OECD publishing

OECD FRAMEWORK FOR THE CLASSIFICATION OF AI SYSTEMS

OECD DIGITAL ECONOMY PAPERS

February 2022 **No. 323**

Federal Ministry
of Labour and Social Affairs

OECD
BETTER POLICIES FOR BETTER LIVES

OECD Digital Economy Papers

This report was approved and declassified by written procedure by the Committee on Digital Economy Policy (CDEP) on 13 January 2022 and prepared for publication by the OECD Secretariat. For more information, please visit www.oecd.ai.

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. The programme aims to help ensure that adoption of AI in the world of work is effective, beneficial to all, people-centred and accepted by the population at large. 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/.

Note to Delegations:

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

DSTI/CDEP(2020)13/FINAL

Abstract

As artificial intelligence (AI) integrates all sectors at a rapid pace, different AI systems bring different benefits and risks. In comparing virtual assistants, self-driving vehicles and video recommendations for children, it is easy to see that the benefits and risks of each are very different. Their specificities will require different approaches to policy making and governance.

To help policy makers, regulators, legislators and others characterise AI systems deployed in specific contexts, the OECD has developed a user-friendly tool to evaluate AI systems from a policy perspective. It can be applied to the widest range of AI systems across the following dimensions: People & Planet; Economic Context; Data & Input; AI model; and Task & Output.

Each of the framework's dimensions has a subset of properties and attributes to define and assess policy implications and to guide an innovative and trustworthy approach to AI as outlined in the OECD AI Principles.

Abrégé

À l'heure où l'intelligence artificielle (IA) gagne rapidement du terrain dans tous les secteurs, il apparaît clairement que différents systèmes d'IA présentent des avantages et des risques divers. Lorsque l'on compare des assistants virtuels, des véhicules autonomes et des recommandations concernant des vidéos pour les enfants, on comprend sans difficulté que les avantages et les risques de chacun varient sensiblement. Leurs spécificités exigent par conséquent des approches différenciées de l'élaboration et de la gouvernance des politiques.

C'est pourquoi l'OCDE a mis au point un outil convivial, afin d'aider les décideurs, les régulateurs, les législateurs et autres acteurs à identifier les caractéristiques des systèmes d'IA déployés dans des contextes particuliers et les évaluer du point de vue de l'action publique. L'outil peut être appliqué à l'éventail complet de systèmes d'IA dans les dimensions suivantes : les individus et la planète ; le contexte économique ; les données et entrées ; le modèle d'IA ; et les tâches et résultats.

À chaque dimension du cadre est associée un sous-ensemble de propriétés et d'attributs servant à définir et évaluer les incidences sur l'action publique et à guider l'adoption d'une approche innovante et fiable de l'IA, telle qu'exposée dans les Principes de l'OCDE sur l'IA.

Übersicht

Da die künstliche Intelligenz (KI) sich in rasantem Tempo in allen Bereichen ausbreitet, bringen verschiedene KI-Systeme unterschiedliche Vorteile und Risiken mit sich. Wenn man virtuelle Assistenten, selbstfahrende Fahrzeuge und Videoempfehlungen für Kinder vergleicht, kann man leicht feststellen, dass die Vorteile und Risiken jedes einzelnen sehr unterschiedlich sind. Ihre Besonderheiten erfordern unterschiedliche Ansätze für Politikgestaltung und Verwaltung.

Um politischen Entscheidungsträgern, Regulierungsbehörden, Gesetzgebern und anderen dabei zu helfen, KI-Systeme, die in bestimmten Kontexten eingesetzt werden, zu charakterisieren, hat die OECD ein benutzerfreundliches Instrument zur Bewertung von KI-Systemen aus politischer Sicht entwickelt. Es kann auf die unterschiedlichsten KI-Systeme in den folgenden Dimensionen angewendet werden: Menschen & Planet, Wirtschaftlicher Kontext, Daten & Eingabe, KI-Modell und Aufgabe & Ausgabe.

Jede der Dimensionen des Rahmens verfügt über eine Untergruppe von Eigenschaften und Attributen, um die politischen Auswirkungen zu definieren und zu bewerten und einen innovativen und vertrauenswürdigen Ansatz für KI gemäß den KI-Grundsätzen der OECD anzuleiten.

Executive summary

The OECD Framework for the Classification of AI Systems helps assess policy opportunities and challenges

AI changes how people learn, work, play, interact and live. As AI spreads across sectors, different types of AI systems deliver different benefits, risks and policy and regulatory challenges. Consider the differences between a virtual assistant, a self-driving vehicle and an algorithm that recommends videos for children.

The OECD developed a user-friendly framework for policy makers, regulators, legislators and others to characterise AI systems for specific projects and contexts. The framework links AI system characteristics with the OECD AI Principles (OECD, 2019), the first set of AI standards that governments pledged to incorporate into policy making and promote the innovative and trustworthy use of AI.

Ways to use the framework

The framework allows users to zoom in on specific risks that are typical of AI, such as bias, explainability and robustness, yet it is generic in nature. It facilitates nuanced and precise policy debate. The framework can also help develop policies and regulations, since AI system characteristics influence the technical and procedural measures they need for implementation. In particular, the framework provides a baseline to:

- **Promote a common understanding of AI**: Identify features of AI systems that matter most, to help governments and others tailor policies to specific AI applications and help identify or develop metrics to assess more subjective criteria (such as well-being impact).
- **Inform registries or inventories**: Help describe systems and their basic characteristics in inventories or registries of algorithms or automated decision systems.
- **Support sector-specific frameworks**: Provide the basis for more detailed application or domain-specific catalogues of criteria, in sectors such as healthcare or in finance.
- **Support risk assessment**: Provide the basis for related work to develop a risk assessment framework to help with de-risking and mitigation and to develop a common framework for reporting about AI incidents that facilitates global consistency and interoperability in incident reporting.
- **Support risk management**: Help inform related work on mitigation, compliance and enforcement along the AI system lifecycle, including as it pertains to corporate governance.

Key dimensions structure AI system characteristics and interactions

The framework classifies AI systems and applications along the following dimensions: People & Planet, Economic Context, Data & Input, AI Model and Task & Output. Each one has its own properties and attributes or sub-dimensions relevant to assessing policy considerations of particular AI systems.

- **People & Planet**: This considers the potential of applied AI systems to promote human-centric, trustworthy AI that benefits people and planet. In each context, it identifies individuals and groups that interact with or are affected by an applied AI system. Core characteristics include users and impacted

stakeholders, as well as the application's optionality and how it impacts human rights, the environment, well-being, society and the world of work.

- **Economic Context:** This describes the economic and sectoral environment in which an applied AI system is implemented. It usually pertains to an applied AI application rather than to a generic AI system, and describes the type of organisation and functional area for which an AI system is developed. Characteristics include the sector in which the system is deployed (e.g. healthcare, finance, manufacturing), its business function and model; its critical (or non-critical) nature; its deployment, impact and scale, and its technological maturity.
- **Data & Input:** This describes the data and/or expert input with which an AI model builds a representation of the environment. Characteristics include the provenance of data and inputs, machine and/or human collection method, data structure and format, and data properties. Data & Input characteristics can pertain to data used to train an AI system ("in the lab") and data used in production ("in the field").
- **AI Model:** This is a computational representation of all or part of the external environment of an AI system – encompassing, for example, processes, objects, ideas, people and/or interactions that take place in that environment. Core characteristics include technical type, how the model is built (using expert knowledge, machine learning or both) and how the model is used (for what objectives and using what performance measures).
- **Task & Output**: This refers to the tasks the system performs, e.g. personalisation, recognition, forecasting or goal-driven optimisation; its outputs; and the resulting action(s) that influence the overall context. Characteristics of this dimension include system task(s); action autonomy; systems that combine tasks and actions like autonomous vehicles; core application areas like computer vision; and evaluation methods.

Applicability of the framework to AI "in the lab" versus "in the field"

Some criteria of the framework are more applicable to AI "in the field" contexts than AI "in the lab" contexts, and vice versa. AI "in the lab" refers to the AI system's conception and development, before deployment. It is applicable to the Data & Input (e.g. qualifying the data), AI Model (e.g. training the initial model) and Task & Output dimensions (e.g. for a personalisation task) of the framework. It is particularly relevant to *ex ante* risk-management approaches and requirements. AI "in the field" refers to the use and evolution of an AI system after deployment and is applicable to all the dimensions. It is relevant to *ex post* risk-management approaches and requirements.

Classification and the AI system lifecycle

The AI system lifecycle can serve as a complementary structure for understanding the key technical characteristics of a system. The lifecycle encompasses the following phases that are not necessarily sequential: planning and design; collecting and processing data; building and using the model; verifying and validating; deployment; and operating and monitoring (OECD, 2019d[1]). The dimensions of the OECD Framework for the Classification of AI Systems can be associated with stages of the AI system lifecycle to identify a dimension's relevant AI actors, which is relevant to accountability.

Résumé

Le Cadre de l'OCDE pour la classification des systèmes d'IA aide à évaluer les opportunités et les défis du point de vue de l'action des pouvoirs publics

L'IA modifie la façon dont les individus apprennent, travaillent, jouent, interagissent et vivent. À l'heure où l'IA gagne du terrain dans tous les secteurs, il apparaît clairement que différents types de systèmes d'IA présentent des avantages, des risques et des défis politiques et réglementaires divers. Les caractéristiques distinctives des assistants virtuels, des véhicules autonomes et des algorithmes recommandant des vidéos pour les enfants en sont une bonne illustration.

L'OCDE a mis au point un cadre convivial que les décideurs, les régulateurs, les législateurs et autres acteurs peuvent utiliser pour identifier les caractéristiques des systèmes d'IA pour des projets et dans des contextes particuliers. Le cadre relie les caractéristiques des systèmes d'IA aux Principes de l'OCDE sur l'IA (OCDE, 2019), premier ensemble de normes relatives à l'IA que les pouvoirs publics se sont engagés à prendre en compte lors de l'élaboration des politiques et qui promeuvent une utilisation innovante et digne de confiance de l'IA.

Usages du cadre

Bien que générique par nature, le cadre permet aux utilisateurs d'examiner en détail des risques propres à l'IA (par exemple en termes de biais, d'explicabilité et de robustesse). Il favorise un débat de fond nuancé et précis. Il peut également aider à l'élaboration de politiques et de réglementations, puisque les caractéristiques des systèmes d'IA influent sur les mesures techniques et procédurales prises pour les mettre en œuvre. Le cadre vise en particulier à fournir une base de référence pour:

- **Favoriser une compréhension commune de l'IA** : identifier les caractéristiques les plus importantes des systèmes d'IA, afin d'aider les pouvoirs publics et autres acteurs à adapter les politiques à des applications spécifiques de l'IA et de permettre l'identification ou la mise au point d'indicateurs pour évaluer des aspects plus subjectifs (tels que l'impact sur le bien-être).
- **Aider à la constitution de registres ou d'inventaires** : aider à décrire les systèmes et leurs caractéristiques fondamentales dans les inventaires ou registres d'algorithmes, ou systèmes de prise de décision automatisée.
- **Étayer les cadres sectoriels** : fournir le socle pour la constitution de catalogues de critères plus détaillés propres à des applications ou des domaines, tels la santé ou la finance.
- **Aider à l'évaluation des risques** : fournir le socle pour des travaux connexes visant à concevoir un cadre d'évaluation des risques pour éliminer ou limiter les risques, et à mettre au point un cadre commun pour le signalement des incidents liés à l'IA, dans l'optique de favoriser la cohérence et l'interopérabilité des signalements à l'échelle mondiale.
- **Aider à la gestion des risques** : apporter de la matière aux travaux connexes sur la réduction des risques, la conformité et le contrôle tout au long du cycle de vie des systèmes d'IA, notamment pour ce qui est de la gouvernance des entreprises.

Cinq dimensions structurent les caractéristiques des systèmes d'IA et leurs interactions

Le cadre classe les systèmes et les applications d'IA selon cinq dimensions : les individus et la planète, le contexte économique, les données et entrées, le modèle d'IA, et les tâches et résultats. Chacune d'elles possède ses propres propriétés et attributs, ou sous-dimensions, qui servent à l'évaluation des incidences des différents systèmes d'IA sur l'action des pouvoirs publics.

- **Individus et planète** : évalue la capacité des applications d'IA de promouvoir une IA centrée sur l'humain, digne de confiance, servant les intérêts des individus et de la planète. Dans chaque contexte, on identifie les individus et les groupes qui interagissent avec une application d'IA ou sont concernés par son utilisation. Les caractéristiques principales sont les utilisateurs et les parties prenantes concernées, le caractère facultatif de l'application et son impact sur les droits humains, l'environnement, le bien-être, la société et le monde du travail.
- **Contexte économique** : décrit l'environnement économique et sectoriel dans lequel un système d'IA appliquée est mis en œuvre. La dimension se rapporte généralement à une application d'IA particulière plutôt qu'à un système d'IA générique, et décrit le type d'organisation et de domaine fonctionnel pour lequel un système d'IA est mis au point. Les caractéristiques sont le secteur dans lequel le système est déployé (santé, finance, industrie manufacturière, par exemple), sa fonction et son modèle économique ; son caractère critique (ou non critique) ; son déploiement, son impact et son échelle, et sa maturité technologique.
- **Données et entrées** : décrit les données et/ou les entrées spécialisées à partir desquelles un modèle d'IA bâtit une représentation de l'environnement. Les caractéristiques ont trait à la provenance des données et des entrées, à la méthode de collecte (automatisée et/ou humaine), à la structure et au format des données, et aux propriétés des données. Elles peuvent se rapporter à des données utilisées pour entraîner un système d'IA (« en laboratoire ») ou à des données utilisées en production (« sur le terrain »).
- **Modèle d'IA** : représentation informatique de tout ou partie de l'environnement externe d'un système d'IA – comprenant par exemple les processus, objets, idées, personnes et/ou interactions propres à cet environnement. Les principales caractéristiques sont le type technique, la façon dont le modèle est construit (à partir de connaissances d'experts, par apprentissage automatique ou les deux) et la façon dont il est utilisé (à quelles fins et avec quelles mesures de performance).
- **Tâches et résultats** : désigne les tâches exécutées par le système (personnalisation, reconnaissance, prévision ou optimisation basée sur des objectifs, par exemple) ; les résultats produits ; et la ou les action(s) qui en découlent et influent sur le contexte général. Les principales caractéristiques de cette dimension sont la ou les tâche(s) du système ; l'autonomie d'action ; les systèmes combinant des tâches et des actions, à l'instar des véhicules autonomes ; les principaux domaines d'application, comme la vision par ordinateur ; et les méthodes d'évaluation.

Applicabilité du cadre à l'IA « en laboratoire » ou « sur le terrain »

Certains critères du cadre s'appliquent davantage à l'IA « en laboratoire », d'autres à l'IA « sur le terrain ». L'IA « en laboratoire » désigne la conception et le développement des systèmes d'IA avant la phase de déploiement. On la retrouve dans les dimensions Données et entrées (qualification des données, par exemple), Modèle d'IA (entraînement du modèle initial) et Tâches et résultats (tâche de personnalisation, par exemple) du cadre. Elle est particulièrement pertinente pour les approches et exigences de gestion du risque *ex ante*. L'IA « sur le terrain » désigne l'utilisation et l'évolution d'un système d'IA après son déploiement. On la retrouve dans toutes les dimensions. Elle est particulièrement pertinente pour les approches et exigences de gestion du risque *ex post*.

Lien entre la classification et le cycle de vie d'un système d'IA

Le cycle de vie d'un système d'IA peut servir de structure complémentaire pour appréhender les principales caractéristiques techniques d'un système. Il comprend les phases suivantes, qui ne suivent pas nécessairement un ordre séquentiel : planification et conception ; collecte et traitement des données ; construction et utilisation du modèle ; vérification et validation ; déploiement ; et exploitation et suivi (OECD, 2019d[1]). Les dimensions du Cadre de l'OCDE pour la classification des systèmes d'IA peuvent être associées à différents stades du cycle de vie des systèmes afin d'identifier les acteurs de l'IA concernés par une dimension, ce qui va dans le sens du principe de responsabilité.

Kurzfassung

Der OECD-Rahmen für die Klassifizierung von KI-Systemen hilft bei der Bewertung politischer Möglichkeiten und Herausforderungen

KI verändert, wie Menschen lernen, arbeiten, spielen, interagieren und leben. Da sich KI in allen Sektoren ausbreitet, bieten verschiedene Arten von KI-Systemen unterschiedliche Vorteile, Risiken und politische und regulatorische Herausforderungen. Denken Sie an die Unterschiede zwischen einem virtuellen Assistenten, einem selbstfahrenden Fahrzeug und einem Algorithmus, der Videos für Kinder empfiehlt.

Die OECD hat einen benutzerfreundlichen Rahmen für politische Entscheidungsträger, Regulierungsbehörden, Gesetzgeber und andere entwickelt, um KI-Systeme für bestimmte Projekte und Situationen zu charakterisieren. Der Rahmen verbindet die Merkmale von KI-Systemen mit den KI-Prinzipien der OECD (OECD, 2019), dem ersten Satz von KI-Standards, zu deren Einbeziehung in die Politik sich die Regierungen verpflichtet haben und die die innovative und vertrauenswürdige Nutzung von KI fördern.

Verwendung des Rahmens und nächste Schritte

Der Rahmen ermöglicht es den Nutzern, sich auf spezifische, für KI typische Risiken wie Verzerrungen, Erklärbarkeit und Robustheit zu konzentrieren, ist aber dennoch allgemeiner Natur. Er erleichtert eine differenzierte und präzise politische Debatte. Der Rahmen kann auch bei der Entwicklung von Strategien und Vorschriften helfen, da die Merkmale von KI-Systemen die technischen und verfahrenstechnischen Maßnahmen zu ihrer Umsetzung beeinflussen. Der Rahmen dient insbesondere dazu:

- **Ein gemeinsames Verständnis von KI zu fördern**: Identifizierung der wichtigsten Merkmale von KI-Systemen, um Regierungen und andere Akteure dabei zu unterstützen, ihre Politik auf spezifische KI-Anwendungen zuzuschneiden und Metriken zur Bewertung subjektiverer Kriterien (z. B. Auswirkungen auf das Wohlbefinden) zu ermitteln oder zu entwickeln.
- **Register oder Verzeichnisse zu informieren**: Hilfe bei der Beschreibung von Systemen und ihren grundlegenden Merkmalen in Verzeichnissen oder Registern von Algorithmen oder automatisierten Entscheidungssystemen.
- **Sektorspezifische Rahmen zu unterstützen**: Die Grundlage für detailliertere anwendungs- oder domänenspezifische Kriterienkataloge bilden, z. B. im Gesundheitswesen oder im Finanzwesen.
- **Die Risikobewertung zu unterstützen**: Schaffung einer Grundlage für die damit verbundenen Arbeiten zur Entwicklung eines Risikobewertungsrahmens, der bei der Risikominderung und -begrenzung hilft, und zur Entwicklung eines gemeinsamen Rahmens für die Berichterstattung über KI-Vorfälle, der die globale Kohärenz und Kompatibilität bei der Berichterstattung über Vorfälle erleichtert.
- **Das Risikomanagement zu unterstützen**: Beitrag zur Information über die damit zusammenhängenden Arbeiten zur Eindämmung, Einhaltung und Durchsetzung von KI-Systemen im gesamten Lebenszyklus, auch im Hinblick auf die Unternehmensführung.

Fünf Dimensionen strukturieren die Merkmale und Interaktionen von KI-Systemen

Der Rahmen klassifiziert KI-Systeme und -Anwendungen anhand von fünf Dimensionen: Menschen & Planet, Wirtschaftlicher Kontext, Daten & Eingabe, KI-Modell und Aufgabe & Ausgabe. Jede hat ihre eigenen Eigenschaften und Attribute oder Unterdimensionen, die für die Bewertung politischer Überlegungen zu bestimmten KI-Systemen relevant sind.

- **Menschen & Planet**: betrachtet das Potenzial von KI-Anwendungen zur Förderung einer menschenzentrierten, vertrauenswürdigen KI, die den Menschen und dem Planeten zugute kommt. In jedem Kontext werden Personen und Gruppen identifiziert, die mit einer KI-Anwendung interagieren oder

von ihr beeinflusst werden. Zu den wichtigsten Merkmalen gehören die Benutzer und die betroffenen Interessengruppen sowie die Wahlfreiheit der Anwendung und ihre Auswirkungen auf die Menschenrechte, die Umwelt, das Wohlergehen, die Gesellschaft und die Arbeitswelt.

- **Wirtschaftlicher Kontext:** beschreibt das wirtschaftliche und sektorale Umfeld, in dem ein angewandtes KI-System eingesetzt wird. Er bezieht sich in der Regel eher auf eine spezifische KI-Anwendung als auf ein allgemeines KI-System und beschreibt die Art der Organisation und den Funktionsbereich, für den ein KI-System entwickelt wird. Zu den Merkmalen gehören der Bereich, in dem das System eingesetzt wird (z. B. Gesundheitswesen, Finanzwesen, Fertigung), seine Geschäftsfunktion und sein Geschäftsmodell, sein kritischer (oder nicht kritischer) Charakter, sein Einsatz, seine Auswirkungen und sein Umfang sowie seine technologische Reife.
- **Daten & Eingabe:** beschrei die Daten und/oder den Experteninput, mit denen ein KI-Modell eine Darstellung der Umgebung erstellt. Zu den Merkmalen gehören die Herkunft der Daten und Eingaben, die maschinelle und/oder menschliche Erfassungsmethode, die Datenstruktur und das Format sowie die Dateneigenschaften. Daten- und Eingabemerkmale können sich auf Daten beziehen, die zum Trainieren eines KI-Systems („im Labor“) verwendet werden, und auf Daten, die in der Produktion („im Feld“) verwendet werden.
- **KI-Modell:** ist eine computergestützte Darstellung der gesamten oder eines Teils der externen Umgebung eines KI-Systems, die z. B. Prozesse, Objekte, Ideen, Menschen und/oder Interaktionen umfasst, die in dieser Umgebung stattfinden. Zu den Hauptmerkmalen gehören der technische Typ, die Art und Weise, wie das Modell erstellt wird (mit Hilfe von Expertenwissen, maschinellem Lernen oder beidem) und wie das Modell verwendet wird (für welche Ziele und mit welchen Leistungskennzahlen).
- **Aufgabe und Ausgabe**: bezieht sich auf die Aufgaben, die das System ausführt, z. B. Personalisierung, Erkennung, Vorhersage oder zielgerichtete Optimierung, seine Ausgabe und die daraus resultierende(n) Aktion(en), die den Gesamtkontext beeinflussen. Zu den Hauptmerkmalen dieser Dimension gehören die Systemaufgabe(n), Handlungsautonomie, Systeme, die Aufgaben und Handlungen kombinieren, wie z. B. autonome Fahrzeuge, Kernanwendungsbereiche, wie z. B. Computer Vision, und Bewertungsmethoden.

Anwendbarkeit des Rahmens auf KI „im Labor“ gegenüber KI „im Feld“

Einige Kriterien des Rahmens sind eher auf KI-Kontexte „im Feld“ anwendbar als auf KI-Kontexte „im Labor“ und umgekehrt. KI „im Labor“ bezieht sich auf die Konzeption und Entwicklung des KI-Systems vor dem Einsatz. Sie ist anwendbar auf die Dimensionen Daten & Eingabe (z.B. Qualifizierung der Daten), KI-Modell (z.B. Training des Ausgangsmodells) und Aufgabe & Output (z.B. für eine Personalisierungsaufgabe) des Rahmens. Sie ist besonders relevant für *Ex-ante*-Risikomanagementkonzepte und -anforderungen. KI „im Feld“ bezieht sich auf die Nutzung und Weiterentwicklung eines KI-Systems nach der Einführung. Sie ist auf alle Dimensionen anwendbar, einschließlich der Dimensionen Menschen & Planet und Wirtschaftlicher Kontext. Sie ist relevant für *Ex-post*-Risikomanagementkonzepte und -anforderungen.

Verbindung zwischen der Klassifizierung und dem Lebenszyklus des KI-Systems

Der Lebenszyklus eines KI-Systems kann als ergänzende Struktur für das Verständnis der wichtigsten technischen Merkmale eines Systems dienen. Der Lebenszyklus umfasst die folgenden Phasen, die nicht notwendigerweise aufeinander folgen: Planung und Entwurf, Sammlung und Verarbeitung von Daten, Aufbau und Nutzung des Modells, Verifizierung und Validierung, Einsatz sowie Betrieb und Überwachung (OECD, 2019d[1]). Die Dimensionen des OECD-Rahmens für die Klassifizierung von KI-Systemen können mit verschiedenen Phasen des Lebenszyklus von KI-Systemen in Verbindung gebracht werden, um die für eine Dimension relevanten KI-Akteure zu ermitteln, was somit dem Grundsatz der Rechenschaftspflicht entspricht.

Acknowledgements

The OECD Framework for the Classification of AI Systems is based on the work of the OECD Network of Experts Working Group on AI Classification & Risk and the OECD Secretariat in 2020 and 2021 and was prepared under the aegis of the OECD Committee for Digital Economy Policy (CDEP). The expert group was co-chaired by Marko Grobelnik, Slovenian Jozef Stefan Institute (JSI); Dewey Murdick, Center for Security and Emerging Technology (CSET) at Georgetown University; and Jack Clark, Anthropic. Karine Perset, OECD Digital Economy Policy Division, led the report development and drafting. Key sections of the publication were researched and drafted by Luis Aranda, Louise Hatem and Nobuhisa Nishigata from the OECD Digital Economy Policy Division; and by Catherine Aiken, CSET; Peter Cihon, GitHub; Sebastian Hallensleben, VDE Association for Electrical, Electronic & Information Technologies; Tim Rudner, University of Oxford; and by the expert group co-chairs Marko Grobelnik, Dewey Murdick, and Jack Clark. Audrey Plonk, Andrew Wyckoff and Sarah Box, OECD, provided oversight.

Over 60 experts participated in the Experts Working Group, which held one physical and regular virtual meetings between February 2020 and December 2022.[1] The framework benefited significantly from the contributions of Peter Addo, Agence Française de Développement; Tatjana Evas, Killian Gross, Juha Heikkila and Irina Orssich, European Commission; Nozha Boujemaa, IKEA; Lord Tim Clement-Jones, United Kingdom House of Lords; Olivia Erdelyi, University of Canterbury School of Law and Soul Machines; Irene Ek, Swedish Agency for Growth Policy Analysis; Theodoros Evgeniou, INSEAD; Jonathan Frankle, Massachusetts Institute of Technology (MIT); Emilia Gómez, DG Joint Research Centre (JRC), European Commission; Yoichi Iida and Yuki Hirano, Japan Ministry of Internal Affairs and Communications (MIC); Katrina Kosa-Ammari, Latvian Ministry of Foreign Affairs; Raj Madhavan, Unites States Department of State; Dunja Mladenić, Slovenian Jožef Stefan Institute; Michel Morvan, Cosmo Tech; Clara Neppel, Institute of Electrical and Electronics Engineers (IEEE) European Business Operations; Igor Perisic, LinkedIn; Sally Radwan, Egyptian Ministry of Communications & Information Technology; Roberto Sanchez, Inter-American Development Bank (IDB); Daniel Schwabe, Catholic University in Rio de Janeiro (PUC-Rio); Olly Salzmann, Deloitte KI GmbH and KIParkDeloitte GmbH; Viknesh Sounderajah, Imperial College London; Elham Tabassi, United States National Institute of Standards and Technology (NIST); and all the members of the Experts Working Group (see Annex C).

The OECD and the Expert Group also wish to acknowledge the contributions made in June 2021 by individuals during the public consultation phase on the preliminary version of the framework. Significant revisions and improvements were made to the report based on the feedback received during those consultations, and the authors sincerely thank all who participated. While the Secretariat is unable to list all those who participated, particularly noteworthy written feedback was provided by Patrick Drake-Brockman and Alana Walsch, Government of Australia; Leandro Vocholko, Government of Brazil; Benoit Deshaies, Government of Canada; the Governments of Chile and Columbia; the national administrations of Brazil, Portugal and Singapore; and Clara Standring at the United Kingdom's Office of National Statistics (ONS). Significant contributions were also made by Emilia Gómez, Isabelle Hupont, Fernando Martínez-Plumed and Songül Tolan, European Commission's Joint Research Centre (JRC); Kristen Little, IEEE; and Leonidas Aristodemou, OECD.

The Experts Working Group also thanks Sean McGregor, Partnership on AI (PAI); Sawyer Bernath, Berkeley Existential Risk Initiative; Patricia Shaw, Beyond Reach Consulting; Nicolas Blanc, Confédération

française de l'encadrement - Confédération générale des cadres (CFE-CGC); Benjamin Larsen, Copenhagen Business School; Naira Hambardzumyan, Deloitte; Pr. Alba Enrique; Lucas Dalmedico Gessoni, Eldorado Research Institute; Ansgar Koene, Ernst and Young (EY); Claire Boine and Richard Mallah, Future of Life Institute (FLI); Samuel Curtis, Nicolas Miailhe and Nicolas Moës, The Future Society; Maria Gonsalves; Isaac Robinson, Harvard University; Alan Chan, Institut Québecois d'Intelligence Artificielle (MILA); Axel Gruvaeus, Kairos; Elena Simperl, Kings College London; Peter Damn, KMD Denmark; David Ellison, Lenovo; Mike Sparling, Multi-Health Systems Inc.; Michaela Regneri, Otto in Germany; Koen Cobbaert, Philips Belgium; Evgenia Vasin, Sberbank Russia; Stuart Elliot, United States National Academy of Sciences (NAS); Ingo Elsen, Aachen University of Applied Sciences Germany; Marcela Distefano, Universidad Argentina de la Empresa; Gehrard Weiss, University of Maastricht; Alessandro Saffiotti, University of Sweden; Narayanan Sundaraparipurnan; Heather Von Stackelberg; Pam Dixon, World Privacy Forum; and Vladimir Sadilovski, YesaulTUbe. The group also thanks the American Property Casualty Insurance Association (APCIA); Association of Test Publishers (ATP); Centre for AI and Digital Policy; Consumer Technology Association (CTA); Data Privacy Brasil Research Association; European Centre for Not-for-Profit Law (ECNL); Electronic Privacy Information Centre (EPIC); French Insurance Federation (FFA); Instituto Atlantico; Software Alliance (BSA); and United Nations Interregional Crime and Justice Research Institute (UNICRI).

The OECD gratefully acknowledges the contributions made by individuals and institutions that took the time to test the framework through the online survey (see Table 7).

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

Table of contents

FIGURES

TABLES

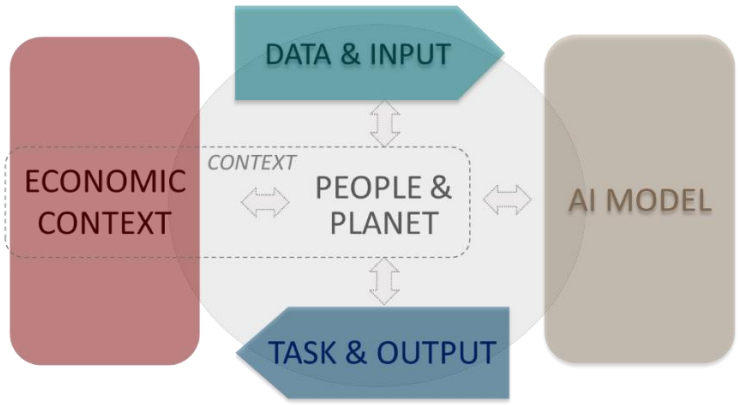
1 Overview and goal of the framework

Different types of AI systems raise different policy opportunities and challenges. Section 2 of this report introduces and describes a framework to assess AI systems' impact on public policy in areas covered by the OECD AI Principles (OECD, 2019d[1]). Section 3 puts the framework into use to classify specific AI systems and applications. Section 4 discusses how the framework could be used to help assess basic social, physical and ethical risks associated with specific types of AI systems.

Introducing the framework and its purpose

The framework primary purpose is to characterise the application of an AI system deployed in a specific project and context, although some dimensions are also relevant to generic AI systems. It classifies AI systems and applications along the following dimensions: People & Planet, Economic Context, Data & Input, AI Model and Task & Output (Figure 1). These dimensions build on the conceptual view of a generic AI system established in previous OECD work (see Box 1 later in this section).

Figure 1. Key high-level dimensions of the OECD Framework for the Classification of AI Systems



The AI Principles as a lens for analysing policy considerations

Each of the framework's dimensions has distinct properties and attributes, or sub-dimensions that are relevant to assessing policy considerations associated with a particular AI system. The 10 OECD AI Principles, adopted in 2019, help structure the analysis of policy considerations associated with each dimension and sub-dimension. The Principles cover the following themes:

Table 1. The OECD AI Principles

Values-based principles for all AI actors	*Recommendations to policy makers for AI policies*
Principle 1.1. People and planet	*Principle 2.1.* Investment in R&D
Principle 1.2. Human rights, privacy, fairness	*Principle 2.2.* Data, compute, technologies
Principle 1.3. Transparency, explainability	*Principle 2.3.* Enabling policy and regulatory environment
Principle 1.4. Robustness, security, safety	*Principle 2.4.* Jobs, automation, skills
Principle 1.5. Accountability	*Principle 2.5.* International cooperation

Source: (OECD, 2019d[1])

Balancing user-friendliness and accuracy

The framework has been designed to be user-friendly (Table 2). It balances simplicity and useful explanations. It includes the basic criteria for which information is likely to be available and that are essential to obtaining relevant results from using the framework. Additional criteria that provide key information on the AI system in question, but for which it has been and may continue to be challenging to find sufficient objective and consistent information are marked as {where objective and consistent information is available}.[2]

To date, the framework does *not* address governance at the corporate, institution or AI systems level; nor does it cover the use of mitigation measures or compliance and enforcement measures along the AI system lifecycle (the subject of a related stream of work).

Uses for the framework

The framework allows users to zoom in on specific risks that are typical of AI, such as bias, explainability and robustness, yet it is generic in nature. It facilitates nuanced and precise policy debate. The framework can also help develop policies and regulations, since AI system characteristics influence the technical and procedural measures they need for implementation. In particular, the framework serves to:

- Promote a common understanding of AI and its most important characteristics among a variety of stakeholders so they can tailor policies for specific types of AI systems.
- Describe AI systems and their basic characteristics in algorithm inventories, or registries of automated decision systems, which are being built in several jurisdictions.
- Provide the basis for more detailed application or domain-specific catalogues of criteria, e.g. in healthcare, finance or industry. For example, the UK Medicines and Healthcare Products Regulatory Agency (MHRA) and the UK National Institute of Health and Care Excellence (NICE) Health Technology Assessment (HTA) programme are using and adapting the classification framework for AI systems to assist in triaging technologies for health technology assessment.[3]
- Provide the basis for a weighed risk-assessment tool that can help measures for mitigation and minimising risk (see Section 4 Next Steps).
- Inform related work on mitigation, compliance and enforcement along AI systems' lifecycles.

Key elements of the framework

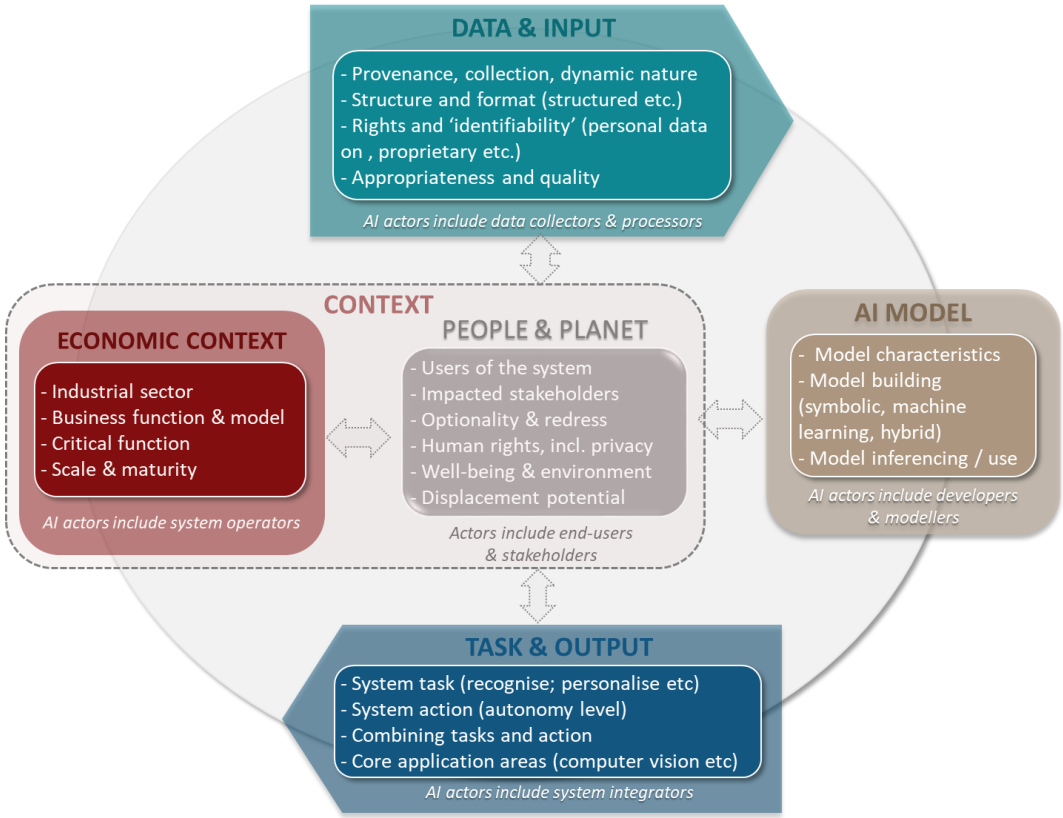
Each of the framework's dimensions has distinct properties and attributes, or sub-dimensions that are relevant to assessing policy considerations associated with different AI systems (Figure 2). Stakeholders include anyone involved in or affected by AI systems. AI actors are stakeholders who play active roles throughout the AI system lifecycle and can vary according to each dimension (OECD, 2019d[1])

Table 2. Classification framework dimensions and criteria at a glance

PEOPLE & PLANET	Criteria	Description
USERS	Users of AI system	What is the level of competency of users who interact with the system?
STAKEHOLDERS	Impacted stakeholders	Who is impacted by the system (e.g. consumers, workers, government agencies)?
OPTIONALITY	Optionality and redress	Can users opt out, e.g. switch systems? Can users challenge or correct the output?
HUMAN RIGHTS	Human rights and democratic values	Can the system's outputs impact fundamental human rights (e.g. human dignity, privacy, freedom of expression, non-discrimination, fair trial, remedy, safety)?
WELL-BEING & ENVIRONMENT	Well-being, society and the environment	Can the system's outputs impact areas of life related to well-being (e.g. job quality, the environment, health, social interactions, civic engagement, education)?
DISPLACEMENT	*{Displacement potential}*	*Could the system automate tasks that are or were being executed by humans?*
ECONOMIC CONTEXT	**Criteria**	**Description**
SECTOR	Industrial sector	Which industrial sector is the system deployed in (e.g. finance, agriculture)?
BUSINESS FUNCTION & MODEL	Business function	What business function(s) is the system employed in (e.g. sales, customer service)?
	Business model	Is the system a for-profit use, non-profit use or public service system?
CRITICALITY	Impacts critical functions / activities	Would a disruption of the system's function / activity affect essential services?
SCALE & MATURITY	Breadth of deployment	Is the AI system deployment a pilot, narrow, broad or widespread?
	{Technical maturity}	*How technically mature is the system (Technology Readiness Level –TRL)*
DATA & INPUT	**Criteria**	**Description**
COLLECTION	Detection and collection	Are the data and input collected by humans, automated sensors or both?
	Provenance of data and input	Are the data and input from experts; provided, observed, synthetic or derived?
	Dynamic nature	Are the data dynamic, static, dynamic updated from time to time or real-time?
RIGHTS & IDENTIFIABILITY	Rights	Are the data proprietary, public or personal data (related to identifiable individual)?
	"Identifiability" of personal data	If personal data, are they anonymised; pseudonymised?
STRUCTURE & FORMAT	*{Structure of data and input}*	*Are the data structured, semi-structured, complex structured or unstructured?*
	{Format of data and metadata}	*Is the format of the data and metadata standardised or non-standardised?*
SCALE	*{Scale}*	*What is the dataset's scale?*
QUALITY AND APPROPRIATENESS	*{Data quality and appropriateness}*	*Is the dataset fit for purpose? Is the sample size adequate? Is it representative and complete enough? How noisy are the data?*
AI MODEL	**Criteria**	**Description**
MODEL CHARACTERISTICS	Model information availability	Is any information available about the system's model?
	AI model type	Is the model symbolic (human-generated rules), statistical (uses data) or hybrid?
	{Rights associated with model}	*Is the model open-source or proprietary, self or third-party managed?*
	{Discriminative or generative}	*Is the model generative, discriminative or both?*
	{Single or multiple model(s)}	*Is the system composed of one model or several interlinked models?*
MODEL-BUILDING	Model-building from machine or human knowledge	Does the system learn based on human-written rules, from data, through supervised learning, through reinforcement learning?
	Model evolution in the field ML	Does the model evolve and / or acquire abilities from interacting with data in the field?
	Central or federated learning ML	Is the model trained centrally or in a number of local servers or "edge" devices?
MODEL INFERENCE	*{Model development / maintenance}*	*Is the model universal, customisable or tailored to the AI actor's data?*
	{Deterministic and probabilistic}	*Is the model used in a deterministic or probabilistic manner?*
	Transparency and explainability	If information available to users to allow them to understand model outputs?
TASK & OUTPUT	**Criteria**	**Description**
TASKS	Task(s) of the system	What tasks does the system perform (e.g. recognition, event detection, forecasting)?
	{Combining tasks and actions into composite systems}	*Does the system combine several tasks and actions (e.g. content generation systems, autonomous systems, control systems)?*
ACTION	Action autonomy	How autonomous are the system's actions and what role do humans play?
APPLICATION AREA	Core application area(s)	Does the system belong to a core application area such as human language technologies, computer vision, automation and / or optimisation or robotics?
EVALUATION	*{Evaluation methods}*	*Are standards or methods available for evaluating system output?*

Note: Criteria and descriptions in grey and marked with an {} symbol = those where objective and consistent information is available. ML = for machine learning AI models.

Figure 2. Characteristics per classification dimension and key actor(s) involved



Note: Actors are illustrative, non-exhaustive and notably relevant to accountability.
Source: Based on the work of ONE AI and the AI system lifecycle work of AIGO (OECD, 2019f[2]).

People & Planet

People & Planet are at the centre of the framework. This dimension considers the potential of AI actors to develop applied AI systems that promote human-centric, trustworthy AI that benefits people and planet. It identifies individuals and groups that interact with or are affected by an applied AI system in a specific context. Core characteristics include users and impacted stakeholders, as well as the application's optionality and how it impacts human rights, the environment, well-being, society and the world of work.

This dimension is important for public policy because, for example, AI user competency varies, which matters for accountability, transparency and explainability. AI systems that impact specific stakeholder groups such as consumers raise consumer protection and product safety considerations. Users and stakeholders often have different degrees of choice in whether to be subject to the effects of an AI system, or varying ability to opt-out of or reverse a system's output. More broadly, accountability and transparency are critical in contexts where the outcomes of an AI system can impact human rights, such as in criminal sentencing or determining educational opportunities.

Actors in this dimension include end-users and stakeholders who use or are impacted by AI systems. Stakeholders encompass all organisations and individuals involved in or affected by AI systems, directly or indirectly.

Economic Context

Economic Context refers to the economic and sectoral environment in which an AI system is implemented. It is usually related to an applied AI system rather than to a generic one, and describes the type of organisation and functional area for which an AI system is developed. Characteristics include the sector in which the system is deployed (e.g. healthcare, finance, manufacturing), its business function and model; if its nature is critical or non-critical; its deployment, impact and scale, and its technological maturity.

This dimension is important for public policy because AI raises sector-specific considerations. These include patient data privacy in healthcare, safety considerations in transportation, transparency and accountability in public services (particularly in areas like security and law enforcement), and security and robustness considerations in critical functions like energy infrastructure. Other characteristics such as AI system maturity are particularly relevant to accountability, R&D investment, safety, robustness and security.

AI actors in this dimension include system operators who plan and design and operate and monitor AI systems.

Data & Input

Data & Input refers to the data and/or expert input with which an AI model builds a representation of the context or environment (both the Economic Context and People & Planet context). Expert input is typically human knowledge that is codified into rules. Characteristics include the provenance of data and inputs, machine and/or human collection method, data structure and format, and data properties. Data & Input characteristics can pertain to data used to train an AI system ("in the lab"; see next section for additional explanation) and data used in production ("in the field"; see next section).

One of the key reasons this dimension is important for public policy is because, for one, systems that process personal or sensitive data generate concerns about privacy, inclusiveness, human rights and bias/fairness. Data that is generated synthetically for scenario simulation (e.g. a car accident) or to supplement non-representative data sets are relevant for safety and inclusiveness. The degree to which data are static or dynamic is relevant for accountability, particularly when assessing AI systems that iterate and evolve over time, or change their behaviour in unforeseen ways. Whether data are proprietary, public or personal is relevant to transparency and explainability; bias; scale-up; economic, social and environmental impacts; research; data availability and computational capacity.

AI actors in this dimension include data collectors and processors who collect and process data, by gathering and cleaning data, labelling, performing checks for completeness and quality, and documenting the characteristics of the dataset.

AI Model

An AI model is a computational representation of all or part of the external environment of an AI system – encompassing, for example, processes, objects, ideas, people and/or interactions that take place in that environment. AI models use data and/or expert knowledge provided by humans and/or automated tools to represent, describe and interact with real or virtual environments. Core characteristics include technical type, how the model is built (using expert knowledge, machine learning or both) and how the model is used (for what objectives and using what performance measures).

The AI Model dimension is important for public policy because key properties of AI models – degree of transparency and/or explainability, robustness, and implications for human rights, privacy and fairness – depend on the type of model as well as the model-building and inferencing processes. For example, systems using neural networks are often seen as having the potential to provide comparatively higher accuracy but less explainability than other types. Explainability is often tied to system complexity; the more

complex a model is, the harder it is to explain. The degree to which a model evolves in response to data is relevant to public policy and consumer protection regimes, especially for AI systems that can learn from iterations and evolve over time. Understanding how a model was developed and/or maintained is another key consideration for assigning roles and responsibilities throughout risk management processes.

AI actors in this dimension include developers and modellers who build and use models and verify and validate them.

Task & Output

The Task & Output dimension refers to the tasks the system performs, e.g. personalisation, recognition, forecasting or goal-driven optimisation; its outputs; and the resulting action(s) that influence the overall context. Core characteristics of this dimension include system task(s); action autonomy; systems that combine tasks and actions like autonomous vehicles; core application areas like computer vision; and evaluation methods.

This dimension is important for public policy because personalisation tasks, for example, generate outputs that could raise bias and fairness issues. Recognition tasks can raise concerns of human rights, robustness and security as well as bias. Moreover, the actions taken based on the outcomes of an AI system, e.g. by autonomous vehicles, generate issues of fairness, safety, security and accountability. More broadly, the level of autonomy in an AI system's actions and the role of humans also raise important questions for human rights and fundamental values, among other concerns.

AI actors in this dimension include system integrators who deploy AI systems.

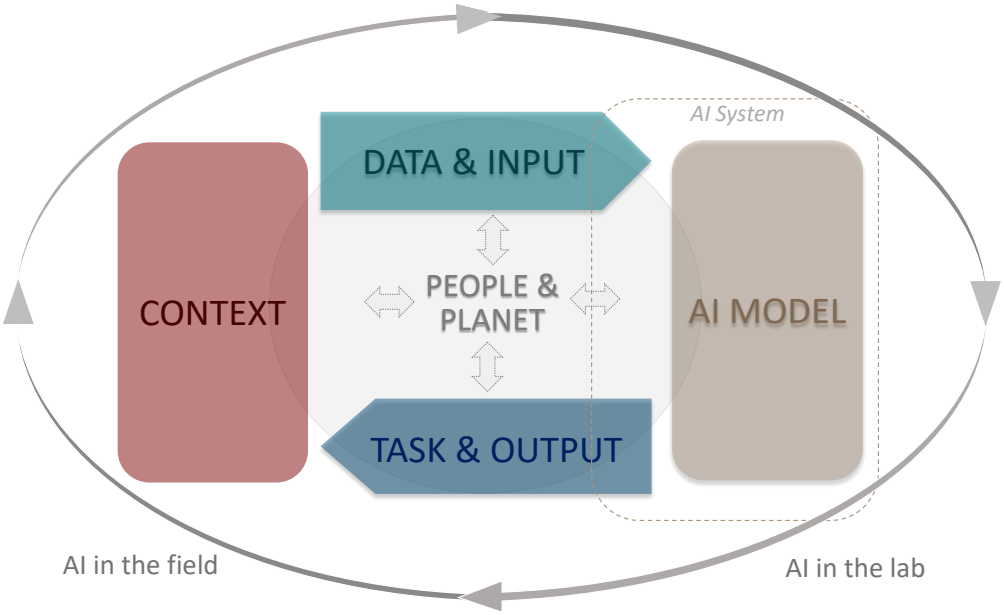
Applying an AI system "in the lab" and/or "in the field"

As already mentioned, some criteria of the framework are more applicable to "AI in the field" contexts than to "AI in the lab" contexts, and vice versa "in the field":

- **AI "in the lab"** refers to the AI system's conception and development, before deployment. It is applicable to the Data & Input (e.g. qualifying the data), AI Model (e.g. training the initial model) and Task & Output dimensions (e.g. for a personalisation task) of the framework. It is particularly relevant to *ex-ante* risk-management approaches and requirements.
- **AI "in the field"** refers to the use and evolution of an AI system after deployment and is particularly relevant to *ex-post* risk-management approaches and requirements. It is applicable to all the dimensions, including the People & Planet and Economic Context dimensions. In addition, it is important to underscore that an AI system "in the field" can change in many significant ways over time, especially with regards to breadth of deployment, technological maturity, users and capabilities. For example, this can happen through improved or different datasets that become available for model-building.

The framework's definitions, concepts and criteria are designed to be dynamic since the classification of an AI system may change. This may happen when a system evolves and incorporates new data and techniques, is deployed more widely, matures or grows in capacity. In addition, the framework may need to be reviewed at regular intervals for continued relevance in view of social, technical and legal developments that may affect AI systems as well as the contexts in which they evolve.

Figure 3. Application of an AI system "in the lab" or "in the field"



Link between the classification and actors in the AI system lifecycle

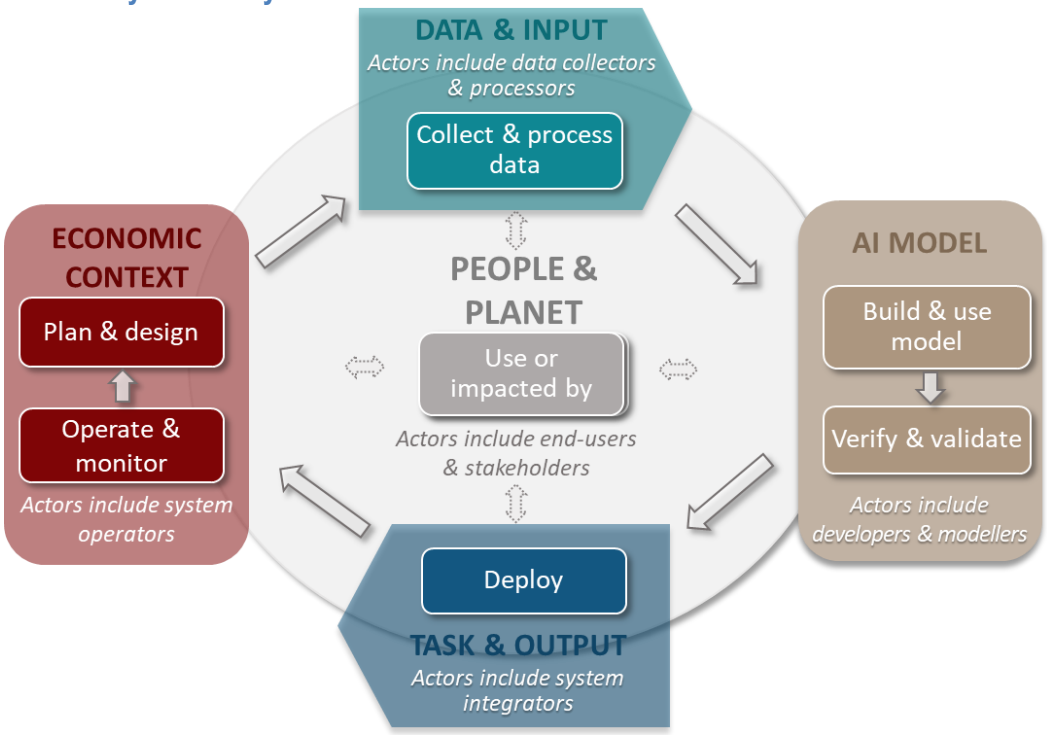
The AI system lifecycle can serve as a complementary tool to understand a system's key technical characteristics and encompasses the following phases: planning and design; collecting and processing data; building and using the model; verifying and validating; deployment; and operating and monitoring (OECD, 2019d[1]). These phases often take place in an iterative manner and are not necessarily sequential. The decision to retire an AI system from operation may occur at any point during the operating and monitoring phase.

The dimensions of the OECD Framework for the Classification of AI Systems can be associated with different stages of an AI system's lifecycle. This is useful for identifying which actors are relevant to each dimension, which matters for accountability and risk management measures (OECD, 2019d[1]).[4]

AI actors are those who play an active role throughout the AI system lifecycle and can include organisations and individuals that deploy or operate AI (OECD, 2019d[1]). AI actors are a subset of stakeholders and may differ depending upon the dimension. As mentioned above, lifecycle actors and stakeholders in the different dimensions are:

- **People & Planet**: End-users and stakeholders who use or are impacted by AI systems. Stakeholders encompass all organisations and individuals involved in or affected by AI systems, directly or indirectly.
- **Economic Context**: System operators who plan and design and operate and monitor applied AI systems.
- **Data & Input**: Data collectors and processors who collect and process data, including gathering and cleaning data, labelling, performing checks for completeness and quality, and documenting the characteristics of the dataset.
- **AI Model**: Developers and modellers who build and use models and verify and validate them.
- **Task & Output**: System integrators who deploy AI systems.

Figure 4. The AI system lifecycle

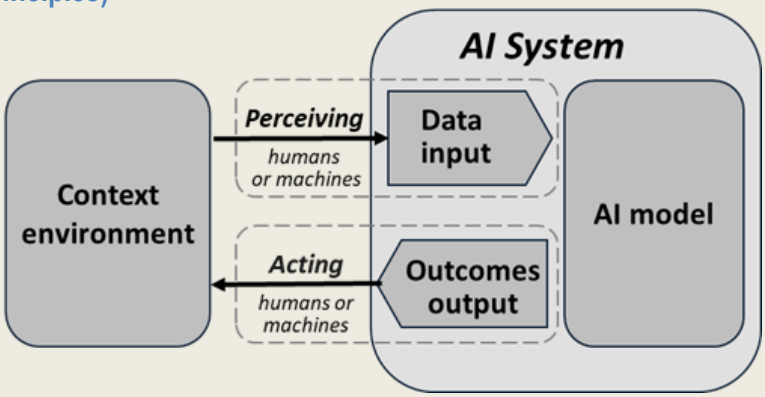


Note: Actors are illustrative and not exhaustive and based on previous OECD work on the AI system lifecycle.
Source: Based on the AI system lifecycle work of AIGO (OECD, 2019f[2]).

Box 1. Characterisation of an AI system based on the OECD AI Principles (2019)

"An AI system is a machine-based system that is capable of influencing the environment by producing recommendations, predictions or other outcomes[5] for a given set of objectives. [6]
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.[7]
AI systems are designed to operate with varying levels of autonomy (OECD, 2019f[2])."

Figure 5. Stylised conceptual view of an AI system (per OECD AI Principles)



Source: (OECD, 2019f[2])

It should be noted that in the present classification framework, the "Perceiving" (data collection) and "data/input" elements – that were separate elements in the original OECD work presented in Figure 5 – have been combined in an effort to simplify the framework (see dotted lines in Figure 4). The "Outcomes" and "Acting" elements have also been combined.

Other important scoping considerations

The framework may need to be reviewed at regular intervals for its dynamic nature, as mentioned above but also for continued relevance in view of social, technical and legal developments that may affect AI systems as well as the contexts in which they evolve.

The framework's dimensions are designed as independent, orthogonal units that affect one another. For example, tasks linked to the Task & Output dimension in the framework impact the Data & Input dimension and how the AI Model dimension is formulated. It also matters whether or not the data was collected for a specific purpose. Collecting data for one purpose and then using it for a purpose that was not intended during the collection stage can cause a misalignment between a task's ideal objective and the approximation of the objective by the data provided.

The *degree of generality* of an AI system refers to its ability to perform several tasks including ones for which it was not initially trained. While there is no single indicator of generality, several criteria in this framework can indicate generality when combined and where objective and consistent information is available. These include: 1) "Scale"; 2) "Model development/maintenance"; and 3) "Combining tasks and actions into multi-task, composite systems".

The framework's dimensions are meant to be used when considering issues in any policy domain that may be affected by an AI system. For example, labour policy makers might take into account: 1) People & Planet: job creation and displacement, access to training; 2) Economic Context: dismissal policy and social dialogue/worker consultation in the industry of deployment; 3) Data & Input: job automation by sensors, new jobs in areas such as data science or data labelling; 4) AI Model: building AI skills and attracting talent; and 5) Task & Output: task automation and impacts on job quality and quantity. Health technology standards might take into account: 1) People & Planet: patients, stakeholders and planet; 2) Economic Context: health technology efficiency; 3) Data & Input: clinical records and trial data; 4) AI Model: transparency and explainability; 5) Task & Output: output description, post deployment change management plan, benchmarking, and oversight.

2 Classification framework

This section describes the characteristics of each of the framework's dimensions, key actors in each dimension and their role in the AI system lifecycle. The roles that actors play are especially relevant to the principle of accountability. The framework provides a structured way to assess AI systems' potential to promote the development of human-centric, trustworthy AI as set out in the OECD AI Principles, i.e. AI systems that benefit people and planet; uphold human rights, democratic values and fairness; are transparent and explainable; are robust, secure and safe; and whose operators are accountable– and to implement policy recommendations in the areas of AI R&D investment; data, compute, technologies; enabling policies and regulation; labour and skills; and international cooperation (OECD, 2019d[1]).

People & Planet

People & Planet are at the centre of the framework (see Figure 1). The framework focuses on human rights and well-being in considering how "people" as a whole interact with and are affected by an AI system throughout that system's lifecycle. People interact with AI systems in many ways – from designing and deciding what data to collect and how to collect it, labelling data, choosing baselines and performance criteria, to putting in place explainability and evaluation mechanisms.

In using the framework to review an AI system for acquisition and deployment, the following definitions apply:

- *Users* of an AI system or application are the individuals or groups that utilise the system for a specific purpose.
- *Impacted stakeholders* can be indirectly or directly affected by the deployment of an AI system or application but do not necessarily interact with the system. An AI system or application can impact several different stakeholder groups.

For example, regarding a credit scoring system, the intended users are typically bank employees who use the system to assess customers' creditworthiness. The impacted stakeholders are the consumers as well as the regulatory body overseeing financial stability.

The following sections describe key criteria of AI users and impacted stakeholders, and how the AI system impacts them indirectly or directly.

Users' AI competency

Often, the users and intended users of an AI system are not those who developed and implemented it; nor do they usually operate it. Users can range in competency from AI experts to amateur end-users. For example, AI systems deployed in sectors such as healthcare or agriculture are often used by practitioners or domain experts who are not typically AI experts. In light of the implications of the wide variety in levels of expertise among end-users, AI systems can be distinguished based on whether their typical users have any systems operation training, according to the following terminology:

- *Amateur*: User who has no training[8]

- *Trained practitioner who is not an AI expert*: User with some specific training on how to use the AI system in question
- *AI expert*: User with specific training and knowledge of how AI works in the application or system considered (an AI expert or system developer)
- Other

Why does this matter? AI system users relate to accountability (Principle 1.5); transparency and explainability (Principle 1.3); and safety, security and robustness (Principle 1.4).

Impacted stakeholders

The following stakeholders may be impacted directly or indirectly, and consciously or unconsciously by the AI system:

- Workers/employees
- Consumers
- Business
- Government agencies/regulators
- Scientists/researchers
- Children or other vulnerable or marginalised groups

Why does this matter? Stakeholders impacted by the system are most relevant to transparency and explainability (Principle 1.3) and to policy and regulatory frameworks (Principle 2.2). Stakeholder groups such as consumers, workers/employees or children are often covered by existing policy and regulatory regimes. In Europe, the General Data Protection Regulation (GDPR) gives data subjects (those whose data is collected, held or processed) the right, under some circumstances, to not be subject to automated decision-making.

Optionality and redress

"Optionality" or "dependence" refers to the degree of choice that users or impacted stakeholders have on whether or not they are subject to the effects of an AI system, whether their involvement is active or passive. Optionality can be understood as the extent to which users can opt out of "the effects" or "the influence" of the AI system, e.g. by switching to another AI system and the societal repercussions of doing so, e.g. for access to healthcare or financial services. This is also referred to as "switchability" (AI Ethics Impact Group, 2020[8]). It is important to consider the human aspect or the degree to which they are involved in developing AI systems and models, of the operation and outputs of the system, and if humans are "in", "on" and "out-of-the-loop". The following are generally considered to be distinct modes of optionality in a given AI system:

- Users cannot *opt out* of the AI system's output.
- Users can *opt out* of the AI system's output.
- Users can *challenge or correct* the AI system's output.
- Users can *reverse* the AI system's output *ex-post*.

Benefits and risks to human rights and democratic values

Some AI systems generate outputs that can impact individuals' human rights, either negatively or positively (see Table 3). Low-risk contexts for individuals, such as a restaurant recommendation may not determine

an individuals' actions in an area of life that is directly related to fundamental human rights, such as health or fair access to employment, for instance, even if they affect certain aspects of self-determination. As a result, systems operating in low-risk contexts may not need multi-layered and costly approaches to verification and can therefore rely mostly on machines, provided there are adequate protections regarding combining different datasets at aggregate levels. Having humans in the loop of certain AI system processes and/or a human appeal process are important to reducing risk.

Table 3. Sample checklist for assessing the potential impact of an AI system on selected human rights and democratic values, direct or indirect

Impact of AI on human rights or democratic values	Outcome-dependent	No impact
Liberty, safety and security		
Physical, psychological and moral integrity		
Freedom of expression, assembly and association		
Freedom of thought, conscience and religion		
Rule of law, absence of arbitrary sentencing		
Equality and non-discrimination		
Social and economic rights (e.g. health, education)		
Quality of democratic institutions (e.g. free elections)		
Right to property		
Aggregate society-level risk (please detail)		
Other (detail)		

Note: International human rights refer to a body of international laws, including the Universal Declaration of Human Rights, as well as regional human rights systems such as that of the Council of Europe. Human rights provide a set of universal minimum standards based on, among others, values of human dignity, autonomy and equality, in line with the rule of law. Human rights overlap with wider ethical concerns and with other areas of regulation relevant to AI, such as personal data protection or product safety law. Risk-assessment frameworks would usually also include the likelihood the risk will occur, its impact and mitigation measures.
Source: Based on (UN General Assembly, 1948[3]) (CoE, 2020[4]), (CoE, 1998[5]) and (OHCHR, 2011[6]).

Why does this matter? Transparency and explainability (Principle 1.3), as well as accountability (Principle 1.5), are widely viewed as having higher importance in contexts where the outcomes of an AI system can impact human rights. Examples include AI used to sentence criminals, recommend decisions about educational opportunities or conduct job screenings. Such high-stakes situations often require formal transparency and accountability mechanisms (Principles 1.3 and 1.5), including transparency about the role of AI and human involvement in the process (e.g. human-in-the-loop), the full consequences of the AI system's action on all stakeholders and the availability of appeals processes, particularly where life and liberty are at stake. Across the spectrum, people broadly agree that AI-based outcomes (e.g. a credit score) should not be the only decisive factor when applications or decisions have a significant impact on people's lives. Such applications may for example require that a human consider the social context, the precise decisions enabled by the AI system as well as its limitations and the variables it uses to help avoid unintended consequences. For example, the GDPR stipulates that a human must be in the loop if a decision has legal or similarly significant effects on people.9

Benefits and risks to the environment, well-being and society

Many AI systems use human data as inputs and generate outputs that can impact individuals' and societies'[10] well-being, either positively or negatively.[11] This impact pertains to different areas of life, such as work and job quality, environment quality, social connections and civic

engagement, among others (Table 4). In addition, AI systems can impact societal well-being in the aggregate.

Table 4. Sample checklist for assessing potential impact of an AI system's outcomes on well-being

Impact of AI on well-being	Outcome-dependent	No impact
Physical and mental health		
Housing		
Income and wealth		
Quality of job		
Quality of environment		
Social connections		
Civic engagement		
Education, knowledge and skills		
Work-life balance		
Aggregate society-level impact (please detail)		
Duration of impact		

Note: Outcomes can be captured by measures of inequality, for example.
Source: (OECD, 2020[7]).

Why does this matter? In the current context of accelerating climate change and loss of biodiversity across the planet, AI brings significant opportunities to mitigate risks and to help adapt. These and other dimensions of universal well-being call for responsible AI that is based on algorithms and optimisation functions that are human-centric, user-defined, guarantee benefits for people and planet (Principle 1.1) and maintain accountability (Principle 1.5). International cooperation (Principle 2.5) is a must in the face of urgent global challenges.

Work, human and employment displacement potential {where objective and consistent information is available}

An AI system's ability to automate tasks previously or currently conducted by humans depends on a variety of task-dependent factors such as perception and manipulation requirements, presence of uncertainty in a task, and creative and social intelligence factors.[12] Historically, automation has been limited to tasks that:

- require perceiving and manipulating homogeneous objects with clearly defined processes and limited uncertainty;
- are conducted within controlled environments; and
- do not require creativity or social interaction.

However, recent innovations in AI are changing the automation landscape and more tasks typically executed by higher-skilled workers are being automated. For simplicity, an AI system's potential to automate tasks can be split into three categories, listed below.

- *High displacement potential*: AI systems that perform tasks that use clearly defined processes and outputs (e.g. tasks performed by clinical lab technicians, optometrists, chemical engineers, actuaries, credit analysts, accountants, operations research analysts, concierges, mechanical drafters, brokerage clerks and quality control inspectors). This does not imply a high likelihood of being replaced by AI. That would require a more complex assessment of the technical feasibility and context of the task to be performed.

- *Low displacement potential*: AI systems perform tasks that require reasoning about novel situations (e.g. research), interpersonal skills (e.g. teachers and managers, some baristas) and physical occupations that require perception and manipulation of a plurality of irregular objects in uncontrolled environments with limited room for mobility (e.g. maids, cleaners, cafeteria attendants, hotel porters, roofers and painters, massage therapists, plasterers and stucco masons). This does not necessarily mean that the occupation will not see significant automation of key tasks.
- *No displacement potential*: Some AI systems execute tasks that could not be performed by humans with the same accuracy, specificity or scale (e.g. AI systems used in cybersecurity and threat detection).

Why does this matter? An AI system's capacity to automate tasks and improve worker productivity can impact the world of work (Principle 2.4). This impact of AI will have implications for education strategies for affected groups as well as potential policies to share the benefits of increased worker productivity.

Key actors in the People & Planet dimension are end-users and stakeholders

The end-users (or intended users) of an AI system or application are the individuals or groups that use the system for a specific purpose. The impacted stakeholders encompass all organisations and individuals involved in or affected by AI systems, directly or indirectly. They do not necessarily interact with the system and can be indirectly or directly affected by the deployment of an AI system or application.

Economic Context

According to this framework, the Economic Context dimension of an AI system represents its socio-economic environment, including its broader natural and physical environment. This dimension is mostly relevant to a specific application of an AI system rather than to a generic AI system. The context is observable and can be influenced through actions resulting from an AI system's outputs (OECD, 2019f[2]). Core characteristics of the Economic Context, described in more detail in the sections that follow, include the sector in which an AI system is deployed, its business function, its critical (or non-critical) nature, and its deployment impact and scale. The key AI actor in this dimension is the system operator.

Industrial sector

AI is diffusing rapidly throughout industrial sectors and is increasingly applied in fields such as finance and insurance, advertising, transport, manufacturing and healthcare. Each industrial sector represents a different context that has different implications, in terms of industry structure, regulation and policy making, for AI systems. In April 2021, the European Commission (EC) proposed an AI-related package of governance guidelines and regulations, including a proposal for AI regulation using a risk-based approach that incorporates several risks linked to other sectors, notably healthcare, transport, energy and parts of the public sector (e.g. law enforcement, migration, border control, judiciary, social security and employment). It should be noted that some applications of AI span multiple industries or even multiple functions within a single industry – for example, recruitment.

This classification framework uses the International Standard Industrial Classification of All Economic Activities (ISIC REV 4), which allows for comparability with other sources of cross-country data on employment, skills, demography of enterprises, value-added and more. The highest sectoral-level categories of economic activities, "sections" include (this document does not address AI systems used in military contexts):

- Section A Agriculture, forestry and fishing
- Section B Mining and quarrying
- Section C Manufacturing
- Section D Electricity, gas, steam and air conditioning supply
- Section E Water supply; sewerage, waste management and remediation activities
- Section F Construction
- Section G Wholesale and retail trade; repair of motor vehicles and motorcycles
- Section H Transportation and storage
- Section I Accommodation and food service activities
- Section J Information and communication
- Section K Financial and insurance activities
- Section L Real estate activities
- Section M Professional, scientific and technical activities
- Section N Administrative and support service activities
- Section O Public administration and defence; compulsory social security
- Section P Education
- Section Q Human health and social work activities
- Section R Arts, entertainment and recreation
- Section S Other service activities

- Section T Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use
- Section U Activities of extraterritorial organizations and bodies

Why does this matter? The policy implications of deploying AI systems vary significantly from one sector to the next. The industrial sector has a particularly high impact on economic and social benefits (Principle 1.1) and on jobs and skills (Principle 2.4).

Business function and model

For cross-functional use of AI in an organisation – such as in recruitment, promotion, training or even dismissal – the business function classification will be very important. Functional areas in organisations in which AI systems can be employed include but are not limited to:

- Human resource management
- Sales
- ICT management and information security
- Marketing and advertisement
- Logistics
- Citizen/customer service
- Procurement
- Maintenance
- Accounting
- Monitoring and quality control
- Production
- Planning and budgeting
- Research and development
- Compliance and justice

Why does this matter? Different AI systems can perform the same task in different functional areas, with different implications for policy making. For instance, a forecasting algorithm used to improve (optimise) logistics may have different implications than a forecasting system designed to support hiring decisions. The business function for which the AI system is used will thus have a specific impact on economic and social benefits (Principle 1.1); fairness and absence of bias (Principle 1.2); security, safety and robustness (Principle 1.4); and jobs and skills (Principle 2.4)

Business model: for-profit use, non-profit use or public service

Operators may use an AI system (OECD, 2011[8]) in the following ways:

- For-profit use – subscription fee model
- For-profit use – advertising model
- For-profit use – other model
- Non-profit use (outside public sector) – voluntary donations and community models
- Public service
- Other

Why does this matter? The use case and business model of the AI system operator can be relevant to determining the objectives that an AI system is optimising for, e.g. maximising engagement in advertising-based business models.

Impacts on critical functions and activities

Critical functions are economic and social activities for which "the interruption or disruption would have serious consequences. They include: 1) the health, safety, and security of citizens; 2) the effective functioning of services essential to the economy and society, and of the government; or 3) economic and social prosperity more broadly" (OECD, 2019a[9]); (OECD, 2019e[10]). Public safety, health and consumer protection are vast domains many professions use AI in assessments (e.g. pilots, nuclear reactor operators, law enforcement). It is important to note that the risks associated with the use of AI in critical functions should be weighed against the benefits and requirements for security and the accuracy of outcomes. Further, not all systems in a critical sector are considered critical. For example, administrative time-tracking systems in hospitals and banks are not considered critical systems. Critical systems and activities are defined as follows:

- *AI system* deployed in a critical sector or infrastructure (e.g. energy, transport, water, health, digital infrastructure and finance).
- *AI system performs or serves a critical function independent from its sector* (e.g. conducting elections, maintaining supply chains, law enforcement, providing medical care, supporting the financial system).

Why does this matter? In some sectors, critical functions are accompanied by heightened risk considerations with ex-ante regulations. The critical function will have a particular impact on security, safety and robustness (Principle 1.4). In the European Union, the Network and Information Security (NIS) Directive mandates the supervision of critical sectors. EU Member states must supervise the cybersecurity of critical market operators ex-ante in critical sectors like energy, transport, water, health, digital infrastructure and finance, and ex-post surveillance is required for critical digital service providers like online market places, cloud services and online search engines. In the United States, national critical functions include conducting elections, maintaining supply chains, law enforcement and providing medical care (CISA, 2019[11]). In the financial sector, banks operate some critical functions, like the SWIFT system for sending payment orders between financial institutions.

Scale of deployment and technology maturity

AI systems' economic and social impact varies depending on the following four factors: breadth of an AI system's deployment; its maturity, from a technology standpoint; stakeholder(s) impacted by the system; and the for-profit, non-profit or public service use of the system.

Breadth of deployment

The breadth of deployment of an AI system relates to the number of individuals affected by a system and is characterised by the following attributes:

- *Pilot project*
- *Narrow deployment*: Deployment is at the level, for example, of one company or of one small, targeted region
- *Broad deployment*: Deployment is at the level, for example, of one sector across different countries
- *Widespread deployment*: Deployment reaches across countries and sectors

Technical maturity {where objective and consistent information is available}

The maturity of deployed AI systems can vary widely. Technology readiness levels (TRLs) can help classify an AI system's technical maturity. The following nine TRL categories are based on Joint Research Centre (JRC) analysis (Martinez Plumed, 2020[12]) building on the NASA TRL framework (Mankins, 1995[13]). It should be noted that TRLs are viewed as a rough measure (Terrile et al., 2015[14])

- *Basic principles observed and reported – TRL 1*: Lowest level of technology readiness, where research begins to be translated into applied R&D. Sample output might be a scientific article on a new technology's principles.
- *Technology concept and/or application formulated – TRL 2*: Speculative practical applications are invented based on assumptions not yet proven or analysed. Sample output might be a publication or reference highlighting the applications of the new technology.
- *Analytical and experimental critical function and/or characteristic proof-of-concept – TRL 3*: Continued research and development efforts include analytical studies and lab studies to physically validate analytical predictions of separate elements of the technology. Sample output might be measurement of parameters in the lab.
- *Component and/or layout in controlled environment – TRL 4*: Basic technological components are integrated to verify they can work together but in a relatively superficial manner. Sample output might be integration of ad hoc software or hardware in the laboratory.
- *Component and/or layout validated in relevant environment – TRL 5*: Reliability is significantly increased and basic technological components integrated with fairly realistic supporting elements that can be tested in a simulated environment. Sample output might be realistic laboratory integration of components.
- *Representative model or prototype system demonstrated in relevant environment – TRL 6*: Sample output might be testing a prototype in a realistic laboratory environment or in a simulated operational environment.
- *System prototype demonstration in operational environment – TRL 7*: Examples include testing the prototype in operational testing platforms (real-world clinical setting, vehicle, etc.).
- *System or subsystem complete and qualified through test and demonstration – TRL 8*: Technology proved to work in its final form and under expected conditions. In most cases, this TRL represents the end of true system development. Examples include developmental testing and evaluation of the system to determine if the requirements and specifications are fulfilled.
- *Actual system or subsystem in final form in operational environment – TRL 9:* Actual application of the technology in conditions such as those encountered in operational conditions Strong monitoring and improvement processes are critical to continue to improve the system.

Why does this matter? AI system maturity is particularly relevant to safety, robustness and security (Principle 1.4); accountability (Principle 1.5); and R&D investment (Principle 2.1).

Key actors in the Economic Context dimension include system operators

AI system operators are key AI actors in the Economic Context dimension, which can be associated with the planning and design or specification stage of the AI system lifecycle as well as, following deployment, with the operating and monitoring phase. Planning and design of an AI system involves defining the system and its objectives, underlying assumptions, context and requirements (OECD, 2019f[2]). The planning and design or specification phase is critical for public policy, and any major public failures and issues related to other sections of the OECD framework can be avoided if first addressed at the specification phase. For example, in the case of a credit system, the operator of the system is likely to be the bank's information

technology department. Planning and design processes for this system, then, might require expertise from data scientists, domain experts and governance experts.

Operating and monitoring an AI system involves continuously assessing its recommendations and impacts (both intended and unintended) in light of the system's objectives as well as the ethical considerations that go into its operation. In this phase, problems are identified and adjustments made by reverting to other phases or, if necessary, retiring an AI system from production. For monitoring purposes, AI system operators can establish:

- *Transparent, accessible information about the AI system's objectives and assumptions*: Provide interested stakeholders with access to useful information.
- *Performance monitoring mechanisms*: Such as metrics to assess the performance and accuracy of the AI system.
- *Tools or processes for developing or maintaining trustworthy AI*: Using tools like guidelines; governance frameworks; product development or lifecycle tools including for model robustness; risk management frameworks; sector-specific codes of conduct; process standards; technical validation approaches; technical documentation; technical standards; toolkits, toolboxes or software tools; educational material; change-management processes; certification (technical and/or process-related); or tools for protection against adverse attacks.

Why does this matter? Key actors in the Economic Context dimension are often AI system operators who plan and design the AI system and, following deployment, operate and monitor it. System operators relate to accountability (Principle 1.5); transparency and explainability (Principle 1.3); and safety, security and robustness (Principle 1.4).

Data & Input

An AI system can be based on expert input or on data, both of which can be generated by humans or automated tools such as machine-learning algorithms. Historically, AI systems were powered by *expert input* in the form of logical representations that formed the basis of early optimisation and planning tools such as those used in medical diagnosis, credit-card fraud detection or in chess playing (e.g. IBM's Deep Blue). They needed researchers to build detailed decision structures to translate real-world complexity into rules to help machines arrive at human-like decisions. Expert input also includes structures such as ontologies, knowledge graphs, decision rules and analytical functions (see section on Structure of data).[13] In recent years, AI systems have become more and more statistical and probabilistic and are increasingly powered by a growing variety of data types.

Data can be collected and processed during development in the lab as well as during production or run time in the field. Most of the characteristics of AI system data are relevant to both training data and data used in the field, except for the dynamic nature (the degree to which it changes or updates) of the data, which is relevant primarily during production, or in the field.

Core characteristics of the Data & Input dimension are data provenance (where the data comes from); data collection and origin (e.g. data collection, origin, dynamic nature and scale); domain (e.g. personal, proprietary or public); data quality and appropriateness; and their technical characteristics (e.g. structure and encoding). The next sections draw from OECD work (OECD, 2019[15]).

Collection method, provenance and dynamic nature

Detection and collection of data and input

Humans or machines can detect and collect ("track") data and input from the context or environment by:

- *Collected by humans*: This takes place when a human is needed to observe and collect information that requires subjective judgment, such as a person's mental state. Other examples of data collected by humans are crowd-sourcing data and human-based computation, where certain steps of the computation process are conducted by humans.
- *Collected by automated sensors*: Devices that automatically monitor and record data include cameras, microphones, thermometers, laboratory instruments and other sensors such as Internet of Things (IoT) devices, but also the automated recording of information from online log files, mobile phones, GPS watches and activity wristbands.
- *Collected by humans and automated sensors*: Some data are collected by humans together with automated tools. In healthcare applications, data from sensors such as heartbeat or blood pressure detectors will often be combined with a doctor's assessment.

Why does this matter? Data and input collection through automated sensing can benefit society in fields such as healthcare and safety (e.g. activity trackers associated with health applications) or environmental applications. Data and input collection can also surface labour-market considerations (Principle 2.4), including the automation of tasks (e.g. security surveillance or maintenance assessments); improving worker safety and satisfaction; measuring worker productivity; and codifying expert knowledge.

Provenance of data and input

The following list draws on the data provenance categorisation made by Abrams (Abrams, 2014[16]) and the OECD (OECD, 2019[15]) of data collected with decreasing levels of awareness. It should be noted that these categories can overlap and most systems will combine data from different sources. Here, we broaden the original categorisation that focused on personal data to also cover expert input and non-personal data, as well as data that are synthetically generated.

- *Expert input*: Human knowledge that is codified into rules and structures such as ontologies (concepts and properties), knowledge graphs and analytical functions (e.g. the objective function or rewards an AI model will optimise for).
- *Provided data*: Data that originate from actions by individuals or by organisations that are aware of the data being provided. They include "initiated" (e.g. a license application), "transactional" (e.g. bills paid) and "posted" (e.g. social networking posts) data.
- *Observed data*: Collected through observation of a behaviour or activity through human observation or the use of automated instruments or sensors. Examples include website visitor provenance and browsing patterns observed by a website administrator. Observed data also include sounds, scents, temperature, GPS position or soil acidity. Observed data about individuals can be "engaged" (e.g. voluntarily accepting cookie tracking on a website), "unanticipated" (e.g. the tracking of seconds spent looking at a specific image online) or "passive" (e.g. CCTV images of individuals).
- *Synthetic data*: Usually generated by computer simulations, including data collected through reinforcement learning. Synthetic data allow for simulation of scenarios that are difficult to observe or replicate in real life (e.g. a car accident) or are otherwise too expensive to collect at scale (e.g. millions of miles of driving time for self-driving cars). They include most applications of physical modelling, such as music synthesisers or flight simulators. AI system output of synthetic data approximates reality but is generated algorithmically.
- *Derived data*: Data taken from other data to become a new data element. Derived data include computational (e.g. a credit score) and categorical data (e.g. age group of a buyer). They can be inferred (e.g. the product of a probability-based analytic process like a fraud score or risk of accident) or aggregated (e.g. abstracted from more fine-grained data). Proprietary data are often characterised as derived data.

Why does this matter? Awareness and consent for the provision of personal data about individuals is a critical focus area for privacy and consumer protection (Principle 1.2). Synthetic data allow for simulation of scenarios that are difficult to observe or replicate in real life (e.g. a car accident) and are relevant to safety (Principle 1.4). Expert input is typically human knowledge that is codified into rules.

Dynamic nature of data

Data can be "static" or "dynamic", to varying degrees:

- *Static data*: These data do not change after they are collected (e.g. a given publication, a product's batch number or the geographic latitudes and longitudes of a fixed element like a building or a mountain).
- *Dynamic data updated from time-to-time*: Dynamic data continually change after they are recorded in order to maintain their integrity. Models relying on dynamic data can leverage "incremental algorithms" that update the model frequently based on incoming data. Dynamic data can be updated from time-to-time without necessarily being real-time data. Examples include timetables of flights' estimated time of arrival using batch processing.
- *Dynamic real-time data*: Dynamic real-time data are delivered immediately after collection with no delay. Examples of systems that use real-time data processing include an alarm system triggered by

an entry signal, a recommender system that evolves in real-time as it is being used (e.g. with a streaming video service like YouTube) and an autonomous driving system that reacts to real-time environmental data.

Why does this matter? The degree to which data are static or dynamic is particularly relevant to public policy and accountability (Principle 1.5) for AI systems that can iterate and evolve over time and may change their behaviour in unforeseen ways.

Scale {where objective and consistent information is available}

The scale of a dataset is a continuous variable that has an ever-increasing upper limit. If real-time, scale can be roughly measured as the order of magnitude of bytes per time unit (e.g. tens of petabytes per second) or the number of requests to the AI system per second. If static, size is measured in bytes (e.g. hundreds of gigabytes). The scale of data continues to change as technology advances. The upper limit is generally reached by very few government and commercial enterprises that accommodate high-velocity, real-time data streams supporting extremely large data volumes.

The scale of data can be:

- *Very large*: One exabyte (one billion gigabytes) or larger. Extremely large volumes of data take time to gather/accumulate and require complex systems to operate and process.
- *Large*: Tens of petabytes (per second if real-time).
- *Medium*: Hundreds of gigabytes.
- *Small*: Tens of gigabytes or smaller. There are no constraints to transferring and processing small amounts of data in current broadband networks and computing environments.

Why does this matter? AI powered by machine-learning technology is known to rely on large volumes of data to function well, based on which patterns are inferred. There is active research on AI systems that use less data, such as one-shot learning (Principle 2.1). These AI systems are learning through self-play – via reinforcement learning – to drastically reduce the scale of the data needed to train a model.

In terms of robustness of AI systems (Principle 1.4), researchers have found that there is a trade-off between the quantity of data and the number of variables in a model. A larger model – one with more parameters – consumes more input resources (e.g. compute capacity, data) than smaller models. However, at large scales, large models can learn to use data more efficiently than smaller models, leading to the counterintuitive result that larger models can match the performance of smaller models while using less data. This has implications for situations where training-data samples are expensive to generate, which likely confers an advantage to large companies entering new domains with models based on supervised learning.[14]

Data size also relates to the efforts to build the technology infrastructure to process, transfer and share large volumes of data for AI (Principle 2.2).

Rights, or domain, and identifiability

Rights associated with data and input

This sub-dimension distinguishes the rights, or domain, associated with data and input used by an AI system, and the policy implications when used in training data and/or in a deployment context. Data domains include the following three categories, which can overlap in certain applications (see Figure 6):

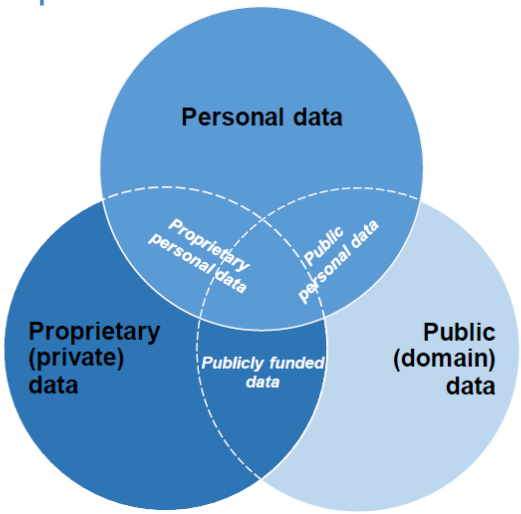
- *Proprietary data*: Data that are privately held, often by corporations, and typically protected by intellectual property rights – including copyright and trade secrets – or by other access and control regimes such as contract and cyber-criminal law. There is typically an economic interest to restrict access to proprietary data. [15] Input into an AI system in the form of rules can be considered a form of proprietary data.
- *Public data*: Data not protected by intellectual property rights or any other rights with similar effects and that in many cases can be shared for access and re-used through open data regimes.
- *Personal data*: Data that "relates to an identified or identifiable individual".

Why does this matter? Proprietary data raise issues such as transparency and explainability (Principle 1.3); bias in AI systems (Principle 1.2); as well as considerations of business scale-up (Principle 2.2).

Public data is relevant to economic, social and environmental impacts (Principle 1.1); research (Principle 2.1); and data availability and compute capacity (Principle 2.2).

Personal data is associated with privacy considerations and legislation and usually requires more restrictive access regimes. Personal data is relevant to issues related to human rights, fairness and privacy (Principle 1.2).

Figure 6. Personal, private and public domains of data



Source: (OECD, 2019[15]).

Identifiability of personal data

Personal data taxonomies differentiate between different categories of personal data. ISO/IEC 19441 (2017) distinguishes five categories, or "states", of data identifiability:

- *Identified data*: Data that can be unambiguously associated with a specific person because they contain personal identifiable information.
- *Pseudonymised data*: Data for which all personal identifiers are substituted by aliases. The alias assignment is such that it cannot be reversed by reasonable efforts, except for the party that performed the assignment.
- *Unlinked pseudonymised data*: Data for which all personal identifiers are irreversibly erased or substituted by aliases. The linkage cannot be re-established by reasonable efforts, including by the party that performed the assignment.

- *Anonymised data*: Data that are not linked to attributes that can be altered (i.e. attributes' values are randomised or generalised) in such a way that there is a reasonable level of confidence that a person cannot be identified, directly or indirectly, by the data alone or in combination with other data.
- *Aggregated data*: Statistical data that do not contain individual-level entries and are combined with information about enough different persons that individual-level attributes are not identifiable.

Why does this matter? The type of personal data used by AI systems has implications for individuals' human rights, fairness and privacy (Principle 1.2). Data identifiability can help assess the level of risk to privacy and inform the need for legal and technical protection and access control. Concerns are raised that even absent personal data, AI systems are able to infer data and correlations from proxy variables that are not personally identified, such as purchasing history or location. In addition, some regimes such as the EU's GDPR, distinguish "sensitive personal data" that consist of racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data, health data or data concerning a person's sex life or sexual orientation. In the United States, personal data considered sensitive include data about children and financial and health information.[16]

Data quality and appropriateness {where objective and consistent information is available}

Data appropriateness (or "qualification") is about defining criteria to ensure that the data are appropriate for use in a project, fit for purpose, and relevant to the system or process following standard practice in the industry sector. For example, in clinical trials to evaluate drug efficiency, criteria for using patient data must include patients' clinical history (previous treatments, surgery, etc.). Data quality also plays a key role for AI systems, as do the standards and procedures to manage data quality and appropriateness.

An AI application or system applies standard criteria or industry-defined criteria in order to assess:

- *Data appropriateness*: Data are appropriate for the purpose for which they are to be used, following standard practice in the industry sector.
- *Sample representativeness*: Selected variables and training or evaluation data accurately depict/reflect the population in the AI system environment.
- *Adequate sample size*: Sample size displays an appropriate level of granularity, coverage and sufficiency of data.
- *Completeness and coherence of sample*: Sample is complete, with minimal missing or partial values. Outliers must not affect the quality of data.
- *Low data "noise"*: Data is infrequently incorrect, corrupted or distorted (e.g. intentional or unintentional mistakes in survey data, data from defective sensors).

Why does this matter? Data appropriateness impacts the accuracy and reliability of the outcome of AI systems and relates to their robustness, security and safety (Principle 1.4.) The use of inappropriate data/input in an AI system can lead to erroneous and possibly dangerous conclusions.[17]

Data quality has important policy implications for human rights and fairness (Principle 1.2), as well as to the robustness and safety of an AI system (Principle 1.4): from both fairness and robustness perspectives, datasets must be inclusive, diverse and representative so they do not misrepresent specific (sub) groups.

Structure and format of data and input

Structure of data and input

This sub-dimension identifies common types of data structures:

- *Unstructured data*: Include data that either do not have a pre-defined data model or are not organised and labelled in a pre-defined manner (e.g. text, image, audio, video and other data types such as sensor data and interlinkages between graph networks, social media or website data). Unstructured data often include irregularities and ambiguities that are difficult for traditional programmes to analyse. Unstructured data are sometimes referred to as "raw" data.
- *Semi-structured data*: In practice, most data combine both unstructured and structured data. For example, a photo taken with a smartphone consists of the image itself (unstructured data), accompanied by structured metadata about the image (when and where it was taken, what device took the picture, the picture format, its resolution, etc.). Similarly, data on a social network such as Twitter include unstructured text alongside structured metadata about the author of the text and his or her networks. In addition to social media, examples of semi-structured data include device or sensor data.
- *Structured data*: Data that are stored in a pre-defined format and are straightforward to analyse. Structured data have labels describing their attributes and relationships with other data. Vast amounts of user data from websites or e-commerce sites are structured, fuelling the development of a wide variety of marketing techniques (e.g. personalised advertisements), recommendation mechanisms (e.g. Amazon products, Netflix content, Spotify recommendations, YouTube's "up next" videos) and engagement systems (e.g. Facebook feeds). Examples of structured data include interlinked tables and databases.
- *Complex structured data*: Data often produced in the form of a model, which is both the output of an AI algorithm and can be used as input to another system. Examples of complex structured data include ontologies (e.g. partial models of the environment), knowledge graphs, rules (e.g. expert systems) and analytical functions (e.g. adversarial learning or reinforcement learning functions).

Each of these structural alternatives can be encoded in various data formats (e.g. binary, numeric, text) and represent different forms of media (e.g. audio, image, video or their combinations). "Data labelling" is the process of tagging data samples, which generally require human knowledge to build training data.

Why does this matter? Data and input structure relates to transparency and auditability (Principle 1.3) and directly impacts the AI model choice. Structured data are easier to document and audit (Principle 1.5) and also influences data-sharing policies (Principle 2.2)

Format of data and metadata {where objective and consistent information is available}

Data format (or encoding) refers to the format of the data themselves. Data format is closely related to data collection (e.g. a camera will produce a specific format of image data) and to data modelling, where different modelling techniques will require specific data formats (e.g. time-series modelling requires temporally sequenced data).

Dataset metadata may include information on how a dataset was created, its composition, its intended uses and how it has been maintained over time. Formats and standards for annotating datasets often need to be developed while taking into account industry sector and use case:

- *Standardised data format*: Standardised data have a format pre-agreed to by the providers of the data, which allows for easier comparability, i.e. for a dataset to be compared to other datasets.
- *Non-standardised data format*: Data can also be in ad hoc formats created for the purpose of particular applications (e.g. video).
- *Standardised dataset metadata*.

- *Non-standardised dataset metadata*.

Why does this matter? Standardisation of data formats facilitates interoperability and data re-use across applications and for accessibility, and can help ensure that data are findable, catalogued, searchable and re-usable. The use of standardised formats may improve an AI system's robustness and security by making it easier to address security vulnerabilities (Principle 1.4).

In some sectors, standardised templates for dataset metadata annotation are being developed. Standardised metadata facilitates the development and sharing of training datasets and, by extension, can help accelerate the development and use of an AI system (Principle 2.2).

Key AI actors in the Data & Input dimension include data collectors and data processors

The Data & Input dimension maps directly to the data collection and processing stage of the AI system lifecycle (Figure 3), where it involves 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 was created, composition, intended uses and how it was maintained over time (OECD, 2019[15]). Key actors also include data rights holders who are impacted by whether these rights are intellectual property (e.g. copyright) or personal data rights such as those recognised by the GDPR in the EU.

Data collection and processing currently involves expertise from actors such as data scientists, domain experts, data engineers and data providers. Actions performed by data collectors and processors include:

- *Performing checks*: For data quality and appropriateness.
- *Transparent information about the data and inputs used in the AI system*: Providing interested stakeholders with access to meaningful information on the data and inputs used in the AI system.
- *Labelling data*: Such as tagging data with informative data.
- *Protecting personal data*.
- *Documenting data and dataset characteristics*.
- *Using tools or processes for trustworthy AI*: Such as guidelines, governance frameworks, product development/lifecycle tools, risk management, sector-specific codes of conduct, process standards, technical validation approaches, technical documentation, technical standards, toolkits/toolboxes/software tools, educational material, change-management processes, and certification (technical and/or process-related).

AI Model

AI model-building is part of a system's development in the lab (especially for non-learning AI) whereas model inferencing, the process of *using* the model, takes place in production in the field.

To accurately classify AI systems, it is helpful to identify the core AI model, or models, around which the system is built as they determine a wide array of characteristics. The AI Model dimension considers AI models as composites of multiple, core, technical components, and analyses the choice of AI models, how the models are built, how the models are interlinked with one another and other sub-systems, and how they are used, also known as model "inferencing". This section discusses how the traits of an AI model relate to policy considerations.

What is an AI model? AI models are actionable representations of all or part of the external context or environment of an AI system (encompassing, for example, processes, objects, ideas, people and/or interactions taking place in context). AI models use data and/or expert knowledge provided by humans and/or automated tools to represent, describe and interact with real or virtual environments (Box 2).

Box 2. OECD characterisation of an AI model, included in the OECD AI Principles (2019)

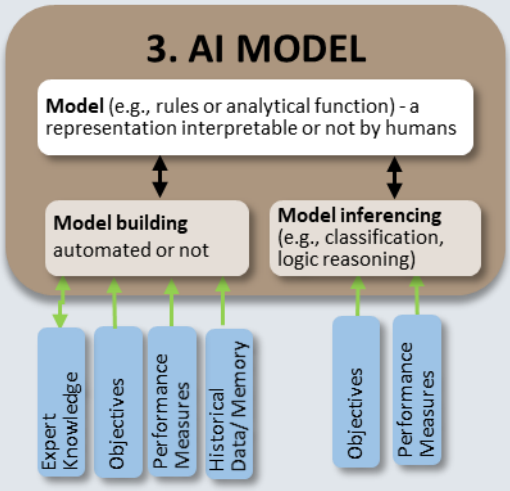
To examine and classify different types of AI systems and different scenarios, it is helpful to identify the following elements of an AI model (Figure 7):

- The ***model*** itself, an object that forms the core of an AI system and represents all or part of the system's external environment.

- The ***model-building*** process, often called "training" or "optimisation," that is part of the system development in the lab.

- The process of ***using the model***, in which model-inferencing algorithms generate outputs for information or action from the model, given specific objectives and performance measures. Model-inferencing tends to take place in production in the field.

Figure 7. Detailed conceptual view of an AI model



Source: (OECD, 2019f[2])

The purpose of an AI model within a deployed system: The model choice and model-building approach depend on the *purpose* of the AI system, i.e. the problem that the AI system is trying to solve. AI models can be used to make recommendations in response to an input, such as answering a question, "what is the best next move in a chess game?", generating data in response to a prompt, "what would a person look like if they were 20 years older?", making predictions about the future courses of action, "will this road become congested?" and a wide range of other processes (see Task of the system sub-section in the Task & Output section).

AI models are not universal: It is important to highlight that AI models include assumptions and biases about the world around them (OECD, 2019f[2]). Many different representations of the same environment can be developed to serve different purposes; there is no "universal" – unique or "correct" – model to represent a given reality.

One model or many models? Most applied AI systems use composite models: In practice, the vast majority of AI systems in deployed, real-world contexts are *composite systems*. They are composed of a variety of interlinked AI sub-models derived from different sources, working together for a specific purpose. For instance, Google Photos – a popular application for storing and searching photos – consists of several systems and subsystems, including those for storing data (photo storage), searching data via words that the user specifies, and interpreting the search terms which leverages additional systems, some of which are also AI models. When a user searches for photos by date, a human-designed system looks up a photo's upload date or the date encoded in the photo metadata. When a user searches for photos by a concept such as "flowers", an AI model carries out a recognition function to match photos with the query.

Why does this matter? Key properties of AI models, such as the degree of transparency and/or explainability (Principle 1.3); the level of robustness (Principle 1.4); and human rights, privacy and fairness implications (Principle 1.2), depend on the type of model as well as the model-building and inferencing processes (see questions in Box 3).

It is important to note that AI models can be built to achieve a specific set of objectives but then used with different objectives, as in the case of transfer learning, for example (Principle 1.4). Some leading experts stress the importance of very carefully specifying AI objectives to be the fulfilment of human goals rather than intelligence and efficiency (Russell, 2019[17]).

Box 3. Transparency and explainability (Principle 1.3) and safety, security and robustness (Principle 1.4) throughout an AI system's lifecycle

Risk assessment and management throughout the AI system lifecycle is the topic of a separate, follow-on facet of research (see section on Refining classification criteria). Possible questions to help determine AI system transparency and explainability (Principle 1.3) might include:

- Is it clear what the objectives of the AI system are, i.e. is it possible to formalise the problem that the system is being asked to solve?
- Does the AI system provide useful and meaningful information for understanding its performance and outputs/decisions?
- Can all of the AI system's outputs – both intermediary and final – for achieving a given goal be explained?
- Can the determinant data or knowledge that an AI system uses to make decisions be identified?
- Do two similar-looking cases verifiably result in similar outcomes, i.e. can the consistency and integrity of AI system outcomes be verified?[18]

Possible questions for policy makers to help determine the safety, security and robustness of AI systems might include:

- Do safety metrics exist that can evaluate the safety of an AI system for a given use case?
- How does the entity deploying the AI system test for safety during development?
- What measures has the entity deploying the AI system taken to do an adversarial evaluation – that is, explore the AI system through the lens of being a "bad actor" and trying to break it?
- Does the AI system change significantly if it is trained with variations of the data available?
- Are there measures in place to validate and verify the AI system's outcomes?
- What measures are in place to facilitate traceability in the AI system, including in relation to datasets, processes and decisions made during the AI system lifecycle?

AI model characteristics

Information availability

The amount of information, organised into the following basic categories available about the model(s) used in an AI system, can provide a first indication of its degree of transparency:

- *Detailed* information about the model(s) used in the system is available.
- *Some* information about the model(s) used in the system is available.
- *No* information about the model(s) used in the system is available.

Licensing rights associated with the model {where objective and consistent information is available}

A model can be available under open-source or proprietary licensing regimes. “Open-source software” (OSS) is software for which source code is public and can be freely copied, shared and modified (OECD, 2019c[18]). Models based on proprietary software aim to protect an organisation’s source code partially or fully, including through intellectual property regimes. The following are common types of licensing rights:

- Self-managed OSS
- Third-party managed OSS
- Self-managed proprietary
- Third-party managed proprietary

AI model types[19]

- *Symbolic AI models:* Symbolic or knowledge-based AI uses human-generated logical representations to infer a conclusion from a set of constraints (variables). These constraints include rules, ontologies and search algorithms and rely on explicit descriptions of variables – agents like humans, entities like factories, objects like machines, variables that can be stock conditions – and descriptions of the inter-relations between these variables. Symbolic models are expressed in languages such as mathematical logic (if/then statements or more abstract ways of representing knowledge via mathematical formulae), agent-based models, event-driven models, etc. (see example in Box 4). Symbolic AI is still in widespread use for optimisation and planning tools.
- *Statistical AI models*: Statistical AI models (e.g. genetic algorithms, neural networks and deep learning) identify patterns based on data rather than expert, human knowledge. They have seen increasing uptake recently. Statistical AI models were previously used primarily for recognition purposes (for instance, translating writing on cheques into machine-readable code). More recently, they are also being used for tasks like generation, such as synthesising and generating images or audio. Models that rely on data are designed to effectively extract and represent knowledge from data rather than to contain “explicit” knowledge – knowledge that is sharable and easily comprehensible.
- *Hybrid AI models*: Many applied AI systems combine symbolic and statistical models into “hybrid” models. For example, NLP algorithms often combine statistical approaches building on large amounts of data and symbolic approaches that consider issues such as grammar rules.

Box 4. Car engine production systems: Example of a symbolic AI model

One real-world example of a symbolic AI model is a car engine production system that involves different factories. In this example, each factory may have different machines assembling parts, operated by teams with different skills who are working on specific shifts. Parts can be sent from one factory to another using different logistics systems. The specificities and processing rules of this heterogeneous system (which comprises humans, factories, machines, stock, finances, etc.) are codified based on expert knowledge that describes each specific part of the system and its interactions with the rest. The AI model in this case is built, validated and calibrated based on this knowledge and can be used to simulate possible demand in order to optimise production mechanisms in response to demand volatility.

Why does this matter? An AI model's degree of "explainability" is determined primarily by the design of the AI model and is linked to the complexity of the system. The more complex the model, the harder it is to explain. Explainability means enabling people affected by the outcome of an AI system to understand how the outcome was arrived at.

Sub-models of rules-based, symbolic models are often easily understood, making it comparatively straightforward to find certain types of errors. By contrast, some machine-learning systems, notably, neural networks, rely on abstract mathematical relationships between factors that can be challenging or even impossible for humans to understand.

Hybrid AI systems that combine models built on both data and human expertise are viewed as a promising alternative that addresses the limitations of both system and machine-learning approaches. They can provide visibility on complex situations or environments with many interactions, help to predict what may happen in the future, and foster inclusive and sustainable growth and well-being (Principle 1.1).

Discriminative or generative models {where objective and consistent information is available}

- *Discriminative model*: Focuses on predicting data labels by learning to distinguish between dataset classes. They are more robust and capable of addressing outlier issues but cannot generate new data and can misclassify data points. Examples of discriminative models include regression analyses, support-vector machines (SVM), traditional neural networks, decision trees and random forests.
- *Generative model*: Involves discovering and learning the patterns and distribution of the input data, enabling the generation of new plausible examples that could be part of the original distribution. Examples of generative models include naïve Bayes models, hidden Markov models, linear discriminant analysis (LDA) and generative adversarial networks (GANs).
- *Models combining both discriminative and generative properties*: Designed to achieve a given goal or generate an output.

Why does this matter? For policy purposes, whether a model is discriminative or generative determines the type of output that it generates: Outputs from discriminative models are predictions whereas outputs from generative models are artefacts {where objective and consistent information is available}.

- *Model ensembles*: In some cases, the system is underpinned by an AI model interacting independently with other AI models. Model ensembles are collections of models that act together in parallel to cooperate on a single task or decision, which increases complexity but often improves accuracy.

- *System is composed of a single AI model*.

Why does this matter? A system involving a high degree of distinct, interacting systems may be more complex than a system composed of a single model, which often increases the probability of failures (Principle 1.4). It should be noted that systems are more and more often multi-tasking systems.[20] It is important to note that as models are combined, accuracy often improves but errors can propagate and multiply more easily, especially if uncertainty characterisations are not properly taken into account by downstream models.

Model-building

Model-building from machine-learned or human-encoded knowledge in the lab

The model-building process is often called "training" or "optimisation". Objectives (e.g. output variables) and performance measures (e.g. accuracy, resources for training and representativeness of the dataset) guide the model-building process. AI systems using machine learning for model-building have seen tremendous uptake over the past few years. Machine learning, as explained earlier in this report, is a set of techniques that allows machines to learn in an automated manner through patterns and inferences rather than through explicit instructions from a human. Machine-learning approaches often teach machines to reach an outcome by showing them many examples of correct outcomes. However, they can also define a set of rules and let the machine learn by trial and error. Machine learning contains numerous techniques that have been used for decades and range from linear and logistic regressions, decision trees and principle component analysis to deep neural networks. A machine-learning AI model can be built from data that is labelled or unlabelled, resulting in different machine-learning paradigms; whether data is labelled or not may already be inferred by the section on Format of data and metadata. When analysing AI systems, they can be roughly grouped into how much emphasis they place on human-encoded knowledge acquisition versus machine-learned knowledge acquisition:

- *Acquisition from human-encoded knowledge* (for example, writing rules): Human-written rules that capture relationships between elements of the environment by logical rules enable an AI model to deduce a conclusion from a set of constraints and data. They require that researchers build detailed and human-understandable decision structures to translate real-world complexity and help machines make decisions.
- *Acquisition from data through supervised learning*: AI models identify a relationship between input dimensions and labelled target dimensions.
- *Acquisition from data through unsupervised learning*: AI models identify a relationship between input data points based on their similarity.
- *Acquisition from data through semi-supervised learning*: AI models use both labelled and unlabelled data to identify a relationship between input dimensions and labelled target dimensions.
- *Acquisition from data through reinforcement-learning*: Does not require labelled input or data, nor suboptimal output to be corrected. Reinforcement learning leverages both systems' ability to explore current knowledge.
- *Acquisition from data, augmented by human-encoded knowledge*: Hybrid AI model systems combining human-encoded knowledge with knowledge acquired from data are common. For example, self-driving cars are frequently built using complex human-encoded rule sets that encode laws about how to drive – acceptable turns, speed limits, aspects related to braking speeds and tolerances, and so on. These rule sets are then combined with vision systems typically based on neural networks, which have acquired their capabilities via supervised learning on datasets annotated by the self-driving car companies.

Why does this matter? AI systems are only as good as the data they are trained on or the expert input they are built on. Machine-learning systems can make predictions about data similar to that on which they were trained, from which they derive associations and patterns that can fail in settings that are meaningfully different from those encountered in training.

Data-labelling, as explained earlier in this report, is the process of tagging data samples, which generally requires human knowledge to build training data. Data-labelling is critical and can itself require some explainability in contexts such as content moderation, where assigning a label such as "misinformation" or "violent" is important. In supervised learning, the label itself represents extra knowledge that is most often provided by "a human in the loop", while in unsupervised learning, a human did not label such content.

Expert systems have their own limitations as they require humans to build detailed decision structures to translate real-world complexity and help machines produce outputs.

Model evolution in the field (applicable only to machine learning systems)

Some machine-learning models can continue to evolve, acquiring abilities from interacting directly with data during the model-building process:

- *No evolution during operation* (no interaction): Dataset is static and does not change over time. An AI model is given a dataset and learns patterns or associations from it.
- *Evolution during operation through active interaction* (including uncontrolled learning): AI model actively interacts with the environment and receives data based on these interactions. An example of such a setting is a robot arm, which learns to perform a task, such as picking up a cup, by repeatedly attempting to perform the task and receiving feedback on which movements were successful and which movements were not.
- *Evolution during operation through passive interaction*: AI model receives a continuous stream of data (for example, stock prices), which the system is unable to affect but to which it needs to adapt.

Why does this matter? The degree to which a model evolves in response to data and input from its environment in the field is particularly relevant to public policy for AI systems that can iterate and evolve over time and may change their behaviour in unforeseen ways. Model evolution and model drift (where a model degrades because of changes in data, input or output) are directly relevant to safety, security and robustness (Principle 1.4) as well as accountability and liability (Principle 1.5).

AI models using static data are comparatively more stable. There may be a trade-off between the adaptive nature of an AI system (i.e. whether the model evolves in the field based on input from its environment) and the quality of its outcomes. This trade-off may be more acute with real-time data, as more conflicting data may arrive faster, creating a further risk of compromising the quality of the outcomes (Principle 1.4).

Central or federated learning (applicable only to machine-learning systems)

Models of machine-learning systems can be trained centrally or via multiple local servers or edge devices such as smartphones:

- *Centralised learning*: Uploads all datasets to a central processing environment to train an algorithm. All datasets are considered local to the training environment. Most current machine learning is centralised.
- *Federated (collaborative) learning*: Trains an algorithm across multiple processing environments, which can include edge devices or different data centres. Data samples are kept locally within each environment and are not copied across environments. There is no centralised, complete dataset with which the algorithm can train.[21]

Why does this matter? Federated learning helps to address critical issues like privacy (Principle 1.2), data security and data access rights (Principle 1.4) by building models without sharing data. It also distributes the computing requirements to train an AI system, although this may actually increase latency (processing delays).

Model development and maintenance {where objective and consistent information is available}

There are multiple approaches to AI model development and maintenance (BSA, 2021[19]), including:

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

Why does this matter? Understanding how an AI system's model was developed and/or maintained is a key consideration for assigning roles and responsibilities throughout a risk-management process. It is also relevant to assessing the system's robustness, security and safety (Principle 1.4) as well as accountability (Principle 1.5). For instance, the developer that trained and maintained the universal model on behalf of others would generally be best positioned to address most aspects of model risk management throughout the system's lifecycle. For customisable models, many key risk-management responsibilities would likely shift to the organisation that re-trained and/or customised the model. The bulk of risk-management responsibilities for tailored models fall on the entity that developed the model.

Model inference, or using a model

An AI model can be used in many different ways, and "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 (see Box 2, Figure 7). Different inference strategies can be used to derive varying results from the same model. These strategies are usually designed to optimise specific objectives and performance measures like robustness, accuracy, speed, business metrics or other criteria.

Deterministic inference is when the model's outcomes can be fully determined by the parameter values and random variation is not possible. Inference strategies include reasoning techniques used in expert systems. If random variation is present in the context in which the model operates, then *probabilistic*

inference may be more appropriate. There are a variety of methods for comparing and selecting different inference strategies.

Deterministic and probabilistic models {where objective and consistent information is available}

- *Deterministic models*: Follow precise rules ("if this, then that") and generate a single outcome.
- *Probabilistic models*: Infer several possible models to explain data and deciding which model to use is uncertain. Outcomes made using these different models are also uncertain. Probabilistic models quantify these uncertainties.[22] The different outcomes are associated with different levels of, for instance, performance measures like level of confidence, robustness or risk that can be optimised with different inferencing techniques.[23]
- *AI systems can combine both deterministic and probabilistic models*.

Why does this matter? For policy purposes, whether a model is probabilistic is relevant to testing and testability (Principle 1.4) as well as to explainability (Principle 1.3). Probabilistic models can generate multiple outcomes with information about their uncertainty. Given the randomness element in probabilistic models, a specific outcome may not easily be reproducible (Principle 1.3 and 1.4).

Model transparency and explainability

Different AI models can exhibit different degrees of transparency and explainability. Among other things, this entails determining whether meaningful and easy-to-understand information is made available to:

- Make stakeholders aware of their interactions with AI systems, including in the workplace
- Enable those affected by an AI system to understand and challenge the outcome and how it was produced by the AI model (e.g. by setting the weights of the AI model's components)

Why does this matter? An AI system's transparency and explainability (Principle 1.3) relates to how the model is used and how easily and thoroughly the structure and outputs of the system – i.e. understanding the link between input and output – can be understood, and by whom.

Key actors in the AI Model dimension include developers and modellers

Model-building and interpretation involve the creation or selection of models/algorithms, their calibration and/or training and inferencing (use). It also involves verification and validation, whereby models are executed and tuned (maximising performance) with tests to assess performance across various dimensions and considerations. Model-building and inferencing involve human experts such as modellers, model engineers, data scientists, developers and domain experts. Currently, model verification and validation involves data scientists, data/model/systems engineers and governance experts. Actions performed by developers and modellers include:

- *Model-building* by selecting and training a model
- *Verification and validation* to execute and tune models, including metrics to authorise the system for broader deployment

When it comes to testing to assess performance across various dimensions and considerations, characteristics of the team of AI system developers – such as gender, country, cultural background – have been shown to impact the way AI systems are built, as developers can incorporate unconscious biases (Freire, 2021[20]). This may result in advocacy for diversity in teams that create AI systems. Actors in the AI Model dimension may also include auditors.

Task & Output

The Task & Output dimension describes what the AI system does – the task it performs and the action that derives from it. The model produces recommendations, predictions or other outcomes for a given set of objectives, and a human or a machine (an "actuator") acts upon those recommendations, predictions or outcomes to influence the environment in which the AI system operates, at varying levels of autonomy.

Task(s) of the system

The task of an AI system drives the choice of AI model and refers to what the system does, i.e. the function that it performs.[24] The following seven categories cover most tasks performed by AI systems (Table 5):

- *Recognition*: Identifying and categorising data (e.g. image, video, audio and text) into specific classifications as well as image segmentation and object detection.
- *Event detection*: Connecting data points to detect patterns, as well as outliers or anomalies.
- *Forecasting*: Using past and existing behaviours to predict future outcomes.
- *Personalisation*: Developing a profile of an individual and learning and adapting its output to that individual over time.
- *Interaction support*: Interpreting and creating content to power conversational and other interactions between machines and humans (possibly involving multiple media such as voice, text and images).
- *Goal-driven optimisation*: Finding the optimal solution to a problem for a cost function or predefined goal.
- *Reasoning with knowledge structures*: Inferring new outcomes that are possible even if they are not present in existing data, through modelling and simulation.

The above categories are subject to change as AI technologies evolve and should be regularly reviewed and broadened to include new types of applications.

Table 5. AI system tasks

	What it does	Type of learning / reasoning	Examples
Recognition	Identifies and categorises data (e.g. image, video, audio and text) into specific classifications. Output is often one label, e.g. "this is a cat".	Supervised classification.	Image & object detection; facial recognition; audio, sound, handwriting and text recognition; gesture detection.
Event detection	Connects data points to detect patterns as well as outliers or anomalies.	Uses non-machine learning cognitive approaches as well as machine learning. Event detection increasingly uses unsupervised and reinforcement learning techniques in which the AI system does not know what it is looking for.	Fraud and risk detection, flagging human mistakes, intelligent monitoring.
Forecasting	Uses past and existing behaviours to predict future outcomes, generally to help make decisions. Contains a clear temporal dimension.	Tends to use machine-learning techniques – such as supervised learning – and is adaptive and helps improve forecasting over time. Some knowledge-based symbolic systems also perform forecasting, mostly based on uncertainty representation. Forecasting is generally used for decision support. It may include descriptive analytics, predictive analytics and projective analytics.	Assisted search, predicting future values for data, predicting failure, predicting population behaviour, identifying and selecting best fit, identifying matches in data, optimising activities, intelligent navigation.
Personalisation	Develops a profile of an individual and then learns and adapts to that individual over time. The output is usually a ranking, e.g. a search engine ranking.	Most personalisation algorithms are based on supervised or reinforcement learning models.	Recommender systems based on search and browsing (Netflix, Amazon), personalised fitness, wellness, finance.

Interaction support	Interprets and creates content to power conversational and other interactions between machines and humans (e.g. involving voice, text, images). Can be real-time or not.	Interaction tends to use semi-supervised or reinforcement learning, enabling models to evolve.	Chatbots, voice assistants, sentiments model and intent analysis.
Goal-driven optimisation	Gives systems a goal and the ability to find the optimal solution to a problem, which can be by learning through trial and error. It assumes a cost function is given.	Goal-driven optimisation is not necessarily a number. It can be called "prescription" when based on an optimisation.	Game playing, resource/logistics optimisation, iterative problem-solving, bidding and advertising, real-time auctions, scenario simulation.
Reasoning with knowledge structures	Infers new outcomes that are possible, even if they are not present in existing data, through modelling and simulation.	This task involves causal reasoning rather than correlation and uses AI techniques beyond machine learning.	Expert systems, legal argumentation, recruitment systems, diagnosis, planning.
Other	Please describe		

Source: OECD, 2022.

Why does this matter? A few policy considerations associated with the tasks performed by AI systems include:

Recognition systems require data that is representative and unbiased to function appropriately. Recognition of people and biometrics, such as facial recognition or voice recognition systems, can raise concerns in relation to human rights (Principle 1.2) and robustness and security in case of adversarial attacks (Principle 1.4).

Event detection can benefit people and planet (Principle 1.1), safety and security (Principle 1.4), yet in some contexts raises human rights concerns (Principle 1.2) when used to monitor individuals' activity.

In **forecasting,** depending on the application, keeping a human in the loop may be important for accountability (Principle 1.5).

Personalisation can impact social structures and well-being positively (Principle 1.1 benefits), but can also conflict with human values and individuals' right to self-determination (Principle 1.2.) as it tends to provide people with content that they have liked before or that similar people have liked; contributing to disinformation and echo-chamber effects.

Interaction support tasks in which an AI system interacts with people may implicate data usage and data privacy (Principle 1.2) and may require higher transparency and disclosure of the fact that one is interacting with a chatbot (Principle 1.3). It may also impact labour markets (Principle 2.4).

Goal-driven optimisation and other similar tasks using machine-learning algorithms that can learn from themselves through trial and error and may require humans in or "on the loop" (Principle 1.5). Limits on the power of this type of system may be needed if exponential growth occurs (e.g. artificial general intelligence). In addition, when goal specification is imperfectly defined, these systems may drift away from intended behaviour, thus potentially raising issues for robustness (Principle 1.5.).

Reasoning with knowledge structures is promising to help inclusive and sustainable growth and well-being (Principle 1.1), by allowing for the simulation of different scenarios considering causal and counterfactual relationships and situations that change with time, such as improving legacy power generation systems.[25]

Combining tasks and actions into multi-task, composite systems {where objective and consistent information is available}

AI systems frequently perform several tasks, such as event detection, forecasting and personalisation (see Table 5 for the full list), before producing an outcome that influences the environment. A composite AI is essentially an interlinked network of agents. Several composite systems that combine different tasks are common and well-known. They may generate specific policy considerations that differ from those produced by single-task systems. However, it may be difficult to anticipate and test composite AIs, which, like any complex system, will have unpredictable boundaries and impacts and will be hard or impossible to fit into a formula. For example, a hybrid AI can be planned and tested by the same developer, but the network of AIs cannot.

The following are examples of composite AI systems or application areas that combine several tasks and, in some cases, actions. The list is not exhaustive. When selecting the right system, users may add other relevant types of composite systems, as appropriate.

- *Content generation* (also referred to as synthesis): Includes generating new images, video, text, assessment and audio. This task combines forecasting and recognition tasks. However, the output often combines several existing elements such as images, text and audio to produce an object that was never seen before. This task tends to use structured learning. Examples of content generation include machine translation, generative art, news stories – including fake news, spam emails and "deep fake" videos.
- *Autonomous systems*: Robots and robotic systems increasingly embed different tasks – such as recognition and goal-driven optimisation – to perform an action in the real world. Similarly, autonomous systems perform a recognition task, which they use to try to find an optimal path to arrive at the best solution, and then act accordingly. In an autonomous vehicle, this could be a recommendation to turn left or right to minimise travel time, followed by the execution of the action. This system is autonomous in the sense that it does not require human supervision to act on its environment, but the way in which it performs its task (recognition) is not autonomous, as it relies on supervised learning.
- *Monitoring and control systems*: These systems manage, command, direct or regulate the behaviour of other devices or systems using control loops. Control systems generally assess environments through recognition, event detection or forecasting and propose a goal-driven action. They range from domestic heating controllers to large industrial control systems that use reinforcement learning to manage processes or machines. Control systems are common in the context of robotics or factories. One example of a control system is a fraud detection system with an event detection task combined with an action (e.g. freezing a bank account).
- *"IoAI", combining AI and Internet of Things (IoT)*: AI is allowing insights to be extracted from data. "Internet of Things" (IoT) refers to the connection of an increasing number of devices and objects over time to the Internet. Following the convergence of fixed and mobile networks, and between telecommunications and broadcasting, the IoT represents the next step in the convergence between Information and Communications Technologies (ICTs) and economies and societies. IoT is becoming common in daily life, with many billions of interconnected "smart" devices, equipment, machines and infrastructure creating opportunities for automation and for interaction in real time.

Why does this matter? From a policy perspective, content generation has relevant implications for human rights and democratic values (Principle 1.2), particularly when the content generated is realistic enough to be confused with "real" content. AI content generation amplifies the need to provide meaningful information to make stakeholders aware of their interactions with AI systems so that those affected understand and challenge the outcome (Principle 1.3; Principle 1.5). AI content generation also raises intellectual property rights questions (e.g. on the patentability of AI-assisted inventions and copyrighting of AI-generated creative work).

Autonomous systems and control systems have received increased attention and have direct implications for human safety (Principle 1.4) and accountability (Principle 1.5) as well as transparency and explainability.

Action autonomy level

A human or machine "actuator" (see Box 1) uses the outcome from the AI system (more specifically, the outcome from the inferencing process) to perform an action – prescribed by humans – that influences the environment in which the system operates. The way in which this action is performed determines the autonomy level of an AI system; that is, the degree to which a system can act without human involvement.

Below are four common variations in the degree of AI system autonomy, using a typology that originated in the field of aviation (Endsley, 1987[21]):

- *No-action autonomy* (also referred to as "human support"): System cannot act on its recommendations or output. The human uses or disregards the AI system's recommendations or output at will.
- *Low-action autonomy* (also referred to as "human-in-the-loop"): System evaluates input and acts upon its recommendations or output if the human agrees.
- *Medium-action autonomy* (also referred to as "human-on-the-loop"): System evaluates input and acts upon its recommendations or output unless the human vetoes.
- *High-action autonomy* (also referred to as "human-out-of-the-loop"): System evaluates input and acts upon its recommendations or output without human involvement.

It should be noted that for certain AI models with no, low or medium autonomy – such as active learning algorithms – the user or operator of the AI system may be contributing to its training and/or to validating its outputs.

Why does this matter? High-action autonomy systems pose important policy considerations, in particular when deployed in critical functions and activities or in contexts that may put human rights or fundamental values (Principle 1.2) at risk.

AI systems in which the user or operator may be contributing to its training and/or to the validation of its outputs raise transparency (Principles 1.3) and accountability (Principles 1.5) considerations.

Core application areas

Below are four core application areas but the list is not exhaustive. When selecting an AI system, users may consider other relevant applications areas, as appropriate.

- *Human language technologies*: Analyse, modify, produce or respond to human text and speech. Human language technologies may combine tasks like recognition, personalisation and interaction support.
- *Computer vision*: Concerned with training computers to interpret and understand the visual world. Feeding digital images and videos into deep learning models, machines can identify and classify objects and react to what they "see." Computer vision may include tasks like object recognition and event detection.
- *Robotics*: A system that contributes to the movement of robots. This involves the mechanical aspects and programme systems that make it possible to control robots, such as recognition and goal-driven optimisation – to perform an action in the real world. Autonomous vehicles are supported by robotic systems.
- *Automation and/or optimisation*: Process automation and/or simulation using structured data such as data mining, pattern recognition, a recommendation system or forecasting/prediction. Numerical

optimisation maximises or minimises a real function to optimise a business process such as scheduling, process controlling or operational research.

Evaluation methods {where objective and consistent information is available}

In some cases, there are agreed standards or general methods to assess an AI system or application within a given industry context and for a type of task(s):

- There are industry standards for evaluating AI systems as applied to this specific task and context.
- There are other methods for evaluating this AI system as applied to this specific task and context.
- There are no methods or industry standards available for evaluating this system.

Why does this matter? The availability of agreed standards or general methods to assess an AI system or application within a given industry context and for a type of task(s) relates to accountability (Principle 1.5) and robustness and safety concerns (Principle 1.4). The ability of operators and in some cases users to evaluate output can help identify incidents or instances of malfunctioning and assess the degree of reliability of the system, with a view to regulate or improve it.

Key actors in the Task & Output dimension include system integrators

The Task & Output dimension is often associated with the deployment stage of the AI system lifecycle (OECD, 2019f[2]). Deployment into live production involves piloting, checking compatibility with legacy systems, ensuring regulatory compliance, managing organisational change and evaluating user experience. Deployment currently involves experts such as system integrators, developers, systems/software engineers, testers and domain experts.

3 Applying the framework

Applying the framework to real-world systems with expert and survey input

The OECD AI Framework is designed to classify the application of AI systems in specific, real-world contexts. The OECD.AI Network Experts has classified four systems – systems 1, 2, 8 and 9 – out of the nine in the list below. The group invited a broad range of stakeholders from government, business, civil society and academia to test the framework's usability and robustness through an online survey and public consultation, whereby respondents used the framework to classify a selection of applied AI systems and AI systems (systems 1 to 7) or create an AI system classification of their choice (systems 8 and 9):

- **System 1 – Credit-scoring system**: Recommendation engine to help gauge a loan applicant's credit-worthiness. It does so by using human-based inputs (e.g. a set of rules) and data inputs (e.g. loan payments histories) to assess whether applicants are repaying loans on a regular basis.
- **System 2 – AlphaGo Zero**: Plays the board game Go better than professional human players. The board game's environment is virtual and player positions are constrained by the rules of the game. AlphaGo Zero uses both human-based inputs, including the rules of Go, and machine-based inputs, primarily data learned through repeated play against itself.
- **System 3 – Recommendation engine**: Assists consumers shopping online by generating personalised suggestions based on users' browsing history and data. For instance, Amazon currently uses item-to-item collaborative filtering, which scales to massive datasets and produces high-quality recommendations in real time. This type of filtering matches each of the user's purchased and rated items to similar items, then combines those similar items into tailored recommendations.
- **System 4 – Automated voice assistant**: Uses Natural Language Processing (NLP) to match user text or voice input to executable commands. Many continually learn using AI techniques including machine learning. Some of these assistants, like Google Assistant (which contains Google Lens) and Samsung Bixby, also have the added ability to do image processing to recognise objects in the image to help the users get better results from the clicked images.
- **System 5 – C-CORE** scans and processes satellite imagery over ocean areas to locate marine environmental structures or objects such as icebergs. The system determines object type, position and size of identified structures and automatically enters that information into a marine safety database.
- **System 6 – CASTER** reviews inputted molecular information of drugs for medical research purposes. The system is trained to recognize possible drug interactions and then models organic chemical reactions to predict drug-to-drug interactions, including potential harmful interactions.
- **System 7 – Face Image Quality**: FIQ is a tool for determining the quality of a digital face image. The system reviews a digital face image and produces a face image quality score, which is used by developers of facial recognition technologies to help determine the reliability of a face image collected through their technology.
- **System 8 –** Manufacturing plant hybrid management system, such as a Qlector.com LEAP system.
- **System 9 –** Generic AI system Generative Pre-trained Transformer 3 (GPT-3), taking into account that most of the functionality will depend on the final application context**.**

151 respondents provided a complete classification of one of the first seven systems on this list and another 18 respondents provided a complete classification of another AI application or system of their choice. They also provided feedback on the framework throughout the survey. Several questions were rephrased following respondents' feedback. Where the responses to criteria lacked consistency, the assumption was that responding required in-depth knowledge of a particular application; these criteria were annotated with "{where objective and consistent information is available}". See Table 6 for a breakdown of the survey responses.

Table 6. Analysis of AI system classification survey results, system examples 1-7

Framework dimension	Criterion	# possible answers	Average consistency	C-CORE	SCORE	CASTER	AlphaGo Zero	Voice assistant	FIQ	Recommendation engine	% uncertain or blank	low	medium	high
People & Planet	Users' AI competency	3	***63%***	Low	High	High	Low	High	Med	High	N/A	2	1	4
	Impacted stakeholders	2	***74%***	High	High	High	High	Low	Med	Med	N/A	1	2	4
	Optionality	2	***78%***	High	High	High	Low	Low	High	Med	25%	2	1	4
	Risks to human rights and democratic values	3	***70%***	High	High	High	High	Med	Med	Med	N/A	0	3	4
	Potential effects on people's well-being	3	***68%***	Med	Med	High	High	High	High	Med	8%	0	3	4
	Potential for human labour displacement	3	***64%***	Med	High	High	High	Med	High	Med	21%	0	3	4
Economic Context	Industrial sector	17	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
	Business function	14	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
	Business model	2	***76%***	High	High	Low	Low	High	Low	High	22%	3	0	4
	Impact on critical activities	2	***70%***	Med	Low	Low	High	Med	High	Med	4%	2	3	2
	Technical maturity (TRL)	3	***68%***	High	High	Low	High	High	High	High	7%	1	0	6
Data & Input	Detection and collection	3	***63%***	High	Low	High	Low	Low	Med	Low	8%	4	1	2
	Provenance of the data	2	***75%***	High	Low	High	Med	Med	High	Med	N/A	1	3	3
	Dynamic nature of the data	3	***60%***	High	High	Med	Low	Med	Low	Low	3%	3	2	2
	Rights	2	***72%***	High	High	Low	Med	Low	Med	Med	N/A	2	3	2
	Identifiability of personal data	2	***85%***	High	High	High	High	High	High	Low	N/A	1	0	6
	Structure of the data	3	***60%***	Low	High	Low	High	Low	Med	High	7%	3	1	3
	Format of the data	2	***66%***	Low	High	Low	Med	Low	Low	Med	N/A	4	2	1
	Scale of the data	3	***59%***	High	Low	Med	Low	Med	Med	Med	25%	2	4	1
	Appropriateness and quality of the data	3	***52%***	Med	Low	Low	High	Med	Med	Low	17%	3	3	1
Model	Information availability	3	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
	Type of AI model	3	***59%***	High	Med	Med	Med	High	Low	Low	15%	2	3	2
	Rights associated with model	4	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
	Single or multiple models	2	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
	Generative or discriminative	2	***80%***	Med	High	Med	Low	High	High	Med	31%	1	3	3
	Model building	3	***56%***	High	Low	High	Low	Med	High	Low	27%	3	1	3
	Model evolution[ML]	2	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
	Federated or central learning[ML]	2	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
	Development and maintenance	3	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
	Deterministic or probabilistic	2	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
	Model transparency	3	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
Task & Output	Task(s) performed by system	7	***78%***	High	High	Med	Med	Med	High	Med	8%	0	4	3
	Combining tasks and actions	2	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
	System's level of autonomy	3	***56%***	Low	Low	Med	High	Med	Med	Low	10%	3	3	1
	Degree of human involvement	3	***58%***	Low	Low	High	High	Med	Med	Low	15%	3	2	2
	Core application	4	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
	Evaluation	3	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
Number of fully completed surveys				**7**	**26**	**10**	**28**	**17**	**12**	**51**				

Note: When there are two possible survey responses (e.g. Yes or No), "High" consistency means that over 75% of responses are the same, while "Medium" consistency means that over 66% of responses are the same. When there are three possible responses (e.g. Increase/Same/Decrease), "High" consistency means that over 65% of responses are the same while "Medium" consistency means that over 50% of responses are the same. When there are multiple questions for one criterion (e.g. impacted stakeholders), "consistency" refers to the average consistency per answer. ML = for machine-learning models.

Source: Based on the 151 surveys that were fully completed, out of a total of over 850 surveys that were received in June 2021 (700 of which were partially completed). Builds on (Aiken, 2019[22]).

These key conclusions from the survey responses are reflected throughout the report:

- The framework is best suited to specific applications of AI systems rather than to generic AI systems; thus, classifying an AI system using the framework requires more information on the specific application, context of use and more.
- The more specific the applications (e.g. credit-scoring system, AlphaGo Zero), the more consistent the survey responses, as many of the criteria pertain to specific application areas. The more general the systems (e.g. voice assistant, recommendation engine) the less consistent the responses.
- Respondents were significantly better at classifying criteria in the People & Planet and Economic Context dimensions consistently than the criteria in the other, more technical dimensions.
- Classifying technical characteristics requires more information than is typically available about an AI system; they tended to generate many uncertain or blank responses.
- The following sections delve deeper into how the four examples – Systems 1, 2, 8 and 9 – align with each of the dimensions of the OECD Framework for the Classification of AI Systems.

System 1: Credit-scoring system

A credit-scoring system is representative of a machine-based system that influences its environment (whether people are granted a loan). It makes recommendations (a credit score) for a given set of objectives (that, together, determine credit-worthiness). It does so by using both machine-based inputs (historical data on people's profiles and on whether they repaid loans) and human-based inputs (a set of rules). With these two sets of inputs, the system perceives real environments (whether people have replayed loans in the past or whether they are repaying loans on an ongoing basis). It transforms such perceptions into models automatically. A credit-scoring algorithm could use a statistical model, for example. Finally, it uses model inferencing (the credit-scoring algorithm) to formulate a recommendation (a credit score) of options for outcomes (providing or denying a loan).

People & Planet

Core characteristic	Survey question	Response
Users of AI system	What is the level of competency of users who interact with the system?	Amateur (bank employee)
Impacted stakeholders	Who is impacted by the system (e.g. consumers, workers, government agencies)?	Consumers, the bank
Optionality and redress	Can users opt out, e.g. switch systems? Can users challenge or correct the output?	Not optional / cannot opt out
Human rights and democratic values	Can the system's outputs impact fundamental human rights?	*Possible impact on*: - rule of law; absence of arbitrary sentencing - equality and non-discrimination - right to property - economic and social rights
Well-being, society and the environment	Can the system's outputs impact areas of life related to well-being (e.g. job quality, the environment, health, social interactions, civic engagement, education)?	*Possible impact on*: - physical and mental health - work and job quality - quality of environment - social connections - civic engagement - education - work-life balance
{Displacement potential}	*Could the system automate tasks that are or were being executed by humans?*	*Depends on deployment context*

Economic Context

Core characteristic	Survey question	Response
Industrial sector	Which industrial sector is the system deployed in (e.g. finance, agriculture)?	Section K: Financial and insurance activities (per ISIC REV 4)
Business function	What business function(s) or functional areas is the AI system employed in (e.g. sales, customer service, human resources)?	Sales, customer service
Business model	Is the system a for-profit use, non-profit use or public service system?	For-profit use – other model
Impacts critical functions / activities	Would the disruption of the system's function or activity affect essential services?	Yes
Breadth of deployment	Is the AI system deployment a pilot, narrow, broad or widespread?	Broad
{Technical maturity}	*How technically mature is the system (Technology Readiness Level –TRL)?*	*Actual system or subsystem in final form in operational environment – TRL 9*

Data & Input

Core characteristic	Survey question	Response
Detection and collection	Are the data and input collected by humans, automated sensors, both?	Humans (set of rules) and automated sensing devices (e.g. loan payments)
Provenance of data and input	Are the data and input from experts; provided, observed, synthetic or derived?	- provided by experts (rules) - provided by loan candidate (e.g. personal information) - observed by the algorithm (e.g. history of payments) - derived data (e.g. credit rating and other scores)
Dynamic nature	Are the data dynamic, static, dynamic updated from time to time or real-time?	- static (e.g. gender); - dynamic data updated from time to time (e.g. salary) - dynamic real-time data (e.g. day-to-day payments
Rights associated with data and input	Are the data proprietary (privately held), public (no intellectual property rights) or personal data (related to identifiable individual)?	Personal and proprietary
Identifiability of personal data	If personal data, are they anonymised or pseudonymised?	Identified data
{Data quality and appropriateness}	*Is the dataset fit for purpose? Is the sample size adequate? Is it representative and complete enough? How noisy are the data?*	*- quality unknown* *- appropriate data*
{Structure of the data and input}	*Are the data structured, semi-structured, complex structured or unstructured?*	*Structured data*
{Format of data and metadata}	*Is the format of the data and metadata standardised or non-standardised?*	*Standardised*
{Scale}	*What is the dataset's scale?*	*Small or medium*

AI Model

Core characteristic	Survey question	Response
Model information availability	Is any information available about the system's model?	Yes
AI model type	Is the model symbolic (human-generated rules), statistical (uses data) or hybrid?	Hybrid
{Rights associated with model}	*Is the model open-source or proprietary, self or third-party managed?*	*Proprietary*
{Discriminative or generative}	*Is the model generative, discriminative or both?*	*Discriminative (score as a probability)*
{Single or multiple model(s)}	*Is the system composed of one model or several interlinked models?*	*Yes*
Model-building from machine or human knowledge	Does the system learn based on human-written rules, from data, through supervised learning or through reinforcement learning?	Acquisition from data, augmented by human-encoded knowledge
Model evolution in the field (applicable only to machine-learning systems)	Does the model evolve and / or acquire abilities from interacting with data in the field?	Evolution during operation through passive interaction
Central or federated learning (applicable only to machine-learning systems)	Is the model trained centrally or in a number of local servers or edge devices?	Central
{Model development and maintenance}	*Is the model universal, customisable or tailored to the AI actor's data?*	*Context-dependent*
{Deterministic and probabilistic}	*Is the model used in a deterministic or probabilistic manner?*	*Deterministic*
Transparency and explainability	Is information available to users to allow them to understand model outputs?	Context-dependent

Task & Output

Core characteristic	Survey question	Response
Task(s) of the system	What tasks does the system perform (e.g. recognition, event detection, forecasting)?	Forecasting, reasoning with knowledge structures
{Combining tasks and actions into composite systems}	*Does the system combine several tasks and actions (e.g. content generation systems, autonomous systems, control systems)?*	*Yes*
Action autonomy	How autonomous are the system's actions and what role do humans play?	Low autonomy
Core application area(s)	Does the system belong to a core application area such as human language technologies, computer vision, automation and / or optimisation or robotics?	Human language technologies
{Evaluation methods}	*Are there standards or methods available for evaluating system output?*	*Yes*

System 2: AlphaGo Zero

AlphaGo Zero is an AI system that plays the board game Go better than any professional, human Go players. The board game's environment is virtual and fully observable. Game positions are constrained by the objectives and the rules of the game. AlphaGo Zero is a system that uses both human-based inputs (the rules of the game) and machine-based inputs (learning based on playing iteratively against itself, starting from completely random play). It abstracts the data into a stochastic (randomly determined) model of actions ("moves" in the game) and is trained via so-called reinforcement learning. Finally, it uses the model to propose a new move based on the state of play.

People & Planet

Core characteristic	Survey question	Response
Users of AI system	What is the level of competency of users who interact with the system?	Expert practitioner (e.g. DeepMind engineers)
Impacted stakeholders	Who is impacted by the system (e.g. consumers, workers, government agencies)?	None for now; if deployed in production, specific communities (e.g. Go players)
Optionality and redress	Can users opt out, e.g. switch systems? Can users challenge or correct the output?	If deployed in production, some optionality
Human rights and democratic values	Can the system's outputs impact fundamental human rights?	No
Well-being, society and the environment	Can the system's outputs impact areas of life related to well-being (e.g. job quality, the environment, health, social interactions, civic engagement, education)?	No
{Displacement potential}	*Could the system automate tasks that are or were being executed by humans?*	*Low displacement potential (TBD)*

Economic Context

Core characteristic	Survey question	Response
Industrial sector	Which industrial sector is the system deployed in (e.g. finance, agriculture)?	Section R: Arts, Entertainment, and Recreation (per ISIC REV 4)
Business function	What business function(s) or functional areas is the AI system employed in (e.g. sales, customer service, human resources)?	N/A
Business model	Is the system a for-profit use, non-profit use or public service system?	Non-profit (outside public sector) or for profit
Impacts critical functions / activities	Would the disruption of the system's function or activity affect essential services?	No
Breadth of deployment	Is the AI system deployment a pilot, narrow, broad or widespread?	Narrow
{Technical maturity}	*How technically mature is the system (Technology Readiness Level –TRL)?*	*System or subsystem complete and qualified through test and demonstration – TRL 8*

Data & Input

Core characteristic	Survey response	Response
Detection and collection	Are the data and input collected by humans, automated sensors, both?	Humans (the rules of the game of Go) and automated sensing devices
Provenance of data and input	Are the data and input from experts; provided, observed, synthetic or derived?	Provided by experts (the rules of the game of Go), observed by the algorithm, and synthetic data
Dynamic nature	Are the data dynamic, static, dynamic updated from time to time or real-time?	Static (human knowledge) and dynamic, real-time data (each move in the game)
Rights associated with data and input	Are the data proprietary (privately held), public (no intellectual property rights) or personal data (related to identifiable individual)?	Public and proprietary
Identifiability of personal data	If personal data, are they anonymised; pseudonymised?	N/A
{Data quality and appropriateness}	*Is the dataset fit for purpose? Is the sample size adequate? Is it representative and complete enough? How noisy are the data?*	*Representative and appropriate, low noise/missing values/outliers*
{Structure of the data and input}	*Are the data structured, semi-structured, complex structured or unstructured?*	*Complex structured*
{Format of data and metadata}	*Is the format of the data and metadata standardised or non-standardised?*	*Standardised and non-standardised*
{Scale}	*What is the dataset's scale?*	*TBD (large or very large)*

AI Model

Core characteristic	Survey question	Response
Model information availability	Is any information available about the system's model?	Yes
AI model type	Is the model symbolic (human-generated rules), statistical (uses data) or hybrid?	Hybrid
{Rights associated with model}	*Is the model open-source or proprietary, self or third-party managed?*	*Proprietary*
{Discriminative or generative}	*Is the model generative, discriminative or both?*	*Generative*
{Single or multiple model(s)}	*Is the system composed of one model or several interlinked models?*	*One model*
Model-building from machine or human knowledge	Does the system learn based on human-written rules, from data, through supervised learning or through reinforcement learning?	Acquisition from data, augmented by human-encoded knowledge
Model evolution in the field ML	Does the model evolve and / or acquire abilities from interacting with data in the field?	Evolution during operation through passive interaction
Central or federated learning ML	Is the model trained centrally or in a number of local servers or edge devices?	Central
{Model development and maintenance}	*Is the model universal, customisable or tailored to the AI actor's data?*	*Context-dependent*
{Deterministic and probabilistic}	*Is the model used in a deterministic or probabilistic manner?*	*Both*
Transparency and explainability	Is information available to users to allow them to understand model outputs?	Context-dependent

Task & Output

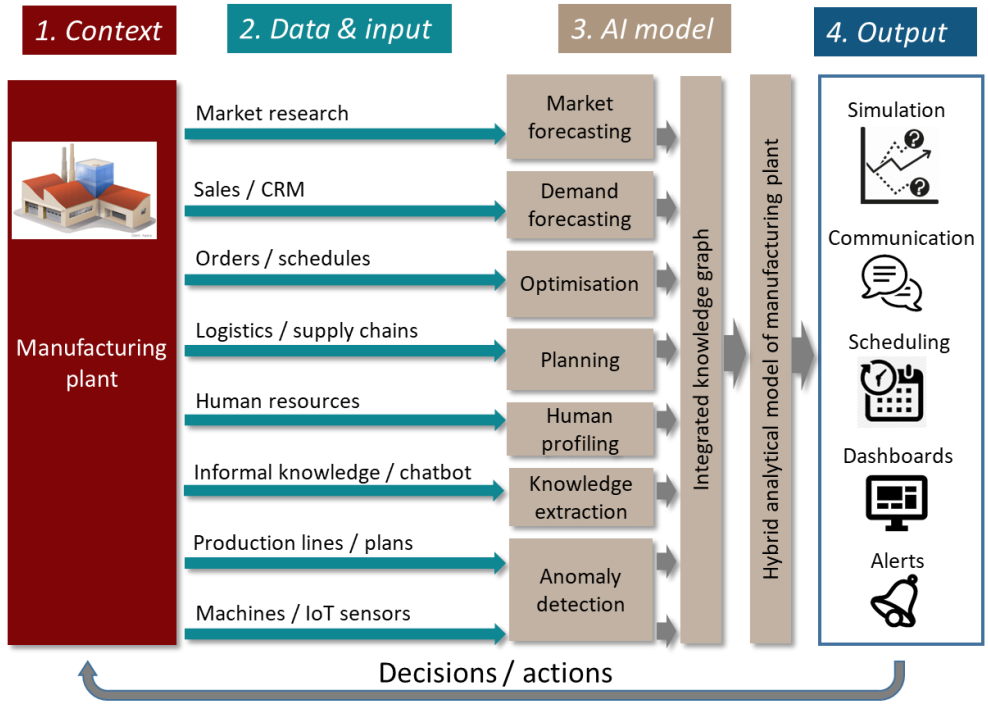
Core characteristic	Survey question	Response
Task(s) of the system	What tasks does the system perform (e.g. recognition, event detection, forecasting)?	Forecasting, goal-driven optimization, reasoning with knowledge structures
{Combining tasks and actions into composite systems}	*Does the system combine several tasks and actions (e.g. content generation systems, autonomous systems, control systems)?*	*Yes*
Action autonomy	How autonomous are the system's actions and what role do humans play?	High
Core application area(s)	Does the system belong to a core application area such as human language technologies, computer vision, automation and / or optimisation or robotics?	Human language technologies and/or computer vision (TBC)
{Evaluation methods}	*Are there standards or methods available for evaluating system output?*	*Yes*

System 3: Qlector.com LEAP system to manage a manufacturing plant

An AI system controlling and running a manufacturing plant is a representative example of a complex hybrid AI system, with the specific context of the physical manufacturing plant factory floor. Different AI models, associated with different data sources perform particular activities for the factory based on different types of data and input. These modelling activities include: event detection (anomaly detection based on data from machines in production lines); goal-driven optimisation (based on orders and schedules, logistics and supply chain data); reasoning with knowledge structures (simulations); interaction support (customer relationship management); demand forecasting (based on sales); and strategic market forecasting (based on market research).

For illustrative purposes, all of these models can be combined into a large, evolving knowledge graph with a symbolic AI type of data structure that interconnects the different tasks performed at the factory (Figure 8). The resulting AI model is a hybrid analytical model of a manufacturing plant that some could refer to as a "digital twin" of the factory. The outputs of the model include: alerts, information on dashboards, scheduling, communications with customers, and simulations of possible futures to inform decisions. While humans are often involved in the actions resulting from the system outputs, factory processes are increasingly autonomous. The output/decision feeds back into the context/physical environment.

Figure 8. AI system to help manage a manufacturing plant



Source: Example taken from Qlector.com LEAP system

People & Planet

Core characteristic	Survey question	Response
Users of AI system	What is the level of competency of users who interact with the system?	Amateurs, non-expert and expert practitioners
Impacted stakeholders	Who is impacted by the system (e.g. consumers, workers, government agencies)?	Consumers, workers / employees, business
Optionality and redress	Can users opt out, e.g. switch systems? Can users challenge or correct the output?	Varies
Human rights and democratic values	Can the system's outputs impact fundamental human rights?	*Possible impact on*: - physical, psychological and moral integrity - equality and non-discrimination
Well-being, society and the environment	Can the system's outputs impact areas of life related to well-being (e.g. job quality, the environment, health, social interactions, civic engagement, education)?	*Possible impact on*: - health - income and wealth - environmental quality - social connections - work and job quality - work-life balance
{Displacement potential}	*Could the system automate tasks that are or were being executed by humans?*	*High displacement potential*

Economic Context

Core characteristic	Survey question	Response
Industrial sector	Which industrial sector is the system deployed in (e.g. finance, agriculture)?	Section C: Manufacturing (per ISIC REV 4)
Business function	What business function(s) or functional areas is the AI system employed in, (e.g. sales, customer service, human resources)?	Many, including sales, customer service, planning and budgeting, procurement, logistics, human resource management, monitoring and quality control, production, maintenance
Business model	Is the system a for-profit use, non-profit use or public service system?	For-profit use
Impacts critical functions / activities	Would the disruption of the system's function or activity affect essential services?	Depends on the type of goods manufactured
Breadth of deployment	Is the AI system deployment a pilot, narrow, broad or widespread?	Narrow
{Technical maturity}	*How technically mature is the system (Technology Readiness Level –TRL)?*	*Actual system or subsystem in final form in operational environment – TRL 9*

Data & Input

Core characteristic	Survey question	Response
Detection and collection	Are the data and input collected by humans, automated sensors, both?	Humans and automated sensing devices
Provenance of data and input	Are the data and input from experts; provided, observed, synthetic or derived?	All (expert input, provided data, observed data, synthetic data and derived data)
Dynamic nature	Are the data dynamic, static, dynamic, updated from time to time or real-time?	Static (human knowledge) and dynamic real-time data (data from machines in production lines)
Rights associated with data and input	Are the data proprietary (privately held), public (no intellectual property rights) or personal data (related to identifiable individual)?	Proprietary
'Identifiability' of personal data	If personal data, are they anonymised, pseudonymised?	N/A
{Data quality and appropriateness}	*Is the dataset fit for purpose? Is the sample size adequate? Is it representative and complete enough? How noisy are the data?*	*Representative and appropriate, noise /missing values/outliers*
{Structure of the data and input}	*Are the data structured, semi-structured, complex structured or unstructured?*	*All (unstructured, semi-structured, unstructured, complex structured)*
{Format of data and metadata}	*Is the format of the data and metadata standardised or non-standardised?*	*Standardised and non-standardised*
{Scale}	*What is the dataset's scale?*	*Medium*

AI Model

Core characteristic	Survey question	Response
Model information availability	Is any information available about the system's model?	Yes
AI model type	Is the model symbolic (human-generated rules), statistical (uses data) or hybrid?	Hybrid
{Rights associated with model}	*Is the model open-source or proprietary, self or third-party managed?*	*Proprietary*
{Discriminative or generative}	*Is the model generative, discriminative or both?*	*Discriminative and generative*
{Single or multiple model(s)}	*Is the system composed of one model or several interlinked models?*	*Yes*
Model building from machine or human knowledge	Does the system learn based on human-written rules, from data, through supervised learning or through reinforcement learning?	Acquisition from data, augmented by human-encoded knowledge
Model evolution in the field ML	Does the model evolve and / or acquire abilities from interacting with data in the field?	Evolution during operation through active and passive interaction
Central or federated learning ML	Is the model trained centrally or in a number of local servers or edge devices?	Federated
{Model development and maintenance}	*Is the model universal, customisable or tailored to the AI actor's data?*	*Context-dependent*
{Deterministic and probabilistic}	*Is the model used in a deterministic or probabilistic manner?*	*Both*
Transparency and explainability	*Is information available to users to allow them to understand model outputs?*	*Context-dependent*

Task & Output

Core characteristic	Survey question	Response
Task(s) of the system	What tasks does the system perform (e.g. recognition, event detection, forecasting)?	All (recognition, event detection, forecasting, interaction support, goal-driven optimization, reasoning with knowledge structures)
{Combining tasks and actions into composite systems}	*Does the system combine several tasks and actions (e.g. content generation systems, autonomous systems, control systems)?*	*Yes*
Action autonomy	How autonomous are the system's actions and what role do humans play?	Medium autonomy
Core application area(s)	Does the system belong to a core application area such as human language technologies, computer vision, automation and / or optimisation or robotics?	Human language technologies, robotics, computer vision, optimisation
{Evaluation methods}	*Are there standards or methods available for evaluating system output?*	*Yes*

System 4: GPT-3

GPT-3 is a large, pre-trained language model that has the capacity to search across, generate and manipulate strings of text. The model can take in arbitrary inputs, in the form of text strings, which lead to it generating an output. GPT-3 can be conditioned with up to 2048 distinct characters, which enable it to learn from the examples it is primed with.

GPT-3 is a general purpose AI system, meaning it can theoretically be used to deploy applications in any sector of the economy. Such applications would need to be considered within their specific socio-economic context; for example, a creative-writing application built with GPT3 should be treated differently from one that seeks to give a user medical advice in response to a query. Examples of GPT3 use cases include text classification activities to search across news articles and generation of emails from a summary sentence. The example of creative writing is applied to the classification framework below.

People & Planet

Core characteristic	Survey question	Response
Users of AI system	What is the level of competency of users who interact with the system?	Amateur
Impacted stakeholders	Who is impacted by the system (e.g. consumers, workers, government agencies)?	Workers (e.g. could lead to automation of some tasks), consumers
Optionality and redress	Can users opt out, e.g. switch systems? Can users challenge or correct the output?	Optional / can opt out
Human rights and democratic values	Can the system's outputs impact fundamental human rights?	*Possible impact on*: - rule of law, absence of arbitrary sentencing - freedom of thought, conscience and religion - equality and non-discrimination - quality of democratic institutions (e.g. free elections)
Well-being, society and the environment	Can the system's outputs impact areas of life related to well-being (e.g. job quality, the environment, health, social interactions, civic engagement, education)?	*Possible impact on*: - work and job quality - education
{Displacement potential}	*Could the system automate tasks that are or were being executed by humans?*	*TBD*

Economic Context

Core characteristic	Survey question	Response
Industrial sector	Which industrial sector is the system deployed in (e.g. finance, agriculture)?	Section J: Information and Communication (per ISIC REV 4)
Business function	What business function(s) or functional areas is the AI system employed in (e.g. sales, customer service, human resources)?	Any
Business model	Is the system a for-profit use, non-profit use or public service system?	For-profit use – other model (e.g. business intelligence) or non-profit use (e.g. research, journalism)
Impacts critical functions / activities	Would the disruption of the system's function or activity affect essential services?	No
Breadth of deployment	Is the AI system deployment a pilot, narrow, broad or widespread?	Narrow deployment
{Technical maturity}	*How technically mature is the system (Technology Readiness Level –TRL)?*	*System prototype demonstration in operational environment – TRL 7*

Data & Input

Core characteristic	Survey question	Response
Detection and collection	Are the data and input collected by humans, automated sensors, both?	Collected by humans and automated sensing devices (e.g. collected by humans with subsequent filtering by machines and humans)
Provenance of data and input	Are the data and input from experts; provided, observed, synthetic or derived?	Observed and derived
Dynamic nature	Are the data dynamic, static, dynamic updated from time to time or real-time?	Dynamic data updated from time to time
Rights associated with data and input	Are the data proprietary (privately held), public (no intellectual property rights) or personal data (related to identifiable individual)?	Public and proprietary
Identifiability of personal data	If personal data, are they anonymised, pseudonymised?	N/A
{Data quality and appropriateness}	*Is the dataset fit for purpose? Is the sample size adequate? Is it representative and complete enough? How noisy are the data?*	*Noisy data, that is, by design, highly representative and diverse with regard to a large part of (predominantly English) text and code found on the Internet; appropriate data*
{Structure of the data and input}	*Are the data structured, semi-structured, complex structured or unstructured?*	*Unstructured data*
{Format of data and metadata}	*Is the format of the data and metadata standardised or non-standardised?*	*Non-standardised*
{Scale}	*What is the dataset's scale?*	*Very large*

AI Model

Core characteristic	Survey question	Response
Model information availability	Is any information available about the system's model?	Yes
AI model type	Is the model symbolic (human-generated rules), statistical (uses data) or hybrid?	Statistical (data-driven)
{Rights associated with model}	*Is the model open-source or proprietary, self or third-party managed?*	*Proprietary*
{Discriminative or generative}	*Is the model generative, discriminative or both?*	*Generative*
{Single or multiple model(s)}	*Is the system composed of one model or several interlinked models?*	*One*
Model-building from machine or human knowledge	Does the system learn based on human-written rules, from data, through supervised learning or through reinforcement learning?	Acquisition from data, augmented by human-encoded knowledge
Model evolution in the field ML	Does the model evolve and / or acquire abilities from interacting with data in the field?	Evolution during operation through passive interaction
Central or federated learning ML	Is the model trained centrally or in a number of local servers or edge devices?	Central
{Model development and maintenance}	*Is the model universal, customisable or tailored to the AI actor's data?*	*Context-dependent*
{Deterministic and probabilistic}	*Is the model used in a deterministic or probabilistic manner?*	*Deterministic*
Transparency and explainability	*Is information available to users to allow them to understand model outputs?*	*Context-dependent*

Task & Output

Core characteristic	Survey question	Response
Task(s) of the system	What tasks does the system perform (e.g. recognition, event detection, forecasting)?	Reasoning with knowledge structures, interaction support, recognition, personalisation
{Combining tasks and actions into composite systems}	*Does the system combine several tasks and actions (e.g. content generation systems, autonomous systems, control systems)?*	*Yes*
Action autonomy	How autonomous are the system's actions and what role do humans play?	Low autonomy
Core application area(s)	Does the system belong to a core application area such as human language technologies, computer vision, automation and / or optimisation or robotics?	Human language technologies
{Evaluation methods}	*Are there standards or methods available for evaluating system output?*	*TBD*

4 Next steps

Refining classification criteria based on real-world evidence

In an effort to add value to the framework and make it more actionable and meaningful to stakeholders looking to incorporate an AI system into their activities, next steps for the OECD Experts Working Group include populating the current classification system with actual AI systems and refining the specific classification criteria based on this evidence.

The refinement process includes identifying or developing metrics or proxies to help assess subjective criteria such as impact on human rights and well-being. It is important to point out that there may be a trade-off between developing a simple, user-friendly assessment (which was the goal of the initial classification framework exercise) and a very accurate assessment, as the latter may require significant in-depth information on an AI system that may be unknown to the average user. Some contexts may require more detailed follow-up questions to assess AI systems that are not relevant in others. If the potential impacts are large, the focus may need to be on potential biases in the data or modelling process and focus, or on data representativeness, that is, whether such data can impact the decisioning system. For example, whether the ethnicity of a car owner might impact an insurance claim evaluation, whether the model should move to an unexamined state, etc.

Tracking AI incidents

A related next step is to develop a common framework for reporting AI incidents, especially those that are negative or harmful, or controversies. The incidents framework would leverage the classification framework and help to ensure global consistency and interoperability in incident reporting. It would be part of a Global AI Incidents Tracker at the OECD, with the contributions of partner institutions, to build the evidence base about risks that have materialised into incidents or near incidents.

Developing a risk assessment framework

The classification framework presented in this report describes key aspects of an applied AI system, including the various contexts in which it impacts the real world, the nature and type of data and input, the various AI models, and the types of tasks it executes and output it produces. Information derived from this framework for a specific use case, augmented by information on governance and risk mitigation processes, could be useful in assessing the associated ethical and societal risks of an applied AI system, which may have considerable practical significance for stakeholders in numerous contexts – policy making, business, etc.

Box 5. AI risk-based approaches to AI system application

Policy makers favour a risk-based approach to regulating AI in order to focus oversight and intervention where it is most needed, while avoiding unnecessary hurdles to innovation. The OECD AI Principles state that *"AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias"*.

The risks in using any AI system strongly depend on the application. Since it is difficult to anticipate and assess every possible use case, applied AI systems should be grouped into a small collection of risk levels. As part of the EC AI Act, the European Commission put forward four levels of risk: unacceptable, high, limited and minimal. Various academic groups as well as expert panels (e.g. the German Data Ethics Commission and IEC SEG10) have proposed four to five risk levels. And the ISO, National Institute of Standards and Technology (NIST), Institute of Electrical and Electronics Engineers (IEEE) and others are working on risk-assessment and risk-management frameworks from different angles and objectives.

Regardless of the number of risk levels or which organisation proposes them, the following are typical criteria for determining the risk level of an AI application or system:

- Scale, i.e. seriousness of adverse impacts (and probability)
- Scope, i.e. breadth of application, such as the number of individuals that are or will be affected
- Optionality, i.e. degree of choice as to whether to be subject to the effects of an AI system

Part of the OECD's value add is its capability to involve and coordinate with several groups working on AI risk assessment and management so as to promote international interoperability in designing technical, policy and governance AI risk frameworks. The OECD Experts Working Group, with members from across sectors and professions, plans to conduct further analysis of the criteria to include in a risk assessment and how best to aggregate these criteria, taking into account that different criteria may be interdependent. The group will use examples of AI systems in clearly different risk categories to assess the usefulness of different criteria and to try to calibrate these criteria in an empirical way where possible, leveraging evidence from its Global AI Incidents Tracker.

The next phase of work is expected to produce an actionable AI system risk methodology, which will build on the current AI system classification framework and on other work streams taking place in partner organisations, as well as, for example, the OECD AI Network of Experts Expert Group on Trustworthy AI. It should be emphasised that discussions about handling AI risks rely on and are complementary to existing, well-established risk assessment frameworks, e.g. for functional and product safety (OECD, 2016[23]) or digital security (OECD, 2015[24]), as well as existing frameworks for quality management systems; hence, these discussions tend to focus on ethical and societal risks. Existing human rights and responsible business-impact assessment guidelines are also directly relevant (OECD, 2018[25]). As of the development of this report, coordination with these groups and with partner organisations has begun.

Annex A. Sample AI applications by sector, ordered by diffusion

Sector (per ISIC REV 4)	Description	Main applications of AI
Information and communication (Section J)	Includes the production and distribution of information and cultural products, the provision of the means to transmit or distribute these products, as well as data or communications, information technology activities and the processing of data and other information service activities.	Advertising Image or text processing Personalised content generation Augmented and virtual reality Customer services Network security Network management, predictive maintenance Software production
Professional, scientific and technical activities (Section M)	Includes specialised activities that require a high degree of training and make specialised knowledge and skills available to users including legal affairs, management, consultancy, architecture, engineering, R&D, advertising and more.	Legal and accounting AI applications Marketing and advertising services (e.g. personalised advertising and pricing, click prediction systems, recommendations based on social media posts, emails, web navigation, psychometric assessment development, etc.) Scientific research and development
Financial and insurance activities (Sector K)	Includes financial service activities, insurance, reinsurance and pension funding and activities to support financial services, funds and holdings.	Credit scoring Financial technology lending Cost reduction in the front and middle office Fraud detection and legal compliance Insurance Algorithmic trading
Administrative and support service activities (Section N)	Includes a variety of activities that support core business functions and of which the primary purpose is not the transfer of specialised knowledge. This includes security services, renting and leasing, office administrative functions and reservation services.	Auditing expense reports Hiring applications Smart contracts Customer relations
Agriculture, forestry and fishing (Section A)	Includes the exploitation of vegetal and animal natural resources such as growing of crops, raising and breeding of animals, harvesting of timber and other plants, animals or animal products from a farm or their natural habitats.	Agricultural robots and drones Crop and soil monitoring Predictive analytics
Manufacturing (Section C)	Includes the physical or chemical transformation of materials, substances or components into new product. The materials transformed are products of agriculture, forestry, fishing, mining or quarrying as well as products of other manufacturing activities. Excludes waste.	Market and domain forecasting Product assembly Asset optimisation Supply-chain management and planning Anomaly detection
Public administration and defence; compulsory social security (Section O)	Includes activities of a governmental nature, normally carried out by the public administration such as public order and safety, legislative activities, foreign affairs, national defence and more.	Predictive algorithms in the legal system Predictive policing Use of AI by the judiciary Use of AI in defence (e.g. drone footage for surveillance, cyberdefence, command and control, autonomous vehicles)

Wholesale and retail trade (Section G)	Includes wholesale and retail sale (i.e. sale without transformation) of any type of goods and the rendering of services incidental to the sale of these goods. Wholesaling and retailing are the final steps in the distribution of goods. Goods bought and sold are also referred to as merchandise.	Customer management (e.g. prediction of customer needs, identification of upsell and cross-sell opportunities, agile response mechanism) Operational efficiency (e.g. just-in-time production/delivery, product categorisation/placement, demand forecasting, check-out free store) Legal efficacy (e.g. compliance systems to predict violations in the supply chain; legal contract translation, cataloguing and implementation - "smart contracts") Customer acquisition (e.g. matching buyers and sellers, personalised ads/referrals – see "Marketing and advertising services") Customer retention (e.g. learning and predicting customers' preferences and needs, tailored offers, dynamic pricing) Customer service (e.g. conversational interfaces, voice and video search, chatbots, mood tracking)
Education (Section P)	Includes private and public education, all levels from pre-school to higher education, adult education, sport education, literacy programmes and more.	Personalising learning with AI (e.g. adaptive tests and learning systems)[26] Supporting students with special needs with AI (e.g. wearables using AI) Reducing dropout rates (e.g. predictive and diagnosis models) Construction of scoring of tests or exams Fraud detection during exams Chatbots
Human health and social work activities (Section Q)	Includes the provision of health and social work activities. Activities include a wide range of activities, starting from health care provided by trained medical professionals in hospitals and other facilities, to residential care activities that still involve a degree of health care activities to social work activities without any involvement of health care professionals.	Detection (e.g. outbreak alerts) Precision medicine (e.g. treatments) Optimise health systems (e.g. resource allocation, workflow management) Facilitating health research (e.g. drug discovery, vaccine development) Preventative / personalised healthcare (e.g. self-monitoring tools, applications and trackers) Nursing and elderly care Diagnosis (e.g. radiology)
Transportation and storage (Section H)	Includes the provision of passenger freight transport, whether scheduled or not, by rail, pipeline, road, water or air. Also includes associated activities such as terminal and parking, cargo handling, storage. Includes rental and postal activities.	Warehouse and supply-chain management Shipping and itinerary route optimisation, including based on traffic data Autonomous driving systems Computer vision technologies that track driver's eyes / focus to assess distraction
Accommodation and food service activities (Section I)	Includes the provision of short-term stay accommodation for visitors, of complete meals and drinks for immediate consumption. The type of supplementary services provided within this section can vary widely. Excludes long-term stay and primary residence.	AI-powered chatbots (e.g. booking, ordering) Face recognition (check-in) Analysis of customer, occupancy and guest feedback data
Construction (section F)	Includes general and specialised construction activities for buildings and dwellings, civil engineering works, new work, additions, repairs and alterations.	3D Building Information Modelling (BIM) Buildings simulators Drones and sensors on construction sites Data analytics based on the real-time data collected on-site

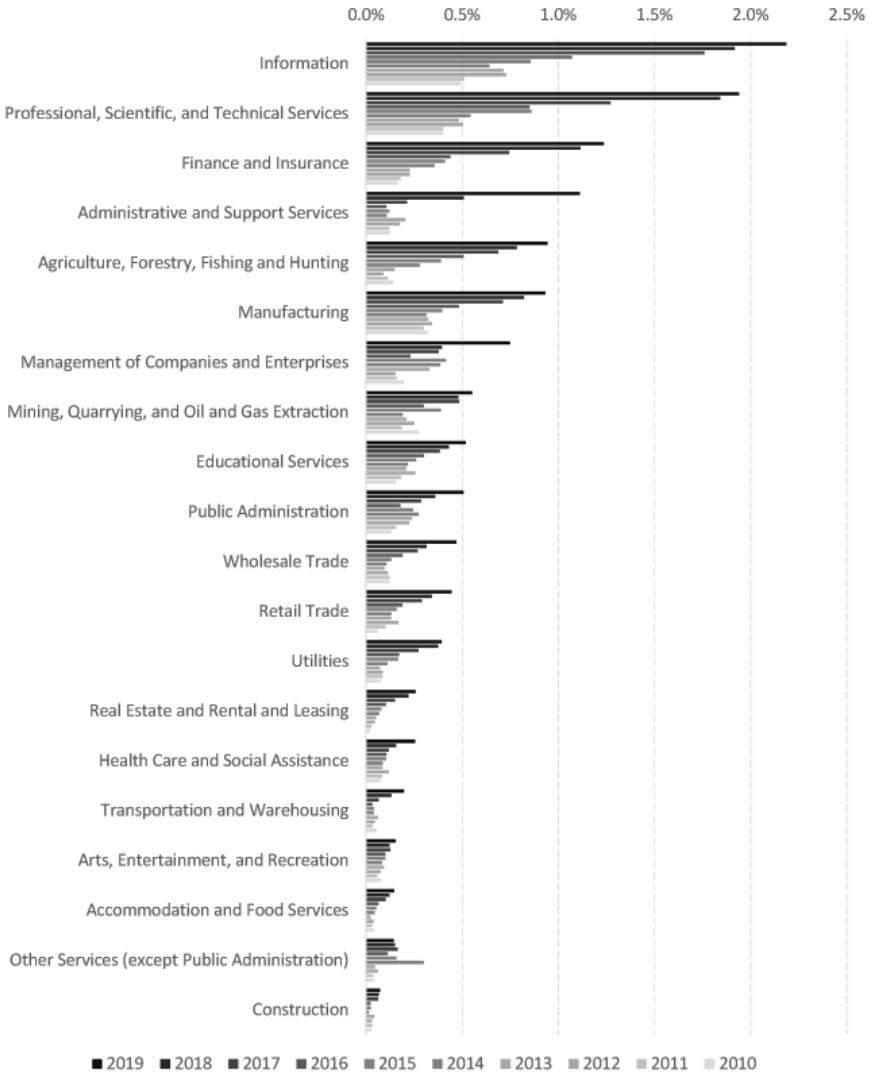
Note: Table is ordered from ISIC REV 4 industry sectors that are seeing the most AI adoption to those that are experiencing the least AI-adoption (see also Annex B). Not all ISIC REV 4 sectors are included.
Source: Various sources including (OECD, 2019a[9]).

Annex B. AI adoption per industry

Some researchers are using AI labour demand – that is, firms' jobs data – as a proxy for AI adoption in different industries. Figure B.1 shows the percentage of AI-related job vacancies – out of total vacancies – by two-digit North American Industry Classification System (NAICS) industries between 2010 and 2019 (Alekseeva, 2019[26]).

As the figure indicates, AI adoption has grown primarily in industries such as information; professional, scientific and technical services; finance and insurance; administrative and support services; agriculture; management; mining, quarrying, and oil and gas extraction; education; public administration; whole and retail trade; and manufacturing. Adoption of AI in fields such as healthcare or transportation seems to be much lower.

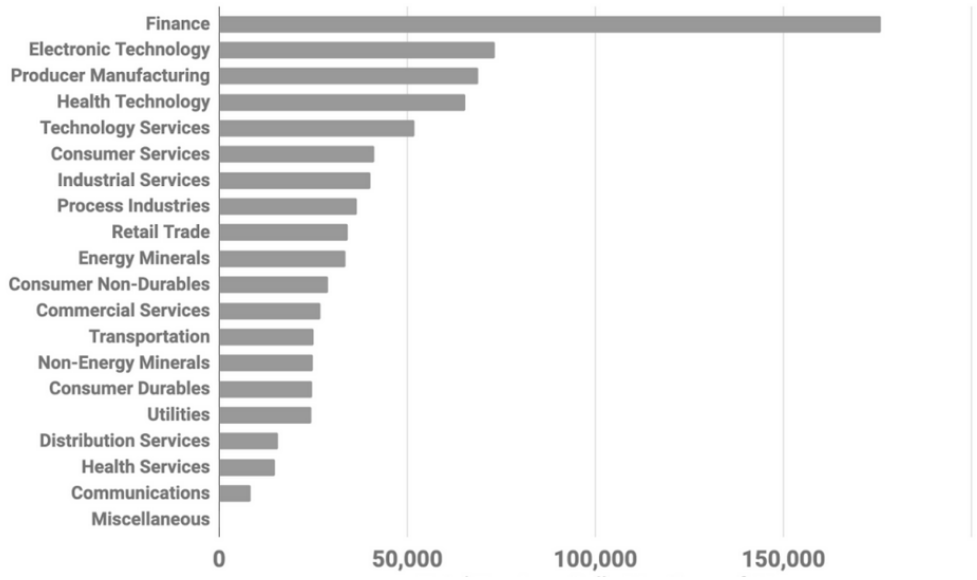
Figure B.1. Share (%) of AI jobs by industry, 2010-2019



Note: Industries are ranked by AI share, from largest to lowest share, in 2019. Data include only job postings with two-digit NAICS codes.
Source: (Alekseeva, 2019[26]).

An alternative measure of AI adoption is mentions of AI in company earnings calls. The share of earnings calls where AI is mentioned has increased substantially in recent years, especially in the finance, electronic technology, and producer manufacturing sectors. While mentions of AI are prevalent in the health technology sector, AI is seldom mentioned in health services. Mentions of AI are the lowest in earnings calls in the communication sector (Figure B.2).

Figure B.2. Mentions of AI in earnings calls by sector, 2018-2019



Source: Stanford AI Index 2021, https://aiindex.stanford.edu/report/.

Annex C. WG CAI membership

The up-to-date list of WG CAI (Working Group on Classification of AI) members and member biographies are available on the OECD.AI Policy Observatory [link].

Name	Title	Organisation	Group / Delegation
Dewey Murdick (Co-chair)	Director of Data Science	Center for Security and Emerging Technology (CSET), School of Foreign Service, Georgetown University	Civil Society and Academia
Marko Grobelnik (Co-chair)	AI Researcher & Digital Champion	Jožef Stefan Institute	Technical
Sebastian Hallensleben (co-Chair)	Head of Digitalisation and AI	VDE Association for Electrical, Electronic & Information Technologies	Technical
Jack Clark (Co-chair in 2020-21)	Co-chair	AI Index, Stanford University	Technical
Jefferson de Oliveira Silva	Web Technologies Study Center / Professor	NIC.br / Pontifical Catholic University	Brazil
Sally Radwan	Minister Advisor for Artificial Intelligence	Ministry of Communications & Information Technology	Egypt
Renaud Vedel	Coordinateur, stratégie nationale en IA	Ministère de l'intérieur	France
Peter Addo	Head of DataLab and Senior Data Scientist	Agence Française de Développement (AFD)	France
Judith Peterka	Head, AI indicators	Policy Lab Digital, Work & Society	Germany
Michael Schoenstein	Head of Strategic Foresight & Analysis	Policy Lab Digital, Work & Society	Germany
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
Yoichi Iida	Chair of the CDEP and Going Digital II Steering Group	Ministry of Internal Affairs and Communications (MIC)	Japan
Tatsuya Akagawa	Deputy Director	Ministry of Internal Affairs and Communications (MIC)	Japan
Yuki Hirano	Deputy Director, Multilateral Economic Affairs Office, Global Strategy Bureau	Ministry of Internal Affairs and Communications (MIC)	Japan
Takayaki Honda	Assistant Director, Multilateral Economic Affairs Office, Global Strategy Bureau	Ministry of Internal Affairs and Communications (MIC)	Japan
Risa Kashiwaze	Official, Multilateral Economic Affairs Office, Global Strategy Bureau	Ministry of Internal Affairs and Communications (MIC)	Japan
Katrina Kosa-Ammari	Counsellor at Foreign Economic Relations Promotion Division	Ministry of Foreign Affairs	Latvia
Andrey Ignatyev	Head of Analytics - Center for Global IT Cooperation – CGITC	Ministry of Economic Development	Russia
Anna Abramova	Head of the Department of Digital Economy and Artificial Intelligence	MGIMO-University	Russia
Dunja Mladenić	Head of Artificial Intelligence Department	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	Turkey
Mehmet Haklidir	Chief Researcher, Scientific and Technological Research Council	Informatics and Information Security Research Center	Turkey
Osman Musa Aydin	Advisor to the Deputy Minister, Defense Industry Expert	Ministry of Industry and Technology	Turkey
Fatma Bujasaim	Head of International Cooperation	Artificial Intelligence Office	United Arab Emirates
Lord Tim Clement-Jones	Lord	House of Lords	United Kingdom
Lynne Parker	Deputy United States Chief Technology Officer	The White House	United States
Elham Tabassi	Chief of Staff, Information Technology Laboratory	National Institute of Standards and Technology (NIST)	United States
Mark Latonero	Senior Policy Advisor on AI	National Institute of Standards and Technology (NIST)	United States
Farahnaaz H Khakoo	Assistant Director	US Government Accountability Office (GAO)	United States
Taka Ariga	Chief Data Scientist | Director, Innovation Lab	US Government Accountability Office (GAO)	United States
Nicholas Reese	Policy expert	Department of Homeland Security	United States
Mohammed Motiwala	Multilateral Engagement Officer	Department of State	United States
Raj Madhavan	Policy Fellow and Program Analyst	Department of State	United States
Kilian Gross	Head of Unit, Artificial Intelligence Policy Development and Coordination	European Commission DG Connect	European Commission
Juha Heikkilä	Adviser for Artificial Intelligence	DG Connect	European Commission
Tatjana Evas	Legal and Policy Officer	DG Connect	European Commission

Irina Orssich	Team Leader AI	DG Connect	European Commission
Emilia Gómez	Lead Scientist, Human behaviour and machine intelligence	DG Joint Research Centre (JRC)	European Commission
Giuditta de Prato	Researcher	DG Joint Research Centre (JRC)	European Commission
Prateek Sibal	AI Policy Researcher, Knowledge Societies Division	UNESCO Communication and Information Sector	IGO
Roberto Sanchez	Advisor - Data Scientist	Inter-American Development Bank	IGO
Denise Vandeweijer	Director of AI Engineering Operations	AppliedAI	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
Marco Ditta	Executive Director, ISP Group Data Officer	Intesa Sanpaolo	Business
Michel Morvan	Co-Founder and Executive Chairman	Cosmo Tech	Business
Nicole Primmer	Senior Policy Director	BIAC	Business
Nozha Boujemaa	Global VP, Digital Ethics and Responsible AI	IKEA Retail (Ingka Group)	Business
Olivia Erdelyi	Director of Ethics and Policy/Lecturer	Soul Machines/University of Canterbury, School of Law	Business
Olly Salzmann	Partner Deloitte/Managing Director	Deloitte KI GmbH and KIParkDeloitte GmbH	Business
Abe Hsuan	Independent Expert	Irwin Hsuan	Technical
Clara Neppel	Senior Director	IEEE European Business Operations	Business
Eric Badique	Independent Expert		Technical
Jonathan Frankle	PhD Candidate	MIT Internet Policy Research Initiative (IPRI)	Technical
Masashi Sugiyama	Director, Center for Advanced Intelligence Project	RIKEN, Japan	Technical
Taylor Reynolds	Technology Policy Director	MIT Internet Policy Research Initiative (IPRI)	Technical
Adriano Koshiyama	Research Fellow in Computer Science,	University College London (UCL)	Civil Society and Academia
Catherine Aiken	Researcher	Center for Security and Emerging Technology (CSET), Georgetown University	Civil Society and Academia
Daniel Schwabe	Professor at the Department of Informatics	Catholic University in Rio de Janeiro (PUC-Rio)	Civil Society and Academia
Eva Thelisson	Researcher	AI Transparency Institute	Civil Society and Academia
Guillaume Chevillon	Professor - Co Director ESSEC	ESSEC Business School, Paris	Civil Society and Academia
Jibu Elias	Research and Content Head	INDIAai	Civil Society and Academia
Jim Kurose	Professor	University of Massachusetts Amherst	Civil Society and Academia
Nicolas Moes	Head of Operations and EU AI Policy	The Future Society (TFS)	Civil Society and Academia
Philip Dawson	Policy Lead	Schwartz Reisman Institute for Technology and Society	Civil Society and Academia
Saurabh Mishra	Researcher and Manager of the AI Index Program	Stanford Institute for Human-Centered Artificial Intelligence (HAI)	Civil Society and Academia
Suso Baleato	Secretary	CSISAC	Civil Society and Academia
Theodoros Evgeniou	Professor, Decision Sciences and Technology Management	INSEAD	Civil Society and Academia
Tim Rudner	PhD Candidate	University of Oxford	Civil Society and Academia
Vincent C. Müller	Professor for Philosophy of Technology	Technical University of Eindhoven	Civil Society and Academia

The AI Secretariat team also wishes to thank the following for their presentations to the Experts Working Group: Gregor Strojin (Council of Europe) on CAHAI's feasibility study; Neville Matthew (OECD Working Party on Consumer Policy) on risk assessment for product safety; Helen Toner (CSET) and Sean McGregor (PAI) on the Partnership on AI (PAI) AI Incident Database (AIID) as well as Irene Kitsara (World Intellectual Property Office (WIPO) on WIPO's approach to AI-based patent taxonomies.

The AI Secretariat team within the OECD Digital Economy Policy division supporting this working group is made up of Karine Perset (Head of the AI Unit of the OECD Division for Digital Economy Policy), Luis Aranda (Policy Analyst, OECD AI Policy Observatory), Louise Hatem (Junior Policy Analyst, OECD.AI) and Orsolya Dobe (Young Associate, OECD.AI). Other secretariat members participating in the working group include Leonidas Aristodemou, Fernando Galindo-Rueda, Marguerita Lane, and Pierre Montagnier.

Annex D. Participants in the public consultation

A public consultation on a preliminary version of the framework was held in June 2021. The OECD would like to thank all those who tested the framework by classifying an AI system of their choice using the online survey. Over 800 surveys were submitted, including 151 complete surveys. While all respondents' names cannot be displayed in this report, the OECD would like to acknowledge in particular those who filled in the survey in its entirety.

Table 7. Respondents to the June 2021 public consultation who completed the online survey

Name	Surname	Institution (where provided)
Enrique	Alba	N/A
Ali	Al-Khulaifi	Gulf Center for development
Fikret	Anil Özörnek	Türkiye Yapay Zeka Inisiyatifi
Mauricio	Araya	Universidad Técnica Federico Santa María, Chile (UTFSM)
Elio	Atenógenes Villaseñor Garcia	National Institute of Statistics and Geography (Mexico).
Alexandre	Barbosa	Cetic.br/NIC.br
Hassan	Bashiri	N/A
Rohan	Baxter	Australian Taxation Office
Sawyer	Bernath	Berkeley Existential Risk Initiative
Victor	Bernhardtz	Unionen
Przemyslaw	Biecek	MI2DataLab at Warsaw University of Technology
Nicolas	Blanc	French Confederation of Management – General Confederation of Executives , CFEN/ACGC
José	Brito de Souza	Escola de Guerra Naval (Naval War College), Brasil
James	Butcher	United Nations Interregional Crime and Justice Research Institute (UNICRI)
Jarbas	Camurça	Instituto Atlântico
Horacio	Canales	Centro de Ingeniería y Desarrollo Industrial, Mexico (CIDESI)
André	Carlos	N/A
Benjamin	Cedric Larsen	Copenhagen Business School
Marcelo	Cezar Pinto	Federal University for Latin American Integration (UNILA)
Alan	Chan	Mila
Koen	Cobbaert	Philips
Fernando	Cormenzana	Agencia para el Desarrollo del Gobierno de Gestión Electrónica y la Sociedad de la Información y del Conocimiento, Uruguay
Jan	Czarnocki	KU Leuven Center for IT & IP Law
Erica	Da Cunha Ferreira	Federal University of Rio de Janeiro
Lucas	Dalmedico Gessoni	Eldorado Research Institute
Peter	Damm	KMD A/S
Eric	Deeben	Office for National Statistics, United Kingdom
Werner	Denzin	SiDi
John	Dickson	Australian Competition & Consumer Commission
Marcela	Distefano	Argentine University of Business (UADE)
Pam	Dixon	World Privacy Forum
Ian	Dowson	William Garrity Associates Ltd
Patrick	DrakeN/ABrockman	Digital Transformation Agency (DTA), Australia
Lars	Eidnes	N/A
Irene	Ek	Swedish Agency for Growth Policy Analysis
Stuart	Elliott	US National Academy of Sciences
Ingo	Elsen	FH Aachen University of Applied Sciences
Marcos	Evandro Cintra	Universidade Federal Rural do Semi-Árido, Brasil (UFERSA)
Manoel	Galdino	Transparência Brasil
Alexander	Galt	Inter IKEA Group
Nicolas	Gaude	N/A
Maria	Gonsalez	N/A
Axel	Gruvaeus	Kairos Future
Daniel	Guerreiro Silva	Universidade de Brasília / Senacon and Ministério da Justiça e Segurança Pública, Brazil
Miguel R.	Guevara	Universidad de Playa Ancha
Rajan	Gupta	Centre for Information Technologies and Applied Mathematics (CITAM)
Naira	Hambardzumyan	Deloitte
Thomas	Hampson	N/A
András	HLÁCS	Permanent Delegation of Hungary to OECD
Arliones	Hoeller Jr.	IFSC

Stefan	Ivanica	N/A
Jhalak M.	Kakkar	Centre for Communication Governance, National Law University, Delhi (India)
Bogumil	Kaminski	Warsaw School of Economics, Poland (SGH)
George	Khoury	Institute for Defense Analyses
Mwendwa	Kivuva	KICTANet
Ludek	Knorr	N/A
Ansgar	Koene	Ernst & Young
Iwona	Kowalska	Vice President at Musimap
Joelma	Kremer	Ministry of Education
Robert	Kroplewski	Plenipotentiary for the Polish Minister of Digital Affairs
François	Laviolette	AI researcher - director of Big Data Research Center at the Unversité of Laval, Canada
Adriano	Leal	Instituto de Pesquisas Tecnológicas
Ioannidis	Leonidas	Plannam
Jacky	Liang	Carnegie Mellon University
Dr Jacques	Ludik	N/A
Maciej	Majewski	PhD
Marcelo	Malheiros	Universidade Federal do Rio Grande, Brasil (FURG)
Richard	Mallah	Future of Life Institute
Jawad	Masood	Automotive Technology Centre of Galicia - Spain (CTAG)
Sean	McGregor	AI Incidents Database
Nikola	Mojic	2021.ai
Erick	Muzart Fonseca dos Santos	TCU/Brazil
Marco	Neves	InteractIdeas
Nobuhisa	Nishigata	Japan Ministry of Internal Affairs and Communication
Anthony	Nolan	N/A
Irene	Olivan Garcia	OECD
Gloria	Ortega	N/A
Liam	PeetN/APare	University of Alberta
Nitendra	Rajput	Mastercard
Ashok	Rao	MaGC
Alexandre	Reeberg de Mello	SENAI Innovation Institute for Embedded Systems
Michaela	Regneri	N/A
Isaac	Robinson	Harvard
Vladimir	Sadilovski	Independent research in AI and ethics
José Lucas	Safanelli	University of São Paulo, Brasil (ESALQ/USP)
Alessandro	Saffiotti	Orebro University
Ricardo	Sanz	Universidad Politecnica de Madrid, Spain (UPM)
Nathalia	Sautchuk	N/A
Calli	Schroeder	Electronic Privacy Information Center (EPIC), USA
Leonardo	Seabra	Empresa Potiguar de Promoção Turística (EMPROTUR), Brasil
Patricia	Shaw	Beyond Reach Consulting Limited
Ricardo	Silva	Terumo BCT
Elena	Simperl	King's College London
Jürgen	Skerhut	N/A
Aliaksandr	Smirnou	Cybr Consult FZ LLE & Locations Solutions Telematics LLC
Mike	Sparling	MultiN/AHealth Systems Inc.
Andreas	Spechte	CEO, Silicon Castles
Clara	Standring	Office of National Statistics, United Kingdom
Basile	Starynkevitc	N/A
JF	SulzerConsultant	N/A
Jose	Teran-Vargas	National Institute of Statistics and Geography (INEGI), Mexico
Sonja	Thiel	BLM
Daniel	Trento	Brazil Ministry of Agriculture
Austin	Tripp	University of Cambridge
Evgeny	Vasin	Sberbank of Russia
Leandro	Volochko	Ministério Público do Estado de Mato Grosso (MPMT) Brasil
Heather	Von Stackelberg	N/A
Pr. Toby	Walsch	University of New South Wales, Sydney (Australia)
Yudhanjaya	Wijeratne	LIRNEasia
Jennifer	Wortman	Microsoft Research
Mac	Yokozawa	Center for International Economic Collaboration (CFIEC), Japan (also on behalf of Business at OECD and Keidanren)
Jay	Yoo	Software Policy & Research Instiute (SPRI)
Danielle	Zaror Miralles	Investigadora en el Centro de Derecho Informático (CEDI) N/A OptIA

Notes

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

[2] Some information may be challenging to obtain at initial stages of deployment (e.g. assessing the impact of an AI system on well-being), on some technical characteristics of the model or when information may be proprietary or constitute "trade secrets".

[3] The National Institute of Health and Care Excellence (NICE) Health Technology Assessment (HTA) programme in the United Kingdom evaluates health technologies for their clinical and cost-effectiveness, following regulatory approval.

[4] AI actors are those who play an active role in the AI system lifecycle. Public- or private-sector organisations or individuals that acquire AI systems to deploy or operate them are also considered to be AI actors. AI actors include, among other roles, technology developers, systems integrators, and service and data providers (OECD, 2019f[2]).

[5] The OECD AI Principles also indicates “or decisions”, which the Experts Working Group decided should be excluded here to clarify that an AI system does not make an actual decision, which is the remit of human creators and outside the scope of the AI system.

[6] In the current classification framework, a “machine-based system” includes but is not limited to software that uses models developed with the following techniques and approaches: 1) Data-driven approaches, including supervised, unsupervised and reinforcement machine learning and other statistical approaches to learning and inference and, 2) Knowledge-driven approaches, including logic and knowledge representation, simulation, inductive programming, knowledge bases, inference and deductive engines, symbolic reasoning and expert systems.

[7] This characterisation of an AI system has been modified by replacing the term “interpret” with “use” to avoid confusion with the term “model interpretability”.

[8] Amateur users may be aware of or conscious or not that they are interacting with an AI system.

[9] For more details on GDPR guidelines, visit: https://gdpr-text.com/read/article-22/#links

[10] The OECD and global organisations like the IEEE have recognised that AI design and implementation impact socio-economic environments in their entirety. For this reason, they emphasise that all elements of Table 3 should be given equal priority in the design phase. IEEE, for example, created the IEEE 7010-2020 standard containing Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being, which uses multiple indicators and allows users to generate their own

"societal impact assessment" to help design responsible AI systems using a "triple bottom line" of people, planet and prosperity.

[11] The OECD promotes holistic indicators to measure well-being that consider long-term environmental sustainability alongside increased human well-being that goes beyond statistics like GDP. See OECD, *Measuring Well-Being and Progress*: https://www.oecd.org/statistics/measuring-well-being-and-progress.htm.

[12] This draws on the ongoing work of the OECD Employment, Labour and Social Affairs Directorate (ELS) and the Global Partnership on AI (GPAI) Working Group on the Future of Work.

[13] Many applications combine symbolic and statistical approaches to inputting data. For example, natural language processing (NLP) algorithms often combine statistical approaches that build on large amounts of data and symbolic approaches that consider issues such as grammar rules. Autonomous driving systems, for example, use both machine-based inputs (historical driving data) and human-based inputs (a set of driving rules). Combining models built on both data and human expertise is considered promising to help address the limitations of both approaches.

[14] See the *State of AI Report 2021*, available at https://www.stateof.ai/ and based on research by OpenAI.

[15] Both data and data format can be proprietary, *i.e.* the data format might only be known by the owner.

[16] See the US Federal Trade Commission's (FTC) information on privacy and data security for more details: https://www.ftc.gov/news-events/media-resources/protecting-consumer-privacy-security/ftc-policy-work.

[17] See, for example, Raoult, D. (2020), "Lancet gate: a matter of fact or a matter of concern", *New Microbes and New Infections*, Vol. 38, Elsevier, Amsterdam,, https://www.sciencedirect.com/science/article/pii/S2052297520301104.

[18] This requirement should not hinder learning. Gradual iterative learning with controllable deviation from previously accepted outcomes can offer manageable risk levels and allows for an AI system's evolution.

[19] The three categories presented in the list are sometimes referred to as logic AI, non-logicist AI and hybrid AI, respectively.

[20] Open AI's CLIP and DALL-E, for example, could be considered as performing both computer vision and NLP. DeepMind's MuZero performs any task given that it can learn as a planning algorithm.

[21] Federated learning is a subset of distributed machine learning. While distributed machine learning runs algorithms on edge devices to split the training workload across different machines, federated learning trains the algorithm on the edge, summarises the changes and returns only this focused update back to the main model.

[22] The uncertainty can be structural (e.g. whether a linear regression or a neural network is more appropriate, and, if the latter, how many layers should it have, etc.), parametric (which values of a model's parameters make the best prediction) or "noise"-related (e.g. pixel noise or blur in images).

[23] While machine learning techniques are usually used in building or adjusting a model, they can also be used to interpret a model's results.

[24] This section draws on some of the research undertaken by Cognilytica, *The Seven Patterns of AI* (https://www.cognilytica.com/2019/04/04/the-seven-patterns-of-ai/), although the Expert Group modified, added and removed some tasks.

[25] See, for example: https://www.3ds.com/sustainability/sustainability-insights/designing-disruption/executive-summary.

[26] http://www.oecd.org/education/trustworthy-artificial-intelligence-in-education.pdf.

References

Abrams, M. (2014), *The Origins of Personal Data and its Implications for Governance*, http://dx.doi.org/10.2139/ssrn.2510927. [16]

Aikens, C. (2019), *Classifying AI Systems*, https://cset.georgetown.edu/publication/classifying-ai-systems/. [22]

Alekseeva, L. (2019), *The Demand for AI Skills in the Labour Market*, Center for Economic Policy Research (CEPR), London, https://repec.cepr.org/repec/cpr/ceprdp/DP14320.pdf. [26]

BSA (2021), *Confronting Bias: BSA's Framework to Build Trust in AI*, The Software Alliance, Washington, DC, https://ai.bsa.org/wp-content/uploads/2021/06/2021bsaaibias.pdf. [19]

CISA (2019), *National Critical Functions Set*, Cibersecurity & Infrastructure Security Agency, Washington, DC, https://www.cisa.gov/national-critical-functions-set. [11]

CoE (2020), *The Feasibility Study on AI Legal Framework Adopted by the CAHAI*, Council of Europe, Strasbourg, https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da. [4]

CoE (1998), *European Convention on Human Rights*, Council of Europe, Strasbourg, https://www.echr.coe.int/documents/convention_eng.pdf. [5]

Endsley, M. (1987), "The application of human factors to the development of expert systems for advanced cockpits", *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 31, Iss. 12, pp. 1388-1392, https://doi.org/10.1177/154193128703101219. [21]

Freire, A. (2021), *Measuring Diversity of Artificial Intelligence Conferences*, http://dx.doi.org/arXiv:2001.07038. [20]

Friedman, B. and H. Nissenbaum (1996), "Bias in Computer Systems", *ACM Transactions on Information Systems*, Vol. 14, No. 3, pp. 330 –347, https://nissenbaum.tech.cornell.edu/papers/Bias%20in%20Computer%20Systems.pdf. [27]

Mankins, J. (1995), *Technology Readiness Levels*, White Paper, https://www.sciencedirect.com/science/article/pii/S0736585320301842#bb0035. [13]

Martinez Plumed, F. (2020), *AI Watch: Assessing Technology Readiness Levels for Artificial Intelligence*, Publications Office of the European Union, Luxembourg, http://dx.doi.org/doi:10.2760/15025. [12]

OECD (2020), *How's Life? 2020: Measuring Well-Being*, OECD Publishing, Paris, https://doi.org/10.1787/9870c393-en. [7]

OECD (2018), *OECD Due Diligence Guidance for Responsible Business Conduct*, OECD Publishing, Paris, http://mneguidelines.oecd.org/OECD-Due-Diligence-Guidance-for-Responsible-Business-Conduct.pdf. [25]

OECD (2016), *Product Risk Assessment Practices of Regulatory Agencies*, OECD Publishing, Paris, https://www.oecd.org/sti/consumer/product-risk-assessment-practices-regulatory-agencies.pdf. [23]

OECD (2015), *Digital Security Risk Management for Economic and Social Prosperity: OECD Recommendation and Companion Document*, OECD Publishing, Paris, https://doi.org/10.1787/9789264245471-en. [24]

OECD (2011), *The Role of Internet Intermediaries in Advancing Public Policy Objectives*, OECD Publishing, Paris, https://doi.org/10.1787/9789264115644-4-en. [8]

OECD (2019a), *Artificial Intelligence in Society*, OECD Publishing, Paris, https://dx.doi.org/10.1787/eedfee77-en. [9]

OECD (2019b), *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. [15]

OECD (2019c), *Measuring the Digital Transformation: A Roadmap for the Future*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264311992-en. [18]

OECD (2019d), *Recommendation of the Council on Artificial Intelligence*, OECD Publishing, Paris, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. [1]

OECD (2019e), *Recommendation of the Council on Digital Security of Critical Activities*, OECD Publishing, Paris, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0456. [10]

OECD (2019f), *Scoping the OECD AI principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO)*, OECD Publishing, Paris, https://doi.org/10.1787/d62f618a-en. [2]

OHCHR (2011), *Guiding Principles on Business and Human Rights*, United Nations Human Rights Office of the High Commissioner, New York and Geneva, https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf. [6]

Russell, S. (2019), *Human-Compatible: Artificial Intelligence and the Problem of Control*, Penguin Books, New York, http://ISBN 9780525558637. [17]

Terrile, R. et al. (2015), "Calibrating the Technology Readiness Level (TRL) scale using NASA mission data", *2015 IEEE Aerospace Conference*, http://dx.doi.org/10.1109/aero.2015.7119313. [14]

UN General Assembly (1948), *Universal Declaration of Human Rights*, New York, https://www.un.org/en/udhrbook/pdf/udhr_booklet_en_web.pdf. [3]