	Low	High	High
	Relevance to: Accountability	High	High	Not relevant	High
	Relevance to: Investing in research	High	Not relevant	Not relevant	Low
	Relevance to: Data, compute and technologies	High	Not relevant	Not relevant	Medium
	Relevance to: Enabling policy environment	Low	Medium	Not relevant	Medium
	Relevance to: Jobs, skills, transitions	High	Low	Not relevant	High
	Relevance to: International cooperation	Medium	High	Not relevant	High
	*Relevance to: Human agency and oversight**				
Potential for adoption	*Maturity of the tool*	Implemented in multiple projects	Implemented in multiple projects	Implemented in multiple projects	In development
	Degree tool is kept up-to-date	Always up to date	Periodic review	Always up to date	Periodic review
	Degree of free use of the tool	Open source	Open source	Open source	Free-to-use (creative commons)
	Required IT skills to implement	High	High	High	Medium
	Required domain expertise to implement	Medium	Medium	Medium	Medium
	Required data to implement	High	Medium	Medium	Medium
	Required IT infrastructure to implement	Medium	Low	Medium	Medium
	Required operational infrastructure to implement	Medium	Low	Low	Medium
	Required financing to implement	Medium	Low	Low	Low
	Stakeholders involved	IT employees, Operations employees	IT employees, Operations employees, Clients, Suppliers	IT employees	IT employees, Operations employees, Business unions, Trade unions / worker representatives, Government agencies
Implementation incentives	*Expected benefits*	Reduction in the risk of AI system failure, Reduction in cost of AI system implementation, Faster implementation of AI system, Increased quality of results of AI system, Improved ability of AI system's implementation to scale	Reduction in the risk of AI system failure, Increased quality of results of AI system	Reduction in the risk of AI system failure, Improved AI system robustness and security	Reduction in the risk of AI system failure, Faster implementation of AI system, Increased quality of results of AI system, Improved ability of AI system's implementation to scale
	Enforcement mechanisms	Depending on use context	Depending on use context	Depending on use context	Certification

		Microsoft Datasheets for Datasets	Danish Data Ethical Dilemma Game	Negotia AI Governance System
Tool origin	*Organisation*	Microsoft Research	Danish Business Authority	Negotia, The Why Not Lab, Finansforbundet, Nito, YA
	Stakeholder group	Private sector	Public sector	Trade union, Private sector, Public sector
	Country/Region	United States of America (USA)	Denmark	Norway
	Date of publication	2018	2020	2021
Tool categorisation	*Type of Approach*	Procedural	Educational	Procedural / Educational
	Type of Tool	Product development tool / Process-related documentation	Educational material	Governance framework / Collective agreement
Scope	*Technology platform*	Platform neutral	Platform neutral	N/A
	Target stakeholder group	Academia, Private sector, Technical community, Public sector, IGO	Private sector (SMEs)	Trade union / worker representative, Private sector, Public sector
	Policy area	All	Digital economy / Corporate governance	All
	Geographical scope	International	Denmark	Norway
	Target users of the tool	AI system technical developers, Data scientists, Project managers, Researchers	AI system operators, Data scientists, All employees, Project managers	AI system business leader, All employees, Project managers, HR managers, Executive management
	Impacted stakeholders	Employees	Employees	Employees, Regulators, Management
	AI system lifecycle stage(s) covered	All stages	Planning & design, Data collection & processing	Planning & design, Data collection & processing, Verification & validation, Deployment, Operation & monitoring
Alignment with AI Principles	*Relevance to: Socio-economic and environmental impacts*	Medium	Low	High
	Relevance to: Human-centered values and fairness	High	High	High
	Relevance to: Transparency and explainability	High	High	High
	Relevance to: Robustness, security and safety	High	Not relevant	Low
	Relevance to: Accountability	High	Moderate	High
	Relevance to: Investing in research	Low	Not relevant	Not relevant
	Relevance to: Data, compute and technologies	High	Low	Low
	Relevance to: Enabling policy environment	Medium	Not relevant	High
	Relevance to: Jobs, skills, transitions	Low	High	High
	Relevance to: International cooperation	Medium	Not relevant	Not relevant
	*Relevance to: Human agency and oversight**			
Potential for adoption	*Maturity of the tool*	Implemented in multiple projects	In development	Project stage
	Degree tool is kept up-to-date	Periodic review	Periodic review	Periodic review
	Degree of free use of the tool	Free-to-use (creative commons)	Free-to-use (creative commons)	Free-to-use (creative commons)
	Required IT skills to implement	Low	High	Medium
	Required domain expertise to implement	Medium	High	Medium
	Required data to implement	Medium	High	Medium
	Required IT infrastructure to implement	Medium	Moderate	Low
	Required operational infrastructure to implement	Medium	Moderate	Medium
	Required financing to implement	Medium	Low	Low
	Stakeholders involved	IT employees	IT employees, Clients, Suppliers, Government agencies	Operations employees, All employees, Business unions, Trade unions / worker representatives, Suppliers
Implementation incentives	*Expected benefits*	Reduction in the risk of AI system failure, Reduction in cost of AI system implementation, Increased quality of results of AI system	Higher awareness among employees and companies on data ethical issues related to AI	Reduction in the risk of AI system failure, Increased quality of results of AI system, Improved ability of AI system's implementation to scale, Responsbile implementation of AI system
	Enforcement mechanisms	Depending on use context	Internal mediation (ombudsman)	Internal mediation (ombudsman), Governmental regulation, Enforcement body, Reporting frameworks, Collective agreements

Note: The tools were categorised by representatives from the organisations that created them.
"Human agency and oversight" was added ex-post from the EC's seven requirements for trustworthy AI (European Commission, 2019). It has been included for completeness.

Next steps

Developing and maintaining an interactive database

The framework provides the structure for the development of a live database of tools for trustworthy AI on the OECD.AI Policy Observatory. The database will provide AI actors and policy makers with information on the latest tools to help ensure that AI systems in different contexts abide by principles of human rights and fairness; transparency and explainability; robustness, security, and safety and accountability. The database will be part of the projects pursued by CDEP and the OECD AI network of experts working group on trustworthy AI over the 2021-2022 biennium.

Box 3. Ensuring database usefulness and relevance

Keeping the database up to date

25. To be useful, the database of tools for trustworthy AI needs mechanisms to ensure it is kept up to date. To this end, the database will be based on an open submission process, where tools are submitted directly by the organisations creating them and vetted by the OECD Secretariat for accuracy and neutrality of information.

26. Additionally, a biannual review and updating process where organisations will be encouraged to submit new initiatives and review or update existing ones will be established. In the case of existing initiatives, no response or updates for two consecutive years will amount to the initiative being removed from the database. Partnerships with the relevant stakeholders are being forged to facilitate this biannual database review. This includes Business at the OECD (BIAC) and the Trade Union Advisory Committee (TUAC) at the OECD.

27. The process aims at creating a virtuous cycle where the quality and usefulness of the information strengthens demand for it, in turn furthering organisations' interest to keep their tools up to date. Methods to automatically retrieve information from the internet – such as crawlers and APIs – are also being considered as a means to prefill information in the database.

Illustrating tool use and implementation

At a later stage, the working group plans to enrich the database with case studies of how the tools are being used and implemented in the real world. These use cases would be linked to the corresponding tool in the database to illustrate its implementation. Some of the elements that the use cases will capture are the name of the tool being implemented; the organisation implementing it; a brief description of the implementation process, including learnings; the relevance of the use case to the Sustainable Development Goals (SDGs); the ease of implementation; the actual required resources and benefits of implementing the tool; the tool's strengths and weaknesses; and the risks attached to the implementation of the tool.

References

European Commission. (2019). *Ethics guidelines for trustworthy AI.* Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Gutierrez, C., Marchant, G., Carden, A., Hoffner, K., & Kearl, A. (2020). *Preliminary Results of a Global Database on the Soft Law Governance of Artificial Intelligence.* Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3756939

Institute of Electrical and Electronics Engineers. (2021). *The Open Community for Ethics in Autonomous and Intelligent Systems - OCEANIS.* Retrieved from https://ethicsstandards.org

OECD. (2019a). *Recommendation of the Council on Artificial Intelligence.* Retrieved from https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

OECD. (2019b). *Scoping the OECD AI Principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO), in OECD Digital Economy Papers, No. 291.* Retrieved from https://doi.org/10.1787/d62f618a-en

The Future Society. (2020). *Areas for Future Action in the Responsible AI Ecosystem*. Retrieved from https://thefuturesociety.org/wp-content/uploads/2021/01/TFS_GPAI-RAI-Final-Report.docx-1.pdf

Annex A. Process to develop the framework of tools for trustworthy AI

Taking stock

A stock-taking of existing initiatives being used by different actors and in varying contexts to design, build and operate trustworthy AI systems was the first step towards the development of a framework to classify tools for trustworthy AI. To this end, the working group designed and launched a survey to identify practical approaches and good practices to help further inform the implementation of trustworthy AI systems. A total of 75 submissions were received from a wide range of organisations, a majority of which (51%) were from the private sector (Table A 1).

Table A 1. Descriptive statistics of survey submissions

		% of all submissions
Stakeholder type		
	Private sector	51%
	Multistakeholder group	13%
	Technical community	11%
	Public sector	7%
	Other	18%
Policy area*		
	Innovation	43%
	Science and technology	38%
	Digital Economy	30%
	Industry and entrepreneurship	23%
AI Principle*		
	Transparency and explainability	61%
	Human-centred values and fairness	60%
	Accountability	43%
	Robustness, security and safety	40%
	Inclusive growth and sustainable development	29%
AI system lifecycle stage*		
	Planning and design	56%
	Model building and interpretation	55%
	Deployment	53%
	Verification and validation	50%
	Operation and monitoring	49%
	Data collection and processing	49%

*Note: Multiple selection was allowed for these questions. Therefore, the numbers for these categories refer to the percentage of submissions classified under that option, not taking into account the others (i.e. a category would reach 100% if all submissions were classified under it).

Organisations submitting the initiatives were directly involved in, or had substantial knowledge of, the work, due to either first-hand usage or research in co-operation with the implementing organisations. It was required that responses to the survey include initiatives already implemented or in deployment, but not necessarily completed.

Analysing submissions

The submissions were analysed by assigning a random subset of 6-8 initiatives to each working group expert for review. The experts were asked to:

- Rank initiatives by their relevance to the AI Principles (from 'not so relevant' to 'very relevant');
- Provide an assessment of common strengths, weaknesses and/or gaps found after reviewing the relevant subset of initiatives; and
- Share ideas for analysis and identify opportunities for international co-operation for a given type of initiative.

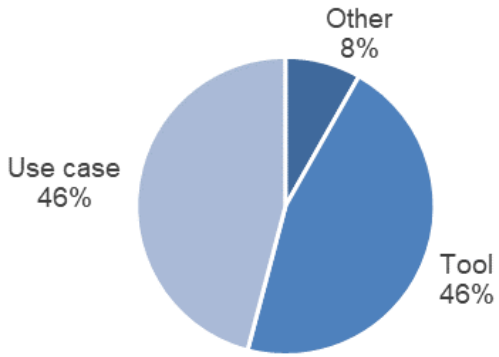
To avoid biased assessments, working group experts were not allowed to review use cases submitted by their organisations. In total, the experts conducted 254 reviews, resulting in an average of 3.4 expert reviews per submission.

A recurring observation from the experts was that the variety of initiatives made it difficult to establish valid comparisons among them. In particular, it was noted that while some initiatives contained tools or frameworks that could be leveraged by others to implement the AI Principles (e.g. toolkits to check for biases or robustness in an AI system), others were merely use cases of AI solutions that aimed to achieve a specific goal (e.g. fraud prevention, disease detection, etc.). Additionally, a few submissions were limited to raising awareness about specific documents or reports on a given AI topic or issue.

Submissions were therefore split into three categories (Figure A 1):

- **Tools** that are already somewhat formalised and can be reused in other use cases;
- **Use cases** that describe how AI is used to achieve specific goals; and
- **Other**, including documents and reports that provide an overview or flag an issue related to a specific AI topic.

Figure A 1. Distribution of submissions by type of initiative



Annex B. Working group members and observers

Table B 1 contains the updated list of members and observers to date of the ONE AI working group on implementing trustworthy AI. Their short bios are available on OECD.AI.

Table B 1. ONE TAI members and observers (March 2021)

Name	Title	Organisation	Group / Delegation
Adam Murray*	[ONE AI Chair] International Affairs Officer	US State Department Office of International Communications and Information Policy	United States
Aishik Gosh	PhD in Artificial Intelligence for Particle Physics in Atlas	European Organization for Nuclear Research (CERN)	Civil Society and Academia
Alana Lomonaco Busto	First Secretary- Cybersecurity, Cybercrime and Digital Affairs	Ministry of Foreign Affairs, International Trade and Worship	Argentina
Alistair Nolan	Secretariat	OECD	OECD
András Hlács	Counsellor	Permanent Delegation of Hungary to OECD	Hungary
Andrey Ignatyev		Ministry of Economic Development	Russia
Angelica Salvi del Pero	Secretariat	OECD	OECD
Anna Byhovskaya	Senior Policy Advisor	Trade Union Advisory Committee (TUAC) to the OECD	Trade Union
Ansgar R. Koene	Global AI Ethics and Regulatory Leader	EY AI Lab, London	Business
Anthony Scriffignano	Chief Data Scientist	Dun & Bradstreet	Business
Balachander Krishnamurthy	Lead Inventive Scientist	AT&T Labs	Business
Barry O'Brien*	Government and Regulatory Affairs Executive	IBM	Business
Barry Smyth	Digital Chair of Computer Science, Director of the Insight Centre for Data Analytics	University College Dublin	Ireland
Ben Macklin	Manager, Global Digital Policy, Digital Economy and Technology Division	Australia's Department of Industry, Innovation & Science	Australia
Carolyn Nguyen*	Director of Technology Policy	Microsoft	Business
Cedric Wachholz	Head of UNESCO's ICT in Education, Science and Culture section	UNESCO	IGO
Christina Colclough	Future of Work and Politics of Technology	Independent Expert	Civil Society and Academia
Clara Neppel	Senior Director	IEEE European Business Operations	Technical
Colin Gavaghan	Director	New Zealand Law Foundation-sponsored Centre for Law and Policy in Emerging Technologies	New Zealand
Cristina Pombo	Principal Advisor and Head of the Digital and Data Cluster, Social Sector	Inter-American Development Bank	IGO
Daniel Faggella	Head of Research, CEO	Emerj AI Research	Business
Danit Gal	Technology Advisor	Independent Expert	Civil Society and Academia
David Sadek	Vice President for Research,	Thales	Business

	Technology & Innovation		
Dino Pedreschi	Professor of Computer Science	University of Pisa	Italy
Dominik Geller	Head of Group Digital Governance	Sanofi	Business
Elettra Ronchi	Secretariat	OECD	OECD
Emilia Gómez	Lead Scientist, Human behaviour and machine intelligence	European Commission DG Joint Research Centre (JRC)	European Commission
Emma Naji	Executive Director	AI Forum New Zealand	New Zealand
Emmanuel Bloch	Director of Strategic Information	Thales	Business
Eric Badique	Adviser for Artificial Intelligence	Independent expert	
Etienne Corriveau-Hebert	Head of partnerships division	Ministère des Relations internationales et de la Francophonie	Other
Eva Thelisson	Researcher	AI Transparency Institute	Civil Society and Academia
Farahnaaz H Khakoo	Assistant Director	US Government Accountability Office	United States
Francesca Sheeka	Secretariat	OECD	OECD
Frederik Weiergang Larsen	Special Adviser	Danish Business Authority	Denmark
Gonzalo López-Barajas Húder	Head of Public Policy and Internet at Telefónica	Telefonica	Business
Grace Abuhamad	Research Program Manager, Trustworthy AI	Service Now	Business
Gregor Strojin	State Secretary	Ministry of Justice	Slovenia
Guillaume Chevillon	Professor - Co Director ESSEC	ESSEC Business School, Paris	Civil Society and Academia
Irene Solaiman	Public Policy	OpenAI	Technical
Heather Benko	Committee Manager, Joint Technical Committee (JTC) 1, Subcommittee 42 on Artificial Intelligence	International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC)	Technical
Jaclyn Kerr	AAAS Science and Technology Policy Fellow	Office of the Science and Technology Advisor to the Secretary	United States
Jennifer Bernal	Lead on Global Policy	Deepmind	Business
Jessica Cussins	Program Lead - AI Security Initiative	Center for Long-Term Cybersecurity (UC Berkeley)	Civil Society and Academia
Jim Kurose	Advisor at the Sorbonne Center for AI	Sorbonne University	Civil Society and Academia
John McCarthy	Global Lead for Shared, Connected and Autonomous Vehicles	Arup	Ireland
Karine Perset	Secretariat	OECD	OECD
Kathleen Walch	Managing partner and principal analyst	Cognilytica	Business
Kerrie Holley	Senior Vice President	United Health Group	Business
Laura Galindo	Secretariat	OECD	OECD
Lisa Dyer	Director of Policy	Partnership on AI	Business
Louise Hatem	Secretariat	OECD	OECD
Luigia Spadaro	Head of the Secretariat of the Undersecretary Mirella Liuzzi	Ministry of the Economic Development	Italy
Luis Aranda	Secretariat	OECD	OECD
Lynette Webb	Senior Manager for AI Policy Strategy	Google	Business
Lynne Parker	Deputy United States Chief Technology Officer	The White House	United States
Marc-Antoine Dilhac	Professor of philosophy	Université de Montréal	Civil Society and Academia
Marek Havrda	AI Policy and Social Impact Director	GoodAI	Czech Republic
Maria Danmark Nielsen	Head of Section	Danish Business Authority	Denmark
Marian Gläser	CEO and Founder	Brighter AI	Business
Marjorie Buchser	Head of Innovation Partnerships and Digital Society Initiative	Chatham House	Civil society and Academia
Marko Grobelnik	AI Researcher & Digital Champion	AI Lab of Slovenia's Jozef Stefan Institute	Technical
Michael Birtwistle	Policy Adviser	Centre for Data Ethics and Innovation (CDEI)	United Kingdom
Najma Bichara	Advisor, Digital Affairs	French Ministry for Europe and Foreign Affairs	France

Nicolas Miailhe	Founder and President	The Future Society (TFS)	Civil Society and Academia
Nicole Primmer	Senior Policy Director	BIAC	Business
Nobu Nishigata	Secretariat	OECD	OECD
Norberto Andrade	Privacy and Public Policy Manager	Facebook	Business
Nozha Boujemaa	Chief Science & Innovation Officer	Median Technologies	Business
Oliver Suchy	Director	Digital World of Work and Future of Work unit of the German Trade Union Confederation (DGB)	Trade Union
Osamu Sudoh	Graduate School of Interdisciplinary Information Studies(GSII)	University of Tokyo	Japan
Peter Cihon	Policy Analyst	Independent Expert	
Philip Dawson	Lead, Public Policy	ElementAI	Business
Raja Chatila	Chair	IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	Technical
Renaud Vedel	Coordonnateur de la stratégie nationale en IA	Ministère de l'intérieur	France
Rosa Meo	Associate Professor of Computer Science	University of Torino	Italy
Ryan Budish	Executive Director, Berkman Klein Center for Internet & Society	Harvard University	Civil Society and Academia
Sally Radwan	Minister Advisor for Artificial Intelligence	Ministry of Communications & Information Technology (Egypt)	Egypt
Sasha Rubel	Programme Specialist, Communication and Information Sector	UNESCO	IGO
Suso Baleato	Secretary	CSISAC	Civil Society and Academia
Sybo Dijkstra	Head of Data Strategy and Artificial Intelligence	Royal Philips	Business
Taka Ariga	Chief Data Scientist | Director, Innovation Lab	US Government Accountability Office	United States
Takahiro Matsunaga	Assitant Director, Multilateral Economic Affairs Office, Global Strategy Bureau	Ministry of Internal Affairs and Communications (MIC-Japan)	Japan
Theodoros Evgeniou	Professor, Decision Sciences and Technology Management	INSEAD	Civil Society and Academia
Tiberio Caetano	Chief Scientist	Gradient Institute (Australia)	Australia
Tim Bradley	Minister-Counsellor (Education and Science)	Australian Government's Department of Industry, Innovation & Science at the Australian Embassy in Washington DC	Australia
Tim Rudner	PhD Candidate	University of Oxford	Civil Society and Academia
Timea Suto	Knowledge Manager, Innovation for All	ICC	Business
Wael Diab	Chair	ISO/IEC JTC 1/SC 42 Artificial intelligence	Technical
Wendell Wallach	Consultant, ethicist, and scholar	Yale University's Interdisciplinary Center for Bioethics	Civil Society and Academia
Wim Rullens	Head of International Organisations	Ministry of Economic Affairs and Communications (Netherlands)	Netherlands
Yeong Zee Kin	Assistant Chief Executive	Infocomm Media Development Authority of Singapore	Singapore
Yoichi Iida	Chair of the CDEP and Going Digital II Steering Group	Ministry of Internal Affairs and Communications (MIC-Japan)	Japan
Yuki Hirano	Deputy Director, Multilateral Economic Affairs Office, Global Strategy Bureau	Ministry of Internal Affairs and Communications (MIC-Japan)	Japan

Note: Working group co-moderators are marked with a * after their last name.