OECD*publishing*

ADVANCING ACCOUNTABILITY IN AI

GOVERNING AND MANAGING RISKS THROUGHOUT THE LIFECYCLE FOR TRUSTWORTHY AI

OECD DIGITAL ECONOMY PAPERS

February 2023 **No. 349**

Federal Ministry
of Labour and Social Affairs

OECD
BETTER POLICIES FOR BETTER LIVES

Foreword

The report illustrates how risk management approaches can enable the implementation of the OECD AI Principles throughout the AI system lifecycle. Notably, this report shows how OECD AI frameworks – including the OECD AI Principles, the AI system lifecycle and the OECD framework for classifying AI systems – can inform accountability in AI.

This report was discussed and reviewed by the OECD.AI expert groups on Tools & Accountability and Classification & Risk in April and September 2022 and by the OECD Working Party on Artificial Intelligence (AIGO) in May and November 2022.

This publication contributes to the OECD's Artificial Intelligence in Work, Innovation, Productivity and Skills (AI-WIPS) programme, which provides policymakers with new evidence and analysis to keep abreast of the fast-evolving changes in AI capabilities and diffusion and their implications for the world of work. For more information, please visit www.oecd.ai/wips. AI-WIPS is supported by the German Federal Ministry of Labour and Social Affairs (BMAS) and will complement the work of the German AI Observatory in the Ministry's Policy Lab Digital, Work & Society. For more information, visit https://oecd.ai/work-innovation-productivity-skills and https://denkfabrik-bmas.de/.

This paper was written by Karine Perset and Luis Aranda, under the supervision of Audrey Plonk, Head of the OECD Digital Economy Policy Division. The report also benefitted from the inputs of delegates for the OECD Working Party on Artificial Intelligence (AIGO), including the Civil Society Information Society Advisory Council (CSISAC) and Business at the OECD (BIAC). Shellie Phillips, Francesca Sheeka, Orsolya Dobe, Mika Pinkashov, John Tarver and Angela Gosmann provided editorial support.

This paper was approved and declassified by the OECD Committee on Digital Economy Policy on 23 December 2022 and prepared for publication by the OECD Secretariat.

Note to Delegations:

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

DSTI/CDEP/AIGO(2022)5/FINAL

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

Acknowledgements

This report is based on the work of the OECD.AI Expert Groups on Tools & Accountability and Classification & Risk. It was prepared under the aegis of the OECD Working Party on AI Governance (AIGO). The expert group on Tools & Accountability was co-chaired by Nozha Boujemaa (IKEA); Andrea Renda (Centre for European Policy Studies); and Barry O'Brien (IBM). The group on Classification & Risk was co-chaired by Marko Grobelnik (Slovenian Jozef Stefan Institute – JSI); Dewey Murdick (Center for Security and Emerging Technology - CSET); and Sebastian Hallensleben (CEN-CENELEC). Luis Aranda and Karine Perset, OECD Digital Economy Policy Division, led the report development and drafting.

Around 200 experts participated in the Expert Groups, which held regular virtual meetings between February 2020 and December 2022 (1Annex B; 1Annex C).[1] Many experts provided invaluable feedback and suggestions, including Barry O'Brien (IBM); Vanja Skoric and Marlena Wisniak (European Centre for Not-for-Profit Law – ECNL); Leon Kester and Eva Thelisson (AI Transparency Institute); Cristina Pombo (Inter-American Development Bank); Nicolas Blanc (CFE-CGC); Lord Tim Clement-Jones (House of Lords of the United Kingdom); Songül Tolan and Emilia Gómez (European Commission Joint Research Centre); Karen McCabe (IEEE), Sean McGregor (Responsible AI Collaborative); Christine Custis and Katya Klinova (Partnership on AI); Olivia J. Erdelyi (University of Canterbury) and Laura Galindo (Meta).

The report benefited significantly from the contributions of Theodoros Evgeniou (INSEAD), Raphaël Rozenberg (École Normale Supérieure) and Fabio Curi, as well as of national delegations to the OECD Working Party on AI Governance, namely: Mohammed Motiwala (US Department of State); Elham Tabassi and Mark Latonero (US National Institute for Standards and Technology - NIST); Ghazi Ahamat (UK Centre for Data Ethics and Innovation until August 2022); Sarah Box (New Zealand Ministry of Business; Innovation and Employment); Christine Hafskjold (Norway's Ministry of Local Government and Regional Development); Carolina von der Weid (Brazil's Ministry of Foreign Affairs); Zee Kin Yeong, Larissa Lim and Angela Tey (Singapore's Infocomm Media Development Authority); Nobuhisa Nishigata and Takayuki Honda (Japan's Ministry of Internal Affairs and Communications); Roxane Sabourin, Allison O'Beirne and Juliet McMurren (Canada's Ministry of Innovation, Science and Economic Development); and Yordanka Ivanova and Salvatore Scalzo (European Commission Directorate-General for Communications Networks, Content and Technology - CNECT).

The Secretariat would also like to thank stakeholder groups at the OECD for their input, including Pam Dixon (Civil Society Information Society Advisory - CSISAC); Christina Colclough (Trade Union Advisory Committee – TUAC); and Nicole Primmer and Maylis Berviller (Business at OECD – BIAC).

Finally, the authors thank Mika Pinkashov and John Tarver for editing this report and Francesca Sheeka, Orsolya Dobe and Angela Gosmann for editorial support. The overall quality of this report benefited significantly from their engagement.

Table of contents

FIGURES

TABLES

Abstract

Trustworthy AI calls for AI actors to be accountable for the proper functioning of their AI systems in accordance with their role, context, and ability to act.

This report presents research and findings on accountability and risk in AI systems by providing an overview of how risk-management frameworks and the AI system lifecycle can be integrated to promote trustworthy AI. It also explores processes and technical attributes that can facilitate the implementation of values-based principles for trustworthy AI and identifies tools and mechanisms to define, assess, treat, and govern risks at each stage of the AI system lifecycle.

This report leverages OECD frameworks – including the OECD AI Principles, the AI system lifecycle, and the OECD framework for classifying AI systems – and recognised risk-management and due-diligence frameworks like the ISO 31000 risk-management framework, the OECD Due Diligence Guidance for Responsible Business Conduct, and the US National Institute of Standards and Technology's AI risk-management framework.

Résumé

Une IA digne de confiance exige que les acteurs de l'IA soient responsables du bon fonctionnement de leurs systèmes d'IA en fonction de leur rôle, du contexte et de leur capacité d'action.

Ce rapport présente des recherches et des résultats sur la responsabilité et le risque dans les systèmes d'IA en donnant un aperçu de la façon dont les cadres de gestion des risques et le cycle de vie des systèmes d'IA peuvent être intégrés pour promouvoir une IA digne de confiance.

Il explore également les processus et les attributs techniques qui peuvent faciliter la mise en œuvre de principes fondés sur des valeurs pour une IA digne de confiance et identifie des outils et des mécanismes pour définir, évaluer, traiter et gouverner les risques à chaque étape du cycle de vie du système d'IA.

Ce rapport s'appuie sur les cadres de l'OCDE dans le domaine de l’IA - notamment les Principes de l'OCDE relatifs à l'IA, le cycle de vie des systèmes d'IA et le cadre de l'OCDE pour la classification des systèmes d'IA - et sur des cadres reconnus de gestion des risques et de diligence raisonnable, tels que le cadre de gestion des risques ISO 31000, le Guide de l'OCDE sur la diligence raisonnable pour un comportement responsable des entreprises et le cadre de gestion des risques liés à l'IA du l’Institut national des normes et de la technologie (NIST) des États-Unis.

Übersicht

Vertrauenswürdige KI erfordert, dass KI-Akteure für das ordnungsgemäße Funktionieren ihrer KI-Systeme entsprechend ihrer Rolle, ihrem Kontext und ihrer Handlungsfähigkeit verantwortlich sind.

Dieser Bericht präsentiert Forschungsprojekte und Ergebnisse zu Verantwortlichkeit und Risiko in KI-Systemen, und gibt einen Überblick darüber, wie Risikomanagement-Rahmenwerke und der Lebenszyklus von KI-Systemen integriert werden können, um vertrauenswürdige KI zu fördern. Dieser Bericht untersucht außerdem Prozesse und technische Eigenschaften, die die Umsetzung von wertebasierten Prinzipien für vertrauenswürdige KI erleichtern können, und identifiziert Werkzeuge und Mechanismen zur Definition, Bewertung, Behandlung und Steuerung von Risiken in jeder Phase des Lebenszyklus von KI-Systemen.

Dieser Bericht nutzt die OECD-Rahmenwerke - einschließlich der OECD-KI-Grundsätze, des Lebenszyklus von KI-Systemen und des OECD-Rahmenwerks für die Klassifizierung von KI-Systemen - sowie anerkannte Rahmenwerke für Risikomanagement und Sorgfaltspflicht wie das ISO 31000-Risikomanagement-Rahmenwerk, die OECD-Leitlinien für die Sorgfaltspflicht bei verantwortungsvollem Geschäftsgebaren und das KI-Risikomanagement-Rahmenwerk des US National Institute of Standards and Technology.

Background and objectives

This report presents research and findings on accountability and risk in AI systems, building on previous OECD work on AI and the work of experts from the OECD.AI network of experts (Annexes A, B and C). This research contributes to understanding the components of the nascent fields of accountability and risk in AI and instigate discussion of their contribution to trustworthy AI. The report leverages OECD frameworks – including the OECD AI Principles, the AI system lifecycle, and the OECD framework for classifying AI systems – and recognised risk-management and due-diligence frameworks like the ISO 31000 risk-management framework, the OECD Due Diligence Guidance for Responsible Business Conduct, and the US National Institute of Standards and Technology's AI risk-management framework.

This report does not provide precise guidance to assess AI risks and impacts, which is the topic of related work undertaken in co-operation with major AI regulatory and standardisation actors. Rather, it:

1. Provides a comprehensive overview of how risk-management frameworks and the AI system lifecycle can be integrated to promote trustworthy AI;
2. Explores processes and technical attributes that can facilitate the implementation of values-based principles for trustworthy AI (such as the OECD AI Principles); and
3. Identifies tools and mechanisms to define, assess, treat, and govern risks at each stage of the AI system lifecycle.

Executive summary

One of the ten OECD AI Principles refers to the **accountability** that AI actors bear for the proper functioning of the AI systems they develop and use. This means that AI actors must take measures ensure their AI systems are trustworthy – i.e. that they benefit people; respect human rights and fairness; are transparent and explainable; and are robust, secure and safe. To achieve this, **actors need to govern and manage risks throughout their AI systems' lifecycle** – from planning and design, to data collection and processing, to model building and validation, to deployment, operation and monitoring.

The following four important steps can help to manage AI risks throughout the lifecycle: (1) Define scope, context, actors and criteria; (2) Assess the risks at individual, aggregate, and societal levels; (3) Treat risks in ways commensurate to cease, prevent or mitigate adverse impacts; and (4) Govern the risk management process. Risk management should be an iterative process whereby the findings and outputs of one step continuously inform the others.

Defining the scope, context, actors and criteria to evaluate is the first step to managing an AI system's risks, which differ depending on the use case and the circumstances. The context of an AI system includes its socioeconomic and physical environment and its potential impacts on people and on the planet. Examining an AI system's context and scope also includes understanding how the system is developed and maintained, including whether a system is built in-house or by a third party. In addition, analysing trade-offs is key to unlocking AI benefits while managing risks and it is also important to consider AI risks against the risks of not using AI in contexts where it can provide key insights.

Assessing AI risks and impacts means identifying, evaluating and measuring the risks that could affect an AI system's ability to function as intended and in a trustworthy manner. Several tools can help assess risks, such as tools to indicate system transparency, detect bias, identify privacy violations and assess a system's security and robustness.

Treating the risks should be commensurate to their potential adverse impacts. Risk treatment refers to the techniques designed to cease, prevent, or mitigate problems identified during the assessment of an AI system, considering the likelihood and impact of each risk. Responses can be technical, such as implementing privacy-preserving machine learning frameworks or de-identifying training data, or process-related, such as requiring documentation of AI model or training data characteristics or ensuring conformity with safety regulations. For risks that cause adverse impacts, redress mechanisms and remedial actions may be required.

Governance underpins the AI risk management process in two ways. First, it provides a layer of scrutiny over the AI risk management process, including through continual monitoring and review, as well as documenting, communicating, and consulting on actions and outcomes. Second, it offers a variety of mechanisms to embed the AI risk management process into broader organisational governance, fostering a culture of risk management both within organisations and across the entire AI value chain.

- *Monitoring and reviewing* is an continual process taking into account the evolving nature of some AI systems and the environments in which they operate. It includes technical components such as verifying that training data is not out-of-date to avoid "data drift". It also includes non-technical components such as monitoring AI incidents, *i.e.* cases where AI risks materialised into harm.

- *Documenting* the steps, decisions, and actions conducted during risk management and explaining their rationale can bolster accountability if it enhances transparency and enables human review. It means keeping a log or audit trail that informs functions like auditing, certification, and insurance. Whether the AI system is built in-house or by a third party, documentation and logs should "follow the system" throughout the AI system lifecycle. That is, each party or actor – AI developer, data processor AI vendor and AI deployer – might need to conduct its own assessment and document actions taken to manage risks.
- *Communicating* that an AI system meets regulatory, governance, and ethical standards is also crucial since the core objective of AI risk management is to ensure AI systems are trustworthy and safe and protect human rights and democratic values. Where appropriate, it is important to verify and communicate that an AI system conforms to and is interoperable with national and international regulations and standards.
- *Consultation* about processes and results is a core element of trustworthy AI because everyone directly or indirectly involved in or affected by the development or use of an AI system plays a role in ensuring accountability in the AI ecosystem. All actors should manage risks based on their roles, the context, and following the state-of-the-art. Actors in the AI ecosystem include: (1) the suppliers of AI knowledge and resources providing the inputs (i.e. "from whom?"); (2) the actors actively involved in the design, development, deployment and operation of the AI system (i.e. "by whom?"); (3) the users of the AI system (i.e. "for whom?"); and (4) the stakeholders affected by the AI system, including vulnerable groups (i.e. "to whom?").
- *Embedding* a culture of risk management in policies and management systems is needed both across organisations operating AI systems and the AI value chain. A culture of risk management requires strong commitment by organisations' leadership teams.

Synthèse

L'un des dix Principes de l'OCDE sur l'IA a trait à la **responsabilité** qui incombe aux acteurs de l'IA de veiller au bon fonctionnement des systèmes d'IA qu'ils mettent au point et utilisent. En d'autres termes, les acteurs de l'IA doivent prendre des mesures afin de faire en sorte que leurs systèmes d'IA soient dignes de confiance – c'est-à-dire qu'ils servent l'intérêt des individus ; respectent les droits humains et l'équité ; garantissent la transparence et l'explicabilité ; et soient robustes, sûrs et sécurisés. Pour ce faire, les **acteurs doivent assurer la gouvernance et la gestion des risques tout au long du cycle de vie de leurs systèmes d'IA** – planification et conception, collecte et traitement des données, construction des modèles, validation, déploiement, exploitation et suivi.

Quatre étapes importantes peuvent aider à la gestion des risques inhérents à l'IA tout au long du cycle de vie des systèmes : (1) définir le champ, le contexte, les acteurs et les critères ; (2) évaluer les risques aux niveaux individuel, global et sociétal ; (3) traiter les risques de manière proportionnée afin de stopper, prévenir ou limiter les effets négatifs ; et (4) assurer la gouvernance du processus de gestion des risques. La gestion des risques devrait être un processus itératif, de sorte que les conclusions et résultats d'une étape étayent en permanence les autres.

Définir le champ, le contexte, les acteurs et les critères en vue de l'évaluation constitue la première étape de la gestion des risques inhérents à un système d'IA, qui varient selon le cas d'utilisation et les circonstances. Le contexte dans lequel s'inscrit un système d'IA comprend l'environnement socio-économique et physique et ses possibles incidences sur les individus et sur la planète. L'examen du contexte et du champ d'un système d'IA a en outre pour objet de comprendre les modalités de mise au point et de maintenance du système, notamment de déterminer s'il est créé en interne ou par une tierce partie. Par ailleurs, il est essentiel d'analyser les arbitrages afin de tirer parti des avantages de l'IA tout en gérant les risques ; il importe également de mettre en balance les risques inhérents à l'IA et ceux qui découleraient d'une non-utilisation de l'IA là où elle peut apporter des éclairages déterminants.

L'*évaluation* des risques et des incidences propres à l'IA consiste à identifier, évaluer et mesurer les risques susceptibles de compromettre la capacité d'un système d'IA de fonctionner comme prévu et en toute fiabilité. Plusieurs outils peuvent aider à l'évaluation des risques, notamment ceux utilisés pour révéler la transparence du système, détecter les biais, repérer les cas de violation de la vie privée et évaluer la sécurité et la robustesse du système.

Le *traitement* des risques doit être proportionné aux conséquences négatives potentielles. Il désigne les techniques destinées à stopper, prévenir ou limiter les problèmes détectés durant l'évaluation d'un système d'IA, en tenant compte de la probabilité et de l'impact de chacun des risques. Les mesures prises peuvent être d'ordre technique (mise en œuvre de cadres d'apprentissage automatique protégeant la vie privée ou dépersonnalisation des données d'entraînement, par exemple), ou porter sur les processus (obligation de préciser les caractéristiques du modèle d'IA ou des données d'entraînement, ou contrôle du respect des réglementations en matière de sécurité). Pour les risques induisant des conséquences néfastes, il pourrait être nécessaire de mettre en place des mécanismes de recours et des mesures correctrices.

La *gouvernance* sous-tend le processus de gestion des risques liés à l'IA de deux façons. D'une part, elle ajoute une couche de contrôle au processus de gestion des risques inhérents à l'IA, par le biais notamment d'une surveillance et d'un examen continus, et d'activités de documentation, de communication et de consultation sur les actions et les résultats. D'autre part, elle offre divers mécanismes permettant d'intégrer le processus de gestion des risques liés à l'IA à la gouvernance organisationnelle de plus grande envergure, favorisant par là même l'instauration d'une culture de la gestion des risques à la fois au sein des organisations et à l'échelle de l'ensemble de la chaîne de valeur de l'IA.

- La *surveillance* et l'*examen* s'inscrivent dans le cadre d'un processus continu tenant compte du caractère évolutif de certains systèmes d'IA et des environnements dans lesquels ils évoluent. Ce volet comprend des aspects techniques, tels que la vérification effectuée pour s'assurer que les données d'entraînement ne sont pas obsolètes afin d'éviter une « dérive des données ». Il comporte également des éléments non techniques, comme la surveillance des incidents liés à l'IA, à savoir les cas dans lesquels les risques se sont concrétisés et ont causé des préjudices.
- Le fait de *documenter* les étapes, les décisions et les actions entreprises dans le cadre de la gestion des risques et d'en exposer les motifs peut aider à l'attribution des responsabilités si la démarche améliore la transparence et permet une vérification humaine. Cela implique de conserver des journaux ou des pistes d'audit destinés à étayer les activités d'audit, de certification et d'assurance. Que le système d'IA ait été mis au point en interne ou par une tierce partie, la documentation et les journaux devraient « suivre le système », tout au long de son cycle de vie. Il se peut, par conséquent, que chaque partie ou acteur – développeur en IA, sous-traitant IA chargé du traitement des données et responsable du déploiement de l'IA – ait besoin de mener sa propre évaluation et de documenter les mesures prises pour gérer les risques.
- La *communication* quant à la conformité d'un système d'IA au regard des normes de réglementation, de gouvernance et d'éthique est également essentielle dans la mesure où l'objectif premier de la gestion des risques en matière d'IA est de garantir que les systèmes sont dignes de confiance, sûrs, et respectent les droits humains et les valeurs démocratiques. Le cas échéant, il importe de vérifier et de faire savoir qu'un système d'IA est conforme aux réglementations et normes nationales et internationales et interopérable avec elles.
- La *consultation* sur les processus et les résultats est un élément fondamental d'une IA digne de confiance car toute personne qui, directement ou indirectement, intervient dans le développement ou l'utilisation d'un système d'IA ou est concernée par ces étapes, joue un rôle dans l'exercice des responsabilités au sein de l'écosystème de l'IA. Tous les acteurs devraient participer à la gestion des risques en fonction de leur rôle, du contexte et de l'état actuel des connaissances. Les acteurs de l'écosystème de l'IA comprennent : (1) les pourvoyeurs de connaissances et de ressources liées à l'IA, qui en fournissent les intrants (« de qui ? ») ; (2) les acteurs qui participent activement à la conception, au développement, au déploiement et à l'exploitation du système d'IA (« par qui ? ») ; (3) les utilisateurs du système d'IA (« pour qui ? ») ; et (4) les parties prenantes affectées par le système d'IA, notamment les groupes vulnérables (« à qui ? »).
- Il est nécessaire d'*intégrer* une culture de la gestion des risques dans les politiques et les systèmes de gestion, à la fois à l'échelle des organisations qui exploitent les systèmes d'IA et à celle de la chaîne de valeur de l'IA. Une telle culture exige un engagement fort de la part des équipes de direction des organisations.

Zusammenfassung

Einer der zehn KI-Grundsätze der OECD bezieht sich auf die Verantwortung, die KI-Akteure für das ordnungsgemäße Funktionieren der von ihnen entwickelten und genutzten KI-Systeme tragen. Dies bedeutet, dass die KI-Akteure Maßnahmen ergreifen müssen, um sicherzustellen, dass ihre KI-Systeme vertrauenswürdig sind, sprich dass sie dem Mensch zugutekommen, Menschenrechte und Fairness achten, sowie transparent, erklärbar, robust, sicher und geschützt sind. Um dies zu erreichen, müssen die Akteure die Risiken während des gesamten Lebenszyklus ihrer KI-Systeme regeln und managen - von der Planung und Konzeption über die Datenerfassung und -verarbeitung, von der Modellbildung und -validierung bis hin zum Einsatz, Betrieb und zur Überwachung.

Die folgenden vier wichtigen Schritte können dabei helfen, KI-Risiken während des gesamten Lebenszyklus zu managen: (1) Definition von Anwendungsbereich, Kontext, Akteuren und Kriterien; (2) Bewertung der Risiken auf individueller, aggregierter und gesellschaftlicher Ebene; (3) Behandlung von Risiken in einer Weise, die geeignet ist, negative Auswirkungen zu verhindern oder abzuschwächen; und (4) Steuerung des Risikomanagementprozesses. Das Risikomanagement sollte ein iterativer Prozess sein, bei dem die Erkenntnisse und Ergebnisse eines Schrittes kontinuierlich in die anderen Schritte einfließen.

Die Festlegung des Umfangs, des Kontexts, der Akteure und der zu bewertenden Kriterien ist der erste Schritt zum Management der Risiken eines KI-Systems, die sich je nach Anwendungsfall und Umständen unterscheiden. Der Kontext eines KI-Systems umfasst sein sozioökonomisches und physisches Umfeld und seine potenziellen Auswirkungen auf Menschen und den Planeten. Zur Untersuchung des Kontexts und des Anwendungsbereichs eines KI-Systems gehört es auch, zu verstehen, wie das System entwickelt und gewartet wird. Dies beinhaltet auch die Frage, ob ein System intern oder von einem Dritten entwickelt wird. Darüber hinaus ist die Analyse von Kompromissen der Schlüssel zur Erschließung der KI-Vorteile bei gleichzeitiger Beherrschung der Risiken. Es ist überdies wichtig, die KI-Risiken gegen diejenigen Risiken abzuwägen, die dann entstehen, wenn KI in Kontexten, in denen sie wichtige Erkenntnisse liefern kann, nicht eingesetzt wird.

Die Bewertung von KI-Risiken und -Auswirkungen bedeutet, diejenigen Risiken zu identifizieren, zu bewerten und zu messen, die die Fähigkeit eines KI-Systems beeinträchtigen könnten, wie beabsichtigt und auf vertrauenswürdige Weise zu funktionieren. Mehrere Instrumente können bei der Risikobewertung helfen, z. B. Instrumente zur Anzeige der Systemtransparenz, zur Erkennung von Verzerrungen, zur Feststellung von Verletzungen der Privatsphäre und zur Bewertung der Sicherheit und Robustheit eines Systems.

Die Behandlung der Risiken sollte ihren potenziellen negativen Auswirkungen angemessen sein. Die Risikobehandlung bezieht sich auf diejenigen Techniken, mit denen Probleme, die bei der Bewertung eines KI-Systems festgestellt wurden, unter Berücksichtigung der Wahrscheinlichkeit und der Auswirkungen der einzelnen Risiken beseitigt, verhindert oder gemildert werden sollen. Die Maßnahmen können entweder technischer Natur, wie z. B. die Implementierung von datenschutzfreundlichen Frameworks für maschinelles Lernen oder die De-Identifizierung von Trainingsdaten, oder prozessbezogen sein, wie z. B. die Forderung nach einer Dokumentation der Merkmale von KI-Modellen oder Trainingsdaten oder die Gewährleistung der Einhaltung von Sicherheitsvorschriften. Bei Risiken, die nachteilige Auswirkungen haben, können Rechtsbehelfsmechanismen und Abhilfemaßnahmen erforderlich sein.

Governance untermauert den KI-Risikomanagementprozess in zweierlei Hinsicht. Erstens bietet sie eine Kontrollebene für den KI-Risikomanagementprozess, u. a. durch kontinuierliche Überwachung und Überprüfung sowie durch Dokumentation, Kommunikation und Konsultation zu Maßnahmen und

Ergebnissen. Zweitens bietet sie eine Reihe von Mechanismen zur Einbettung des KI-Risikomanagementprozesses in eine umfassendere organisatorische Governance, die eine Kultur des Risikomanagements sowohl innerhalb von Organisationen als auch in der gesamten KI-Wertschöpfungskette fördert.

- Die Überwachung und Überprüfung ist ein fortlaufender Prozess, der die evolvierende Art einiger KI-Systeme und die Umgebungen, in denen sie arbeiten, berücksichtigt. Er umfasst technische Komponenten wie die Verifizierung, dass die Trainingsdaten nicht veraltet sind, um eine "Datendrift" zu vermeiden. Es umfasst auch nicht-technische Komponenten wie die Überwachung von KI-Zwischenfällen, d. h. von Fällen, in denen sich KI-Risiken in Schäden niederschlagen.
- Die Dokumentation der Schritte, Entscheidungen und Maßnahmen, die während des Risikomanagements durchgeführt werden, und die Erläuterung der Gründe dafür können die Verantwortlichkeit stärken, wenn sie die Transparenz erhöhen und eine menschliche Überprüfung ermöglichen. Dies bedeutet, dass ein Protokoll oder ein Audit Trail geführt werden muss, der Funktionen wie Prüfung, Zertifizierung und Versicherung unterstützt. Unabhängig davon, ob das KI-System intern oder von einem Dritten entwickelt wurde, sollten die Dokumentation und die Protokolle dem System während des gesamten Lebenszyklus des KI-Systems "folgen". Das heißt, dass jede Partei bzw. jeder Akteur - KI-Entwickler, Datenverarbeiter, KI-Anbieter und KI-Anbieter - möglicherweise seine eigene Bewertung durchführen und die Maßnahmen dokumentieren muss, die zum Risikomanagement ergriffen wurden.
- Die Mitteilung, dass ein KI-System regulatorische, Governance- und ethische Standards erfüllt, ist ebenfalls von entscheidender Bedeutung, da das Kernziel des KI-Risikomanagements ja darin besteht, sicherzustellen, dass KI-Systeme vertrauenswürdig und sicher sind und die Menschenrechte und demokratischen Werte schützen. Gegebenenfalls ist es wichtig, zu überprüfen und zu kommunizieren, dass ein KI-System mit nationalen und internationalen Vorschriften und Standards konform und interoperabel ist.
- Die Konsultation über Prozesse und Ergebnisse ist ein Kernelement vertrauenswürdiger KI, da jeder, der direkt oder indirekt an der Entwicklung oder Nutzung eines KI-Systems beteiligt oder davon betroffen ist, eine Rolle bei der Gewährleistung der Verantwortlichkeit im KI-Ökosystem spielt. Alle Akteure sollten Risiken auf der Grundlage ihrer Rollen, des Kontexts und nach dem Stand der Technik managen. Zu den Akteuren im KI-Ökosystem gehören: (1) die Anbieter von KI-Wissen und -Ressourcen, die die Inputs liefern (d. h. "woher?"); (2) die Akteure, die aktiv am Entwurf, der Entwicklung, dem Einsatz und dem Betrieb des KI-Systems beteiligt sind (d. h. "von wem?"); (3) die Nutzer des KI-Systems (d. h. "für wen?"); und (4) die vom KI-System betroffenen Interessengruppen, einschließlich gefährdeter Gruppen (d. h. "an wen?").
- Die Verankerung einer Kultur des Risikomanagements in Strategien und Managementsystemen ist sowohl in Organisationen, die KI-Systeme betreiben, als auch in der KI-Wertschöpfungskette erforderlich. Eine Kultur des Risikomanagements erfordert ein starkes Engagement der Führungsteams von Organisationen.

1. Introduction

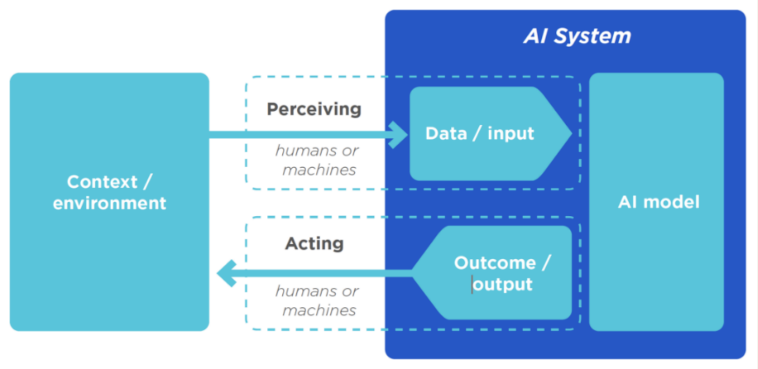
1.1 The need for trustworthy AI

With the rise of artificial intelligence (AI), the legal, ethical, and safety implications of its development and use are becoming increasingly pivotal for governments, businesses, and society (Box 1.1).

AI systems are being developed and used in a range of industries, from transportation and healthcare to agriculture and employment. Governments, businesses, and societies increasingly require assurance that AI systems are trustworthy, as in the financial sector and other domains that leverage risk management and auditing to ensure that their processes abide by certain standards. As such, accountability in AI is an important topic in AI policy discussions.

Box 1.1. What is AI?

The OECD defines an AI system as "a machine-based system that is capable of influencing the environment by producing recommendations, predictions or other outcomes for a given set of objectives. It uses machine and/or human-based inputs/data to: 1) perceive environments; 2) abstract these perceptions into models; and 3) use the models to formulate options for outcomes. AI systems are designed to operate with varying levels of autonomy."



Source: OECD (2022[1]).

1.2 What is trustworthy AI?

The research challenges of identifying, prioritising, and satisfying the attributes that make an AI system trustworthy remain open (NSF, 2022[2]). Several organisations have principles on how to develop AI that benefits all stakeholders (Fjeld et al., 2020[3]; Jobin, Ienca and Vayena, 2019[4]). For example, the values-based OECD AI Principles promote AI that is innovative and trustworthy, and respects human rights and democratic values. These Principles focus on how governments and other actors can shape a human-centric approach to trustworthy AI (Box 1.2).

Box 1.2. Trustworthy AI per the OECD AI Principles

In this report, “trustworthy AI” refers to systems that embody the OECD’s values-based AI Principles:

- **Benefiting people and the planet.** Stakeholders, including civil society and affected communities, should engage in creating credible AI that can contribute to inducing inclusive growth, sustainable development, and wellbeing.
- **Human-centred values and fairness:** the values of human rights, human agency, democracy, and the rule of law should be incorporated throughout an AI system’s lifecycle, while allowing human intervention through safeguard mechanisms.
- **Transparency and explainability:** those who play an active role in the AI system lifecycle (AI actors), including organisations and individuals that deploy or operate AI, should provide information to foster stakeholders’ understanding of the systems, such that people affected by AI systems can comprehend the outcome and challenge the decision when needed.
- **Robustness, security, and safety:** AI systems need to function appropriately while ensuring traceability, and AI actors need to apply systematic risk-management approaches to mitigate, safety and security risks, among others.
- **Accountability:** AI actors should respect the principles and be accountable for the proper operation of AI systems.

The OECD AI Principles also contain five recommendations for national policies and international co-operation. These are: (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.

Also core to other organisations’ principles on AI, the OECD AI Principles constitute the first such standard at the intergovernmental level. Since May 2019, they were adopted by 46 countries and endorsed by the G20.

Source: OECD (2019[5]; 2019[6])

1.3 What is accountability in AI?

The OECD AI Principles state that “AI actors should be accountable for the proper functioning of AI systems and for the respect of the [first four] principles, based on their roles, the context, and consistent with the state-of-the-art” (OECD, 2019[5]). This means AI actors should design, install, and monitor processes that include documenting AI system decisions, conducting or allowing auditing, and providing adequate response to risks and redress mechanisms where justified (OECD, 2019[5]).

Demand is growing in the public and private sectors for tools and processes to help document AI system decisions and to facilitate accountability throughout the AI system lifecycle. The field includes major AI standardisation initiatives, including by the International Organization for Standardization (ISO), Institute of Electrical and Electronics Engineers (IEEE), International Telecommunication Union (ITU), National Institute of Standards and Technology (NIST), European Telecommunications Standards Institute (ETSI), Internet Engineering Task Force (IETF), and European Committee for Electrotechnical Standardization (CEN-CENELEC), with specific strands focusing on AI design (e.g. trustworthiness by design); AI impact, conformity, and risk assessments; and risk-management frameworks for AI. It also includes governmental and intergovernmental initiatives such as the EU’s proposal for a horizontal AI Regulation, the UK’s AI Standards Hub, the European AI Alliance, the Council of Europe’s Committee on Artificial Intelligence (CAI), and the EU-US Trade and Technology Council; certification schemes such as that of the

Responsible AI Institute (RAII), the IEEE CertifAIEd, and Denmark's D-Seal; and risk-management work to provide assurances for trustworthy AI through verification, validation, and auditing.

Such efforts to systematise accountability in AI could improve reliability and enhance trust in the technology and associated practices. Since the magnitude of the challenge will grow, the need for human capital in the AI-accountability industry is expected to increase.

The OECD AI Principles recognise the potential risks[2] AI systems pose to human rights, privacy, fairness, and equality; robustness and safety; and the need to address these, such as by building transparency, accountability, and security into AI systems and enabling continuous monitoring and improvement. The Principles also recognise that different uses of AI present different risks, some of which require higher standards of prevention or mitigation than others.

Risk-management approaches applied throughout the AI system lifecycle can identify, assess, prioritise, and resolve situations that could adversely affect a system's behaviour and outcomes (OECD, 2019[6]). Four steps to manage AI risks while ensuring respect for human rights and democratic values can be identified based on NIST's AI Risk Management Framework, the ISO 31000 risk-management framework (ISO, 2018[7]), and OECD Due Diligence Guidance (OECD, 2018[8]):

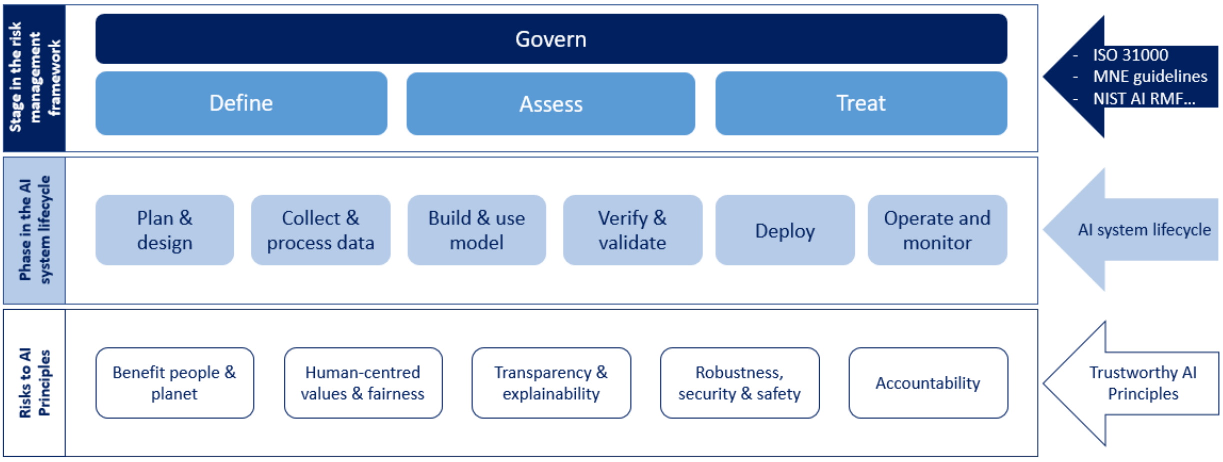
- **Define** scope, context, and criteria, including the relevant AI principles, stakeholders, and actors for each phase of the AI system lifecycle and for the lifecycle itself.
- **Assess** the risks to trustworthy AI by identifying and analysing issues at individual, aggregate, and societal levels and evaluating the likelihood and level of harm (e.g. small risks can add up to a larger risk).
- **Treat** risks to cease, prevent, or mitigate adverse impacts, commensurate with the likelihood and scope of each.
- **Govern** the risk management process by embedding and cultivating a culture of risk management in organisations; monitoring and reviewing the process in an ongoing manner; and documenting, communicating and consulting on the process and its outcomes.

Providing accountability for trustworthy AI requires that actors leverage processes, indicators, standards, certification schemes, auditing, and other mechanisms to follow these steps at each phase of the AI system lifecycle (Figure 1.1). This should be an iterative process where the findings and outputs of one risk-management stage feed into the others.

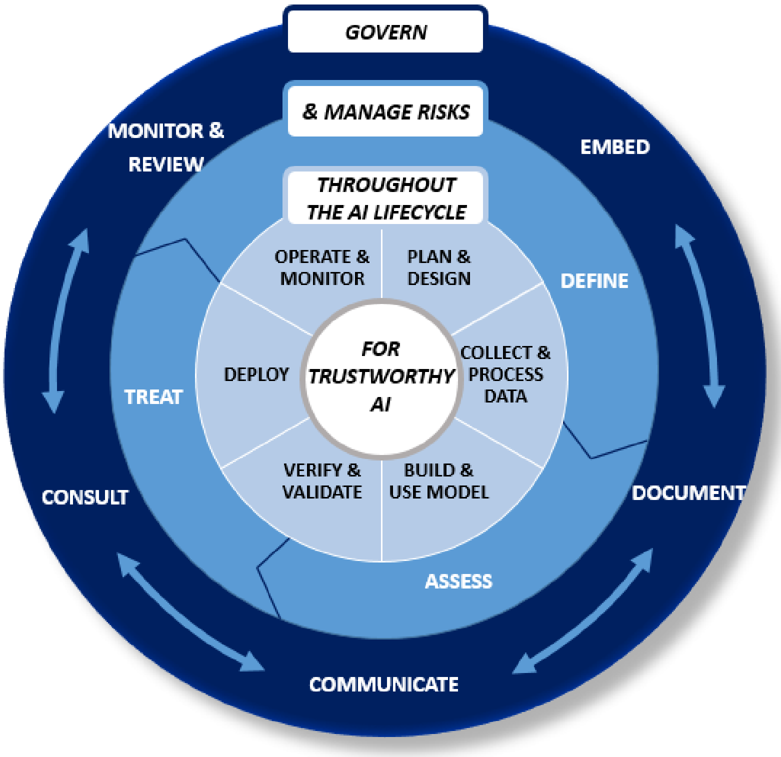
Figure 1.1. High-level AI risk-management interoperability framework

Governing and managing risks throughout the lifecycle for trustworthy AI.

a) Structural view

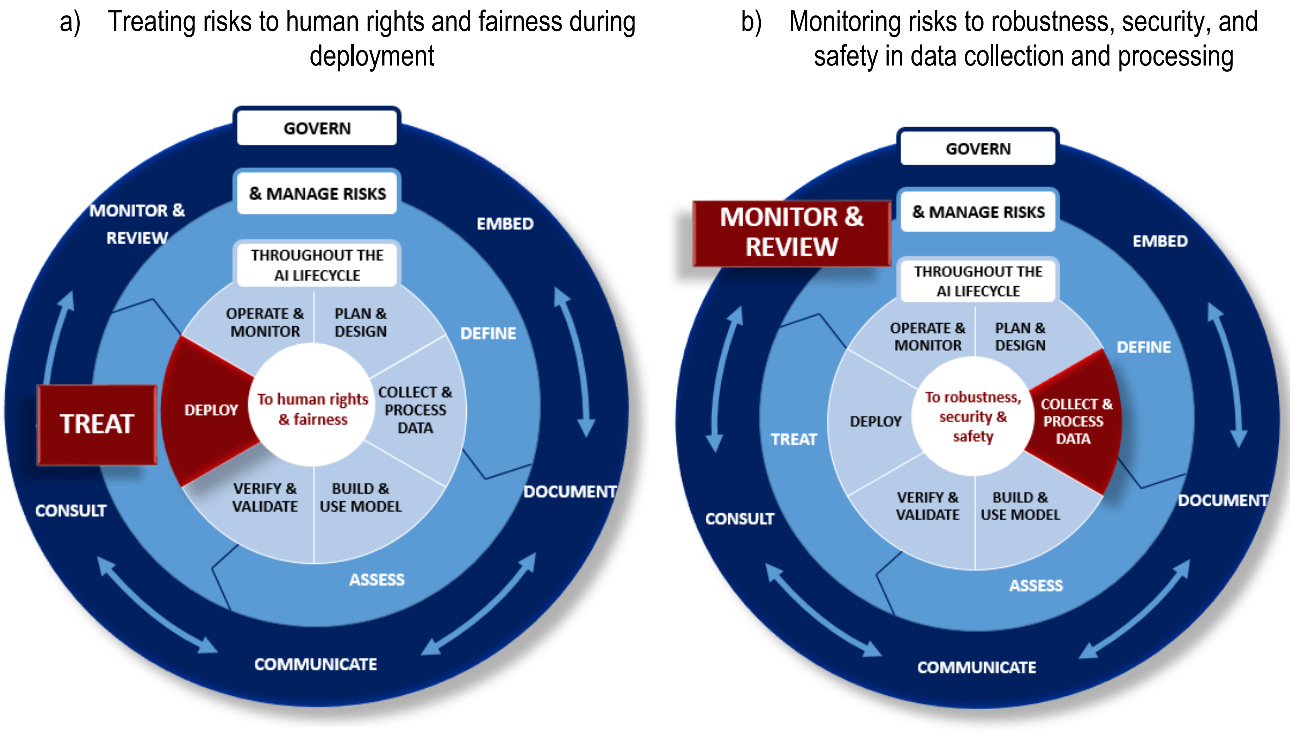


b) Functional view



The high-level AI risk management framework offers a systematic way to govern and manage risks to trustworthy AI at each phase of the AI system lifecycle. The framework's graphical representation enables the users to situate themselves visually at different steps of the process. For example, treating risks to human rights, values, and fairness (e.g. bias and discrimination) during deployment, or monitoring risks to robustness, security, and safety (e.g. adversarial attacks) during data collection and processing would each receive a different graphical representation (Figure 1.2).

Figure 1.2. Sample uses of the high-level AI risk management interoperability framework



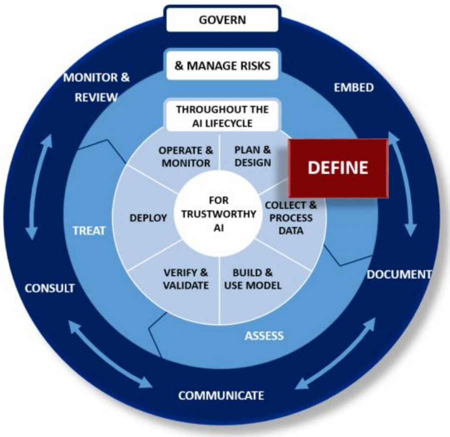
a) Treating risks to human rights and fairness during deployment

b) Monitoring risks to robustness, security, and safety in data collection and processing

The following sections illustrate the use of the high-level AI risk management interoperability framework to define (Section 2), assess (Section 3), treat (Section 4), and govern (Section 5) AI risks. Section 6 presents next steps.

2. DEFINE: Scope, context, actors, and criteria

2.1 Scope



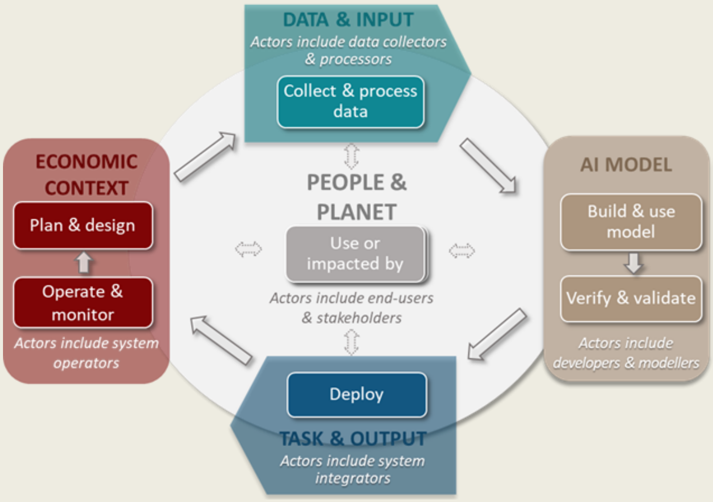
The AI system lifecycle can serve as a tool to understand and analyse a system's scope and its characteristics. It encompasses the following phases: (a) plan and design; (b) collect and process data; (c) build and use[3] the model; (d) verify and validate the model; (e) deploy (including "putting into service" and "placing the AI system on the market")[4]; and (f) operate and monitor the system (OECD, 2019[5]).

These phases often take place in an iterative manner but are not necessarily sequential. The decision to retire an AI system from operation can occur at any point during the operating and monitoring phase. The lifecycle phases can be associated with the key dimensions of an AI system (Box 2.1).

Box 2.1. Mapping the lifecycle phases to the dimensions of an AI system

The OECD Framework for the Classification of AI Systems characterises AI systems and applications along the following dimensions: (1) People & Planet; (2) Economic Context; (3) Data & Input; (4) AI Model; and (5) Task & Output. Each dimension has its own properties, attributes, or sub-dimensions relevant to assessing policy considerations of AI systems.

The phases of the AI system lifecycle can be associated with the dimensions of the OECD Framework for the Classification of AI Systems. For example, the "collect & process data" phase can be associated with the "data & input" dimension, while the "deploy" phase can be associated with the "task & output" dimension. This mapping is useful to identify AI actors in each dimension, with accountability and risk-management implications.


DATA & INPUT
Actors include data collectors & processors
Collect & process data
ECONOMIC CONTEXT
Plan & design
Operate & monitor
Actors include system operators
PEOPLE & PLANET
Use or impacted by
Actors include end-users & stakeholders
AI MODEL
Build & use model
Verify & validate
Actors include developers & modellers
Deploy
TASK & OUTPUT
Actors include system integrators

Note: The actors included in the visualisation are illustrative, not exhaustive and based on previous OECD work on the AI system lifecycle.

Source: OECD (2022[1]).

2.2 Context

Risk management practices differ by context and use case. The context of an AI system represents its socioeconomic environment – including its natural and physical environment. Context is more relevant to the specific application of an AI system than to an AI system in general. Context is observable and can be influenced by actions that result from an AI system's outputs (OECD, 2019[6]).

The OECD Framework for the Classification of AI Systems could be useful to define a system's context and scope. For instance, according to this Framework, core characteristics of the economic context include the sector in which an AI system is deployed, its business function, its critical (or non-critical) nature, and its deployment scale and impact on critical functions and activities. The system operator plays a role in determining and analysing the context of an AI system (OECD, 2022[1]). Using an AI system in ways that differ from its intended use (e.g. secondary uses or misuses) might require a re-assessment of the context and risks.

People interact with AI systems in many ways, including designing the system, defining the task it will perform, deciding what data to collect and how to collect it, labelling data, deciding what model to use, choosing baselines and performance criteria, putting in place evaluation mechanisms, deploying and using the system or being affected by its outputs. It is crucial to consider a system's potential impacts on human rights, well-being, and sustainability throughout its lifecycle, including downstream effects and negative externalities. Additional elements to consider from the Classification Framework include: the AI competencies of the system's end-users; the rights of all stakeholders (e.g. workers, consumers, vulnerable populations, etc.); the system's optionality and redress mechanisms; its benefits and risks to human rights, democratic values, the environment, well-being, and society; and its employment displacement potential (Charisi et al., 2022[9]; OECD, 2022[1]).

Examining a system's context and scope also includes understanding whether the system is built in-house or by a third party, and defining the system development and maintenance schemes. In the case of third-party AI systems, three set ups were identified (BSA, 2021[10]; OECD, 2022[1]):

- **Universal:** the system developer provides AI actors, users or stakeholders with access to a single, pre-trained model.
- **Customisable:** the system developer provides a model that can be customised and/or re-trained by other AI actors, for example, by using different data.
- **Tailored:** the system developer trains a model on behalf of an AI actor or stakeholder using the AI actor or stakeholder's data.

Understanding how an AI system was developed and is maintained is key for AI governance and accountability, and facilitates the allocation of roles and responsibilities throughout the risk management process. It helps define the rights and responsibilities the deployer and user of an AI system has vis-à-vis its developers and vendors. For instance, a developer that trained and maintained a universal AI system on behalf of others would be best-placed to address most risk-management aspects throughout its lifecycle. The same is true for tailored AI systems, where the bulk of risk-management responsibilities fall on the entity that develops and trains the model. However, in the case of customisable AI systems (including general-purpose AI systems), many risk-management responsibilities would likely shift to the organisation that re-trains or customises the model.

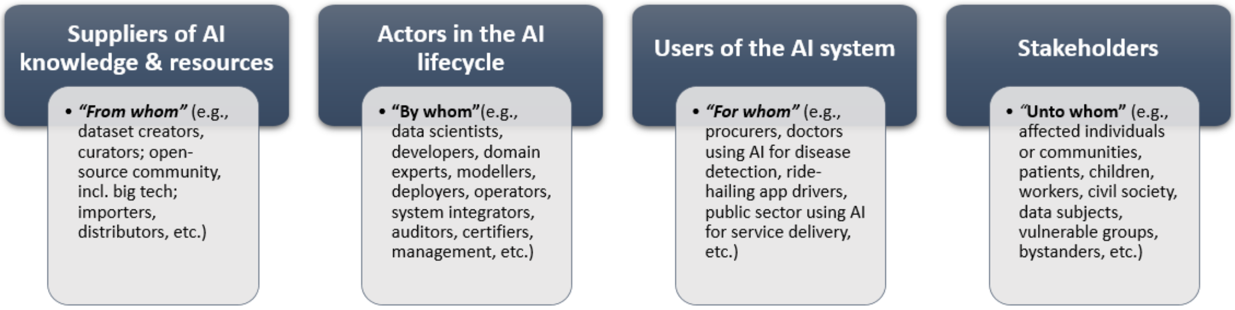
Taking a balanced approach to managing risks without violating human rights or stifling innovation is key. The risks of AI should be balanced against the risks of not using AI in contexts where it can provide crucial benefits and insights.

2.3 Actors

Accountability in the AI ecosystem should be shared by everyone directly and indirectly involved in or affected by the development or use of an AI system. All actors should manage risks based on their roles, the context, and following the state-of-the-art (OECD, 2019[5]). Four questions help identify the actors in the AI ecosystem (Figure 2.1): (1) From whom? – the suppliers of AI knowledge providing the inputs; (2)

By whom – the actors actively involved in the design, development, deployment, and operation; (3) For whom? – the users of the AI system; and (4) Unto whom? – the stakeholders affected by the AI system. An analysis of the actors that should be involved in the risk management process – as well as their roles and responsibilities – should be undertaken during the define stage.[5]

Figure 2.1. Actors in an AI accountability ecosystem



Suppliers of AI knowledge & resources

- *"From whom"* (e.g., dataset creators, curators; open-source community, incl. big tech; importers, distributors, etc.)

Actors in the AI lifecycle

- "By whom"(e.g., data scientists, developers, domain experts, modellers, deployers, operators, system integrators, auditors, certifiers, management, etc.)

Users of the AI system

- *"For whom"* (e.g., procurers, doctors using AI for disease detection, ride-hailing app drivers, public sector using AI for service delivery, etc.)

Stakeholders

- "Unto whom" (e.g., affected individuals or communities, patients, children, workers, civil society, data subjects, vulnerable groups, bystanders, etc.)

Suppliers of AI knowledge and resources ("from whom")

AI knowledge refers to the skills and resources, such as data, code, algorithms, models, research, know-how, training programmes, governance, processes and good practices required to understand and participate in the AI system lifecycle (OECD, 2019[5]).

Such inputs to an AI system can come from external actors or suppliers not directly involved in the development of the system.[6] This includes dataset creators and curators; the open-source community (including big tech companies) that develop pre-packaged code and AI applications for secondary public use; and importers and distributors of AI system components, among others.

Where appropriate, these actors should provide metadata and instructions for how the inputs should (and should not) be used to prevent risks and avoid malicious secondary uses or misuses of their inputs. Actors leveraging these inputs should be accountable for their appropriate, legal, and ethical use.

Actors in the AI system lifecycle ("by whom")

One benefit of leveraging the AI system lifecycle as a framework to manage AI risks is that it allows the identification of actors in each of its phases (OECD, 2019[6]) (OECD, 2022[1]). AI actors that play a role in the AI system lifecycle include data collectors, developers, modellers, deployers, and system integrators (OECD, 2019[5]). Auditors, certifiers, and supervisory authorities also play a role.

AI actors in the "Plan and design" phase are managers, end-users, and stakeholders involved in the definition of a system's objectives, intended purpose, its end-users and underlying assumptions, and the assessment of the context's legal and ethical requirements (OECD, 2019[6]). The planning and design phase is critical as failures and issues related to other lifecycle phases can be avoided if addressed at this phase. Planning and design might require expertise from data scientists, and domain and governance experts. Moreover, participation of a diverse group of actors and stakeholders – including civil society and affected communities – is desirable and "ethics-by-design" and "inclusive AI" approaches could be leveraged in this phase (European Commission, 2020[11]; Park, 2022[12]).

Actors in the "Collect and process data" phase include data collectors and data processors involved in gathering and cleaning data, labelling, performing checks for completeness and quality, and documenting the characteristics of the dataset. Dataset characteristics include information on how a dataset is created, its composition, intended uses, and how it is maintained over time (OECD, 2019[13]). Data collection and processing currently involve expertise from actors such as data scientists, domain experts, data engineers, and data providers (OECD, 2022[1]).

Actors in the "Build and use the model" phase include developers and modellers involved in creating or selecting algorithms and models, their calibration, training, and inferencing or use. Model building and inferencing involve human experts such as modellers, developers, model engineers, data scientists, and domain experts (OECD, 2022[1]).

Actors in the "Verify and validate" phase include developers, modellers, and auditors involved in the testing and tuning of models to assess and improve performance across various dimensions and considerations. Currently, model verification and validation involve data scientists; data, model and systems engineers; and governance experts (OECD, 2022[1]). This phase can also include internal and external auditors and certifiers. Characteristics of the team that builds, tests, and assesses performance across various dimensions and considerations of an AI system – such as gender, country, and cultural background – have been shown to impact the fairness of the system's outputs, as developers can incorporate unconscious biases (Freire, 2021). This may result in advocacy for diversity in teams that build and validate AI systems (OECD, 2022[1]).

Actors in the "Deploy" phase include providers placing the systems on the market or system integrators involved in piloting the deployment of the AI system, checking compatibility with legacy systems, ensuring regulatory compliance, managing organisational change, and evaluating user experience. Other experts involved in the deployment of an AI system include developers, systems and software engineers, testers, managers, and domain experts (OECD, 2022[1]).

Actors in the "Operate and monitor" phase include system operators in the continuous assessment of the system's outputs and impacts (both intended and unintended) against its objectives and the ethical considerations of its operation. Among other things, system operators identify and address issues directly, revert to other phases of the lifecycle or, if necessary, retire an AI system from production. In some cases, the deployer also performs operating and monitoring functions. Human-in-the-loop approaches are key in this phase. Operation and monitoring currently involves expertise such as governance experts, managers, domain experts, and systems and software engineers (OECD, 2022[1]). This phase can also include internal and external auditors, certifiers, supervisory authorities, and organisations that "monitor the monitors".

Users of the AI system ("for whom")

The users of an AI system or application are individuals or organisations that use it to achieve a specific task or objective. They include procurers who acquire AI models, products, or services from a third party, developer, vendor, or contractor (NIST, 2022[14]). Examples include doctors that use AI for disease detection, public sector agencies embedding AI in administrative or security tasks, and financial institutions deploying AI to improve fraud detection. Users should be accountable for the legal and ethical use of an AI system, and have a role in monitoring and reporting its risks and impacts. In some cases, users also perform operating and monitoring functions. Continuous re-skilling and up-skilling is crucial in this regard.

Stakeholders ("unto whom")

Stakeholders encompass all organisations and individuals affected by AI systems, directly or indirectly. Stakeholders do not necessarily interact with the system (OECD, 2022[1]) and may include "bystanders" (e.g. pedestrians affected by self-driving cars). This broad group encompasses civil society, the technical and academic communities, industry, governments, labour representatives and trade unions, and as workers or data subjects. AI suppliers, actors and users can in some cases also be affected by the AI system and thus belong to the stakeholders group (OECD, 2019[5]). Incident-reporting mechanisms and awareness-raising campaigns can help stakeholders monitor downstream risks, negative externalities, and risks that materialise despite a system working as intended. They can also identify secondary uses or misuses of an AI system – or parts of it – for malicious purposes. Regulators play a key role in protecting stakeholders' rights.

2.4 Criteria

AI risks can be evaluated at different levels, including at a governance and process level – focusing on risks to values-based principles (e.g. accountability) – and at a technical level, focusing on technical risks (e.g. robustness and performance), and underlying sub-risks (e.g. statistical accuracy).

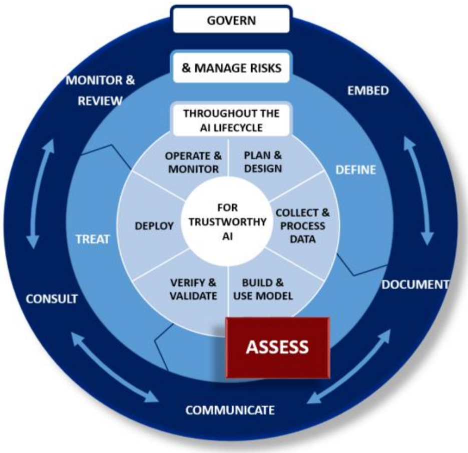
A step towards ensuring accountability in AI is linking the Principles with specific procedural and technical attributes (Table 2.1). While some existing frameworks provide AI actors with substantial guidance – such as the taxonomy of AI trustworthiness in Newman (2023[15]) – turning value-based principles into specific technical requirements and attributes is a rapidly evolving field.

Table 2.1. Sample processes and technical attributes per OECD AI Principle

Principle	***Sample processes and technical attributes***
Benefiting people and the planet	Performance, energy consumption, environmental impact
Human-centred values and fairness	Bias and discrimination, privacy and data governance
Transparency and explainability	Interpretability, documentation, traceability, disclosure, redress mechanisms (including the ability to opt out, when appropriate)
Robustness, security, and safety	Reliability, reproducibility, safety, and vulnerability to tampering
Accountability	Roles and responsibilities, risk management, ongoing processes for continuous improvement

The next section describes some of these processes and technical attributes in detail and explores illustrative tools to assess risks in AI systems.

3. ASSESS: Identify and measure AI risks



Once the scope, context, actors and criteria for an AI system are defined, it is important to assess the risks it poses and which could result in the AI system failing to meet its trustworthiness objectives. This process consists of identifying or discovering risks, analysing the mechanisms by which those risks may occur, and evaluating their likelihood of occurring as well as their severity.

This section provides an overview of concepts, processes, and measures that can help assess risks to trustworthy AI. Some are relevant to multiple Principles in varying degrees. For example, "accuracy" could be relevant to benefiting people and planet through its connection with productivity, and to robustness, security and safety as a system-centric evaluation criterion. To avoid repetition, such multifaceted concepts and measures appear under their most relevant Principle.

3.1 Benefiting people and the planet

Guiding the development and use of AI toward benefiting people and the planet is imperative. Trustworthy AI can advance inclusive growth, sustainable development, and well-being and global development objectives. AI can be leveraged for social good and contribute to achieving the Sustainable Development Goals (SDGs) in education, health, transport, agriculture, environment, and sustainable cities, among other areas (OECD, 2022[16]).

Throughout the AI system lifecycle, AI actors and stakeholders can and should encourage the development and deployment of AI with appropriate safeguards for beneficial outcomes. Multidisciplinary, multi-stakeholder collaboration and social dialogue can help define these beneficial outcomes and how best to achieve them (OECD, 2022[16]).

Ensuring that AI systems benefit people and the planet entails assessing and improving their performance, accuracy, and sustainability. It also entails assessing downstream risks to economic inclusion and well-being. Concepts include:

- **Accuracy**: an AI system's ability to perform the task for which is was developed, such as classifying information into correct categories or making predictions and recommendations that are then verified. Accuracy can be quantified by estimating how well the system works through error rates or metrics like the "expected generalisation performance" (Arlot and Celisse, 2010[17]). Improving systems' accuracy and performance can enhance productivity and economic growth, and could thus enhance well-being (e.g. by improving health-related outcomes) and decrease financial and environmental costs.[7]
- **Sustainability:** the computing power ("compute") used to train AI models has grown exponentially in recent years, affecting workloads and energy consumption at data centres. On the one hand, advances in data science and AI chips manufacturing as well as novel computing architectures are enabling more efficient AI models that leverage smaller training datasets and perform fewer training runs. This leads to a more sustainable use of computational resources, also aided by growth in clean energy to power data centres (Strier, Clark and Khareghani, 2022[18]).

On the other hand, there is ongoing debate about the trade-off between general-purpose AI, including large language models (LLMs), and purpose-specific AI. LLMs require large volumes of training data and computational power, and consume more energy than purpose-specific AI. Higher energy consumption should thus be weighed against the benefits of these systems (Bender et al., 2021[19]).

- **Well-being and economic inclusion:** where appropriate, AI actors should assess the possible downstream impacts of their AI systems on people's well-being and economic inclusion, including impacts and negative externalities for vulnerable populations (in particular children and disadvantaged groups) as well as impacts on job quality and the potential for automation.

3.2 Human-centred values and fairness

AI should be developed based on human-centred values, including human rights, fundamental freedoms, equality, fairness, the rule of law, social justice, data protection and privacy, and consumer rights and commercial fairness (OECD, 2022[20]).

Some uses of AI systems have implications for human rights, including risks that these (as defined in the Universal Declaration of Human Rights) and human-centred values can be deliberately or accidently infringed. It is therefore important to promote "rights and values alignment" in AI systems (i.e. their design with appropriate safeguards), including capacity for human intervention, oversight, and redress, as appropriate to the context. This approach can ensure that AI systems' behaviours protect and promote human rights and align with human-centred values throughout their operation. Remaining true to democratic values can strengthen public trust in AI and support its use to, for example, reduce discrimination or other unfair and/or unequal outcomes (OECD, 2022[20]).

Measures such as human-rights impact assessments (HRIAs), human-rights due diligence, human determination, and human involvement in the AI process (i.e. "human-in-the-loop" approaches), codes of ethical conduct, and quality labels and certifications play a role in promoting human-centred values and fairness (OECD, 2022[20]). Three categories of AI risks to these are: (1) risks of bias and discrimination; (2) risks to privacy and data governance; and (3) risks to other human rights and democratic values.

Bias and discrimination

AI systems can perpetuate bias, exclusion and have a disparate impact on vulnerable and underrepresented populations, such as ethnic minorities, children, the elderly, and the less educated or lower skilled. The underrepresentation of women in some training datasets may also result in biased outputs. Disparate impact is a particular risk in low- and middle-income countries, given the lack of data from these countries to train AI systems in their specificities, and their lower representation overall in the AI industry. Fairness implies that AI can and should empower all members of society and help reduce biases and exclusion.

Identifying biases is a challenge for actors throughout the AI system lifecycle (Box 3.1). Sources of bias include (IDB-OECD, 2021[21]; Barocas and Selbst, 2016[22]):

- **Historical bias**: Pre-existing patterns in the training data, e.g. societal biases.
- **Representation bias (and limited features):** Incomplete information due to missing attributes, inadequate sample size, or total or partial absence of data from sub-populations.
- **Measurement bias:** Omission (or inclusion) of variables that should (or not) be in the model, including proxies for protected attributes or groups (e.g. neighbourhood as a proxy for race).
- **Methodological and evaluation bias**: Errors in the definition of metrics (e.g. erroneous assumptions about a target population), model validation and calibration, and evaluation of results.

- **Monitoring bias and skewed samples**: Inappropriate interpretation of a system's results during monitoring, initial biases that compound over time and skew training data, or temporary changes in the way data is captured.
- **Feedback loops and popularity bias**: Recommendation algorithms suffer from popularity bias, where a few popular items are recommended frequently to users. This creates a feedback loop where frequent recommendations get more reactions and are thus recommended more frequently.

Box 3.1. Errors, biases, and noise: a technical note

System error is the difference between a value predicted by the model and the real value of the variable that is being estimated. **Bias** is when an error systematically favours a specific subset of data or a specific subpopulation. For example, if a variable's predicted value is consistently lower for one subgroup in the data, such as the salary of women with respect to equally qualified men for an equivalent job, the salary variable is biased. Conversely, **noise** is when the error is random.

Source: IDB-OECD (2021[21]).

Different stakeholders have different perspectives on fairness and equity, and, as socio-technical systems, AI applications require expertise beyond that of technologists. To diagnose and mitigate biases in AI, it is important to differentiate between individual and group-level fairness (OECD, 2022[20]). *Individual fairness* means that similar individuals should be treated similarly; and *group fairness* means that the outcomes of an AI system should not vary if the population is split into groups (e.g. by protected attributes).

An important strand of literature seeks to implement mathematical fairness metrics to assess a model's impartiality towards subgroups (IDB-OECD, 2021[21]; Chouldechova, 2017[23]; Kleinberg, Mullainathan and Raghavan, 2016[24]; Corbett-Davies et al., 2017[25]; Koshiyama et al., 2021[26]). Different fairness metrics lead to varying ways to assess bias in a system. Examples include:

- **Equality of opportunity:** Belonging or not to a protected group should not influence an AI system's output. A mathematical definition often used is the Average Odds Difference (Bellamy et al., 2018[27]).
- **Equality of outcome or statistical parity:** Each segment of a protected group (e.g. gender or race) must obtain an output in the same proportion. Statistical Parity Difference is the mathematical representation generally accepted for this concept (Bellamy et al., 2018[27]).
- **Counterfactual justice:** An AI system is considered fair if its outcomes remain the same when the value of the protected attribute is modified, such as when there is a change in race or gender.

The decision about which AI fairness metric to use should consider the context, and its rationale should be documented (IDB-OECD, 2021[21]). In practice, no single AI fairness measure works for all problems and complying with one definition usually means not fully complying with the others (Chouldechova, 2017[23]). Therefore, complying with a given definition of fairness might not guarantee that the outcomes of an AI system are fair.

Privacy and data governance

Unless they are consistent with human rights, and fundamental and democratic values, AI systems can cause or exacerbate impacts of asymmetries in power and access to information, such as between employers and employees, businesses and consumers, or governments and citizens (EU-HLEG, 2019[28]).

When an AI system includes intellectual property, the rights to the model and its parameters must be preserved. In addition, in cases such as medical applications and others, the privacy of the training data needs to be preserved. Data protection in AI systems relates to preventing the exposure of the model and its training data (De Cristofaro, 2020[29]). Data-governance mechanisms should be in place to ensure the quality and integrity of the data used to train the model; its relevance in the system's deployment context; its access protocols; and the model's capacity to process data in a manner that protects privacy and sensitive information. Issues include:

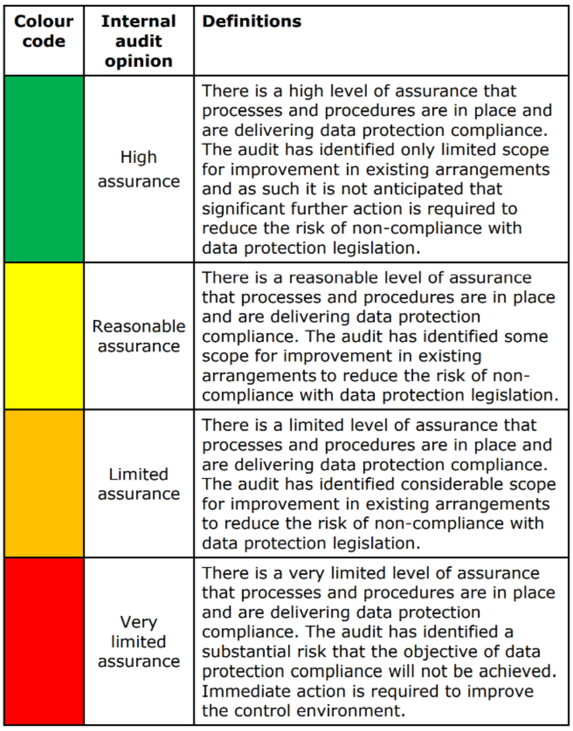
- **Privacy and data protection:** AI systems should respect privacy and data protection throughout their lifecycle (OECD, 2019[5]). This includes information provided by users and user data generated through interaction with the system. Data access and disposal protocols outlining who can access and delete data, and under which circumstances, should also be put in place (Butterworth, 2018[30]).
- **Model security:** The security and privacy of an AI model can be assessed based on: (1) the access level a malicious actor might have, from "black box" (e.g. no knowledge about the model)[8] to "full transparency" (e.g. full information about the model and its training data); (2) the phase in which an attack might happen (e.g. during training or inference); and (3) whether passive (e.g. "honest but curious") or active (e.g. fully malicious) attacks are possible (De Cristofaro, 2020[29]).

Risks to privacy and data governance can arise at the data and the model levels, at their intersection, as well as during the interaction between the human and the AI system. Methods to assess these risks include:

- **At the data level:** Data protection impact assessment is the standard procedure to assess risks (Bieker et al., 2016[31]). This procedure is legally formalised in some jurisdictions, including the EU and the UK (Figure 3.1). These assessments should account for risks of data poisoning, where the training data is maliciously manipulated to affect a model's behavior (Tan and Shokri, 2019[32]).
- **At the model level:** Risks to privacy and data protection at the model level include attempts to infer model parameters and build "knockoff" versions or copies of the model. Techniques that aim to extract a full copy or equivalent version of a model, or to copy some of its functionalities could help AI actors assess vulnerability at the model level (Ateniese et al., 2015[33]; Tramèr et al., 2016[34]; Orekondy, Schiele and Fritz, 2019[35]).
- **At the intersection of data and model levels:** Risks include making inferences about certain members of the population or of the training dataset through its interactions with the model. Techniques to assess vulnerability levels include: statistical disclosure (Dwork and Naor, 2010[36]); model inversion (Fredrikson, Jha and Ristenpart, 2015[37]); inferring class representatives (Hitaj, Ateniese and Perez-Cruz, 2017[38]); and membership and property inference (Shokri et al., 2017[39]; Ganju et al., 2018[40]; Melis et al., 2019[41]).
- **At the human-AI interaction:** Training, checklists and verification processes could help identify risks to privacy and data governance arising from the interaction between the human and the system (e.g. unintentional actions – or lack of action – by developers or users that compromise the privacy or data governance of an AI system).

Figure 3.1. UK Information Commissioner's Office (ICO) qualitative rating for data protection

Colour-coded assessment of an AI system's risks to privacy and data governance at the data level.



Colour code	Internal audit opinion	Definitions
🟩	High assurance	There is a high level of assurance that processes and procedures are in place and are delivering data protection compliance. The audit has identified only limited scope for improvement in existing arrangements and as such it is not anticipated that significant further action is required to reduce the risk of non-compliance with data protection legislation.
🟨	Reasonable assurance	There is a reasonable level of assurance that processes and procedures are in place and are delivering data protection compliance. The audit has identified some scope for improvement in existing arrangements to reduce the risk of non-compliance with data protection legislation.
🟧	Limited assurance	There is a limited level of assurance that processes and procedures are in place and are delivering data protection compliance. The audit has identified considerable scope for improvement in existing arrangements to reduce the risk of non-compliance with data protection legislation.
🟥	Very limited assurance	There is a very limited level of assurance that processes and procedures are in place and are delivering data protection compliance. The audit has identified a substantial risk that the objective of data protection compliance will not be achieved. Immediate action is required to improve the control environment.

Source : ICO (2022[42]).

Advanced privacy-enhancing technologies (e.g. homomorphic encryption, secure multi-party computation, and differential privacy) and novel approaches to training (e.g. using data combined from multiple organisations, federated machine learning) can be leveraged to protect an AI system and increase its privacy (OECD, 2022[20]; De Cristofaro, 2020[29]). The impacts of such mitigations differ according to context and other variables, to be tested on a system-by-system basis. Additionally, emerging data governance models (such as data sharing pools, data cooperatives, and data trusts) could help prevent risks to privacy and promote the democratisation of data governance (Micheli et al., 2020[43]).

Human rights and democratic values

AI "may have disparate effects within and between societies and economies, notably regarding economic shifts, competition, transitions in the labour market, inequalities, and implications for democracy and human rights, privacy and data protection, and digital security" (OECD, 2019[5]).

AI can support the fulfilment of human rights or create new risks that human rights might be deliberately or accidently violated. Human rights law, together with other legal and institutional structures, could serve as a tool to help ensure human-centred AI (Box 3.2).

Box 3.2. Human rights and AI

International human rights refer to a body of international laws, including the International Bill of Rights, and regional human rights systems developed around the world over the past 70 years. Human rights set universal minimum standards based on, among other things, values of human dignity, autonomy, equality, and the rule of law. These standards and the legal mechanisms linked to them create legally enforceable obligations for countries to respect, protect, and fulfil human rights. They also require that those whose rights have been denied or violated be able to obtain remedy.

Recent intergovernmental instruments such as the United Nations (UN) Guiding Principles on Business and Human Rights (OHCHR, 2011[44]) also address private actors in the context of human rights. They confer upon private actors the responsibility to respect human rights. In addition, the 2011 update of government-backed recommendations to business in the OECD Guidelines for Multinational Enterprises (OECD, 2011[45]) contains a chapter on human rights.

Specific human rights include equality, non-discrimination, freedom of expression and association, privacy, and economic, social and cultural rights such as education or health. Human rights also overlap with wider ethical concerns and other areas of regulation relevant to AI, such as personal data protection or product safety law. However, these often have different scope.

Source: OECD (2019[46]).

Human rights frameworks create obligations for the identification and management of AI risks to human rights, including those of marginalised and vulnerable groups. This can be done through human-rights due diligence such as human-rights impact assessments (HRIAs). HRIAs can identify risks that AI system lifecycle actors might not otherwise envisage. To that end, they focus on incidental human rights impacts rather than optimisation of the technology or its outputs. HRIAs or similar risk management processes could help ensure respect for human rights by design throughout the lifecycle of the AI system (OECD, 2019[46]). Regular HRIAs should be carried out at each phase of the lifecycle and when there are changes in context, scope, nature, and purpose of the systems (Council of Europe, 2019[47]).

Examples of HRIAs include Canada's Algorithmic Impact Assessment (AIA), a mandatory tool to support Canada's Treasury Board's Directive on Automated Decision-Making, and the Dutch Impact Assessment *Mensenrechten en Algoritmes* (IAMA), made mandatory by the Dutch Parliament for any algorithm designed to support decision-making in the public and private sectors. Additionally, the "Benefitting people and the planet" dimension of the OECD framework for classifying AI systems includes a sample checklist for assessing the potential impact of an AI system on well-being, and selected human rights and democratic values (OECD, 2022[1]).

HRIAs should also assess risks arising from data labelling and enrichment practices. Multiple studies (Gray and Suri, 2019[48]; PAI, 2021[49]) highlight that the conditions under which data-enrichment labour is sourced and performed are often non-transparent, enabling poor treatment of workers. This is an issue with direct relevance to the accountability of AI actors and one that impacts training data quality. The guidance and toolbox from the Danish Institute for Human Rights illustrates the use of HRIAs for assessing and addressing the adverse impacts of business activities on human rights.

Other risks related to human rights and democratic values to consider include how an AI system's deployment affects the distribution and balance of power across stakeholder groups, and its impacts on human agency through, for instance, manipulation and polarisation of opinions at scale. Although identification of such macro-level risks can be challenging, it is key to accountability in the AI ecosystem.

Trade-offs might exist between different human rights. To reconcile them, international human rights law allows restrictions on different rights and freedoms if these pass the "three-part test" of legality, legitimacy, and proportionality and necessity. HRIAs could facilitate balancing such trade-offs in the design and development stages of an AI system. For example, they could examine if a design choice is in compliance with applicable legislation and proportionate and necessary given its potential impacts on various human rights and the interests of different stakeholders (Arai-Takahashi, 2002[50]; Greer, 2004[51]).

3.3 Transparency and explainability

Explainability and interpretability

Being able to provide clear and meaningful explanations of an AI system's outcomes is crucial to building and maintaining users' trust (Longo et al., 2020[52]). Explainability implies that an AI system should provide plain and easy-to-understand information on the factors and decision processes that serve as the basis for its prediction, recommendation, or decision (OECD, 2022[53]).

Users of explainable AI systems benefit from being able to understand and challenge or contest an outcome, seek redress, and learn through human-computer interfaces. Developers and other AI actors also benefit by being able to identify issues, de-bug the system and learn more about the problem, including understanding causal relationships. Lack of explainability can result in stakeholders not being able to challenge an AI system's output and actors not being able to de-bug a system. Explainability requirements may vary by location or use case, taking into account also applicable legislation. Thus, the same technique or approach might not be applicable in all contexts for a given system. The concept of interpretability is closely related to explainability (Box 3.3).

Box 3.3. Explainability vs interpretability

The definitions of explainability and interpretability in the AI context have been evolving. According to current trends, *explainability* refers to the ability to accurately describe the mechanism, or implementation, that led to an algorithm's output. In turn, *interpretability* refers to whether a human can derive meaning from a system's output for a specific use case.

Source : Schwartz et al. (2021[54]).

There are multiple ways to generate and provide interpretations and explanations of an AI system's output. Explainability and interpretability tools and techniques can be either model-specific or model-agnostic, and either local or global (Hall, 2019[55]; Molnar, Casalicchio and Bischl, 2020[56]):

- **Model-specific vs. model-agnostic tools:** model-specific techniques can be applied to a single class or type of algorithm (e.g. p-values in a linear model), while model-agnostic techniques apply to multiple types of algorithms (e.g. local interpretable model-agnostic explanations "LIME").
- **Local vs. global techniques:** local interpretability techniques detail how a model arrived at a specific prediction, such as showing the subset of pixels that had the biggest impact on the classification of an image (e.g. using techniques such as "Shapley values"); global techniques detail what features are important to the model overall (e.g. using techniques to assess feature or variable importance).

Transparency and traceability

Transparency describes responsible disclosure to ensure people are aware that AI is being used in a prediction, recommendation or decision, or in an interaction (e.g. a chatbot). The growing ubiquity of AI applications could influence the desirability, effectiveness, or feasibility of disclosure in some cases (OECD, 2022[53]). Transparency also means enabling people to understand how an AI system is developed, trained, operated, and deployed in the application domain, so that, for example, users and consumers can make more informed choices. Transparency also refers to the ability to provide meaningful information and clarity about what data and outputs are provided and why, including to regulators and auditors. Thus, transparency need not extend to the disclosure of the source or other proprietary code or datasets, all of which might be too technically complex to be useful for understanding an outcome. Source code and datasets might also be subject to intellectual property regulations, including trade secrets (OECD, 2022[53]).

Traceability in AI describes the need to maintain a complete account of the provenance of data, processes, code, and other elements in the development of an AI system. Traceability often captures granular information about an element or component of an AI system, such as the input data or model, and is essential to enable the auditing of a system.

In sum, transparency can be understood as providing information and disclosure around an AI system, while traceability is the ability to follow elements of the AI system before, during and after deployment (OECD, 2022[53]; IDB-OECD, 2021[21]). Lack of transparency and traceability can hinder trust in AI systems and their use, and dilute accountability for their outputs.

Documenting the risk management process and decisions at each lifecycle phase contributes to transparency, traceability, and accountability (Table 3.1).[9]

Table 3.1. Examples of documentation to assess transparency and traceability at each phase of the AI system lifecycle

AI system lifecycle phase	Sample documentation
Plan and design	Information about the objectives of the AI system, the expected users and potential stakeholders affected by its use and foreseeable misuse
Collect and process data	Data sources, including dataset metadata, data collection processes, and data processing information
Build and use the model	Complete, documented code, including necessary libraries and their appropriate versions
Verify and validate	Information on how the code should be executed to guarantee reproducibility of outputs, including detailed documentation of the parameters and computing requirements
Deploy	Information on how the outputs of the model are used
Operate and monitor	Information about the monitoring strategy, including performance metrics, thresholds, expected model behaviour, and mitigation actions; information about the deficiencies, limitations, and biases of the model, as well as if and how they are communicated to the relevant stakeholders

Source: adapted from IDB-OECD (2021[21]).

3.4 Robustness, security, and safety

Addressing the safety and security challenges of AI systems is critical to fostering trust in AI. In this context, robustness means the ability to endure or overcome adverse conditions, including digital security risks, and maintain its level of performance. AI systems should also not generate unreasonable safety risks, including to physical security, in conditions of either normal use or foreseeable misuse throughout their lifecycle. Laws and regulations in areas such as consumer protection identify what constitutes unreasonable safety risks. Governments, in consultation with stakeholders, must determine to what extent these laws and regulations apply to AI systems (OECD, 2022[57]).

Issues of robustness, security, and safety of AI are interlinked. For example, digital security can affect the safety of connected products such as automobiles and home appliances when risks are not appropriately managed.

Technical concepts related to AI robustness, security, and safety include:

- **Resilience against attack:** the level of protection against software and hardware vulnerabilities (such as data poisoning e.g. tampering with training data to produce undesirable outcomes) and problematic practices (such as data leakage e.g. inclusion of data from the test or validation set in the training dataset; or dual use e.g. misusing the system). "Adversarial robustness" measures how an AI system would perform in a worst-case scenario (Carlini et al., 2019[58]; IDB-OECD, 2021[21]).
- **General safety and fall-back plans:** safeguards that enable a back-up plan in case of problems. The level of safety required depends on the magnitude of the risk posed by an AI system. "Formal verification" (Qin et al., 2019[59]) is relevant as it aims to mathematically check that the behaviour of a system satisfies a given property or specification (e.g. safety).
- **Reliability** (consistent intended behaviour and results), **repeatability** (the same results can be obtained by the same team using the same experimental setup), **replicability** (the same results can be obtained by a different team using the same experimental setup), **reproducibility** (closeness between the results of two actions, such as two outputs of a model, that are given the same input and use the same methodology) and **predictability** (enables reliable assumptions by stakeholders about the output of the system) (Almenzar et al., 2022[60]).

3.5 Interactions and trade-offs between the values-based Principles

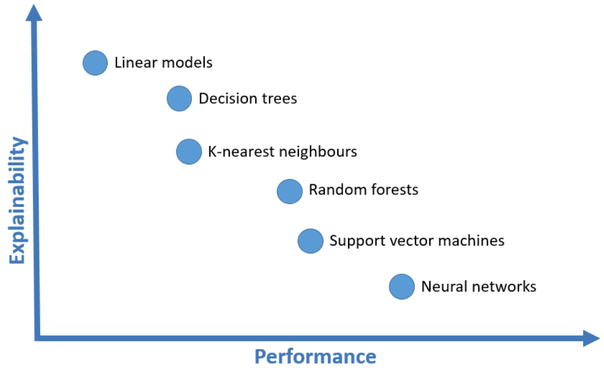
There is growing recognition of trade-offs and interactions between procedural and technical attributes associated with values-based AI Principles. For example: removing bias might cause a loss of accuracy, one component of performance; making a model more explainable could impact system performance and privacy; and improving privacy might limit the capacity to assess adverse impacts of AI systems.

Optimisation of trade-off decisions depends on multiple factors, notably the use-case domain, the regulatory environment, and the values and risk tolerance of the organisation implementing the AI system. In this context, risk tolerance refers to the "organisation's or stakeholder's readiness or appetite to bear the risk in order to achieve its objectives" (NIST, 2022[14]). Trade-offs need to be analysed and balanced within the context at hand.

Some of the most common trade-offs and interactions between procedural and technical attributes associated with values-based Principles for trustworthy AI include:

- **Explainability vs. performance**: The trade-off between explainability and performance of the model[10] has been explored extensively (Goethals, Martens and Evgeniou, 2022[61]; Koshiyama, Firoozye and Treleaven, 2020[62]; ICO-Alan Turing Institute, 2020[63]; Babic et al., 2019[64]; OECD, 2022[1]). "Explainability-by-design" tools and methods are being developed to address this trade-off. Figure 3.2 maps algorithms by their expected explainability and performance levels. Exceptions exist, such as when the explainability of a linear model suffers when the data is pre-processed and includes non-linear features.

Figure 3.2. Illustrative mapping of algorithms by explainability and performance

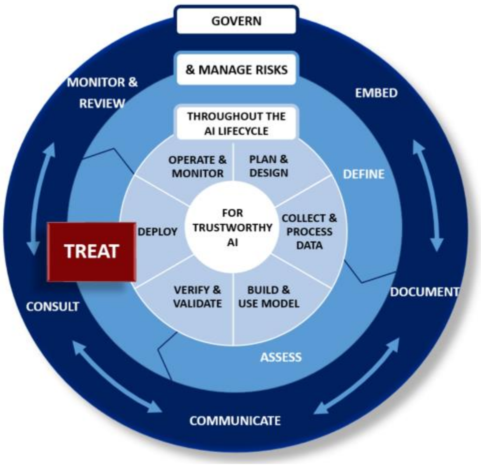


Note: Approximation based on literature. For illustrative purposes only.

- **Fairness vs. performance:** The trade-off between fairness/bias and performance is the subject of significant debate (Feldman et al., 2015[66]; Kleinberg, Mullainathan and Raghavan, 2016[24]; Zafar et al., 2019[67]). Model designers and developers can define acceptable boundaries of bias and performance, for example, by adopting metrics like statistical parity and accuracy. These boundaries can be identified by liaising with business and end-users, and by analysing best practices, standards or regulations commonly adopted in the field of application.
- **Explainability vs. privacy:** AI models are increasingly expected to be both explainable and privacy-preserving. Techniques like feature-importance charts can move in this direction by explaining a model's internal workings while minimising the amount of personal data needed, for example, by identifying unnecessary variables (Goldsteen et al., 2020[68]).
- **Privacy vs. fairness:** A related concern is the trade-off between privacy and fairness. To demonstrate equal performance for all protected groups or attributes, a fair AI system requires a high degree of transparency and explainability, which could come at the expense of privacy. The opposite is also true: the greater the level of privacy, the more difficult it can be to scrutinise an AI system and ensure its fairness. Emerging data governance methods and privacy techniques could help mitigate this trade-off.
- **Transparency vs. security:** There is a well-recognised trade-off between transparency and security: the more transparent a system is, the easier it would be to attack it (Erdélyi and Goldsmith, 2022[69]).
- **Sustainability vs performance:** large AI models (in terms of parameters and computational load) generally perform better than smaller[11] ones but require more energy.

Mapping trade-offs between all the procedural and technical attributes associated with the AI Principles is often difficult and not always desirable. Trade-off analysis aims to optimise the balance for an application, its use-case, and its legal and ethical contexts.

4. TREAT: Prevent, mitigate, or cease AI risks



Once identified in the assessment phase, AI risks must be treated. Risk treatment refers to techniques to prevent, mitigate or cease risks, considering their likelihood and impact. Risk treatment strategies can be grouped into two complementary approaches:

- **Process-related:** how AI actors collaborate, design, and develop AI systems, based on procedural, administrative, and governance mechanisms.
- **Technical**: relate to the technological specifications of a system (e.g. issues related to the AI model, its development and use). Treatment of this type of risk might require re-training and subsequent re-assessment of the AI model.

This section outlines technical and process-related risk-treatment approaches for each AI Principle. As in the assessment section, some of these concepts and measures are relevant to multiple Principles and phases of the AI system lifecycle to different degrees. To enhance clarity and avoid repetition, these multifaceted approaches are included under their most relevant Principle. Process-related and technical approaches should be linked to specific outcomes and measurable metrics when possible.

4.1 Risks to people and the planet

AI systems that improve productivity and respect the natural environment can advance well-being. Table 4.1 illustrates technical and process-related concepts pertaining to human and planetary risks.

Table 4.1. Approaches to treating risks to people and the planet

AI system lifecycle phase	Technical approaches	Process-related approaches
Plan and design	• Devise system architecture according to stakeholders' participation in the AI system design, development, and maintenance (Delgado et al., 2021[70])	• Analyse system's impact on the natural environment and well-being (Xu, Baracaldo and Joshi, 2021[71]) • Understand the economic policies that would mitigate adverse effects of AI on developing economies (Korinek and Stiglitz, 2021[72]) • Follow a framework to evaluate AI's impact on labour demand (Klinova and Korinek, 2021[73]) • Assess the need to engage people who are not on AI development teams (i.e. "inclusive development) (Park, 2022[12])
Collect and process data	• Select features to reduce computational efforts on unrelated features: e.g. mutual information and Monte Carlo-based feature-selection (MIMCFS) (Manikandan and Abirami, 2021[74]) • Treat imperfect and poisoned data from different sources (Wang et al., 2020[75])	• Consider narrowing the data required to train the models, avoiding streams of data unrelated to the task (Spracklen, 2021[76]) and prioritising ethically sourced data.

Build and use the model	• Select appropriate model architecture: e.g. sparse vs dense models (Patterson et al., 2021[77]). • Transfer learning to re-use pre-trained weights from other tasks (Kocmi, 2020[78]) • Code reuse: assess tasks that can be accomplished with existing open-source code (Paleyes, Urma and Lawrence, 2020[79])	• Decide on the model to be used according to its environmental impact (Patterson et al., 2021[77])
Verify and validate	• Cross-validate to assure model robustness and reduce the risk of overfitting on training data: K-fold-CV; leave-one-out (Arlot and Celisse, 2010[17]) • Use combination of covariance-penalty methods and cross-validation to estimate error prediction (Efron and Hastie, 2016[80]) • Avoid concept drift: develop methodologies and techniques for drift detection, understanding and adaptation, e.g. gradual mitigation, abrupt correction, and pre-emptive detection (Escovedo et al., 2018[81])	• Compare overall carbon emissions of the proposed model vis-à-vis “green” algorithms (Patterson et al., 2021[77]) • Bridge the gap between “in the lab” and “in the field” validations; that is, validate taking into account the environment in which the system will be deployed.
Deploy	• Manage configuration of the deployment environment, including the size of the compute resources and enabling auto-scaling (Lindkvist, Stasis and Whyte, 2013[82])	• Calculate operational costs: energy cost of operating AI system hardware, including data-centre overheads (Patterson et al., 2021[77]) • Measure Power Usage Effectiveness (PUE): industry standard metric of data-centre efficiency (Patterson et al., 2021[77]) • Monitor carbon intensity: cleanliness of a data-centre’s energy (Patterson et al., 2021[77])
Operate and monitor	• Monitor logs and metrics using dashboards to indicate system failure and compute usage, e.g. Kibana, Grafana and Zeppelin (Nurgaliev, Karavakis and Aimar, 2016[83])	• Understand the expected generalisation performance of the model on future data, considering the economic and social context (Arlot and Celisse, 2010[17])

4.2 Risks to human-centred values and fairness

Bias and discrimination

Bias should be addressed early in the AI system design and development process, by implementing review points at each lifecycle phase. The outputs of the model should be verified and validated at each review point (IDB-OECD, 2021[21]). Table 4.2 illustrates technical and process-related approaches to treat risks of bias in AI.

Table 4.2. Approaches to treating bias and discrimination

AI system lifecycle phase	Technical approaches	Process-related approaches
Plan and design	• Establish a plan to mitigate proxy discrimination of all stakeholders involved: e.g. prohibit the use of proxies that can lead to discrimination, mandate the collection and disclosure of data about impacted, legally protected classes, without violating privacy rights (Prince and Schwarcz, 2020[84])	• Define protected and non-protected subgroups, and consider possible impacts on them; analyse tool capabilities of mitigating intrinsic data bias (Schwartz et al., 2021[54]) • Employ statistical models that isolate only the predictive power of non-suspect variables (Prince and Schwarcz, 2020[84])

Collect and process data	• Reweigh subjects: remove discrimination without relabeling instances (Kamiran and Calders, 2012[85]) • Oversample minority group and undersample major classes (Iosifidis and Ntoutsi, 2018[86]; Tripathi et al., 2021[87]) • Learn fair representations: encoding the data as well as possible, concealing any information about membership in the protected groups (Zemel et al., 2013[88])	• Understand the sources of bias throughout the AI system lifecycle, such as group attribution, historical, omitted-variable and, selection bias (Suresh and Guttag, 2021[89]) • Data enrichment: incorporate impact on worker well-being into decision-making processes about data enrichment (PAI, 2021[49])
Build and use the model	• Adversarial debiasing: e.g. include a variable for the subject of interest and simultaneously learn a predictor and an adversary (Zhang, Lemoine and Mitchell, 2018[90]) • Fairness constraint: e.g. create a measure of decision-boundary unfairness as a proxy for bias (Zafar et al., 2019[67]; Donini et al., 2018[91]) • Counterfactual fairness: e.g. define that a decision is fair towards an individual if it is the same both in the actual world and in a counterfactual world where the individual belonged to a different demographic group (Kusner et al., 2017[92]) • Remove of disparate impact when a selection process has widely different outcomes for different groups, even as it appears to be neutral (Feldman et al., 2015[66])	• Prioritise context over optimisation: selecting models based solely on accuracy is not the best approach for bias reduction as context should be considered (Schwartz et al., 2021[54])
Verify and validate	• Calibrate equality of odds: minimise error disparity across different population groups while maintaining calibrated probability estimates (Pleiss et al., 2017[93]) • Classify reject-options: instances belonging to deprived and favored groups are labeled with desirable and undesirable labels, respectively (Kamiran and Calders, 2012[85])	• Prevalence at threshold: disentangle normative questions of product and policy design from empirical questions of system implementation (Bakalar, Barreto and Bergman, 2021[94]) • Create a reference dataset serving as “ground truth” to AI developers for testing (Schwartz et al., 2021[54])
Deploy	• Process fairness: reduce the dependency of models on sensitive features, e.g. LimeOut and FixOut (Alves et al., 2021[95])	• Analyse tool performance and, if needed, inform the need for retraining as a redress mechanism to reduce algorithmic discrimination (Schwartz et al., 2021[54])
Operate and monitor	• Correlate mean contribution of potentially biased inputs to the overall model predictions: LIME and SHAP[12] (Alves et al., 2021[95])	• Compare the intended vs. actual context based on the impact on the actors that are affected by the technology to expose early design and development decisions that were poorly or incompletely specified, or based on narrow perspectives (Schwartz et al., 2021[54])

Privacy and data-governance risks

There are several technical and process-related approaches to treating privacy and data-governance risks (Table 4.3). In particular, privacy-enhancing techniques to mitigate risks to personal or sensitive data can be applied at different phases of the AI system lifecycle, for example, during data collection and processing (e.g. emerging data governance models, feature selection and dataset pseudo-anonymisation); during model building and use (e.g. federated learning and differential privacy); and during deployment, monitoring and operation (e.g. rate-limiting and users’ queries management). As stated in section 3.5, trade-offs exist and privacy-preserving approaches might come at the expense of explainability, transparency, and fairness.

Table 4.3. Approaches to treating risks to privacy and data governance

AI system lifecycle phase	Technical approaches	Process-related approaches
Plan and design	• Implement a privacy-preserving machine learning (PPML) framework (Xu, Baracaldo and Joshi, 2021[71])	• Survey the legal and regulatory environments that restrict access to and use of privacy-sensitive data from the stakeholders involved (Xu, Baracaldo and Joshi, 2021[71])
Collect and process data	• Data minimisation by dimensionality: transform data as a tool to increase privacy, e.g. encoders, PCA (Tipping and Bishop, 1999), T-SNE (Van der Maaten and Hinton, 2008[96]) • Make datasets (pseudo)-anonymous (Neubauer and Heurix, 2011[97]), e.g. k-NDDP (Shakeel et al., 2021[98]) • Anonymise/de-identify the data while preserving relevant attributes (Fernandez Llorca and Gomez, 2021[99]) • Add multiplicative and coloured noise – an alternative to the classical data perturbation techniques (Kargupta et al., 2005[100]) • Combine datasets: abolish the distinction between personal/non-personal data and develop a risk-based regulatory approach to data processing (Erdélyi and Goldsmith, 2022[69])	• Identify sensitive and personal data, either in the dataset used for training or accessible by end-users • Ensure the PPML is as robust as possible from the data-owners' standpoint (Xu, Baracaldo and Joshi, 2021[71]) • Create/leverage emerging data governance models: data sharing pools, data cooperatives, and data trusts (Micheli et al., 2020[43])
Build and use the model	• Federated learning: distribute training data across devices and learn on a shared model of locally computed aggregates (McMahan and Ramage, 2017[101]; Kim et al., 2019[102]) • Differential privacy: train deep neural networks with non-convex objectives and under a modest privacy budget (Abadi et al., 2016[103]; Dwork and Naor, 2010[36]) • Defend against data poisoning: techniques to protect the models from fake data injection (Steinhardt, Koh and Liang, 2017[104]) • Private aggregation of teacher ensembles (PATE): transfer the knowledge of an ensemble of "teacher" models to a "student" model (Papernot et al., 2017[105]) • MiniONN: using a privacy-preserving framework, transform a neural network into an oblivious neural network (Liu et al., 2017[106])	• Ensure the PPML has adequate privacy protection in accordance with the trust assumption and threat model settings, incorporating representative architectures such as federated learning (Xu, Baracaldo and Joshi, 2021[71])
Verify and validate	• Analyse the privacy of Python machine-learning frameworks: Privacy Lint (Meta) • Model inversion mitigation to prevent malicious users from attempting to recover the private dataset used to train a model (Fredrikson, Jha and Ristenpart, 2015[37])	• Ensure the PPML is as accurate as the standard model without using privacy-preserving settings (Xu, Baracaldo and Joshi, 2021[71])
Deploy	• Rate-limiting: use strategies for limiting network traffic (Google, 2019[107])	• Ensure the PPML is communicating and computing as effectively as the standard machine learning system (Xu, Baracaldo and Joshi, 2021[71])
Operate and monitor	• Automate compliance-verification and auditability (Chhetri et al., 2022[108])	• Monitor the storage and privacy of sensitive information (U.S. Department of Health and Human Services, 2020[109])

Risks to human rights and democratic values

Procedural approaches exist to treat risks to human rights and democratic values, identified using HRIAs and other tools. Notably, these include contingency plans; support policies for data enrichment teams; engagement and consultation with relevant stakeholders; and remedial actions for those whose rights are violated (e.g. cessation of activity, development of new processes or policies, monetary compensation, etc.) (Table 4.5).

Table 4.4. Approaches to treating risks to human rights and democratic values

AI system lifecycle phase	Process-related approaches
Plan and design	• Develop contingency plans to be included in ethics-by-design approaches • Update company policies against developing potentially harmful AI systems, e.g. deepfake technology (OECD, 2021[110])
Collect and process data	• Use data enrichment: living wage calculators, quality assurance for crowdsourcing, improvements in task design, support policies for workers exposed to harmful content (PAI, 2021[49])
Build and use the model	• Engagement and consultation with external experts, stakeholders, civil rights groups, and oversight bodies (OECD, 2022[111]; FRA, 2020[112])
Verify and validate	
Deploy	• Take remedial actions, including arbitration, cessation of activity, apology, development of new processes or policies, monetary compensation, judiciary action, etc. (OECD, 2019[46]) • Refer to consumer protection and responsible business conduct frameworks (OECD, 2021[110]) • Leverage transparent grievance mechanisms, public reporting, and public oversight (OECD, 2021[110]) • Restrict sale and product support to certain groups, e.g. governments (OECD, 2021[110])
Operate and monitor	

4.3 Risks to transparency and explainability

Technical approaches to treat risks to transparency and explainability in AI include model-specific and model-agnostic approaches. Model-specific techniques can be applied to a single class or type of algorithm (e.g. p-values in a linear model), while model-agnostic techniques apply to different types of algorithms (e.g. local, interpretable, model-agnostic explanations "LIME"). Process-related approaches include documentation tools and deciding on the type of model to use, given possible trade-offs such as between accuracy and explainability (Table 4.5).

Table 4.5. Approaches to treating risks to transparency and explainability

AI system lifecycle phase	Technical approaches	Process-related approaches

Plan and design	• Design an end-to-end explainable AI (XAI) framework, from DataOps to inference: XAI provides information to help users to de-bug models, improve decision-making, and improve trust in automation (Palacio et al., 2021[113])	• Establish processes to document the entire AI system lifecycle to enhance transparency (Raji et al., 2020[114]) • Consider using existing documentation tools, which may be relevant to one or several lifecycle phases: e.g. Google Model Cards, Microsoft Datasheets for Datasets, Meta System Cards, etc. • Consider documenting use cases (including foreseeable misuses of the system) to mitigate "by-design" use-related risks (Hupont and Gomez, 2022[115]; Hupont et al., 2022[116])
Collect and process data	• Perform exploratory data analysis using visualisation tools to understand datasets, e.g. Google Facets • Standardise dataset and model descriptions: frameworks to drive higher data quality standards, e.g. dataset nutrition label framework (Hupont et al., 2022[117]; Holland et al., 2020[118]) • Summarise datasets: explain data through clusters, e.g. K-medoid clustering (Kaufmann and Rousseeuw, 1987[119]) • Engineer explainable features: unsupervised automated discovery of interpretable representations of data, e.g. ß-VAE (Higgins et al., 2017[120])	• Document model inputs in design documentation (U.S. Government Accountability Office, 2021[121])
Build and use the model	• Draw explanations from rule-based approaches: decision trees, rule-induction methods, etc. • Model coefficients from linear models: linear regression, linear discriminant analysis, etc. • Use nearest-prototype: K-nearest-neighbour, Naïve-Bayes. • Use interpretable tree-based models: e.g. explainable boosting machines (Nori et al., 2019[122]) • Use explainable reinforcement learning: PIRL (Puiutta and Veith, 2020[123])	• Decide on self-explainable and interpretable (white box) or complex (black box) solutions (Molnar, Casalicchio and Bischl, 2020[56])
Verify and validate	• Local surrogate explanations: explain individual predictions of black box machine learning models, e.g. LIME (Ribeiro, Singh and Guestrin, 2016[124]) • Apply perturbation techniques: e.g. gradient-based attribution methods (Ancona et al., 2017[125]); permutation importance (Breiman, 2001[126]); SHAP (Lundberg and Lee, 2017[127])	• Assess trade-offs based on model choice, e.g. between accuracy and explainability, according to the application domain and end-users (Veer et al., 2021[128])
Deploy	• Perform simulation (what-if?) analysis: recourse through minimal interventions, moving the focus from explanations to recommendations, e.g. counterfactual explanations and algorithmic recourse (Wachter, Mittelstadt and Russell, 2017[129]; Karimi, Schölkopf and Valera, 2020[130])	• Ensure that model explanations include, at a minimum, the type and source of model input data, the high-level data transformation process, the decision-making criteria and rationale, risks and mitigation measures, and a disclosure about using AI.
Operate and monitor	• SHAP plots: understand feature importance and feature effects, e.g. using summary and force plots • Provide technical documentation and user manuals for operators and users of the system	• Ensure that insights and disclosures are directed to the end-users affected by the model, and not only to machine-learning engineers who use explainability for de-bugging purposes (Bhatt et al., 2019[131])

4.4 Risks to robustness, security, and safety

AI systems should be robust, secure, and safe throughout their lifespan so that they function appropriately in conditions of normal use, foreseeable misuse, or other adverse conditions, and do not pose unreasonable safety risk (OECD, 2019[5]). Approaches to treating risks related to robustness, security, and safety include conformity assessments relative to consumer-safety regulations and secure-by-design approaches that embed security in the system from the planning and design phase (Table 4.6).

Table 4.6. Approaches to treating risks to robustness, security, and safety

AI system lifecycle phase	Technical approaches	Process-related approaches
Plan and design	• DevOps: combine software development (Dev) and IT operations (Ops) (Ghantous and Gill., 2017[132]) • CI/CD: practice continuous development, integration, delivery, and deployment (Shahin, Babar and Zhu, 2017[133]) • Secure-by-design: use good design principles, tools, and mindsets that make security an implicit result (Deogun, Johnsson and Sawano, 2019[134])	• Assess conformity with consumer safety regulations (The European Consumer Organisation, 2021[135])
Collect and process data	• Label-smoothing: use soft targets to reduce overfitting (Müller, Kornblith and Hinton, 2019[136]) • Thermometer-encoding: modify standard neural-network architectures to significantly increase robustness to adversarial examples (Buckman et al., 2018[137]) • Propagate bounds to achieve verified robustness to symbol substitutions (Huang et al., 2019[138])	• Incorporate a data protection and secure integration plan into technical design documentation (U.S. Department of Health & Human Services, 2021[139])
Build and use the model	• Satisfiability modulo theories: determine whether a first-order formula is satisfiable with respect to some logical theory (Bunel et al., 2018[140]) • Evasion attacks: generate adversarial examples and quantify the robustness of the models, e.g., fast gradient sign method (Huang et al., 2017[141]) and DeepFool (Moosavi-Dezfooli, Fawzi and Frossard, 2016[142]) • Mixed integer programming: verify piecewise-linear neural networks as a mixed integer program for model robustness evaluation (Tjeng and Tedrake, 2017[143]) • Variance minimisation: remove adversarial perturbations via a compressed sensing approach (Rudin, Osher and Fatemi, 1992[144]; Guo et al., 2017[145])	• Propose a provable guarantee of robustness as an alternative to heuristic defences: e.g., GloRo Nets (Leino, Wang and Fredrikson, 2021[146]) • Review vendor documentation and rigorously scan for vulnerabilities (U.S. Department of Health & Human Services, 2021[139])
Verify and validate	• Lagrangian relaxation: obtain provable guarantees that neural networks satisfy specifications relating their inputs and outputs • Dataset shift: analyse model robustness keeping the original data (Subbaswamy, 2020[147]) • Area under the receiver operating characteristic (ROC) curve: chart the performance of a binary classifier system as its discrimination threshold is changed (Hanley and Mcneil, 1982[148]) • Reliability metrics: consider the cost of errors or inaccurate predictions (Lhoest et al., 2021[149]) • Dataset shift monitoring: detect unexpected inputs and firing off warnings (Rabanser, Gunnemann and Lipton, 2018[150])	• Decide on the preferred visualisation of the system's outputs, as well as the corresponding validation metrics (Goodfellow, Bengio and Courville, 2016[151])

Deploy	• Monitor situations of possible AI service misuse by costumers (Javadi et al., 2020[152])	• Obtain an Authority to Operate (ATO), a formal declaration by a Designated Approving Authority (DAA) that authorises operation of a Business Product and explicitly accepts the risk to agency operations (U.S. Department of Health & Human Services, 2012[153])
Operate and monitor	• Code versioning: e.g., Git (Github); Mercurial (BitBucket). • Reproducibility: tools that allow reproducibility of models, e.g., Binder; Docker; Kubernetes. • Automated testing: e.g., Travis CI; Scrutinizer CI. • Trap-based monitoring sensors: an efficient way to infer Internet threat activities (Fachkha, 2016[154]). • EFK stack using Kubernetes: monitoring of system logs, performance and storage, e.g., Elasticsearch, Fluentd, Kibana.	• Use of dashboards to monitor performance, errors and suggest courses of action.

Note: Safety of deployment is a different and more demanding task than fail safety.

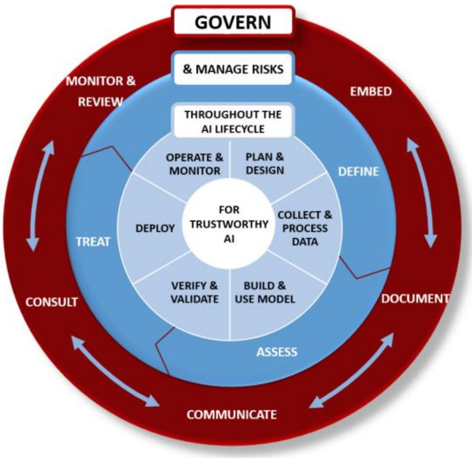
4.5 Anticipating unknown risks and contingency plans

Both known and unknown AI risks should be anticipated to prevent harm. Unknown risks might include risks to robustness (e.g. breakdown); security (e.g. hacks); secondary uses or misuses of a system, including use of pre-packaged coding for malicious purposes; psychological and social impact; and reputational risks.

Risk and impact assessments can be conducted to identify risks and design mitigation strategies before, during, and after deployment. One approach to identifying unknown risks is known as “red teaming”, which refers to systematic and controlled attempts to probe and expose flaws and weaknesses in a system, process, or organisation to identify and mitigate unknown risks (Brundage et al., 2020[155]). Another approach to identify unknown risks is to engage “challengers” – stakeholders likely to oppose the development, operation, or use of the AI system – to provide insights at early phases of the lifecycle regarding potential risks, harmful impacts, or negative effects. Additionally, incident databases (such as the OECD AI Incidents Monitor and the Responsible AI Collaborative’s AI Incident Database) are being developed to identify previously unknown risks posed by one system as realised by similar systems.

Contingency plans should be in place explaining the steps to reduce negative impacts after identified risks occur. Their aim is to lessen the damage of the risk when it materialises. Contingency plans are usually the last line of defence against a risk.

5. GOVERN: Monitor, document, communicate, consult and embed



Governing the risk management process is key to achieving trustworthy AI. Governance is a cross-cutting activity which consists of two main elements. The first element concerns the governance of the risk management process itself and includes monitoring and reviewing, documenting, communicating and consulting on the process and its outcomes. The second governance element ensures the effectiveness of the risk management process by embedding it into the culture and broader governance processes of organisations.

5.1 Monitor, document, communicate and consult

Monitoring and reviewing the risk management process, documenting the steps, options and decisions, as well as communicating and consulting on its results should be a core component of an organisation's governance systems.

Monitor and review

Monitoring and reviewing risks and steps taken to treat them contributes to the correct functioning of an AI system. Given the evolving nature of AI systems and the environments in which they operate (Babic et al., 2019[64]), monitoring should be continuous, rather than a one-off activity, and happen at all stages of the risk management process.

Components

Monitoring and reviewing a model in operation (i.e. once it is live or "in production") is necessary to check that its accuracy and overall behaviour do not deteriorate when exposed to real-world data (Sculley et al., 2015[156]). Several analyses can detect performance drops or malfunctions. Widely used statistical and technical mechanisms in this area include:

- **Data drift:** The environment in which an AI system is deployed will likely evolve over the system's lifespan (Babic et al., 2019[64]) and cause the data distribution to drift. Data drift might decrease the quality (accuracy, fairness, etc.) of a model's predictions. The two main categories of data drift are *covariate shift* and *label shift*. Covariate shift happens when the distribution of the input data changes between the training environment and the live environment. For example, when an AI-based medical exam is run on a group of patients who exhibit symptoms the AI system has not previously encountered. In contrast, label shift in supervised learning models means that the distribution of labels in the training set is different from the distribution of labels in deployment. For example, when a significantly higher proportion of patients with a certain condition is encountered by a deployed AI system compared to the proportion in the training data. Data drift can be detected by statistical methods and addressed through domain adaptation.[13]
- **Prediction drift:** In addition to data drift, changes in the environment in which the AI system operates can alter the conditional distribution between inputs and outputs that the model has learned. Prediction drift, or *concept shift*, means that the model's representations of the world are outdated. This can be addressed by re-training the model or ensuring it learns continuously over

its lifespan (Lu et al., 2019[157]). Continuous re-training and learning can, however, make auditing and certification more challenging. Therefore some regulators, such as the US Food and Drug Administration, have only approved "locked" AI systems – AI models that do not change with use and provide the same result given the same input (Babic et al., 2019[64]). Software applications for managing regulatory compliance in an automated manner (i.e. "RegTech") are also being explored. Finding the right balance between continuous learning and ensuring AI systems behave as designed and/or approved (e.g. by a company or a regulator) is key.

- **Input data quality:** Issues related to the input data of an AI model include changes in data schemes, increased frequency of outliers, missing data or corrupted data. Data quality deterioration can be assessed by checking for data completeness, establishing appropriate data governance mechanisms and verifying statistical properties such as variance, quantiles or the presence of extreme values in the data. Additionally, data generated from the operations and use of AI systems may become input to the system, introducing possible feedback loops that should to be addressed.
- **Computational performance monitoring:** Processing speed and use of computing resources can evolve during an AI system's lifespan. Usage and performance of computing resources should be monitored to guarantee reasonable processing times and keep an eye on costs and energy consumption. The increasingly central role of environmental considerations in companies, governments, and societies makes an AI system's energy consumption and carbon footprint evaluations essential (OECD, 2022[111]).

In addition to technical tools and quantitative tests, non-statistical processes and tools can also monitor and review the behaviour of AI models. For example:

- **Incident reporting:** While AI-specific regulations are under development, it is critical that reporting systems and frameworks are in place and interoperable across jurisdictions. Users of AI systems should be able to report controversies, incidents, or issues, either regarding inaccurate predictions, unfair outcomes, or unexpected or undesirable behaviour. Therefore, incidents-based monitoring and oversight by users and impacted stakeholders should be incentivised in addition to tools-based monitoring. Setting up reporting channels for AI incidents can facilitate monitoring, and provide valuable information and data on how to improve AI systems and design regulatory frameworks. The OECD, along with several partners, is attempting to do so.
- **Human-in-the-loop:** Human-in-the-loop mechanisms monitor AI models at different stages of their development and use, including by testing and validating outputs, responding to system alerts during deployment and, if appropriate, retiring a model from production. Human-in-the-loop approaches are important for monitoring the ethical considerations of an AI system, which should not be delegated to another automated system.
- **Regulatory sandboxes:** Sandboxes are frameworks set up by a supervisory authority to allow companies, researchers and other actors to conduct live experiments in a controlled environment and under the regulator's supervision. Sandboxes can provide means to enhance accountability by monitoring and reviewing AI systems in real-life scenarios before they are deployed in production. AI sandboxes are being implemented in several countries, including Norway's data protection sandbox for AI and Spain's pilot for an AI sandbox to facilitate the implementation of the proposed EU AI Act.
- **Continuous re-skilling and up-skilling:** The monitoring and review of AI systems requires that users, operators and stakeholders be knowledgeable and aware of the systems' objectives, potential benefits and risks, and their legitimate usage.

Incident reporting, human-in-the-loop approaches, sandboxes, and continuous re-skilling and up-skilling can help monitor risks, negative externalities, and risks that materialise despite a system working as intended (e.g. AI systems measuring employee performance, which may work properly but result in unintended risks to workers, including their physical or mental health). They can also identify secondary uses or misuses of an AI system or its parts for malicious purposes.

Frequency

Beyond establishing monitoring and review processes and tools, it is important to set up communication channels and regular reviews to ensure that information about undesirable model behaviour or incidents is shared with stakeholders.

Monitoring and review frequency should be appropriate to the application and context of each AI system. Approaches include: continuous monitoring (e.g. where mechanisms and application programming interfaces (APIs) are put in place to continuously assess changes in the data, environment or model behaviour); pre-defined review processes by developers, operators, auditors, or users (e.g. regular checks to discuss model health and potential upgrades, replacements, or withdrawals of the AI system); and stakeholder reporting as needed (e.g. incident reporting). Additionally, regulation might require systematic reporting of AI incidents (e.g. “transparency reports”) or the auditing of AI models and processes in organisations.

Metrics

Monitoring and review processes and their outputs should be timely and accurately documented. Indicators should cover all relevant known – and unknown, if possible – technical and non-technical risks that fall under each of the AI Principles. Known risks include:

- **Accuracy metrics:** model performance, model quality and prediction drift. These can be used to assess whether a model remains valid or needs to be retrained, replaced, or retired from operation.
- **Data metrics to assess data quality and potential data drift:** these evaluate discrepancies between the original and current distributions of the data and assess data integrity.
- **Fairness, transparency, explainability, and privacy metrics:** measures of bias according to a given fairness paradigm (e.g. equality of opportunity, statistical parity, etc.); model documentation; interpretability measures; and data protection and digital security indicators (Section 3.1).
- **Non-technical metrics:** measures of non-statistical processes and tools, including user skill levels, stakeholder awareness, and incident reporting.

It is also important to monitor unknown risks, including AI system breakdown, hacks, psychological and social impact, and reputational risks. Findings about unknown risks should iteratively inform the other AI system lifecycle phases, including planning and design.

Interfaces

Accountability and risk-management in AI pre-suppose ongoing monitoring and improvement. Systematic and iterative improvement processes could be considered, such as the “Plan-Do-Study-Act” cycle (Deming, 1968[158]). Intuitive and user-friendly interfaces, including “traffic light” indicators and dashboards, can help (Figure 3.1) (Brundage et al., 2020[155]).

Ways to present metrics from an AI system’s monitoring and review process include:

- **Logs:** basic logging of metrics helps create conditional workflows, for example, by setting thresholds for accuracy or fairness outside of which alerts are triggered and automated or manual actions taken, such as retraining a model.
- **Visual reports:** logs can be represented visually through tables, charts or diagrams. Visual reports provide a user-friendly approach to inspect the health of an AI model and detect malfunctions. Visual reports can go beyond plotting. For example, they can provide explanations.

- **Live dashboards:** visual reports can be automated into live or real-time dashboards to enable users to interactively explore model capabilities and shortcomings. Live monitoring dashboards typically allow manual editing of specific variables or data points to assess the influence of changes on a model's outputs.

Colour coding can be used in monitoring and review to indicate that a system is "high-performing/compliant" (e.g. green), "low-performing/compliant" (amber), or "non-compliant" (red).

Access levels for auditing and review

Reviewing and auditing AI systems after development can verify that they function properly and the necessary risk assessment and treatment mechanisms are in place.

Seven access levels enable auditing and review at varying degrees of scrutiny (Koshiyama et al., 2021[26]). They range from "process access", which only allows indirect observation of a system, to "development access", in which all the system's details are disclosed with full transparency (Table 5.1). The intermediate levels describe configurations that limit access to certain components of the system (e.g. knowledge of the objectives, model architecture, input data, etc.).

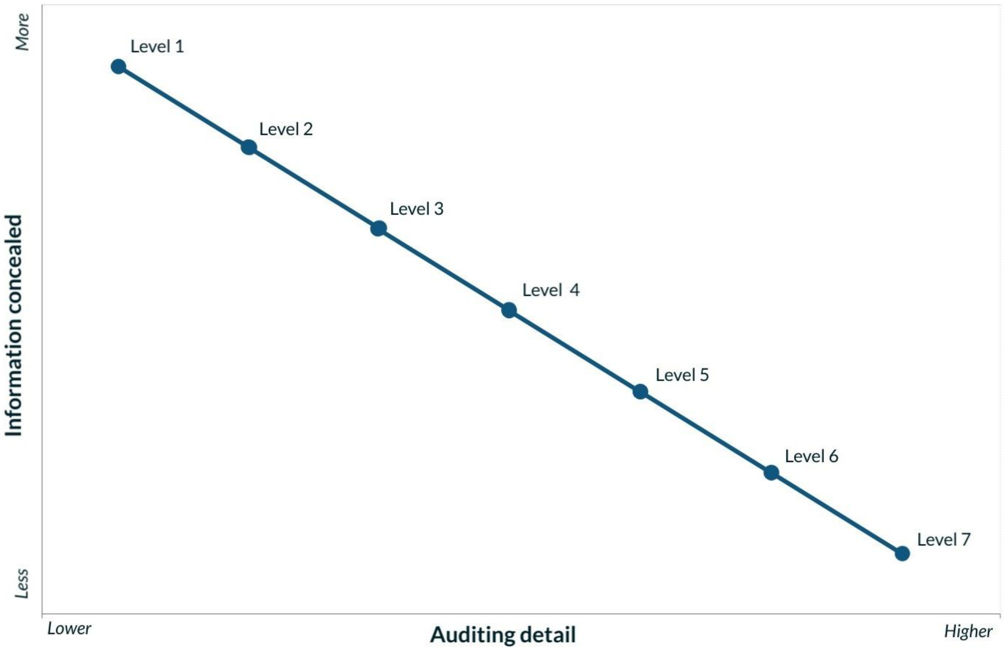
Table 5.1. Characteristics of AI auditing and review access levels

	Level 1 Process access	**Level 2** Model access ("black box")	**Level 3** Input access	**Level 4** Outcome access ("blurry box")	**Level 5** Parameter control	**Level 6** Learning objective	**Level 7** Development Access ("full transparency")
Concealed information	Very High	High	High	High/Medium	Medium	Medium	Low
Feedback detail	Low	Medium	Medium	High/Medium	High	High	Very High
Typical application	Sales forecasting	Digital security	Recruitment	Credit-scoring	Facial recognition	Algorithmic trading	Self-driving vehicle
Potential oversight	Guidelines	External auditing/ certification	External auditing	External auditing	External auditing	Internal/ external auditing	Internal auditing

Source: adapted from Koshiyama et al., 2021.

Different access levels could allow for auditing and review tailored to a specific AI application and its context, including commercial sensitivities and legal and ethical requirements. Oversight mechanisms for the access levels include guidelines, certifications, and internal or external assessments and audits. The accuracy and completeness of auditing and review processes depends on the access level: higher access levels to information enable greater auditing detail and accuracy (Figure 5.1).

Figure 5.1. Trade-off between information concealed and auditing detail by access level



Source: adapted from Koshiyama et al., 2021.

Level 1: Process access

In process access, the reviewer has no direct access to the model or its algorithms. Scrutiny is limited to the model development process. The review process relies on checklists and documentation that can include both qualitative and quantitative information. General and sector-specific guidelines issued by regulators and other government bodies could inform the assessment.

This level of disclosure and feedback detail might be appropriate for AI applications considered low-risk or low-stakes. Nevertheless, comprehensive technical documentation can contain substantial information about the system, including details about how risks are being assessed, treated and monitored and how different trustworthiness requirements are being fulfilled. Such documentation could be a valuable source of information for external stakeholders and auditors.

Level 2: Model access

In model access ("black box"), the reviewer has access to some input data metadata (e.g. the name, types, and ranges of the variables) and the ability to run the model. However, other information, such as the distributions of input variables, is not made available. Therefore, the reviewer depends on artificial inputs to run the model.

This level of access entails the least amount of information disclosed to the reviewer, since no data-sharing agreements are needed. A high level of automation can be achieved, since only API access is needed to perform the analyses. Analyses that can be performed with this access level include adversarial attacks, statistical disclosure, adversarial evaluation of bias and discrimination, feature relevance extraction, and partial dependency explanations.

Level 3: Input data access

In input data access, the reviewer can run the model with the inputs used to train and validate it. However, outputs obtained from the review process cannot be compared with actual system outputs.

Assessing model performance is challenging in the absence of information about the outputs of the system. Some analyses that can be performed with this access level include bias estimations (e.g. from an equality-of-outcome perspective); training data membership inference; inference of model properties; and the creation of surrogate models (i.e. interpretable models that mimic the behaviour of the original model). Synthetic data, mirroring the distribution of the input data, could be used to investigate a model's robustness to gradual changes in the distribution of the training data.

Level 4: Outcome access

In outcome access ("blurry box"), the reviewer can run the model using actual input data and compare actual outputs. Therefore, beside the ability to run the model, the reviewer has access to the output and input data used to train and validate it.

This access level can be seen as a "blurry box", as the reviewer has no access to model parameters and architecture. Techniques available to assess a model under these conditions include model-agnostic procedures (e.g. cross-validation, Shapley values, and feature importance); concept drift analysis; estimation of the accuracy of explanations; and bias estimations (e.g. from an equality-of-opportunity perspective). The reviewer can build baseline or competitor models to assess performance.

Depending on the specifics, this access level yields a medium to high level of detail in the final feedback resulting from the review or audit. Until this access level, apart from data sharing agreements, there is minimal need to share intellectual property or model development information.

Level 5: Parameter manipulation

In parameter manipulation, in addition to access to the output and input data used to train and validate the model, and the ability to run it, the reviewer has access to the model parameters and can thus re-calibrate and re-parametrise the model. However, no information about model type, architecture, or objective function is shared.

This access level allows the reviewer to assess how stable system performance is and evaluate the quality of explanations being provided. From a privacy perspective, this access level allows the reviewer to assess the risk of functionality stealing.

This access level is relatively straightforward to implement via an API and can be automated for external review or auditing. The level of information shared about the model is relatively low, implying regard for intellectual property and other commercial considerations.

Based on certain assumptions, the reviewer could retrain the model through re-parametrisation.

Level 6: Learning objective

In the learning objective level, the reviewer can run a model and directly access most of the related information, including the model's learning procedure, tasks, objectives, parameters, output, and input data used to train and validate the model.

The reviewer is allowed to re-train the model using the actual objective function that the model was initially trained on. However, the reviewer has no access to the model's type (e.g. kernel method) or components (e.g. number of neurons).

This access level allows the reviewer to investigate an almost complete picture of the system, without infringing on its intellectual property. Feedback from the review or auditing process has the potential to be very detailed, including information about the model's complexity and robustness to stress-testing. This access level is enough to perform automated internal and external review and auditing, since human involvement after setting up the relevant APIs is considerably low.

Level 7: Development access

In development access ("full transparency"), the reviewer can run the model and access all the related information, including the model's type and architecture, the learning procedure, tasks performed, action autonomy, parameters, and output and input data used to train and validate the model.[14]

This level of access, which is equivalent to the system developer's access, allows the reviewer to provide richer and more accurate feedback on the model, identifying risks and assessing mitigation strategies in a more thorough manner.

This level of access is often granted to internal reviewers or in-house consultants and might require the direct involvement and collaboration of internal developers as well as contractual agreements concerning non-disclosure, intellectual property sharing, and data-sharing issues, among others.

Document

A trail or log documenting the steps, decisions and actions, and their rationale during the risk management process provides the basis for communication and consultation on the processes and its results, and helps inform functions like auditing, certification, and insurance. Whether the AI system is built in-house or by a third party, documentation and logs should follow the system throughout the supply chain; that is, each involved party or actor – from the developer to the vendor to the deployer – might need to conduct their own assessments and document the actions taken to manage risks.

Documenting risk management around an AI system and its outputs can improve the process itself and enhance communication and interactions with stakeholders inside and outside the organisation. Where appropriate, risk management documentation should be made publicly available. Documentation bolsters accountability by enhancing transparency and enabling human review processes (NIST, 2022[14]).

Communicate

The broader outcome of a risk assessment and management process is to protect human rights and democratic values, improve confidence in the AI system, and ensure it is trustworthy. Verifying and communicating publicly whether an AI system conforms to regulatory, governance, and ethical standards after assessing and treating risks is crucial. It facilitates understanding of risks and of the rationale behind decisions or actions. Communication can include:

- **How the AI system is assured in accordance with general and sector-specific regulations and standards to enable interoperability:** this includes compliance with broad national or regional regulation and standards, provided by agents such as the US National Institute of Standards and Technology (NIST), the UK Information Commissioner's Office (ICO), and European Union regulations; and with sector-specific standards, such as in financial services and healthcare, which can help assess and manage AI risks in different contexts and applications. Specific standards for AI are also being developed by the International Organization for Standardization (ISO), European Committee for Electrotechnical Standardization (CEN-CENELEC), and the Institute of Electrical and Electronics Engineers (IEEE), among others, to help develop and implement systems that are trustworthy and comply with legal and ethical frameworks.

Regulation and standards can be *general* (e.g. set by organisations or government agencies with remits not sector-specific, and encompassing broad areas like privacy, explainability, fairness, robustness, safety, and security) or *sector-specific* (e.g. specific to the use of AI systems in an industrial sector or policy area, such as financial services or defence). Application-specific standards – e.g. facial recognition – are also being developed (NIST, 2022[14]).

Communicating about the system's compliance with national and international standards and regulation is key as cross-border product ecosystems emerge. Governments are exploring ways to facilitate interoperability assessments with existing legal frameworks and standards, including using regulatory experimentation frameworks such as sandboxes.

Conversely, it is crucial to explain when the system (or some of its parts) are not compliant and disclose the harm prevention and remediation measures being taken to achieve compliance.

- **What governance mechanisms are in place:** this includes *technical* mechanisms for tracing and tracking decisions and processes to implement the AI Principles (e.g. robustness, explainability, etc.), and *non-technical* mechanisms to oversee the human and procedural considerations of an AI model – including social, legal, and environmental impact and conformity assessments; human rights impact assessments (HRIAs); user and workforce training and education – and to define roles and responsibilities. There is growing research on algorithmic and data protection impact assessments, including issues related to human rights, and social and environmental concerns (EU-HLEG, 2019[28]; OECD, 2022[1]; Reisman et al., 2019[159]; Koshiyama and Engin, 2019[160]; Kaminski and Malgieri, 2020[161]).
- **How risks are monitored and reviewed, and what mechanisms exist for redress:** the existence, frequency, functionality, and effectiveness of monitoring and review interfaces to track and trace risks (e.g. the number of attacks blocked, risks prevented, etc.). Redress mechanisms – such as processes that enable stakeholders to raise grievances or complaints – should also be established and communicated clearly.
- **Whether the AI system is certified:** certifications confirm that a system, process, or organisation satisfies a standard or regulatory requirement and does what it was designed to do. Certifications can be granted by governments, industry bodies, or other authorities. Certification can be general or sector-specific, and can be granted to all or parts of an AI system (e.g. the model, the data, etc.); to an actor (e.g. a user, a developer, an organisation); and to specific aspects of the system (e.g. explainable, fair, etc.).
- **Whether the AI system is insured:** insurance provides protection against unexpected risks or events. AI insurance programmes are starting to emerge to ensure redress and compensation in cases of unexpected damages or incidents (Kumar and Nagle, 2019[162]). Developing and pricing insurance contracts will require understanding the risks that an AI system faces.

Consult

Consultation with internal and external stakeholders – including civil society and affected communities – involves seeking feedback and insights to inform impact and risk assessments as well as to manage risks at each step of the process. Communication and consultation should take place at all phases of the AI system lifecycle and play a crucial role early in the design phase.

Adapting communications to enable understanding by external stakeholders – including those without a technical background – will facilitate meaningful dialogue and consultation. The format, cost, and frequency of communications and consultations should be assessed based on the context.

5.2 Embed a culture of risk management

A culture of risk management should be cultivated and embedded at all levels of organisations and across the AI value chain, with strong commitment by organisations' leadership teams (NIST, 2022[14]); (ISO, 2018[7]). The risk management process should be integrated into organisational quality and management systems and policies.

Organisations should devise, adopt and disseminate a combination of risk management policies that articulate an organisation's commitments to trustworthy AI. These policies should be embedded into an organisation's oversight bodies.

Risk management expectations and policies should be incorporated into engagement with suppliers and other stakeholders along the value chain (OECD, 2018[8]).

6. Next steps and discussion

This report illustrates how risk management approaches can enable the implementation of the OECD AI Principles throughout the AI system lifecycle. Notably, this report shows how OECD AI frameworks – including the OECD AI Principles, the AI system lifecycle and the OECD framework for classifying AI systems – can inform accountability in AI.

The report is a first step towards defining key components of the AI accountability ecosystem. Its objective is to trigger discussion, including at the OECD Working Party on AI Governance (AIGO) and the OECD.AI expert group on risk and accountability. Next steps include adding a policy layer to feed into on-going risk assessment work in co-operation with the European Commission, the US National Institute of Standards and Technology (NIST), the International Organization for Standardization (ISO), and others as appropriate.

In parallel, the OECD.AI expert group on risk and accountability developed a catalogue of tools and metrics for trustworthy AI to provide an interactive collection of resources to develop and implement AI systems that respect human rights and are fair, transparent, explainable, robust, secure, and safe. These tools, mapped to the OECD AI Principles and the phases of the AI system lifecycle, are expected to facilitate accountability in AI, from documenting and monitoring risks to certification and assurance.

Finally, there is growing demand for risk-assessment tools to calibrate treatment and mitigation strategies to the level of risk of an AI system. In this context, the OECD.AI expert group on risk and accountability is developing a risk-assessment framework, building on the criteria outlined in the classification framework (OECD, 2022[1]), to facilitate global interoperability for assessing and reporting risk. A “global AI incidents monitor” under development is expected to provide the evidence-base to inform this framework. The risk-assessment framework and the AI incidents monitor will be informed by the findings in this report.

Annex A. Presentations relevant to accountability in AI from the OECD.AI network of experts

Since July 2021, the OECD.AI expert group on tools and accountability has taken stock of initiatives and mechanisms in the AI accountability ecosystem. To identify those that exist or are in development, and possible gaps and areas for improvement, the Secretariat invited experts to present various accountability mechanisms for AI (Table A.1, panel a). In addition, since January 2022, the OECD.AI expert group on classification and risk has taken stock of key standards and initiatives in AI risk assessment and management (Table A.1, panel b).

Table A.1. OECD.AI expert presentations

a) OECD.AI Expert Group on Tools & Accountability, June 2021 - September 2022

Name and date	Organisation	Presentation theme
Nozha Boujemaa, 25 June 2021 (11th meeting)	IKEA Retail (Ingka Group)	Algorithmic accountability, technical tools for accountability and value by-design models
Adriano Koshiyama and Emre Kazim, 16 July 2021 (12th meeting)	University College London (UCL)	Auditing algorithms from a technical perspective, including managing legal, ethical, and technological risks of AI, machine learning and associated algorithms
Philipp Slusallek, 16 July 2021 (12th meeting)	Confederation of Laboratories for Artificial Intelligence Research in Europe (CLAIRE); German Research Center for Artificial Intelligence (DFKI)	Introduction to the AI projects at the Confederation of Laboratories for Artificial Intelligence Research in Europe (CLAIRE), including the “Trusted AI Initiative” that uses AI to optimise & certify AI
Ashley Casovan, 31 August 2021 (13th meeting) and 29 April 2022 (17th meeting)	Responsible AI Institute (RAI)	The Responsible AI Institute’s work to design and develop a certification programme for responsible AI
Andrea Renda, 31 August 2021 (13th meeting)	Centre for European Policy Studies (CEPS)	Overview of CEPS’ Study to Support an Impact Assessment of Regulatory Requirements for Artificial Intelligence in Europe and the FCAI Brookings/CEPS Forum for Cooperation on Artificial Intelligence
Craig Shank, 18 October 2021 (14th meeting)	Independent expert	The credibility of soft law to ensure accountability for artificial intelligence
Tyler Gillard and Rashad Abelson, 18 October 2021 (14th meeting)	OECD Centre for Responsible Business Conduct (RBC)	Responsible business conduct and accountability in AI and the link between the OECD AI Principles and the OECD Due Diligence Guidance
Clara Neppel, 13 January 2022 (15th meeting)	Institute of Electrical and Electronics Engineers (IEEE)	The IEEE 7000 Global Standard for addressing ethical concerns during system design
Vanja Skoric, 13 January 2022 (15th meeting)	The European Center for Not-for-Profit Law (ECNL)	Socio-legal architectures for sustainable AI development and the significance of human rights impact assessments (HRIA) as an instrument for accountability and trust
Stephanie Ifayemi, 13 January 2022 (15th meeting)	Digital Standards Policy, UK Department for Digital (DCMS)	The role of digital technical standards in the UK’s National AI Strategy and the framework for G7 collaboration on digital technical standards
Jenny Brennan, 29 April 2022 (17th meeting)	Ada Lovelace Institute	An Ada Lovelace Institute project on algorithmic impact assessment in healthcare
Yordanka Ivanova, 12 July 2022 (18th meeting)	DG CONNECT, European Commission	An update on the EU AI Act
Yeong Zee Kin, 12 July 2022 (18th meeting)	Infocomm Media Development Authority of Singapore	An overview of Singapore’s AI Verify initiative

Mikael Jansen, 16 September 2022 (19th meeting)	D-Seal - Danish Industry Foundation	D-Seal – a labelling program for IT security and responsible use of data in the EU
Kolja Verhage, 16 September 2022 (19th meeting)	Deloitte Risk Advisory	Lessons learned from implementation of values-based AI principles in the private sector

b) OECD.AI Expert Group on Classification & Risk, January - May 2022

Name and date	Organisation	Presentation theme
Viknesh Sounderajah, 2 February 2022 (19th meeting)	Imperial College London	Forming AI Evidence Standards for Health Technology Assessment Programmes, presentation on the study of using the OECD framework for the classification of AI systems in the healthcare sector.
Mark Latonero and Elham Tabassi, 2 February 2022 (19th meeting)	National Institute of Standards and Technology (NIST)	Update on the development of the NIST AI Risk Management Framework.
Sebastian Hallensleben, 2 February 2022 (19th meeting)	CEN-CENELEC	Current European regulation/standardization aspects on AI risk assessment.
Kai Zenner, 24 March 2022 (20th meeting)	European Parliament	Overview of the JURI report's key proposed amendments to the EU AI Act.
Peter Deussen, 24 March 2022 (20th meeting)	ISO	Overview of ISO/IEC 23894's relevance for Artificial Intelligence risk management.

Annex B. Participation in the OECD.AI Expert Group on Classification and Risk

Table B.1. Participation in the OECD.AI Expert Group on Classification & Risk (as of December 2022)

Name	Title	Organisation	Group / Delegation
Golo Rademacher	Policy Lab Digital, Work & Society	German Federal Ministry of Labour and Social Affairs	Germany
Judith Peterka	Head, AI indicators	Policy Lab Digital, Work & Society	Germany
Michael Schoenstein	Head of Strategic Foresight & Analysis	Policy Lab Digital, Work & Society	Germany
Jibu Elias	Research and Content Head	INDIAai	India
Barry O'Sullivan	Chair of Constraint Programming, the School of Computer Science & IT	University College Cork	Ireland
David Filip	Research Fellow, ADAPT Centre	Dublin City University (DCU)	Ireland
Takayuki Honda	Assistant Director, Multilateral Economic Affairs Office, Global Strategy Bureau	Ministry of Internal Affairs and Communications (MIC)	Japan
Yoichi Iida	Chair of the CDEP and Going Digital II Steering Group	Ministry of Internal Affairs and Communications (MIC)	Japan
Katrina Kosa-Ammari	Counsellor at Foreign Economic Relations Promotion Division	Ministry of Foreign Affairs	Latvia
Dunja Mladenić	Head of Artificial Intelligence Department	Slovenian Jožef Stefan Institute	Slovenia
Irene Ek	PhD and leader of the AI portfolio	Swedish Agency for Growth Policy Analysis	Sweden
Bilge Miraç	Advisor	Presidency of Digital Transformation Office	Türkiye
Mehmet Haklidir	Chief Researcher, Scientific and Technological Research Council.	Türkiye Informatics and Information Security Research Center	Türkiye
Fatma Bujasaim	Head of International Cooperation	Artificial Intelligence Office	United Arab Emirates
Lord Tim Clement-Jones	Lord at the UK Parliament	House of Lords	United Kingdom
Farahnaaz H Khakoo	Assistant Director	US Government Accountability Office	United States
Mark Latonero	Senior Policy Advisor on AI	National Institute of Standards and Technology	United States
Mohammed Motiwala	"Multilateral Engagement Officer, OECD and GPAI, Office of Multilateral Affairs International Communications & Information Policy, Bureau of Economic and Business Affairs	Department of State	United States
Elham Tabassi	Chief of Staff, Information Technology Laboratory	National Institute of Standards and Technology	United States
Taka Ariga	Chief Data Scientist | Director, Innovation Lab	US Government Accountability Office	United States
Nicholas Reese	Policy expert	Department of Homeland Security	United States
Giuditta de Prato	Researcher	European Commission Joint Research Centre (JRC)	European Commission
Juha Heikkilä	Head of Unit, Robotics	DG CONNECT	European Commission
Kilian Gross	Head of Unit, Artificial Intelligence Policy Development and Coordination	DG CONNECT	European Commission
Emilia Gómez	Lead Scientist, Human behaviour and machine intelligence	European Commission Joint Research Centre (JRC)	European Commission
Eric Badique	Adviser for Artificial Intelligence	European Commission	European Commission
Irina Orssich	Team Leader AI, DG CONNECT	DG CONNECT	European Commission
Tatjana Evas	Legal and Policy Officer	DG CONNECT	European Commission
Prateek Sibal	AI Policy Researcher, Knowledge Societies Division, Communication and Information Sector	UNESCO	IGO
Roberto Sanchez	Advisor - Data Scientist	Inter-American Development Bank	IGO
Kai Zenner	Head of Office	Axel Voss MEP	IGO
Ghazi Ahamat	Executive Manager - Health and Government	Quantium	Business
Gonzalo López-Barajas Húder	Head of Public Policy and Internet at Telefónica	Telefonica	Business
Igor Perisic	Vice President of Engineering and Chief Data Officer	LinkedIn	Business
Ilya Meyzin	Vice President, Data Science Strategy & Operations	Dun & Bradstreet	Business
Kathleen Walch	Managing partner and principal analyst	Cognilytica	Business
Kuansan Wang	Managing Director	Microsoft Research Outreach Academic Services	Business

Nicole Primmer	Senior Policy Director	BIAC	Business
Marco Ditta	Executive Director, ISP Group Data Officer	Intesa Sanpaolo	Business
Michel Morvan	Co-Founder and Executive Chairman	Cosmo Tech	Business
Olly Salzmann	Partner Deloitte/Managing Director	Deloitte KI GmbH and KIParkDeloitte GmbH	Business
Clara Neppel	Senior Director	IEEE European Business Operations	Technical
Daniel Schwabe	Professor at the Department of Informatics	Catholic University in Rio de Janeiro (PUC-Rio)	Technical
Jack Clark	Co-Founder	Anthropic	Technical
Jonathan Frankle	PhD Candidate	MIT Internet Policy Research Initiative (IPRI)	Technical
Marko Grobelnik	AI Researcher & Digital Champion	AI Lab of Slovenia's Jozef Stefan Institute	Technical
Masashi Sugiyama	Director, Center for Advanced Intelligence Project	RIKEN, Japan	Technical
Peter Addo	Head of DataLab and Senior Data Scientist	Agence Française de Développement (AFD)	Technical
Sebastian Hallensleben	Head of Digitalisation and AI	VDE Association for Electrical, Electronic & Information Technologies	Technical
Taylor Reynolds	Technology Policy Director	MIT Internet Policy Research Initiative (IPRI)	Technical
Abe Hsuan	Independent Expert	Irwin Hsuan	CSO/academia
Aurelie Jacquet	Consultant		CSO/academia
Catherine Aiken	Researcher	Center for Security and Emerging Technology (CSET), Georgetown University	CSO/academia
Eric Badique	Independent consultant	GPAI	CSO/academia
Eva Thelisson	Researcher	AI Transparency Institute	CSO/academia
Guillaume Chevillon	Professor - Co Director ESSEC	ESSEC Business School, Paris	CSO/academia
Jim Kurose	Professor	University of Massachusetts Amherst	CSO/academia
Nicolas Moes	Head of Operations and EU AI Policy	The Future Society	CSO/academia
Olivia Erdélyi	Lecturer	University of Canterbury, School of Law	CSO/academia
Sally Radwan	Independent consultant		CSO/academia
Suso Baleato	Secretary	CSISAC	CSO/academia
Theodoros Evgeniou	Professor, Decision Sciences and Technology Management	INSEAD	CSO/academia
Tim Rudner	PhD Candidate	University of Oxford	CSO/academia
Till Klein	Team lead for Trustworthy AI	appliedAI	CSO/academia
Vincent C. Müller	Professor for Philosophy of Technology	Technical University of Eindhoven	CSO/academia
Fernando Galindo-Rueda	Secretariat	OECD	OECD
Karine Perset	Secretariat	OECD	OECD
Pierre Montagnier	Secretariat	OECD	OECD
Luis Aranda	Secretariat	OECD	OECD
Leonidas Aristodemou	Secretariat	OECD	OECD
Annelore Verhagen	Secretariat	OECD	OECD
Orsolya Dobe	Secretariat	OECD	OECD

Annex C. Participation in the OECD.AI Expert Group on Tools and Accountability

Table C.1. Participation in the OECD.AI Expert Group on Tools & Accountability (as of December 2022)

Name	Title	Organisation	Group / Delegation
Alana Lomonaco Busto	First Secretary- Cybersecurity, Cybercrime and Digital Affairs	Ministry of Foreign Affairs, International Trade and Worship	Argentina
Ben Macklin	Manager, Global Digital Policy, Digital Economy and Technology Division	Australia's Department of Industry, Innovation & Science	Australia
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
Andrejs Vasiljevs	Co-founder and Executive Chairman	Tilde	Business
Angelica Biard	Attachée aux affaires multilatérales	Délégation aux Affaires Francophones et Multilatérales, Gouvernement du Québec	Canada
Etienne Corriveau-Hebert	Head of partnerships division	Ministère des Relations internationales et de la Francophonie	Canada
Matthew Smith	Senior Program Specialist, Education and Science Division	International Development Research Centre	Canada
Marek Havrda	AI Policy and Social Impact Director	GoodAI	Czech Republic
Frederik Weiergang Larsen	Special Advisor	Danish Business Authority	Denmark
Maria Danmark Nielsen	Head of Section	Danish Business Authority	Denmark
Elisa Fromont	Professor	Université de Rennes 1	France
Guillaume Avrin	Evaluation of Artificial Intelligence department	LNE, French National Laboratory for Metrology and Testing	France
Renaud Vedel	Coordonnateur de la stratégie nationale en IA	Ministère de l'intérieur	France
Najma Bichara	Advisor, Digital Affairs	French Ministry for Europe and Foreign Affairs	France
László Boa	General Manager	AI Coalition of Hungary	Hungary
Barry Smyth	Digital Chair of Computer Science, Director of the Insight Centre for Data Analytics	University College Dublin	Ireland
John McCarthy	Global Lead for Shared, Connected and Autonomous Vehicles	Arup	Ireland
Dino Pedreschi	Professor of Computer Science	University of Pisa	Italy
Rosa Meo	Associate Professor of Computer Science	University of Torino	Italy
Osamu Sudoh	Graduate School of Interdisciplinary Information Studies (GSII)	University of Tokyo	Japan
Takayuki Honda	Assistant Director, Multilateral Economic Affairs Office, Global Strategy Bureau	Ministry of Internal Affairs and Communications (MIC)	Japan
Yoichi Iida	Chair of the CDEP and Going Digital II Steering Group	Ministry of Internal Affairs and Communications (MIC)	Japan
Saïd El Haroui	Head of International Organisations	Ministry of Economic Affairs and Communications	Netherlands
Vanja Skoric	Program Director	European Center for Not-for-profit Law	Netherlands
Colin Gavaghan	Director	Law Foundation-sponsored Centre for Law and Policy in Emerging Technologies	New Zealand
Emma Naji	Executive Director	AI Forum New Zealand	New Zealand
Yeong Zee Kin	Assistant Chief Executive	Infocomm Media Development Authority	Singapore
Gregor Strojin	State Secretary	Ministry of Justice	Slovenia
Francois Ortolan	Digital Standards Technical Lead	Department for Digital, Culture, Media and Sport	United Kingdom
Michael Birtwistle	Policy Adviser	Centre for Data Ethics and Innovation (CDEI)	United Kingdom
Farahnaaz H Khakoo	Assistant Director	US Government Accountability Office	United States
Grace Abuhamad	Chief of Staff	U.S. Department of Commerce's National Telecommunications and Information Administration (NTIA)	United States

Jaclyn Kerr	AAAS Science and Technology Policy Fellow	Office of the Science and Technology Advisor to the Secretary	United States
Mark Latonero	Senior Policy Advisor on AI	National Institute of Standards and Technology	United States
Mohammed Motiwala	Multilateral Engagement Officer, OECD and GPAI, Office of Multilateral Affairs International Communications & Information Policy, Bureau of Economic and Business Affairs	Department of State	United States
Taka Ariga	Chief Data Scientist | Director, Innovation Lab	US Government Accountability Office	United States
Juha Heikkilä	Head of Unit, Robotics	DG CONNECT	European Commission
Kilian Gross	Head of Unit, Artificial Intelligence Policy Development and Coordination	DG CONNECT	European Commission
Emilia Gómez	Lead Scientist, Human behaviour and machine intelligence	European Commission DG Joint Research Centre (JRC)	European Commission
Eric Badique	Adviser for Artificial Intelligence	European Commission	European Commission
Salvatore Scalzo	Policy and legal officer, Artificial Intelligence Policy Development and Coordination Unit	DG CONNECT	European Commission
Irina Orssich	Team Leader AI, DG CONNECT	European Commission	European Commission
Cedric Wachholz	Head of UNESCO's ICT in Education, Science and Culture section	UNESCO	IGO
Cristina Pombo	Principal Advisor and Head of the Digital and Data Cluster, Social Sector	Inter-American Development Bank	IGO
Alice Munyua	Director, Africa Innovation and public policy program	Mozilla Africa	Business
Andrejs Vasiljevs	Co-founder and Executive Chairman	Tilde	Business
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
Carolyn N'Guyen	Director of Technology Policy	Microsoft	Business
Gonzalo López-Barajas Húder	Head of Public Policy and Internet at Telefónica	Telefonica	Business
Kathleen Walch	Managing partner and principal analyst	Cognilytica	Business
Craig Shank	Independent Advisor, Consultant, and Speaker	Independent expert	Business
Daniel Faggella	Head of Research, CEO	Emerj AI Research	Business
David Sadek	Vice President for Research, Technology & Innovation	Thales	Business
Dominik Geller	Head of Group Digital Governance	Sanofi	Business
Nicole Primmer	Senior Policy Director	BIAC	Business
Nozha Boujemaa*	Global Vice President, Digital Ethics and Responsible AI	IKEA Retail (Ingka Group)	Business
Philip Dawson	Policy Lead	Schwartz Reisman Institute for Technology and Society	Business
Emmanuel Bloch	Director of Strategic Information	Thales	Business
Emmanuel Kahembwe	CEO	VDE Association for Electrical, Electronic & Information Technologies	Business
Jennifer Bernal	Lead on Global Policy	Deepmind	Business
Lynette Webb	Senior Manager for AI Policy Strategy	Google	Business
Marc-Etienne Ouimette	Global Leader for AI Policy	Amazon Web Services	Business
Marian Gläser	CEO and Founder	Policy Lead	Business
Navrina Singh	CEO	Credo AI	Business
Norberto Andrade	Privacy and Public Policy Manager	Facebook	Business
Peter Cihon	Policy Analyst	Github	Business
Sasha Rubel	Public Policy Lead, AI / ML, EMEA	Amazon Web Services	Business
Will Carter	Global Policy Lead for Responsible AI	Google	Business
Peter Cihon	Policy Analyst	Github	Business
Craig Shank	Independent Advisor, Consultant, and Speaker	Independent expert	Business
Marko Grobelnik	AI Researcher & Digital Champion	AI Lab of Slovenia's Jozef Stefan Institute	Technical
Clara Neppel	Senior Director	IEEE European Business Operations	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
Irene Solaiman	Public policy	OpenAI	Technical
Raja Chatila	Chair	IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	Technical
Sebastian Hallensleben	Head of Digitalisation and AI	VDE Association for Electrical, Electronic & Information Technologies	Technical
Aishik Ghosh	PhD in Artificial Intelligence for Particle Physics in Atlas	European Organization for Nuclear Research (CERN)	Technical
Wael Diab	Chair	ISO/IEC JTC 1/SC 42 Artificial intelligence	Technical
Christine Custis	Director of Programs and Research	Partnership on AI	CSO/academia
Danit Gal	Independent consultant		CSO/academia
Eric Badique	Independent consultant	GPAI	CSO/academia
Eva Thelisson	Researcher	AI Transparency Institute	CSO/academia

Guillaume Chevillon	Professor - Co Director ESSEC	ESSEC Business School, Paris	CSO/academia
Jim Kurose	Professor	University of Massachusetts Amherst	CSO/academia
Philipp Slusallek	Scientific Director and Member of the Executive Board	German Research Center for Artificial Intelligence (DFKI)	CSO/academia
Stephanie Ifayemi	Head of Policy	Partnership on AI	CSO/academia
Suso Baleato	Secretary	CSISAC	CSO/academia
Theodoros Evgeniou	Professor, Decision Sciences and Technology Management	INSEAD	CSO/academia
Tim Rudner	PhD Candidate	University of Oxford	CSO/academia
Mikael Jensen	Director of the D-mærket/D-seal	The Danish Industry Foundation	CSO/academia
Yolanda Lannquist	Head of Research & Advisor	The Future Society	CSO/academia
Andrea Renda*	Senior Research Fellow and Head of Global Governance, Regulation, Innovation and the Digital Economy (GRID)	Centre for European Policy Studies at Duke University	CSO/academia
Ashley Casovan	CEO	RAI	CSO/academia
Carlos Ignacio Gutierrez	Artificial intelligence (AI) policy researcher	Future of Life Institute	CSO/academia
Catherine Régis	Professor & holder of a Canada Research Chair	University of Montreal Law Faculty	CSO/academia
Christina Colclough	Future of Work and Politics of Technology	Independent Expert	CSO/academia
Jessica Newman	Program Lead - AI Security Initiative	Center for Long-Term Cybersecurity (UC Berkeley)	CSO/academia
Marc-Antoine Dilhac	Professor of philosophy	Université de Montréal	CSO/academia
Marjorie Buchser	Head of Innovation Partnerships and Digital Society Initiative	Chatham House	CSO/academia
Nicolas Miailhe	Co-Founder	The Future Society	CSO/academia
Pam Dixon	Founder and Executive Director	World Privacy Forum	CSO/academia
Ryan Budish	Executive Director, Berkman Klein Center for Internet & Society	Harvard University	CSO/academia
Sally Radwan	Independent consultant		CSO/academia
Wendell Wallach	Consultant, ethicist, and scholar	Yale University's Interdisciplinary Center for Bioethics	CSO/academia
Andrew Pakes	Deputy general secretary and research director	Prospect Union	Trade Union
Nicolas Blanc	Délégué national au numérique	CFE-CGC	Trade Union
Oliver Suchy	Director	Digital World of Work and Future of Work unit of the German Trade Union Confederation (DGB)	Trade Union
Victor Bernhardtz	Ombudsman for Digital Labour Markets	Unionen	Trade Union
Anna Byhovskaya	Senior Policy Advisor	TUAC	Trade Union
Angelica Salvi del Pero	Secretariat	OECD	OECD
Alistair Nolan	Secretariat	OECD	OECD
Karine Perset	Secretariat	OECD	OECD
Luis Aranda	Secretariat	OECD	OECD
Francesca Sheeka	Secretariat	OECD	OECD

References

Abadi, M. et al. (2016), *Deep learning with differential privacy*. [103]

Almenzar, M. et al. (2022), *JRC science for policy report*, https://publications.jrc.ec.europa.eu/repository/bitstream/JRC129614/JRC129614_01.pdf. [60]

Alves, G. et al. (2021), *Reducing Unintended Bias of ML Models on Tabular and Textual Data*, https://arxiv.org/abs/2108.02662. [95]

Ancona, M. et al. (2017), "Towards better understanding of gradient-based attribution methods for deep neural networks", *arXiv preprint*, https://doi.org/arXiv:1711.06104. [125]

Arai-Takahashi, Y. (2002), *The Margin of Appreciation Doctrine and the Principle of Proportionality in the Jurisprudence of the ECHR,*, Intersentia nv. [50]

Arlot, S. and A. Celisse (2010), "A survey of cross-validation procedures for model selection", *Statistics surveys*, pp. 40-79. [17]

Ateniese, G. et al. (2015), "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers", *International Journal of Security and Networks*, Vol. 10/3, pp. 137-150. [33]

Babic, B. et al. (2019), *Algorithms on Regulatory Lockdown in Medicine*. [64]

Bakalar, C., R. Barreto and S. Bergman (2021), *Fairness on the ground: Applying algorithmic fairness approaches to production systems.*, arXiv preprint arXiv:2103.06172, https://arxiv.org/abs/2103.06172. [94]

Barocas, S. and A. Selbst (2016), "Big Data's Disparate Impact", *CALIF. L. REV.*, Vol. 104/671. [22]

Bellamy, R. et al. (2018), *AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias*. [27]

Bender, E. et al. (2021), *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*. [19]

Bhatt, U. et al. (2019), *Explainable Machine Learning in Deployment*, arXiv:1909.06342, https://arxiv.org/abs/1909.06342. [131]

Bieker, F. et al. (2016), *A process for data protection impact assessment under the european general data protection regulation*, Springer, Cham. [31]

Breiman, L. (2001), “Random forests”, *Machine learning*, Vol. 45(1), pp. 5-32. [126]

Brundage, M. et al. (2020), *The Malicious Use of AI: Forecasting, Prevention, and Mitigation*, https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf. [155]

BSA (2021), “Confronting Bias: BSA’s Framework to Build Trust in AI”, https://ai.bsa.org/wp-content/uploads/2021/06/2021bsaaibias.pdf. [10]

Buckman, J. et al. (2018), *Thermometer encoding: One hot way to resist adversarial examples*. [137]

Bunel, R. et al. (2018), *A unified view of piecewise linear neural network verification*. [140]

Butterworth, M. (2018), “The ICO and artificial intelligence: The role of fairness in the GDPR framework”, *Computer Law & Security Review*, Vol. 34/2, pp. 257-268. [30]

Carlini, N. et al. (2019), *On evaluating adversarial robustness*, https://arxiv.org/pdf/1902.06705.pdf. [58]

Charisi, V. et al. (2022), “Artificial Intelligence and the Rights of the Child: Towards an Integrated Agenda for Research and Policy”, https://doi.org/10.2760/012329, JRC127564. [9]

Chhetri, T. et al. (2022), “Data Protection by Design Tool for Automated GDPR Compliance Verification Based on Semantically Modeled Informed Consent”, *Sensors*, Vol. 22/7, p. 2763, https://doi.org/10.3390/s22072763. [108]

Chouldechova, A. (2017), “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments”, *Big data*, Vol. 5/2, pp. 153-163. [23]

Ciatto, G. et al. (2020), *Agent-Based Explanations in AI: Towards an Abstract Framework.*, Springer, https://doi.org/10.1007/978-3-030-51924-7_1. [65]

Corbett-Davies, S. et al. (2017), *Algorithmic decision making and the cost of fairness*. [25]

Council of Europe (2019), *Unboxing Artificial Intelligence: 10 steps to protect Human Rights, Recommendation of the Council of Europe*, https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64. [47]

De Cristofaro, E. (2020), “An Overview of Privacy in Machine Learning”. [29]

Delgado, F. et al. (2021), “Stakeholder Participation in AI: Beyond “Add Diverse Stakeholders and Stir””, https://arxiv.org/pdf/2111.01122.pdf. [70]

Deming, E. (1968), *Out of the crisis*, Massachusetts Institute of Technology. [158]

Deogun, D., D. Johnsson and D. Sawano (2019), *Secure by Design*, Manning Publications. [134]

Donini, M. et al. (2018), “Empirical Risk Minimization Under Fairness Constraints”, *Advances in Neural Information Processing Systems*, pp. 2791-2801, https://papers.nips.cc/paper/2018/file/83cdcec08fbf90370fcf53bdd56604ff-Paper.pdf. [91]

Dwork, C. and M. Naor (2010), “On the difficulties of disclosure prevention in statistical databases or the case for differential privacy”, *Journal of Privacy and Confidentiality*, Vol. 2/1. [36]

Efron, B. and T. Hastie (2016), *Computer age statistical inference*, Cambridge University Press. [80]

Erdélyi, O. and J. Goldsmith (2022), *Regulating artificial intelligence: Proposal for a global solution*, https://www.sciencedirect.com/science/article/pii/S0740624X22000843?dgcid=author. [69]

Escovedo, T. et al. (2018), "DetectA: abrupt concept drift detection in non-stationary environments", *Applied Soft Computing* 62, pp. 119-133. [81]

EU-HLEG (2019), "Ethics guidelines for trustworthy AI", https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. [28]

European Commission (2020), "White Paper on Artificial Intelligence: A European approach to excellence and trust", https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf. [11]

Fachkha, C. (2016), *Security Monitoring of the Cyber Space*. [154]

Feldman, M. et al. (2015), *Certifying and removing disparate impact*. [66]

Fernandez Llorca, D. and E. Gomez (2021), "Trustworthy Autonomous Vehicles", https://doi.org/10.2760/120385. [99]

Fjeld, J. et al. (2020), *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*, Berkman Klein Center. [3]

FRA (2020), *Getting the future right: artificial intelligence and fundamental rights*, European Union Agency for Fundamental Rights, https://fra.europa.eu/sites/default/files/fra_uploads/fra-2020-artificial-intelligence_en.pdf. [112]

Fredrikson, M., S. Jha and T. Ristenpart (2015), *Model inversion attacks that exploit confidence information and basic countermeasures*. [37]

Ganju, K. et al. (2018), *Property inference attacks on fully connected neural networks using permutation invariant representations*. [40]

Ghantous, G. and A. Gill. (2017), *DevOps: Concepts, Practices, Tools, Benefits and Challenges*. [132]

Goethals, S., D. Martens and T. Evgeniou (2022), "The non-linear nature of the cost of comprehensibility.", *Journal of Big Data* 9, https://doi.org/10.1186/s40537-022-00579-2. [61]

Goldsteen, A. et al. (2020), "Data Minimization for GDPR Compliance". [68]

Goodfellow, I., Y. Bengio and A. Courville (2016), *Deep Learning (Adaptive Computation and Machine Learning series)*, The MIT Press. [151]

Google (2019), *Rate-limiting strategies and techniques*, https://cloud.google.com/architecture/rate-limiting-strategies-techniques#techniques-enforcing-rate-limits (accessed on 26 April 2022). [107]

Gray, M. and S. Suri (2019), *Ghost work: How to stop Silicon Valley from building a new global underclass*, Eamon Dolan Books. [48]

Greer (2004), ""Balancing" and the European court of Human Rights: a contribution to the [51]

Habermass-Alexy debate", *The Cambridge Law Journal*, Vol. 63/2.

Guo, C. et al. (2017), *Countering adversarial images using input transformations*. [145]

Hall, P. (2019), *An introduction to machine learning interpretability*, O'Reilly Media, Incorporated. [55]

Hanley, J. and B. Mcneil (1982), *The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve*. [148]

Higgins, I. et al. (2017), "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework", *ICLR*, https://openreview.net/forum?id=Sy2fzU9gl. [120]

Hitaj, B., G. Ateniese and F. Perez-Cruz (2017), *Deep models under the GAN: information leakage from collaborative deep learning*. [38]

Holland, S. et al. (2020), "The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards", *Data Protection and Privacy*, Vol. 12, https://doi.org/10.48550/arXiv.1805.03677. [118]

Huang, P. et al. (2019), *Achieving verified robustness to symbol substitutions via interval bound propagation*. [138]

Huang, S. et al. (2017), *Adversarial attacks on neural network policies*. [141]

Hupont, I. and E. Gomez (2022), *Documenting use cases in the affective computing domain using Unified Modeling Language*, https://arxiv.org/pdf/2209.09666.pdf. [115]

Hupont, I. et al. (2022), "Documenting high-risk AI: an European regulatory perspective,", *Computer Magazine*, https://www.techrxiv.org/articles/preprint/Documenting_high-risk_AI_an_European_reg. [117]

Hupont, I. et al. (2022), "The landscape of facial processing applications in the context of the European AI Act and the development of trustworthy systems", *Sci Rep* 12, p. 10688, https://doi.org/10.1038/s41598-022-14981-6. [116]

ICO (2022), *Audits*, UK Information Commissioner's Office, https://ico.org.uk/for-organisations/audits/ (accessed on 30 November 2022). [42]

ICO-Alan Turing Institute (2020), *Explaining decisions made with AI Information*, https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/. [63]

IDB-OECD (2021), *Responsible use of AI for public policy: Data science toolkit*, https://publications.iadb.org/publications/english/document/Responsible-use-of-AI-for-public-policy-Data-science-toolkit.pdf. [21]

Iosifidis, V. and E. Ntoutsi (2018), "Dealing with bias via data augmentation in supervised learning senarios", *Jo Bates Paul D. Clough Robert Jäschke*, https://www.bibsonomy.org/bibtex/2631924ce7d73cd8e3bb6477e84d408fa/entoutsi. [86]

ISO (2018), *ISO 31000 - Risk Management*, https://doi.org/[online] Available at: <https://www.iso.org/obp/ui#iso:std:iso:31000:ed-2:v1:en>. [7]

Javadi, S. et al. (2020), *Monitoring Misuse for Accountable 'Artificial Intelligence as a Service'*, [152] https://doi.org/10.1145/3375627.3375873.

Jobin, A., M. Ienca and E. Vayena (2019), "The global landscape of AI ethics guidelines", *Nat Mach Intell 1*, pp. 388-399, https://doi.org/10.1038/s42256-019-0088-2. [4]

Kaminski, M. and G. Malgieri (2020), *Multi-layered explanations from algorithmic impact assessments in the GDPR In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. [161]

Kamiran, F. and T. Calders (2012), "Data preprocessing techniques for classification without discrimination", *Knowledge and Information Systems* 33, pp. 1-33, https://link.springer.com/article/10.1007/s10115-011-0463-8. [85]

Kargupta, H. et al. (2005), "Random-data perturbation techniques and privacy-preserving data mining", *Knowledge and Information Systems*, Vol. 7/4, pp. 387-414, https://doi.org/10.1007/s10115-004-0173-6. [100]

Karimi, A., B. Schölkopf and I. Valera (2020), "Algorithmic Recourse: from Counterfactual Explanations to Interventions", *arXiv preprint*, https://doi.org/arXiv:2002.06278. [130]

Kaufmann, L. and P. Rousseeuw (1987), "Clustering by Means of Medoids, Data Analysis based on the L1-Norm and Related Methods", pp. 405-416, https://wis.kuleuven.be/stat/robust/papers/publications-1987/kaufmanrousseeuw-clusteringbymedoids-l1norm-1987.pdf. [119]

Kim, H. et al. (2019), "Blockchained on-device federated learning", *IEEE Communications Letters*, https://arxiv.org/pdf/1808.03949.pdf. [102]

Kleinberg, J., S. Mullainathan and M. Raghavan (2016), *Inherent trade-offs in the fair determination of risk scores*. [24]

Klinova, K. and A. Korinek (2021), *AI and Shared Prosperity*. [73]

Kocmi, T. (2020), "Exploring Benefits of Transfer Learning in Neural Machine Translation", https://arxiv.org/abs/2001.01622. [78]

Korinek, A. and J. Stiglitz (2021), "Artificial Intelligence, Globalization, and Strategies for Economic Development", *Institute for New Economic Thinking Working Paper Series* 146, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3812820. [72]

Koshiyama, A. and Z. Engin (2019), *Algorithmic Impact Assessment: Fairness, Robustness and Explainability in Automated Decision-Making*. [160]

Koshiyama, A., N. Firoozye and P. Treleaven (2020), "Algorithms in Future Capital Markets". [62]

Koshiyama, A. et al. (2021), *Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms*, http://dx.doi.org/10.2139/ssrn.3778998. [26]

Kumar, R. and F. Nagle (2019), *The Case for AI Insurance*, Harvard Business Review, https://hbr.org/2020/04/the-case-for-ai-insurance. [162]

Kusner, M. et al. (2017), "Counterfactual fairness", *Advances in Neural Information Processing Systems*, pp. 4066-4076, https://papers.nips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf. [92]

Leino, K., Z. Wang and M. Fredrikson (2021), *Globally-Robust Neural Networks*, https://doi.org/10.48550/arXiv.2102.08452. [146]

Lhoest, Q. et al. (2021), *Datasets: A Community Library for Natural Language Processing*. [149]

Lindkvist, C., A. Stasis and J. Whyte (2013), “Configuration Management in Complex Engineering Projects”, *Procedia CIRP*, Vol. 11, pp. 173-176. [82]

Liu, J. et al. (2017), *Oblivious neural network predictions via MiniONN transformations*. [106]

Longo, L. et al. (2020), *Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions*. [52]

Lu, J. et al. (2019), *Learning under Concept Drift: A Review*, IEEE Transactions on Knowledge and Data Engineering. [157]

Lundberg, S. and S. Lee (2017), “A unified approach to interpreting model predictions.”, *Advances in neural information processing systems*, pp. 4765-4774. [127]

Manikandan, G. and S. Abirami (2021), “An Efficient Feature Selection Framework Based on Information Theory for High Dimensional Data”, *Applied Soft Computing*, Vol. 111. [74]

McMahan, B. and D. Ramage (2017), *Federated learning: Collaborative machine learning without centralized training data*, https://ai.googleblog.com/2017/04/federated-learning-collaborative.html. [101]

Melis, L. et al. (2019), *Exploiting unintended feature leakage in collaborative learning*, IEEE. [41]

Micheli, M. et al. (2020), “Emerging models of data governance in the age of datafication”, *Big Data & Society*, Vol. 7/2, https://doi.org/10.1177/2053951720948087. [43]

Molnar, C., G. Casalicchio and B. Bischl (2020), *Interpretable Machine Learning - A Brief History, State-of-the-Art and Challenges*, Springer. [56]

Moosavi-Dezfooli, S., A. Fawzi and P. Frossard (2016), *Deepfool: a simple and accurate method to fool deep neural networks*. [142]

Müller, R., S. Kornblith and G. Hinton (2019), *When does label smoothing help?*. [136]

Neubauer, T. and J. Heurix (2011), “A methodology for the pseudonymization of medical data”, *International Journal of Medical Informatics*, Vol. 80/3, pp. 190-204, https://doi.org/10.1016/j.ijmedinf.2010.10.016. [97]

Newman, J. (2023), *A Taxonomy of Trustworthiness for Artificial Intelligence: Connecting Properties of Trustworthiness with Risk Managment and the Lifecycle*, UC Berkeley, https://cltc.berkeley.edu/wp-content/uploads/2023/01/Taxonomy_of_AI_Trustworthiness.pdf. [15]

NIST (2022), *AI Risk Management Framework: Second Draft*, https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf (accessed on 18 August 2022). [14]

Nori, H. et al. (2019), “Interpretml: A unified framework for machine learning interpretability”, *arXiv preprint*, https://doi.org/arXiv:1909.09223. [122]

NSF (2022), *National Artificial Intelligence (AI) Research Institutes Accelerating Research, Transforming Society, and Growing the American Workforce, Program Solicitation NSF 22-502*, https://www.nsf.gov/pubs/2022/nsf22502/nsf22502.htm#pgm_intr_txt. [2]

Nurgaliev, I., E. Karavakis and A. Aimar (2016), "Kibana, Grafana and Zeppelin on Monitoring data", https://doi.org/10.5281/zenodo.61079. [83]

OECD (2022), "Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint"*, OECD Digital Economy Papers*, No. 341, OECD Publishing, Paris, https://doi.org/10.1787/7babf571-en. [111]

OECD (2022), "OECD Framework for the Classification of AI systems"*, OECD Digital Economy Papers*, No. 323, OECD Publishing, Paris, https://doi.org/10.1787/cb6d9eca-en. [1]

OECD (2022), *Rationale for the OECD AI Principle on "Accountability"*, https://oecd.ai/en/dashboards/ai-principles/P9 (accessed on 9 April 2022). [164]

OECD (2022), *Rationale for the OECD AI Principle on "Human-centred values and fairness"*, https://oecd.ai/en/dashboards/ai-principles/P6 (accessed on 9 April 2022). [20]

OECD (2022), *Rationale for the OECD AI Principle on "Inclusive growth, sustainable development and well-being"*, https://oecd.ai/en/dashboards/ai-principles/P5 (accessed on 9 April 2022). [16]

OECD (2022), *Rationale for the OECD AI Principle on "Robustness, security and safety"*, https://oecd.ai/en/dashboards/ai-principles/P9 (accessed on 9 April 2022). [57]

OECD (2022), *Rationale for the OECD AI Principle on "Transparency and explainability"*, https://oecd.ai/en/dashboards/ai-principles/P7 (accessed on 9 April 2022). [53]

OECD (2021), *OECD Business and Finance Outlook 2021: AI in Business and Finance*, OECD Publishing, Paris, https://doi.org/10.1787/ba682899-en. [110]

OECD (2021), "Recommendation of the Council on Broadband Connectivity", https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0322 (accessed on 24 June 2022). [163]

OECD (2019), *Artificial Intelligence in Society*, OECD Publishing, Paris, https://doi.org/10.1787/eedfee77-en. [46]

OECD (2019), *Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies*, OECD Publishing, Paris, https://doi.org/10.1787/276aaca8-en. [13]

OECD (2019), *Recommendation of the Council on Artificial Intelligence*, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. [5]

OECD (2019), "Scoping the OECD AI principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO)"*, OECD Digital Economy Papers*, No. 291, OECD Publishing, Paris, https://doi.org/10.1787/d62f618a-en. [6]

OECD (2018), *OECD Due Diligence Guidance for Responsible Business Conduct*, http://mneguidelines.oecd.org/OECD-Due-Diligence-Guidance-for-Responsible-Business-Conduct.pdf. [8]

OECD (2011), “OECD Guidelines for Multinational Enterprises”, https://dx.doi.org/10.1787/9789264115415-en. [45]

OHCHR (2011), *Guiding Principles on Business and Human Rights*, United Nations Human Rights Office of the High Commissioner, https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf. [44]

Orekondy, T., B. Schiele and M. Fritz (2019), *Knockoff nets: Stealing functionality of black-box models*. [35]

PAI (2021), *Responsible Sourcing of Data Enrichment Services*, https://partnershiponai.org/paper/responsible-sourcing-considerations/. [49]

Palacio, S. et al. (2021), *XAI Handbook: Towards a Unified Framework for Explainable AI*, https://doi.org/10.48550/arXiv.2105.06677. [113]

Paleyes, A., R. Urma and N. Lawrence (2020), “Challenges in Deploying Machine Learning: a Survey of Case Studies”, *ACM Computer Surveys*, https://arxiv.org/abs/2011.09926. [79]

Papernot, N. et al. (2017), “Semi-Supervised Knowledge Transfer for Deep Learning from Private Training Data”, https://arxiv.org/pdf/1610.05755.pdf. [105]

Park, T. (2022), *Making AI Inclusive: 4 Guiding Principles for Ethical Engagement*, https://partnershiponai.org//wp-content/uploads/dlm_uploads/2022/07/PAI_whitepaper_making-ai-inclusive.pdf. [12]

Patterson, D. et al. (2021), *Carbon Emissions and Large Neural Network Training*, https://arxiv.org/pdf/2104.10350.pdf. [77]

Pleiss, G. et al. (2017), “On fairness and calibration”, *Advances in Neural Information Processing Systems*, pp. 5680-5689, https://papers.nips.cc/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf. [93]

Prince, A. and D. Schwarcz (2020), “Proxy Discrimination in the Age of Artificial Intelligence and Big Data”, *Iowa Law Review*, Vol. 105/3, pp. 1257-1318, https://ilr.law.uiowa.edu/print/volume-105-issue-3/proxy-discrimination-in-the-age-of-artificial-intelligence-and-big-data. [84]

Puiutta, E. and E. Veith (2020), *Explainable Reinforcement Learning: A Survey*, arXiv preprint arXiv:2005.06247. [123]

Qin, C. et al. (2019), “Verification of non-linear specifications for neural networks”, http://arXiv:1902.09592. [59]

Rabanser, S., S. Gunnemann and Z. Lipton (2018), *Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift*, https://doi.org/10.48550/arXiv.1810.11953. [150]

Raji, D. et al. (2020), *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*, https://doi.org/10.48550/arXiv.2001.00973. [114]

Reisman, D. et al. (2019), *Algorithmic impact assessment: a practical framework for public agency accountability*, AI Now Institute. [159]

Ribeiro, M., S. Singh and C. Guestrin (2016), *“Why should I trust you?” Explaining the* [124] *predictions of any classifier*.

Rudin, L., S. Osher and E. Fatemi (1992), *Nonlinear total variation based noise removal algorithms*, https://doi.org/10.1016/0167-2789(92)90242-F. [144]

Schwartz, R. et al. (2021), *Proposal for Identifying and Managing Bias in Artificial Intelligence (SP 1270)*, https://www.nist.gov/artificial-intelligence/proposal-identifying-and-managing-bias-artificial-intelligence-sp-1270. [54]

Sculley, D. et al. (2015), *Hidden technical debt in machine learning systems*. [156]

Shahin, M., M. Babar and L. Zhu (2017), *Continuous Integration, Delivery and Deployment: A Systematic Review on Approaches, Tools, Challenges and Practices*, https://doi.org/10.1109/ACCESS.2017.2685629. [133]

Shakeel, S. et al. (2021), *k-NDDP: An Efficient Anonymization Model for Social Network Data Release*, p. 2440, https://doi.org/10.3390/electronics10192440. [98]

Shokri, R. et al. (2017), *Membership inference attacks against machine learning models*, IEEE. [39]

Spracklen, L. (2021), *Sparse Models are Fast Models: Improving DNN Inference Performance by over 10X*. [76]

Steinhardt, J., P. Koh and P. Liang (2017), “Certified defences for data poisoning attacks”, *Advances in Neural Information Processing Systems*, https://arxiv.org/pdf/1706.03691.pdf. [104]

Strier, K., J. Clark and S. Khareghani (2022), *Measuring compute capacity: A critical step to capturing AI’s full economic potential*, https://oecd.ai/en/wonk/ai-compute-capacity. [18]

Subbaswamy, A. (2020), *Evaluating Model Robustness to Dataset Shift*, https://doi.org/ArXiv, abs/2010.15100. [147]

Suresh, H. and J. Guttag (2021), “A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle”, https://arxiv.org/pdf/1901.10002.pdf. [89]

Tan, T. and R. Shokri (2019), “Bypassing backdoor detection algorithms in deep learning”. [32]

The European Consumer Organisation (2021), *Regulating AI to Protect the Consumer*. [135]

Tjeng, V. and R. Tedrake (2017), *Verifying neural networks with mixed integer programming*, https://doi.org/arXiv:1711.07356. [143]

Tramèr, F. et al. (2016), *Stealing machine learning models via prediction apis*. [34]

Tripathi, S. et al. (2021), “Ensuring the Robustness and Reliability of Data-Driven Knowledge Discovery Models in Production and Manufacturing”, *Frontiers in Artifical Intelligence*, Vol. 4, p. 22, https://www.frontiersin.org/articles/10.3389/frai.2021.576892/full. [87]

U.S. Department of Health & Human Services (2021), *Trustworthy AI (TAI) Playbook*. [139]

U.S. Department of Health & Human Services (2012), *Enterprise Performance Life Cycle Framework*. [153]

U.S. Department of Health and Human Services (2020), *HHS Policy for Preparing for and Responding to a Breach of Personally Identifiable Information (PII), Version 2.0*, [109] https://www.hhs.gov/web/governance/digital-strategy/it-policy-archive/hhs-policy-preparing-and-responding-breach.html.

U.S. Government Accountability Office (2021), *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities*. [121]

Van der Maaten, L. and G. Hinton (2008), “Visualizing data using t-SNE.”, *Journal of Machine Learning Research*, Vol. 9/11, https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf. [96]

Veer et al. (2021), “Trading Off Accuracy and Explainability in AI Decision-Making: Findings from 2 Citizens’ Juries”, *Journal of the American Medical Informatics Association.*. [128]

Wachter, S., B. Mittelstadt and C. Russell (2017), “Counterfactual explanations without opening the black box: Automated decisions and the GDPR”, *Harv. JL & Tech*, Vol. 31, p. 841. [129]

Wang, Y. et al. (2020), *Generalizing from a few examples: A survey on few-shot learning*, https://arxiv.org/abs/1904.05046. [75]

Xu, R., N. Baracaldo and J. Joshi (2021), “Privacy-Preserving Machine Learning: Methods, Challenges and Directions”, https://arxiv.org/pdf/2108.04417.pdf. [71]

Zafar, M. et al. (2019), “Fairness Constraints: A Flexible Approach for Fair Classification”, *Journal of Machine Learning Research*, Vol. 20/75, pp. 1-42, https://jmlr.org/papers/v20/18-262.html. [67]

Zemel, R. et al. (2013), *Learning fair representations*, https://proceedings.mlr.press/v28/zemel13.html. [88]

Zhang, B., B. Lemoine and M. Mitchell (2018), *Mitigating Unwanted Biases with Adversarial Learning*, https://arxiv.org/abs/1801.07593. [90]

Notes

[1] A smaller steering group composed of the co-chairs, the Secretariat and consultants met regularly between Expert Group sessions.

[2] According to ISO 31000, risk is the “effect of uncertainty on objectives” and “an effect is a deviation from the expected. It can be positive, negative or both, and can address, create or result in opportunities and threats.” This report is concerned with the negative effects of risk.

[3] “Inference” is the process of *using* an AI model – trained from data or manually encoded – to derive a prediction, recommendation or other outcome based on new data that the model was not trained on (OECD, 2022[1]).

[4] As defined in the EU AI Act proposal.

[5] This remains without prejudice to legal frameworks possibly establishing legal responsibility primarily for certain actors (e.g. providers, users). For example, legal responsibility for placing compliant systems on the market in the proposed EU AI Act is assigned to providers. Providers are also responsible for post-market monitoring.

[6] Some AI systems may be fed with data or inputs derived from other AI systems. In such cases, an assessment of the chain of AI knowledge would be necessary to identify the relevant actors or suppliers.

[7] Accuracy could also relate to other Principles, such as robustness and fairness.

[8] In this case the term "black box" is used to refer to the degree of access to information about a model. However, the term “black box” is also commonly used to refer to non-interpretable AI systems. For example, some "black box" models could be considered to be fully transparent but not interpretable (e.g. a complex but fully accessible deep learning model with billions of parameters).

[9] Hupont et al. (2022) provide an overview of documentation obligations to satisfy transparency requirements included in the proposed EU AI Act.

[10] Some research argues that the explainability vs. performance trade-off is not so relevant in cases where the objective function is explainable. A well-defined objective function (e.g. in symbolic terms) would result in better model performance (Aliman et al., 2019).

[11] The term "lightweight models" is increasingly being used to refer to models with less parameters.

[12] It could be argued that these explainability-oriented techniques are appropriate for AI experts (e.g. developers) as a test and validation tool, but are not accessible or mature enough to be used in operation by users without sufficient AI expertise or familiarity with the inner workings and design of the system.

[13] A subset of transfer learning where an algorithm trained in one source domain is applied to a different – but related – target domain (Sugiyama and Kawanabe, 2012).

[14] Access level 7 (development access) assumes that the objective function for an AI system can be learned, which may not always be the case (Wernaart, 2022).