TECHNICAL REPORT

ISO/IEC TR 24368

First edition
2022-08

Information technology — Artificial intelligence — Overview of ethical and societal concerns

Reference number
ISO/IEC TR 24368:2022(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives or www.iec.ch/members_experts/refdocs).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents) or the IEC list of patent declarations received (see https://patents.iec.ch).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html. In the IEC, see www.iec.ch/understanding-standards.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 42, *Artificial intelligence*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html and www.iec.ch/national-committees.

Introduction

Artificial intelligence (AI) has the potential to revolutionise the world and carry a plethora of benefits for societies, organizations and individuals. However, AI can introduce substantial risks and uncertainties. Professionals, researchers, regulators and individuals need to be aware of the ethical and societal concerns associated with AI systems and applications.

Potential ethical concerns in AI are wide ranging. Examples of ethical and societal concerns in AI include privacy and security breaches to discriminatory outcomes and impact on human autonomy. Sources of ethical and societal concerns include but are not limited to:

— unauthorized means or measures of collection, processing or disclosing personal data;

— the procurement and use of biased, inaccurate or otherwise non-representative training data;

— opaque machine learning (ML) decision-making or insufficient documentation, commonly referred to as lack of explainability;

— lack of traceability;

— insufficient understanding of the social impacts of technology post-deployment.

AI can operate unfairly particularly when trained on biased or inappropriate data or where the model or algorithm is not fit-for-purpose. The values embedded in algorithms, as well as the choice of problems AI systems and applications are used for to address, can be intentionally or inadvertently shaped by developers' and stakeholders' own worldviews and cognitive bias.

Future development of AI can expand existing systems and applications to grow into new fields and increase the level of automation which these systems have. Addressing ethical and societal concerns has not kept pace with the rapid evolution of AI. Consequently, AI designers, developers, deployers and users can benefit from flexible input on ethical frameworks, AI principles, tools and methods for risk mitigation, evaluation of ethical factors, best practices for testing, impact assessment and ethics reviews. This can be addressed through an inclusive, interdisciplinary, diverse and cross-sectoral approach, including all AI stakeholders, aided by International Standards that address issues arising from AI ethical and societal concerns, including work by Joint Technical Committee ISO/IEC JTC 1, SC 42.

Information technology — Artificial intelligence — Overview of ethical and societal concerns

1 Scope

This document provides a high-level overview of AI ethical and societal concerns.

In addition, this document:

— provides information in relation to principles, processes and methods in this area;

— is intended for technologists, regulators, interest groups, and society at large;

— is not intended to advocate for any specific set of values (value systems).

This document includes an overview of International Standards that address issues arising from AI ethical and societal concerns.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 22989, *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 22989 and the following apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
agency
ability to define one's goals and act upon them

[SOURCE: ISO/TR 21276:2018, 3.6.2]

3.2
bias
systematic difference in *treatment* (3.13) of certain objects, people, or groups in comparison to others

[SOURCE: ISO/IEC TR 24027:2021, 3.2.2, modified — Removed Note to entry.]

3.3
data management
process of keeping track of all data and/or information related to the creation, production, distribution, storage, disposal and use of e-media, and associated processes

[SOURCE: ISO 20294:2018, 3.5.4, modified — Added "disposal" to definition.]

3.4
data protection
legal, administrative, technical or physical measures taken to avoid unauthorized access to and use of data

[SOURCE: ISO 5127:2017, 3.13.5.01, modified — Removed Note to entry.]

3.5
equality
state of being equal, especially in status, rights or opportunities

[SOURCE: ISO 30415:2021, 3.9, modified — Removed "outcome" from definition.]

3.6
equity
practice of eliminating avoidable or remediable differences among groups of people, whether those groups are defined socially, economically, demographically or geographically

3.7
fairness
treatment (3.13), behaviour or outcomes that respect established facts, societal norms and beliefs and are not determined or affected by favouritism or unjust discrimination

Note 1 to entry: Considerations of fairness are highly contextual and vary across cultures, generations, geographies and political opinions.

Note 2 to entry: Fairness is not the same as the lack of *bias* (3.2). Bias does not always result in unfairness and unfairness can be caused by factors other than bias.

3.8
cognitive bias
human cognitive bias
bias (3.2) that occurs when humans are processing and interpreting information

Note 1 to entry: Human cognitive bias influences judgement and decision-making.

[SOURCE: ISO/IEC TR 24027:2021, 3.2.4, modified — Added "cognitive bias" as preferred term.]

3.9
life cycle
evolution of a system, product, service, project or other human-made entity from conception through retirement

[SOURCE: ISO/IEC/IEEE 12207:2017, 3.1.26]

3.10
organization
company, corporation, firm, enterprise, authority or institution, person or persons or part or combination thereof, whether incorporated or not, public or private, that has its own functions and administration

[SOURCE: ISO 30000:2009, 3:10]

3.11
privacy
rights of an entity (normally an individual or an organization), acting on its own behalf, to determine the degree to which the confidentiality of their information is maintained

[SOURCE: ISO/IEC 24775-2:2021, 3.1.46]

3.12
responsibility
obligation to act or take decisions to achieve required outcomes

Note 1 to entry: A decision can be taken not to act.

[SOURCE: ISO/IEC 38500:2015, 2.22, modified — Changed “and” to “or” and added Note to entry.]

3.13
treatment
kind of action, such as perception, observation, representation, prediction or decision

[SOURCE: ISO/IEC TR 24027:2021, 3.2.2, modified — Changed Note to entry to term and definition.]

3.14
safety
expectation that a system does not, under defined conditions, lead to a state in which human life, health, property, or the environment is endangered

[SOURCE: ISO/IEC/IEEE 12207:2017, 3.1.48]

3.15
security
aspects related to defining, achieving, and maintaining confidentiality, integrity, availability, accountability, authenticity, and reliability

Note 1 to entry: A product, system, or service is considered to be secure to the extent that its users can rely that it functions (or will function) in the intended way. This is usually considered in the context of an assessment of actual or perceived threats.

[SOURCE: ISO/IEC 15444-8:2007, 3.25]

3.16
sustainability
state of the global system, including environmental, social and economic aspects, in which the needs of the present are met without compromising the ability of future generations to meet their own needs

[SOURCE: ISO/Guide 82:2019, 3.1, modified — Removed Notes to entry.]

3.17
traceability
ability to identify or recover the history, provenance, application, use and location of an item or its characteristics

3.18
value chain
range of activities or parties that create or receive value in the form of products or services

[SOURCE: ISO 22948:2020, 3.2.11]

4 Overview

4.1 General

Ethical and societal concerns are a factor when developing and using AI systems and applications. Taking context, scope and risks into consideration can mitigate undesirable ethical and societal outcomes and harms. Examples of areas where there is an increasing risk for undesirable ethical and societal outcomes and harms include the following[24]:

— financial harm;

— psychological harm;

— harm to physical health or safety;

— intangible property (for example, IP theft, damage to a company's reputation);

— social or political systems (for example, election interference, loss of trust in authorities);

— civil liberties (for example, unjustified imprisonment or other punishment, censorship, privacy breaches).

In the absence of such considerations, there is a risk that the technology itself can levy significant social or other consequences, with possible unintended or avoidable costs, even if it performs flawlessly from a technical perspective.

4.2 Fundamental sources

Various sources address ethical and societal concerns specifically or in a general way. Some of these sources are identified.

Firstly, ISO Guide 82 provides guidance to standards developers in considering sustainability in their activities with specific reference to the social responsibility guidance of ISO 26000. This document therefore describes social responsibility in a form that can inform activities related to standardising trustworthy AI.

ISO 26000 provides organizations with guidance concerning social responsibility. It is based on the fundamental practices of:

— recognizing social responsibility within an organization;

— undertaking stakeholder identification and engagement.

Without data, the development and use of AI cannot be possible. Therefore, the importance of data and data quality makes traceability and data management a pivotal consideration in the use and development of AI. The following data-oriented elements are at the core of creating ethical and sustainable AI:

— data collection (including the means or measures of such data collection);

— data preparation;

— monitoring of traceability;

— access and sharing control (authentication);

— data protection;

— storage control (adding, change, removal);

— data quality.

These elements impact explainability, transparency, security and privacy, especially in cases of personal identifiable information being generated, controlled or processed. Traceability and data management are essential considerations for an organization using or developing AI systems and applications.

ISO/IEC 38505-1 considers data value, risks and constraints in governing how data are collected, stored, distributed, disposed of, reported on and used in organizational decision-making and procedures. The results of data mining or machine learning activities in reporting and decision-making are regarded as another form of data, which are therefore subject to the same data governance guidelines.

Furthermore, the description of ethical and societal concerns relative to AI systems and applications can be based on various AI-related International Standards.

ISO/IEC 22989 provides standardized AI terminology and concepts and describes a life cycle for AI systems.

ISO/IEC 22989 also defines a set of stakeholders involved in the development and use of an AI system. ISO/IEC 22989 describes the different AI stakeholders in the AI system value chain that include AI provider, AI producer, AI customer, AI partner and AI subject. ISO/IEC 22989 also describes various sub-roles of these types of stakeholders. In this document we refer to all of these different stakeholder types collectively as stakeholders.

ISO/IEC 22989 includes "relevant regulatory and policy making authorities" as a sub-role of AI subject. Regulatory roles for AI are currently not yet widely defined, but a range of proposals has been made including organizations appointed by individual stakeholders; industry-representative bodies; self-appointed civic-society actors; or institutions established through national legislation or international treaty.

All of these features of ISO/IEC 22989 assist in the description of AI-specific ethical and societal concerns.

As AI has the potential to impact a wide range of societal stakeholders, including future generations impacted by changes to the environment (indirectly affected stakeholders). For example, images of pedestrians on a sidewalk can be captured by autonomous vehicle technology, or innocent persons can be subject to police surveillance equipment designed to survey suspected criminals.

ISO/IEC 23894 provides guidelines on managing AI-related risks faced by organizations during the development and application of AI techniques and systems. It follows the structure of ISO 31000:2018 and provides guidance that arises from the development and use of AI systems. The risk management system described in ISO/IEC 23894 assists in the description of ethical and societal concerns in this document.

ISO/IEC TR 24027 describes the types and forms of bias in AI systems and how they can be measured and mitigated. ISO/IEC TR 24027 also describes the concept of fairness in AI systems. Bias and fairness are important for the description of AI-specific ethical and societal concerns.

ISO/IEC TR 24028 provides an introduction to AI system transparency and explainability, which are important aspects of trustworthiness and which can impact ethical and societal concerns.

ISO/IEC TR 24030 describes a collection of 124 use cases of AI applications in 24 different application domains. The use cases identify stakeholders, stakeholders' assets and values, and threat and vulnerabilities of the described AI system and application. Some of the use cases describe societal and ethical concerns.

ISO/IEC 38507 provides guidance on the governance implications for organization involved in the development and use of AI systems. This guidance is in addition to measures defined in existing International Standards on governance, namely:

— ISO 37000;

— ISO/IEC 38500;

— ISO/IEC 38505-1.

Governance is a key mechanism by which an organization is able to address the ethical and societal implications of its involvement in AI systems and applications.

ISO/IEC 27001 specifies the requirements for establishing, implementing, maintaining and continually improving an information security management system within the context of the organization. It also includes requirements for the assessment and treatment of information security risks tailored to the needs of the organization. The requirements set out in ISO/IEC 27001 are generic in nature and can serve as a foundation for systematic information security management within the context of AI. This, in turn, can have downstream impacts on ethical and societal issues in AI systems and applications. ISO/IEC 27001 is supplemented by ISO/IEC 27002:2022, which provides guidelines for organizational

information security standards and information security management practices including the selection, implementation and management of controls.

ISO/IEC 27701 provides guidance for establishing, implementing, maintaining and continually improving a Privacy Information Management System in the form of an extension to ISO/IEC 27001 and ISO/IEC 27002:2022 for data privacy management within an organization. ISO/IEC 27701 can serve as a foundation for privacy information management within the context of AI.

4.3 Ethical frameworks

4.3.1 General

AI ethics is a field within applied ethics. This means that principles and practices are rarely the result of applying ethical theories. Nevertheless, many of the challenges are closely related to traditional ethical concepts and problems – for example privacy, fairness and autonomy that can be addressed in existing ethical frameworks. See Reference [25] for more possible ethical frameworks. This list of ethical frameworks is neither collectively exhaustive nor mutually exclusive. Hence, ethical frameworks beyond those listed can be considered[26].

4.3.2 Virtue ethics

Virtue ethics is an ethical framework that specifies sets of virtues, which are intended to be pursued (e.g. respect, honesty, courage, compassion, generosity), and sets of vices (e.g. dishonesty, hatred), which are intended to be avoided. Virtue ethics has the strength of being flexible and aspirational. Its primary disadvantage is that it does not offer any specific implementation guidelines. Saying that an AI system is designed to be “honest” is only meaningful if provided with a mechanism by which that virtue is operationalized. However, so long as its technical limitations are kept in mind, virtue ethics can serve as a useful tool for determining whether or not an application of AI is a reflection of human virtues.

4.3.3 Utilitarianism

Utilitarianism is an ethical framework that maximizes good and minimizes harm. A utilitarian choice is one that produces the greatest good and does the least amount of harm to all stakeholders involved. Once the ethical aspects of a problem are explained logically, utilitarian approaches have the strength of being universally understandable and intuitive to implement. Utilitarianism’s primary disadvantage is that utilitarian frameworks permit harming some for the good of the whole. Examples include the Trolley Problem[27] where utilitarianism supports murder, or the example of transplant patients at a hospital, where utilitarianism supports the dissection of a healthy donor to transplant their organs into multiple patients.[28] In addition, many moral considerations are difficult to quantify (e.g. dignity) or are subjective - what is good for one person might not be good for another. Moral considerations vary enough that they are difficult to weigh against each other, for example environmental pollution versus societal truthfulness. Further, utilitarianism as a framework is a form of consequentialism - the doctrine that "the ends support the means". Consequentialism supports creating solutions that offer net benefit, but does not require that those solutions function ethically, e.g. in an unbiased way[29].

4.3.4 Deontology

Deontology is an ethical framework which assesses morality by a set of predefined duties or rules. The specific mechanism for making this determination is a set of rules or codified norms which can be analysed in the moment without needing to calculate what the consequences of those actions can be. An example of such a rule is “equality of opportunity” within fairness. Equality of opportunity dictates that the people who qualify for an opportunity are equally likely to do so regardless of their social group membership. The main disadvantage of deontology is that such universal rules can be difficult to derive in practice and can be brittle when deployed in cross-contextual settings or highly variable environments.

5 Human rights practices

5.1 General

International Human Rights, as outlined in the Universal Declaration of Human Rights, see Reference [30], the UN Sustainable Development Goals, see Reference [31] and UN Guiding Principles on Business and Human Rights, see Reference [32], are fundamental moral principles to which a person is inherently entitled, simply by virtue of being human. They can serve as a guiding framework for directing corporate responsibility around AI systems and applications with the benefit of international acceptance as a more mature framework for assessments of policy and technology. International Human Rights can also provide established process for performing due diligence and impact assessments. The implications of human rights on the governance of AI in organizations are discussed in ISO/IEC 38507.

Frameworks, such as care ethics or social justice, support many of the themes presented in 6.2, including privacy, fairness and non-discrimination, promotion of human values, safety and security and respect for international norms of behaviour. In addition, many sources of international law and legal principles can individually complement several of the themes. They include, but are not limited to the following:

— the Universal Declaration on Human Rights, see Reference [30];

— the UN Guiding Principles on Business and Human Rights, see Reference [32];

— the International Convention on the Elimination of All Forms of Racial Discrimination, see Reference [33];

— the Declaration on the Rights of Indigenous People, see Reference [34];

— the Convention on the Elimination of Discrimination against Women, see Reference [35];

— the Convention of the Rights of Persons with Disabilities, see Reference [36].

These sources can be understood in terms of their objective of enhancing standards and practices with regard to business and human rights, and to achieve tangible results for affected individuals and communities. Relevant issues include due diligence by an organization to identify and mitigate human rights impacts. Where human rights are impacted by an organization's AI activities, clear, accessible, predictable and equitable mechanisms can address and solve grievances.

Some examples of potential impacts of AI on civil and political human rights include:

— right to life, liberty and security of person (e.g. the use of autonomous weapons or AI-motivated intrusive data collection practices);

— right to opinion, expression and access to information (e.g. the use of AI-enabled filtering or synthesizing of digital content);

— freedom from discrimination and right to equality before the law (e.g. impacted by the use of AI-aided judicial risk assessment algorithms, predictive policing tool for forward thinking crime prevention or financial technology);

— freedom from arbitrary interference with privacy, family, home or correspondence (e.g. unauthorized, AI-based means and measures to collect sensitive biometric and physiological data);

— right to education and desirable work (e.g. the use of AI in recruiting people for employment or providing access to education and training).

6 Themes and principles

6.1 General

In addition to the Human Rights practices referred to in Clause 5, AI principles can help guide organizations develop and use AI in responsible ways. The purpose of these principles is to support organizations beyond nonmaleficence and to focus on beneficence of technology. For example, designing AI that is intended to promote social good and that serves that specific function rather than simply aiming to avoid harm.

These principles do not only cover AI providers and producers and their intended use of the AI systems. When making AI systems available to AI customers and other stakeholders, it is important to also examine their potential misuse, abuse and disuse. As emphasized in ISO/IEC TR 24028:2020, 9.9.1, this includes:

— over-reliance on AI systems leading to negative outcomes (misuse);

— under-reliance on AI systems leading to negative outcomes (disuse);

— negative outcomes resulting from using or repurposing AI systems in an area for which it was not designed and tested (e.g. abuse).

AI systems are particularly susceptible to disuse and misuse because of the way in which they mimic human capabilities. When a system seems human-like yet lacks the context that humans would take into account, users can misuse or disuse it. Such misuse or disuse can arise from trusting it more or less than warranted. For example, with autonomous driving, medical diagnosis or loan approvals.

In response to these concerns, in anticipation of government regulation, or in an attempt, through industry self-regulation, several sets of principles for AI have emerged out of the international community. These have been documented in various publications, see Annex A.

This clause follows the structure laid out by the Berkman Klein Center report, see Clause A.1 by grouping AI principles into themes. The themes emerged from the ethical concern that principles attempt to address. Principles within these thematic groups can vary widely and can even contradict each other. AI-specific themes complement those featured in ISO 26000, which sets out principles for an organization to consider when aiming to behave in a socially responsible manner.

6.2 Description of key themes and associated principles

6.2.1 Accountability

Accountability[84] occurs when an organization accepts responsibility for the impact of its actions on stakeholders, society, the economy and the environment. Accountability means that an organization accepts appropriate scrutiny and accepts a duty to respond to this scrutiny. Hence, accountability for AI decisions means ensuring the organizations are capable of accepting responsibility for decisions made on its behalf and understand that it is not absolved of responsibility for erroneous decisions based on, for example, AI machine learning output.

Accountability specifies that the organization and its representatives are responsible for the impact of negative consequences resulting from the AI systems' and applications' design, development and use or misuse by anyone deploying AI technologies. Accountability also provides focus and attention to consider the unintended consequences that can arise due to the evolutionary nature of AI systems and applications, and difficulty predicting how AI systems and applications can be used and repurposed once deployed. Without clear requirements for accountability, constraints and boundaries are unfettered, and potential harms can go unnoticed.

Accountability for the organization's decisions is ultimately the responsibility of the group of people who direct and control an organization. However, accountability is often delegated to the appropriate responsible parties. Employees, therefore, can be trained to understand the implications of their work

in developing, deploying or using AI tools and to be accountable for their area of responsibility. They can also understand what actions to take to ensure appropriate decisions are being made, whether in an organizational or engineering context. For example, it is the organization's responsibility to establish non-discriminatory and transparent policies. It is the engineer's responsibility to develop AI systems and applications that follow those policies by ensuring the development and use of non-discriminatory, transparent, and explainable algorithms.

Accountability provides necessary constraints to help limit potential negative outcomes and establish realistic and actionable risk governance for the organization. Combined, they help to define how to prioritize responsibilities. Some aspects that are covered by this theme are:

— working with stakeholders to assess the potential impact of a system early on in the design;

— validating that stakeholder needs have actually been met;

— verifying that an AI system is working as intended;

— ensuring the traceability of data and algorithms throughout the whole AI value chain;

— enabling a third-party audit and acting on its findings;

— providing ways to challenge AI decisions;

— remedying erroneous or harmful AI decisions when challenge or appeal is not possible.

6.2.2 Fairness and non-discrimination

The theme of fairness and non-discrimination[85] [86] aims to ensure that AI works well for people across different social groups, notably for those who have been deprived of social, political or economic power in their local, national and international contexts. These social groups differ across contexts and include but are not limited to those that require protection from discrimination based on sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation. Some aspects that are covered by this theme are:

— mitigating unwanted bias against members of different groups;

— ensuring that training data and user data are collected and applied in a way reflective of members of different groups;

— treating members of different groups with fairness, equity and equality;

— considering how AI can impact members of different groups differently;

— ensuring equal possibilities for human development and training to all members of different groups;

— ensuring that the impact of human cognitive or societal bias is mitigated during the data collection and processing, system design, model training, and other development decisions that individuals make;

— allowing stakeholders to appeal a decision if they find it unfair.

6.2.3 Transparency and explainability

The transparency principle in ISO 26000 is extended to include explainability of AI systems to ensure that when stakeholders interact with an AI system and application, its decisions are both transparent and explainable, thereby ensuring that its operations are understandable.

The theme of transparency and explainability aims to ensure that people understand when they are interacting with an AI system, how it is making its decisions, and how it was designed and tested to ensure that it works as intended. The ISO 26000 principle focuses on making sure an organization is

transparent in its purposes and processes, whereas AI-specific principles focus on making sure that an AI system is understandable in how it works. Some aspects that are covered by this theme are:

— disclosing traceability, information about algorithms, training data and user data, including how it was collected;

— disclosing the evaluation methods and metrics used to validate how a system works.

— explaining to stakeholders the inputs that were used to reach a decision;

— explaining to stakeholders, as much as is possible, how an AI system arrived at a decision;

— notifying stakeholders when a decision about them is made by an AI system;

— notifying stakeholders when they are interacting with an AI system;

— consider allowing stakeholders to submit test cases to see how the AI system and application reacts to different situations.

Similarly to accountability, transparency and explainability are applied to decisions made at the organizational level as well as at the algorithmic level. For more information on transparency and explainability, see References [86], [87], [88], [89], [90] and [91].

6.2.4 Professional responsibility

Accountability and ethical behaviour are both strongly associated with professional responsibility. The theme of professional responsibility aims to ensure that professionals who design, develop or deploy AI systems and applications or AI-based products or systems, recognize their unique position to exert influence on people, society and the future of AI - especially since policies, norms and principles often lag behind new and emerging technologies. It calls on their professionalism and integrity to ensure the responsible development of AI systems and applications and direct their expertise towards public discourse, education and governance. This also applies to subject matter experts and other professionals who use AI technology to infer or derive decisions for future action (e.g. hiring representatives, academic admissions officials, scientists, judges and law enforcement personnel). This theme is closely aligned with virtue ethics (see 4.3.2), which focuses on the traits, characteristics and virtues of an AI professional. Some aspects covered by this theme are:

— rigorous methods for evaluating the quality of an AI system, understanding the harms that can be caused by quality problems, as well as mitigating against them;

— paying deliberate attention to the likely impacts of an AI system during the development stage, including long term, global effects as well as those that can come from disuse, misuse and abuse;

— consultation of relevant stakeholder groups while developing and managing the use of AI systems and applications, including policymakers, subject matter experts, and interest organizations, especially when determining fitness for purpose;

— ethical guidelines during development, including inclusive design, diversity, transparency and privacy protection;

— professional values and practices of scientific integrity, including scientifically rigorous methods, a commitment to open inquiry, multidisciplinary collaboration. This includes assessing the scientific validity of AI systems and applications making predictions, recommendations or decisions on the basis of correlated data, within the specific context in which the system can be deployed.

6.2.5 Promotion of human values

The theme of promotion of human values aims to ensure that AI is deployed and utilized in a way that maximizes benefit to society, promote humanity's wellbeing and encourage human flourishing. There is literature indicating that some human values are potentially universal.[92] Some uniquely human values include (but are not limited to) integrity, freedom, and respect. AI can be deployed in a way

that respects and aligns with social norms, humanity's best interests, core cultural beliefs and values, and stakeholder values.[93] This theme extends beyond nonmaleficence to focus on the technology's beneficence. For example, designing AI intended to promote social good and serves that specific function in addition to aiming to avoid harm. Promoting human values throughout development, implementation, and deployment of AI is critical.

Particular applications of AI that aim to promote human values include (but are not limited to):

— improving health and healthcare;

— improving living situations;

— improving working conditions;

— environmental and sustainability efforts.

6.2.6 Privacy

Privacy aims to ensure that AI systems and applications are developed and implemented with natural persons' right to privacy in mind, as well as deceased persons' (through the executor of their estate or nominee as applicable). Right to privacy has become one of the most prominent themes in AI development, due in large part, to the General Data Protection Regulation in the European Union, see Reference [94]. An organization can ensure it employs mechanisms to implement and demonstrate compliance with legal requirements concerning privacy and data protection regulations[94]. Procedures, methods and criteria are needed to ensure that AI algorithms are sourced, designed and used to treat the individual in a representative and transparent manner, ensuring the respect of privacy. Common dimensions of privacy include:

— limiting data sourced, collected, used or disclosed to that which is necessary for accomplishing the intended purposes and tasks;

— communication of the purpose of the processing of personal identifiable information and any sharing of it;

— consent: transparency on the data held on a natural or a deceased person, natural or deceased persons' data not to be collected or used without their knowledge or permission;

— control over the use of data: natural or deceased persons' control of the use of its personal identifiable information;

— natural or deceased persons' degree of influence over how and why their information is used;

— ability to restrict data processing: natural or deceased persons' power to have data restricted from collection or use in connection with AI technology;

— rectification: enable natural or deceased persons to modify information if it is incorrect;

— erasure: enable natural or deceased persons to remove personal data from an AI system and application;

— enabling natural or deceased persons to view personal data used by an AI system and application;

— privacy by design: integrating considerations of data privacy into the development of AI systems and applications and throughout the overall life cycle of data use[95];

— dispute resolution: offer mechanisms for resolving disputes in relation to these features.

6.2.7 Safety and security

The theme of safety and security are distinct, yet closely related concepts[96]. The safety and security ethical theme, which can be called "loss prevention", focuses on the intended performance of AI systems and applications and their resistance to being compromised. Safety aims to avoid reducing injury and

damage to the health of people, property and the environment. Safety is often an important aspect of product-specific regulation. Security aims to reduce the probability of unauthorized and inappropriate access to data, assets or property since AI can introduce new security threats like adversarial attacks, data poisoning and model stealing[97]. Of particular importance is systems that interact with the physical world that can affect persons, communities or both. This theme is related to other themes that are concerned with human control of technology and accountability since the themes' implementation mechanisms can also help improve the safety and security of AI systems and applications. Some aspects that are covered by this theme and are often part of a security by design approach to AI are:

— reliability: AI systems and applications that perform consistently as intended;

— testing and monitoring can help minimize downstream misuse and abuse of an AI system and application. Testing and monitoring also help determine whether systems are robust and designed to withstand unforeseen events and adversarial attacks that can damage or manipulate;

— predictability: improving predictability in AI systems and applications is generally presented as a key mechanism to help mitigate the risk of AI systems being compromised;

— assessment and mitigation of safety related risks considering the complexity of the environment in which AI systems and applications can be used (e.g. autonomous driving);

— ensuring that AI systems and applications work as intended (robustness) in all scenarios. This applies in particular to continuous learning systems.

6.2.8 Human control of technology

The theme of human control of technology aims to ensure that important decisions remain subject to human review, see References [59], [85, page 5], [92] and [98, point 1]. The theme refers to the importance of respecting the autonomy of users affected by automated decisions. Human control can be achieved by designing AI systems and applications in such a way that those impacted by automated decisions are able to request and receive human review of those decisions. For high-risk systems, this can also take the form of a requirement to have a qualified human in the loop to provide authorisation of automated decisions. Some aspects that are covered by this theme are:

— design AI systems and applications that enable human operators to review or authorize automated decisions;

— allow for the ability to opt in or opt out of automated decisions;

— critically evaluate how and when to delegate decisions to AI systems and applications, and how such systems and applications can transfer control to a human in a manner that is meaningful and intelligible.

AI systems and applications can support human autonomy and decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems and applications enable a democratic, flourishing and equitable society by supporting user agency, fostering fundamental rights and allowing human oversight.

6.2.9 Community involvement and development

Social responsibility related to community involvement and development implies identifying stakeholders affected by an AI system and application, and the solicitation and satisfaction of their concerns in a transparent manner. Special attention is given to vulnerable and less developed communities living in geographical areas where AI systems and applications operate. Because many AI-based products and services are virtual and transnational, identifying such communities can represent a challenge. AI systems can have potential impacts on lives of all peoples globally, regardless of whether they are direct consumers of AI-based services or indirectly affected stakeholders. Therefore, an organization can consider not only identifying and engaging with stakeholders, but also recognizing (and building) a relationship with a community, the organization's place in that community and its responsibility for the social, political, economic and cultural development of that community.

A further consideration for organizations aiming to involve affected communities is to include community contribution to the innovation process when creating and operating AI-based goods and services. This is especially relevant when introduction of AI is disruptive to existing patterns of community life, its cohesion and employment opportunities. In particular, impact on employment through AI-driven automation raises issues of community education and culture, employment creation, skills development, and investment in community social and health measures to mitigate negative impacts of its negative impacts. Organizations can involve communities in multiple ways, for example in interviews, focus groups, workshops, conferences, questionnaires, surveys, studies and stakeholder advisory boards. For more information see References [19] [86, end of requirement V] and.[93] Reference [87] emphasizes stakeholder consultation, which for some projects can require community involvement, and possibly citizen panels.

6.2.10 Human centred design

The theme of human centred design in ISO 9241-210 aims to ensure that people who use the AI system and application as well as other stakeholder groups, including those who can be affected (directly or indirectly) by their use, are taken into account. Ensuring that an AI system and application is human-centred[86] includes that:

a) the design is based upon an explicit understanding of users, tasks and environments;

b) users are involved throughout design and development;

c) the design is driven and refined by user-centred evaluation;

d) the process is iterative;

e) the design addresses the whole user experience;

f) the design team includes multidisciplinary skills and perspectives.

One suggestion of how to do design satisfying these goals is in Reference [93].

6.2.11 Respect for the rule of law

The rule of law demands, inter alia, that even powerful organizations and systems comply with the law. Compliance with legal requirements, including recourse to judicial redress as appropriate against decisions rendered by AI systems and applications, is an established aspect of information and communication technology, data governance and risk management. The Organisation for Economic Co-operation and Development [OECD] states the importance of respecting the rule of law throughout the AI system life cycle[98]. The High-Level Expert Group of the European Union highlights the rule of law and the need for organizations not to use an individual's data unlawfully in relation to privacy legislation[59]. Further, the Ethically Aligned Design initiative by the Institute of Electrical and Electronic Engineers [IEEE EAD] encourages stakeholders to engage in educating government, lawmakers and enforcement agencies about the potential for the misuse of AI[25].

According to ISO 26000, following the rule of law in each jurisdiction in which it operates can include:

— complying with laws and regulations even when they are not enforced in that jurisdiction;

— abiding by all legal obligations throughout the whole AI value chain and periodically reviewing compliance of the stakeholder's activities and relationships;

— ensuring the purposes for which AI is developed and used to be lawful and specified.

6.2.12 Respect for international norms of behaviour

ISO 26000 advises that in situations or in jurisdictions where the rule of law does not provide strong safeguards for social and environmental impact, organizations can strive to respect international norms of behaviour and avoid complicity with organizations that do not respect such norms. Further,

it advises organizations to review the nature of their activities in jurisdictions where legislation or its implementation departs from international norms in a way that has significant consequences. IEEE EAD advises stakeholders to engage in establishing new norms related to AI systems.[25]

6.2.13 Environmental sustainability

An organization involved in products or services can undertake to minimize impact to the environment, the health and well-being of stakeholders and the impact on wider society.[99] An organization can commit to substantially reduce greenhouse gas emissions, to limit the global average temperature increase,[100] abide by the sustainable development goals of the United Nations, such as combatting climate change and its impacts, committing to clean water, sanitation, affordable and clean energy, responsible consumption and production – both on life below water and life on land.[31] One sustainability dilemma challenging data- and computing-intensive technologies, such as AI, is the ever-increasing need for energy resources as large data sets and algorithms require consumption of even greater amounts of processing power. This increased need is occurring even as global sustainable development goals call for energy efficiency and lowering non-renewable consumption. Hence, the importance to offer transparent information to stakeholders about energy consumption, climate change and the mitigation of adverse impacts across the AI-based service value chain, in order to enable stakeholders to make sustainable decisions. An organization can also utilise AI systems and applications to foster sustainability and manage environmental impacts and climate change, through a life cycle approach aimed at reducing waste, reusing products and components, and recycling materials. Examples include energy grid optimisation, precision agriculture, sustainable supply chains, climate monitoring, and environmental disaster prediction.

6.2.14 Labour practices

Labour practices are addressed in ISO 26000. Furthermore, the International Labour Organization, and the UN tripartite agency, brings together governments, employers and workers to set labour standards, develop policies and devise programmes that are adopted by consensus, to promote decent work. Of particular interest are fundamental conventions covering what are considered fundamental rights of work[101]:

— freedom of association and the effective recognition of the right to collective bargaining;

— the elimination of all forced or compulsory labour;

— the effective abolition of child labour and adherence to minimum working age legislation;

— equal remuneration and the elimination of discrimination.

Potential considerations regarding the role AI plays in labour relations include:

— ensuring that the rules regarding employment and employment relationships are understood and that humans are involved in the type of decisions that require effective human oversight and empathy (for example the use of AI in managing workers, including gig-economy workers, avoiding discrimination between workers, preventing disproportionate and undue surveillance at work, particularly in remote work, protecting worker privacy, eliminating all forms of forced or compulsory labour and the effective abolition of child labour);

— assuring fair remuneration, working conditions and health and safety, workers protection and other concerns are addressed, (for example, crowd sourced and outsourced workers preparing training data or content moderators exposed to AI-mediated social media content);

— issues of human development and training, especially in a setting where the introduction of AI eliminates work roles or changes their nature in a major fashion (for example, retraining);

— anticipating the consequences of the introduction of AI and reskilling of the workforce;

— assurances that respect for human life and human dignity are maintained and that AI and big data systems do not negatively affect human agency, liberty and dignity;

— providing for rules on businesses' and developers' liability;

— making AI the subject of social dialogue and collective bargaining according to the rules and practices in place in each organization.

7 Examples of practices for building and using ethical and socially acceptable AI

7.1 Aligning internal process to AI principles

7.1.1 General

This clause contains a non-exhaustive list of practices for building and using ethical and socially acceptable AI. Such considerations would typically be a part of a broader organization's management process.

7.1.2 Defining ethical AI principles the organization can adopt

In addition to creating an initial set of principles that align with organizational values, it can be of interest also to create a process to evaluate and update these principles on a regular basis. Key AI-related themes and associated principles are described in 6.2.

Principles can be more impactful if they are communicated externally and written as clear, unambiguous statements of what an organization can or cannot do. Clear statements can help an organization assess whether it is adhering to its principles or not.

7.1.3 Defining applications the organization cannot pursue

In addition to establishing AI principles, the organization can also list application domains for which it categorically cannot design or deploy AI. AI principles tend to be utilitarian, which means that if the benefits are substantial enough, such benefits can always outweigh potential harms. Consequently, the decision about application areas the organization cannot pursue is driven by deontology, defining areas of application that can never be pursued, without exception. Such a list can be grounded in widely accepted human rights and can reflect and project an organization's commitment to particular core values.

7.1.4 Review process for new projects

Key AI-related themes and associated principles as described in 6.2 are intentionally high-level and require careful and nuanced consideration when being applied to concrete projects. A robust, repeatable, and trusted review process can include several elements:

— a review body encompassing the range of perspectives, disciplines and experience required for a thorough review. This can be divided into one team that handles day-to-day operations and initial assessments, and one more senior decision body that handles the most complex and difficult issues, including decisions that affect multiple products and technologies, or issues where conflicts between multiple ethical themes and principles exist. Having a dedicated review team can help ensure that the AI ethics expertise in the organization grows over time, and that people throughout the organization are able to draw on their experience;

— a procedure for flagging projects to the review body, and awareness of this procedure throughout the organization;

— guidelines, frameworks and protocols for the review body to consistently and reliably assess a project and reach a verdict on its alignment with the AI principles, other ethical considerations or both;

— teams or individuals that can dedicate their time to develop technical mitigation strategies to the ethical issues identified by the review body, for example related to unfair bias, explainability and human control;

— procedures for recording case law and ensuring that precedents set can be carried to relevant cases;

— follow-up workflows to ensure that the recommended or required mitigations are adequately implemented;

— regular monitoring to ensure that the AI system and application is working as intended throughout the life cycle. This can include monitoring for concept and distribution drift, periodic retraining to ensure model fairness, updating language models to reflect changes to linguistic practice, and assessing misuse or disuse that can occur over time for example due to automation bias.

These elements, depending on the size and nature of the organization, can be made up of both internal and external stakeholders, the latter including external agencies, advisory boards and invited subject matter experts. Reviewing a project can also include using a formal method for risk assessment, such as the method for risk assessment established in ISO/IEC 23894 (based on ISO 31000).

7.1.5 Training in AI ethics

Training on ethics, principles and organizational values for all members of the organization is key to establishing a culture of ethical AI development, deployment and use. The training can emphasize why ethical development of AI is important, how to align with principles and values, how to enact an ethical decision-making process in the product life cycle, how to apply ethical frameworks, methodologies and impact assessments and how processes and governance structures relate to principles work in the organization. The training can also uncover the ethical issues that arise when developing and deploying AI, going a level deeper than the business rationale for developing AI systems and applications, and drilling down into what values are at stake when implementing a new system. Training can also be customized based on specific roles, with emphasis on how different stakeholders are in a position to spot issues depending on their expertise, area of responsibility, and position in the life cycle. Finally, training can include AI digital literacy schemes for all relevant stakeholders.

7.2 Considerations for ethical review framework

7.2.1 Identify an ethical issue

The first step when making a decision is to recognize the ethical issues in the situation. This requires decision-makers to go beyond considering the business justifications (what is legal, what is efficient, etc.) to develop, deploy and implement AI and to drill down into the ethical issues (what is good for humanity, what respects the user, etc.). Training in AI ethics mentioned in 7.1.5 can help decision-makers build this ethical identification skillset. Identifying ethical issues can occur throughout the AI value chain and life cycle. This ranges from the ideation stage ("can we build this AI system and application?"), the development stage ("what data can we use?", "how can we mitigate against potential risks?"), the deployment stage ("how can the AI system and application be used for purposes other than intended?") and post-launch ("how is the AI system and application being used?").

7.2.2 Get the facts

Decision-makers benefit from having access to as many relevant facts and much relevant information as is practical. Gathering additional information as it relates to AI ethics-related decisions can involve asking relevant stakeholders (product managers, technical leads, researchers, etc.) for more documentation and information. Relevant stakeholders can also include those outside the immediate team developing the technology and can extend to additional internal and external stakeholders (members of a demographic likely to be impacted by the technology, a non-governmental organization with expertise in this area, etc.).

7.2.3 List and evaluate alternative actions

Different stakeholders are likely to propose different actions and to focus on different concerns, for example business, legal, and ethical. Decision-makers can ideally enumerate both the proposed actions and the list of concerns that they are using to evaluate each action. For the purpose of AI ethics, it is important that one of the concerns is adherence to the AI principles that the organization has chosen to adopt.

7.2.4 Make a decision and act on it

Once a decision is made, the deciding person or body (such as those who are accountable to the decision) can explicitly outline any recommendations and mitigations that are part of the decision. The decision can designate ownership over its implementation, to establish clear roles of accountability, including ongoing monitoring of the decision (who is monitoring outcomes of the decision, the effectively use of feedback from external stakeholders or how information is communicated to those affected etc.). The decisions can be documented with proper explanation and justification of the decision. Proper documentation helps catalogue precedents set and mitigations outlined so as to increase the effectiveness of decision-making going forward and to establish a clear baseline for future similar cases.

7.2.5 Act and reflect on the outcome

At the end of the process, it can be valuable to look back on how to notice ethical issues earlier, and whether any of the mitigating actions can be reused in a broader context. A post-mortem can help the organization to learn and to improve its processes, particularly if the organization is responding to an ethical issue that arose unexpectedly, and that was not caught by the existing review processes.

8 Considerations for building and using ethical and socially acceptable AI

8.1 General

This document contains a non-exhaustive list of considerations for building and using ethical and socially acceptable AI. Such considerations would typically be a part of an organization's broader management process.

As discussed in 6.2, ethical issues concerning AI systems and applications, often require context awareness. That is to say, the presence, nature, extent and severity of an ethical concern with an AI system and application often depends upon the particular socio-political, economic, physical context of its development, implementation, audience or use. Therefore, the ethical use of AI systems and applications is related to a sufficiently rich awareness and understanding of the relevant context, since what can be ethically and socially acceptable in one context (sector, geographical region, intended purpose, demographic audience, political or physical environment, etc.) is not acceptable in another context.

Considerations in this document can be complemented considering type of AI system and application, its design and intended functions. Furthermore, a non-exhaustive list of use cases is referred to in Annex B.

8.2 Non-exhaustive list of ethical and societal considerations

8.2.1 General

This clause contains a non-exhaustive list of considerations for building and using ethical and socially acceptable AI systems and applications. Such considerations would typically be a part of a broader organization's management process.

8.2.2 International human rights

International human rights considerations include the following:

— In relation to AI systems and applications, are international human rights fully respected?

— Are any processes in place for end users, or other persons directly or indirectly affected by the AI system, to report potential human rights issues, and for the organization to respond and take such feedback into account in order to improve the AI system?

— Are any processes in place to assess and report to the relevant stakeholders' instances of short term or lasting long-term impact on a data subject's autonomy and agency?

— Are processes in place to provide stakeholders a right of redress, relief or remedy or a means of seeking relief or remedy?

8.2.3 Accountability

Accountability includes considerations of appropriate processes in place in relation to the following:

— Is there an end to end trail of decisions made in relation to the AI system?

— Can the AI system and application be disabled in case of deviation from intended outcomes?

— Can abuse, misuse or disuse of the AI system and application be detected or predicted?

— Are testing and monitoring strategies and processes in place to avoid creating or reinforcing bias in data and algorithms?

— Are the implemented algorithms tested with regards to their reliability, correctness and reproducibility and are the reliability and correctness measures and reproducibility conditions under control?

— How and where can the test methodology, results, and changes based on results be documented?

— Are there clearly defined duties and responsibilities for stakeholders involved in the development or deployment of AI systems and applications?

— Is the output of AI systems and applications transparent and explainable?

— Is the application of AI systems and applications made in compliance with applicable regulations and standards?

8.2.4 Fairness and non-discrimination

Fairness and non-discrimination considerations include the following:

— Which actions have been taken to identify and mitigate bias in the dataset used for model training?

— Is the data source likely to contain historical bias that risks being perpetuated?

— Is the training data representative of the affected population?

— Are there clearly communicated reporting mechanisms on how to raise exclusion issues, especially by users of, or others affected by, the AI system and application?

8.2.5 Transparency and explainability

Transparency and explainability considerations include the following:

— Is it clear who or what can benefit or come to harm from the intended function of the AI system?

— Is the nature of the AI system, and are the potential or perceived risks communicated in a clear and understandable way to the intended users and people potentially impacted by the system?

— Are traceability functions (e.g. logging of activities) of the AI system available to make it auditable, particularly in critical situations?

— Is there a stakeholder reporting and communication strategy regarding the use, collection and maintenance of training datasets?

— Is there an 'end to end' reporting and communication strategy regarding collection, maintenance and use of data?

— Is there a stakeholder reporting and communication strategy with respect to how any data collected by the deployed system are collected, used and maintained?

— Does the strategy for reporting and communication accommodate communications to relevant stakeholders including management, government regulators, civil society, journalists and other watchdogs?

— Were employees and employees' representatives consulted about the nature and scope of the AI system?

— Does the organization communicate the AI system's and application's purpose and data flow processes to stakeholders, and any others whose information is collected, inferred, used, maintained or some combination thereof (for example through a notice placed near the deployment or a QR code with a policy link)?

— Can the system clearly convey benefits, risks, and how benefits outweigh potential risks to all relevant stakeholders?

— Are potential sources of variability in decision-making that occur at prediction time identified, and is it possible to measure this variability?

— When deploying an AI system created by a third party, what kind of documentation, training and support is required by the third party to prevent misuse, disuse or abuse?

— Are processes in place to assess the environmental impact of the development and deployment of the AI system (e.g. electricity, water, land use)?

8.2.6 Professional responsibility

Professional responsibility considerations include the following:

— Are the correlations upon which the AI system makes predictions, decisions and recommendations supported by appropriate scientific research?

— Do measures exist to ensure the security, integrity, validity and accuracy of the data used by the AI system and application?

— In addition to accuracy of data, what system accuracy testing is necessary, prior to deployment, in an environment that controls known and reasonably foreseeable risks?

— Are processes in effect to take into account known scientific flaws and factors related to deployment in problematic settings that can contribute to harmful outcomes?

— Is technology fit for purpose, especially for deployment in the specific context?

— Are models meant for deployment sufficiently robust and of adequate quality and have been built on comprehensive scenarios?

8.2.7 Promotion of human values

Promotion of human values considerations include the following:

— Is information provided in the case of possible harm to humans, as a result of the AI systems' and applications' predictions, recommendations or decisions?

— Have experts, representatives of impacted groups (including workers), or other stakeholders been consulted in determining whether the AI system and application, as deployed, is legitimate in the eyes of the persons who would be most affected?

— If applicable, is there a description of the stakeholders of the application?

— Are processes in place to disclose known vulnerabilities, risks or bias and is there a process in place for reporting them?

— Are processes in place to assess the AI systems' and applications' impacts on employment and the future of work, and to report such instances to relevant stakeholders?

8.2.8 Privacy

Privacy considerations include the following:

— Are communication mechanisms available to indicate where issues related to privacy can be raised?

— Are processes in place for end users, or other stakeholders directly or indirectly affected by the AI system, to report privacy concerns?

— Are processes in place for relevant stakeholders to respond to privacy concerns and to take such feedback (when justified) into account to improve the AI system?

— When a user is unwilling to share all data required to buy, rent or otherwise access a service or product, is the user refused the services or products in their entirety or are alternatives offered?

— How carefully have risks associated with the collection and storage of personal data been assessed?

— Are there internal data governance mechanisms in place that include traceability of how the personal data was obtained?

— Has the internal data governance and handling of personal data been respected, and is the data traceable or re-identifiable?

— Has informed consent been properly collected?

— Can the user or other stakeholder fully exercise their rights on its personal identifiable data?

— Is the data subject informed on how valid consent is given, and if needed or requested by the data subject, on how such consent can be revoked?

— Has privacy definitions applicability been assessed, considered and acted upon in the context (sector, geographical area, etc.) where the AI system is being developed and deployed?

— What data flow and verification capabilities are being provided to ensure that the collected personal information is used as intended?

— Are there features to support the ability of data subjects to review the presence, relevance and accuracy of their personal information?

8.2.9 Safety and security

Safety and security considerations include the following:

— Are communication mechanisms available to raise issues related to safety and security?

— Are there any known safety and security impacts in the intended use of the AI system and application or in the event of a failure of its intended functions (i.e. providing wrong results, being unavailable, or used for a task for which it has not been tested)?

— What technical measures can be taken to protect the collected data against loss and unauthorized access, destruction, use, modification and disclosure?

8.2.10 Human control of technology

Human control of technology considerations include the following:

— Does the AI system require a qualified human in the loop to minimize risk?

— Do measures exist to prevent abuse, misuse or disuse of the AI system?

— Is there documentation that specifies contexts and persons for which the technology is (and is not) expected to perform optimally, given data set composition and other limitations?

— Are there appropriate measures in place to ensure that control over the AI system can safely be transferred to and performed by a human operator when needed?

8.2.11 Community involvement and development

Community involvement and development considerations include the following:

— When AI systems and applications are used to manage public benefits and healthcare, is there sufficient oversight in making eligibility determination and allocating government benefits?

— Is human assistance available to end users engaging with AI systems and applications?

— Is all relevant information on AI systems and applications, which that are used in connection with public services, made public and accessible?

8.2.12 Human centred design

Human-centred design considerations include the following:

— Does the organization acknowledge that people differ in their abilities and needs, uses ergonomics and social data on the nature and extent of these differences?

— Does the organization make usability and accessibility strategic business objectives?

— Does the organization adopt a total system approach?

— Are health, safety and well-being business priorities?

— Does the organization value its employees and create meaningful work?

— Is the organization open and trustworthy?

— Does the organization act in socially responsible ways?

8.2.13 Respect for the rule of law

Respect for the rule of law[102] considerations include the following:

— Does the judiciary have a sufficient level of understanding about AI system and applications to ensure respect for the rule of law?

— In relation to AI systems and applications, is a person ensured due and diligent judicial process and fair and equal treatment under the rule of law?

— Does the judiciary act as a guarantor of recourse whenever international human rights or freedoms are breached by AI systems and applications?

— Is a person that is subject to an algorithmic decision able to legally examine and contest its reasoning?

— Is the rule of law safeguarded by the accurate handling of data throughout the whole AI value chain?

8.2.14 Respect for international norms of behaviour

Respect for international norms of behaviour considerations include the following[102]:

— Are international norms of behaviour respected when using or promoting the use of AI systems and applications?

— Are legitimate interests (e.g. security), rights (e.g. intellectual property rights) or relevant stakeholder values[93] safeguarded?

8.2.15 Environmental sustainability

Environmental sustainability considerations in relation to AI systems and applications include the following:

— Has the carbon footprint been evaluated taking into account its benefits, any reduced impact, and any extra consumption generated?

— Is there a policy to use renewable energy and to avoid (or use) heat dissipation, especially in data centres?

— Have studies been conducted to evaluate and reduce emissions of pollutants into the air, water and soil (wherever relevant) as much as possible?

— Is there a green procurement practice to evaluate suppliers of goods and services on their environmental impacts?

— Is there a life-cycle approach (including for disposal) to reduce waste, re-use products or components, and recycle materials?

8.2.16 Labour practices

Labour practices considerations include the following:

— Have the bodies in charge of social dialogue in the organization (whatever their form; trade unions, staff representatives, health and safety committees, etc.) been informed and, when relevant, consulted on the implementation of an AI system and application?

— Has the company staff been directly informed of the implementation of an AI system and application?

— When AI involves surveillance of workers, are there steps taken to check that the surveillance is limited to its sole objective and without disproportionate and undue effects? Is there a procedure to report related problems?

— Has the possibility of discriminatory treatments of workers on the basis of biased algorithms been anticipated?

— Are steps taken to prevent abuse of data protection and privacy in the workplace?

— Has the impact on jobs been anticipated with workers reskilling when necessary?

— Is there a plan to provide AI and digital literacy schemes for the affected workers?

— Are the rules on the organization's or developers' responsibility and liability clearly understood and communicated?

Annex A
(informative)

AI principles documents

A.1 Berkman Klein report

Mapping consensus in ethical and rights-based approaches to principles for AI[37].

Several sets of principles or themes for AI have emerged, and keep emerging, out of the international community. 36 international AI principles documents published between 2016 to 2019 are analysed in the survey by the Berkman Klein Center (see 6.1), based on input from different industries and different geographies. More recent international publications, of comparable status, which also list several sets of AI principles or themes, are set out in Clause A.2 to A.9.

A.2 Global

Ten principles for ethical AI[38]

After reviewing 90 organizations PWC™1) identified ten core principles to help defining ethical AI. Based on its own work, both internally and with clients, the article presents a few ideas for how to put these principles into practice. The principles are Interpretability, Reliability and robustness, Security, Accountability, Beneficiality, Privacy, Human agency, Lawfulness, Fairness and Safety.

Incompleteness of moral choice and evolution towards fully autonomous AI[39]

The problem of ethical decision-making, viewed from the perspective of computer, technical and natural sciences, lies only in the complexity of the topic. AI scientists and developers basically proceed from the Turing machine model, assuming that a machine can be constructed to resolve any problems (including ethical decision-making issues) that can mechanically calculate a particular function if this function can be put into an algorithm. Thus, ethical decision-making is conceived as an abstract concept whose manifestation does not depend on the particular physical organism in which the algorithm takes place, nor on what it is made of. If in practice, a sufficiently complex algorithm is built, it will also exhibit sufficiently complex behaviour that can be characterized as ethical in the full sense of the word. This article reflects the main argument that if a task requires some form of moral authority when it is performed by humans, its full automation, transferring the same task to autonomous machines, platforms, and AI algorithms, necessarily implies the transfer of moral competence. The question of what this competence includes presupposes empirical research and reassessing purely normative approaches in AI ethics.

Hewlett Packard Enterprise™2) (Belgium). *AI Ethics and principles*[40]

The organisation states it considers AI to be a powerful, transformative technology that can amplify human capabilities, but also presents risks. It invests in ethical AI principles and believes in the following ethical and responsible AI principles: AI privacy-enabled security, AI human-focused principle, AI inclusivity principle, AI robust principle and Responsible AI.

UNESCO. *Resource Guide on Artificial Intelligence (AI) strategies*[41]

1) PricewaterhouseCoopers (PwC) is a trademark of PwC Business Trust. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO or IEC.

2) Hewlett Packard Enterprise is a registered trademarks of Hewlett Packard Enterprise Company and/or its affiliates. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO or IEC.

Artificial Intelligence (AI) has rapidly emerged in a context in which sustainable development has been the overarching goal of the international community. The United Nations has called upon governments to develop national strategies for sustainable development, incorporating policy measures to achieve the 2030 Agenda for Sustainable Development. While AI technologies can support breakthroughs in achieving the Sustainable Development Goals (SDGs), they can also have unanticipated consequences that will exacerbate inequalities and negatively impact individuals, societies, economies and the environment. AI implementation will need to be supported by a multi-disciplinary review to steer AI in a direction that will respect human rights and human dignity. This reference guide is a collection of key references that can provide a global overview of discussions on AI Ethics, technical standards, and examples of national strategies. It comprises three main chapters on AI: Ethical Principles and Impacts; Technical Standards and International strategies; and National Strategies. There are no specific AI Ethics principle definition in this document, instead, the reader is encouraged to familiarize themselves with the work of international and national governments, private organisations, civil society and resources from Europe, the Americas and Asia.

Unite Paper. *A Framework for Ethical AI at the United Nations*[42]

The paper not only provides an attempt to define "Ethical AI", it also considers how an organisation such as the UN would go about implementing Ethical AI. It proposes to view ethical values "as being composed of different layers, going from global to group to individual: Core ethical values based on inalienable human rights (human dignity, autonomy); constitutional values (rule of law, equity, privacy); group specific values (beliefs or cultural norms); and individual ethical values (personal convictions)." It recommends that the UN develop a framework consisting of ethical principles, architectural standards, assessment methods, tools and methodologies, and a policy to govern the implementation and adherence to this framework, accompanied by an education program for staff.

UNESCO AD Hoc Expert Group. *Outcome document: first draft of the Recommendation on the Ethics of Artificial Intelligence*[43]

The document contains a distinction between values and ethics principles. Values include Respect, protection and promotion of human dignity, human rights and fundamental freedoms; Environment and ecosystem flourishing; Ensuring diversity and inclusiveness; and Living in harmony and peace. The document identifies the following AI ethics principles: Proportionality and do no harm; Safety and security; Fairness and non-discrimination; Sustainability; Privacy; Human oversight and determination; Transparency and explainability; Responsibility and accountability; Awareness and literacy; and multi-stakeholder and adaptive governance and collaboration.

AI4DA. *Ethics in Artificial Intelligence*[44]

The article contains a good primer and points to up to date resources on the topic. The foundation also provides AI readiness assessments for countries outside the so-called first and second world. Recent reports include the Philippines, El Salvador, Nigeria and Myanmar. In its analysis, it cites a common top five of ethical principles: transparency, justice and fairness, non-maleficence, responsibility and privacy. It furthermore states that "no (document) examined discussed AI within care, nurture, in terms of help, welfare, social responsibility, ecological networks, nor dignity or solidarity. Moreover, the documents mostly came from economically developed countries, while Central and South America, Africa and Central Asia were very underrepresented and thus excluded from the ethics discourse."

A.3 North America

Government of Canada. *Directive on Automated Decision-Making*[45]

The Government of Canada is increasingly looking to utilize artificial intelligence to make, or assist in making, administrative decisions to improve service delivery. The Government is committed to doing so in a manner that is compatible with core administrative law principles such as transparency, accountability, legality, and procedural fairness. This technology is changing rapidly and this Directive will continue to evolve to ensure that it remains relevant. Appendix A contains definitions, including a definition of procedural fairness. Appendix B contains four Impact Assessment Levels that range from impacts that are reversible and brief (Level I) to decisions that are irreversible and perpetual (Level

IV); although they are unrelated to risk. Appendix C lists procedural requirements for each Impact Level; ranging from Peer Reviews to Notices posted through service delivery channels to Human-in-the-loop for decisions to Explanation requirements to Training to Contingency planning to Approval for the system to operate.

Information Technology Industry Council. *AI Policy principles, Executive summary*[46]

A broad-ranging set of topics covering AI, industry and government policy is covered in this document. Regarding values and ethics, the document calls attention to industry's responsibility in promoting responsible development and use, which includes: responsible design and deployment, safety and controllability, robust and representative data, interpretability and liability of AI systems due to autonomy. The latter a topic few other publications have identified as a topic of interest.

IBM. *From Roadblock to Scale: The Global Sprint Towards AI Study*[47]

AI is embedded in everyday life, business, government, medicine and more. By embedding ethical principles into AI applications and processes, systems can be built based on trust. Topics discussed include: trust and transparency principles, everyday ethics for AI, AI fairness and AI explainability, human intelligence augmentation and data ownership.

A.4 South America

Principles and axes of Chile's AI Policy[48]

On October 27, 2021, Chile published its National Policy on Artificial Intelligence which addresses the benefits of AI technology and the goal of positioning the country as a leader in innovation, research, and development, and of democratising technology solutions. The AI policy identifies four transversal principals: AI with a focus on people's welfare, respect for human rights, and security; AI for sustainable development; Inclusive AI as well as Globalized and evolutive AI.

Artificial Intelligence in Brazil: the Brazilian Strategy for Artificial Intelligence (BSAI/EBIA) and Bill No. 21/2020 - October 2021[49]

This article analyses Bill No. 21/2020 in the Chamber of Deputies and discover some important points of the Brazilian Strategy for Artificial Intelligence (BSAI/EBIA) in greater detail. Uniquely among the articles and documents in this annex, it considers the acceleration and strengthening of Information and Communication Technologies (ICT) as a consequence of the COVID19 pandemic. The article links the BSAI/EBIA objective of "to contribute to the elaboration of ethical principles for the development and use of responsible AI" and criticisms of the law as being "generic" and lacking "a technical basis".

Legal framework for artificial intelligence advances in Brazil[50]

Brazil's Congress has passed a bill that creates a legal framework that outlines rules around the use of AI in the country. The bill covers the rights of users of AI systems, such as being informed about the institution responsible for the development of the AI system they are interacting with, as well as the right to access clear and adequate information about the criteria and procedures used by the system. It contains rules around ensuring respect for human rights and democratic values. Ethical concerns frequently raised are listed. Brazil adheres to the Organization for Economic Co-operation and Development (OECD)'s human-centred AI Principles, which provide for recommendations around areas such as transparency and explainability.

Argentina and Uruguay[51]

The article provides an introduction to the AI strategies in Argentina and Uruguay and is a pointer to a variety of resources otherwise not available in English. The Argentinian strategy (published in 2019) seeks to promote AI in the private sector, minimize ethical risks, and develop talent, amongst other objectives. The Argentinian document acknowledges the challenge on how to face the ethical and social challenges that the technological transformation imposes while promoting its development and implementation that favours economic growth and social development. In the course of this development, it calls on the application of AI techniques in which anthropological, cultural and

psychological aspects can play a key role, emphasising to include experts in these areas in the research team. This will allow not only the training of professionals from different disciplines in AI techniques, but also the development of solutions with implementation potential in various areas.

Mexico: the story and lessons behind Latin America's first AI strategy[52]

In 2017 the government of Mexico, together with strategic actors from civil society, private sector, academia, and international cooperation, decided to develop the first AI strategy for the country and Latin America. It sought to answers major AI-related questions including "How will jurisdictions address ethical issues around algorithmic accountability, openness, and transparency?" The policy brief contains the recount of the leaders of the initiative from government and civil society, along with recommendations for other governments seeking to develop AI strategies in the future.

A.5 Africa

Defining what's ethical in artificial intelligence needs input from Africans[53]

"Applying a principle of explicability to AI research in Africa: should we do it?"[53] argues that "Given that values vary across cultures, an additional ethical challenge is to ensure that these AI systems are not developed according to some unquestioned but questionable assumption of universal norms but are in fact compatible with the societies in which they operate. This is particularly pertinent for AI research and implementation across Africa, a ground where AI systems are and will be used but also a place with a history of imposition of outside values." Explicability can assist to "contribute to responsible and thoughtful development of AI that is sensitive to African interests and values".

Top AI challenges in Africa: Compute[54]

The article focuses on the top challenges facing accessing compute in AI in Africa. A diverse group of AI community leaders from Ghana, Nigeria, Kenya, Tanzania Morocco and others talked about nuances in their societies and suggested solutions. Key issues, which need to be overcome before a practical debate over AI Ethics is possible are: Low access to GPUs, Storage capacity challenges, Lack of understanding as to which GPU to select for a desired purpose.

Africa, Artificial Intelligence, & Ethics[55]

The resource is a transcript of a Carnegie Council podcast. Key barriers to AI adoption and the discussion of Ethics concern: the inability of most AI to work with African languages (be that written or spoken); accountability for foreign domiciled organisations operating in Africa; protecting inexperienced consumers of social media platforms driven by algorithm from toxic content; nuances in culture between African countries; access to toolsets from major companies that can be used for local development; trust in technology; the risk of the concept of AI administrative universality across all African nations. It discusses how conversations in communal gatherings called kgotla are transcribed using natural language and incorporated in Botswana's legal and legislative process; thus delivering greater social inclusion.

Towards an ethics of AI in Africa: rule of education[56]

The wide-ranging survey of resources by authors of the African continent is a starting point for a comprehensive survey of AI and AI Ethics in Africa. Specifically, it discusses Ubuntu lessons in Africa and continues to state that "The development of AI must be in accordance with African values about man. Those that put human relationships at the centre and not individualism. Unfortunately, a rational view of people has always been marked by contradictions, exclusions, and inequalities. The development of AI based on data generated by such a vision will contribute to increase the already existing inequalities."

Top AI challenges in Africa: Ethics[57]

Bias due to poor representation, job losses, the application of AI in general, what lives AI should improve, what actions AI should be allowed to take, open data access to especially African data are presented as of specific concern to African communities. In addition to covering Ethics in AI in the African setting,

the article draws attention to the work of the Lacuna Fund, which provides funding to collect data from the underserved population in the fields of Agriculture, Health and Language to overcome fundamental barriers to achieving fair and unbiased outcomes.

A.6 Europe

European Commission. *Proposal for regulation on a European approach on artificial intelligence*[58]

The regulatory proposal aims to provide AI developers, deployers and users with clear requirements and obligations regarding specific uses of AI. At the same time, the proposal seeks to reduce administrative and financial burdens for business, in particular small and medium-sized enterprises.

High-Level Expert Group on Artificial Intelligence. *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*[59].

The document sets out the final assessment list for trustworthy AI (ALTAI) presented by the High-Level Expert group on Artificial intelligence (AL HELG) set up by the European Commission, to help assess whether the AI system that is being developed, deployed, procured or used, adheres to the following seven requirements; (i) human agency and oversight, (ii) technical robustness and safety, (iii) privacy and data governance, (iv) transparency, (v) diversity, non-discrimination and fairness, (vi) societal and environmental well-being, and (vii) accountability.

Addressing AI ethics through codification[60]

The article sets out several approaches based on reviewed material that can be used to codify ethical issues in practical applications of AI systems. The first part of the paper is primarily focusing on two groups of problems considered significant for practical applications of AI systems, whereas the second part of the article considers some of the current discussions on AI ethics, commonly used ethical principles, international regulation of shared interests in AI development, as well as related issues in which AI ethics issues play an important role. The third and fourth parts of the article consider practical approaches used by professional communities to regulate ethical issues.

European Parliament. *Resolution with recommendations to the Commission on a civil liability regime for artificial intelligence*[61]

The suggested creation of a future-oriented and a unified approach, setting standards for citizens and businesses to ensure the consistency of rights and legal certainty, thereby fostering digital innovation and ensuring a high-level protection of citizen and consumer rights, by harmonizing the liability regimes for AI-systems.

European Parliament. *Data subjects, digital surveillance, AI and the future of work*[62]

The report provides an in-depth overview of the social, political and economic urgencies in identifying the so-called 'new surveillance workplace'. The report assesses the range of technologies that are being introduced to monitor, track and, ultimately, watch workers, and looks at the immense changes they imbue in several arenas. A number of principles are set out in the report that apply to the right to a private life and to protections around how workers' data are collected and used. The report looks at some of the tensions across principles.

StandICT.eu - Focus Group report, Road Map on Artificial Intelligence (AI)[63]

The Road Map creates an overview of standardization activities in IEEE, ETSI, ISO/IEC, ITU-T and CEN-CENELEC (chapter 1.3 and Annex A). The Focus Group has identified 13 themes among which the following seven have been addressed for European standardization (chapter 3): Accountability, Quality, Data for AI, Security and privacy, Ethics, Engineering of AI systems and Safety of AI systems.

A.7 Asia

China's New AI Governance Initiatives Shouldn't Be Ignored[64]

The meta review of Chinese AI governance provides an introduction and references policies and documents that consider ethics and governance from Cyberspace Administration of China (CAC), China Academy of Information and Communications Technology (CAICT) and the Ministry of Science and Technology. It points out that CAC is the most mature and its draft set of thirty rules for regulating internet recommendation algorithms break international ground in areas such as explainability and remediation of harm. It argues that these rules can break international ground.

What you need to know about China's AI ethics rules[65]

The article proves an introduction to China's proposed governance and regulations titled "New Generation Artificial Intelligence Ethics Specifications" from a Western perspective. The specifications aim to integrate ethics and morals into the entire lifecycle of AI; promote fairness, justice, harmony, and safety; and avoid issues such as prejudice, discrimination, and privacy/information leakage. The specification applies to natural persons, legal persons, and other related institutions engaged in activities connected to AI management, research and development, supply, and use. Six basic ethical norms are specified: enhance human well-being, promote fairness and justice, protect privacy and safety, ensure controllability and credibility and improve ethical literacy. Additionally, there are goals for managing AI-related projects including; promote agile governance and development sustainability, actively integrate the ethical standards into management processes, ensure the orderly development and sustainability of AI, clarify responsibilities across different stakeholders and standardize management operating conditions and procedures, strengthen risk prevention by improving risk assessment in AI development, encourage the use and diversification of AI technologies to solve economic and social problems and encourage cross-disciplinary, cross-field, cross-regional and cross-border exchanges and cooperation.

Principles and Practice for the Ethics Use of AI in Hitachi's Social Innovation Business[66]

The governance and ethics of AI are key issues and recognizes the significant societal impact associated with the use of AI. Many different aspects of AI ethics are highlighted, among them safety, privacy, fairness, transparency and accountability, and security. The page documents: Hitachi[TM3]'s AI ethical principles; standards of conduct in the planning phase, the social implementation phase and the maintenance and management phase; as well safety, privacy, fairness, equality and prevention of discrimination, transparency, explainability and accountability, security and compliance.

Malaysia AI Ethics Maturity Report 2021 – AI Doctrine[67]

The document summarises Malaysia's AI Ethics National Framework exploring the dichotomy caused by promoting the innovation capacity of AI on the one hand, on the other hand the development and uptake of ethical and trustworthy AI has become central to the development and deployment of AI. The growth of AI raises several questions and challenges that require careful consideration, particularly when the development and deployment of AI raise ethical dilemmas. The document focuses on the distrust AI has caused. It surveys principles that can counter these issues such as explainability, disclosure, privacy and data governance, quality and integrity of data, access to data, and utility. It includes an AI Ethics Maturity index that is representative of the extent of awareness of the utility of an AI system; the degree of transparency and explainability; and the degree of respect for privacy and adoption of data governance measures.

AIDA - AI Readiness in the Philippines[68]

An overview of the AI readiness in the Philippines. This is but one example of how the Artificial Intelligence 4 Development Agency on a monthly basis highlights the adoption and readiness in a non-western country. Other countries-of-the-month have included: Mexico, Argentina, Myanmar, South Korea (calling out the "ppalli-ppalli" culture, literally translated as "quickly, quickly"), Nigeria, Egypt, Columbia, Uzbekistan, Zimbabwe and Singapore.

International Research Center for AI Ethics and Governance. *The Ethical Norms for the New Generation Artificial Intelligence, China*[69]

3) Hitachi is a trademark of Kabushiki Kaisha Hitachi Seisakusho. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO or IEC.

On September 25th 2021, the National Governance Committee for the New Generation Artificial Intelligence published the "Ethical Norms for the New Generation Artificial Intelligence". It aims to integrate ethics into the entire lifecycle of AI, to provide ethical guidelines for natural persons, legal persons, and other related organizations engaged in AI-related activities. The document contains the full set of norms. The set of ethical norms considers the current ethical concerns of privacy, prejudice, discrimination, and fairness from all walks of life in the society. The set of ethical norms includes general principles, ethical norms for specific activities, organization and implementation guidelines. The set of ethical norms puts forward six fundamental ethical requirements, such as enhancing the well-being of humankind, promoting fairness and justice, protecting privacy and security, ensuring controllability and trustworthiness, strengthening accountability, and improving ethical literacy. At the same time, 18 specific ethical requirements for specific activities such as management, research and development, supply, and use of AI are put forward.

Expert Group on How AI Principles Should be Implemented - AI Governance Guidelines working group[70]

Japan has published the document called "Social Principles of Human-centric AI" adopted by the Integrated Innovation Strategy Promotion Council, which contributes to the formulation of the OECD's recommendations on Artificial Intelligence. The social principles for AI are comprised of seven principles: (1) Human-centric, (2) Education/Literacy, (3) Privacy Protection, (4) Ensuring Security, (5) Fair Competition, (6) Fairness, Accountability, and Transparency, and (7) Innovation. The document presents action targets to be implemented by an AI company, with the aim of supporting the implementation of the AI principles that is required for the facilitation of deployment of AI.

What Buddhism can do for AI ethics[71]

The publication explores what Buddhism can offer to the ethics of AI. It proposes that "any ethical use of AI must strive to decrease pain and suffering" and the "Do no harm principle: the burden of proof would be with those seeking to show that a particular application of AI does not cause harm".

Re-imagining Algorithmic Fairness in India and Beyond[72]

The study highlights ways in which issues surrounding AI in India differ from Western countries and can call for different approaches to achieve fairness.

The Threat of AI and Our Response: The AI Charter of Ethics in South Korea[73]

The charter starts by observing that changes due to Artificial Intelligence (AI) are ongoing, and that there is little refutation of the effectiveness of AI. It goes on to observe that South Koreans have been active in discussions to minimize the side effects of AI and use it responsibly, and that the publication of the Korean AI Charter of Ethics (AICE) is one result of it. The study examines how the South Korean society is responding to threats from AI that can emerge in the future by examining various AIECs in the Republic of Korea. In addition, seven AI threats are classified into three categories: AI's value judgment (Human discrimination in AI, AI's weighing of human value), Malicious use of AI (Lethal AI weapons, AI-based cyber attacks, Excessive privacy intrusion) and Human alienation(AI's usurpation of human occupations, Deepening the alienation of the digitally vulnerable), the authors draw fourteen topics based on three categories: Protection of social values (including Prevention of social discrimination, Social inclusion as a whole, Respect for human dignity, Pursuit of human benefit and happiness), AI control (Explainable algorithm, Use of data based on social ethics, Prepare for malfunctions and hazardous situations, Ultimately human controlled, Limiting the purpose of using AI, Activating the post-management system, Clear division of responsibility and Possible to check whether AI is applied) and Fostering digital citizenship (Culture of continuous multilateral communication, Enhancement of AI utilization capabilities). The analysis indicates that Korea has not yet been able to properly respond to the threat of AI's usurpation of human occupations (jobs). In addition, although Korea's AICEs present appropriate responses to lethal AI weapons, these provisions will be difficult to realize because the competition for AI weapons among military powers is intensifying. The article contains a useful overview of Korea's seven AI Charters of Ethics, the first one (the Draft of the Robot ethics Charter - DREC) having been published as far back as 2007.

A.8 Eurasia

First code of ethics of artificial intelligence signed in Russia[74]

The Code will become part of the Artificial Intelligence federal project and the Strategy for the Development of the Information Society for 2017-2030. It establishes general ethical principles and standards of conduct to guide those involved in activities using artificial intelligence. The Code applies to relations involving ethical aspects of the creation (design, construction, piloting), implementation and use of AI technologies at all stages of the life cycle, which are currently not regulated by Russian law or other regulatory acts.

Sber among pioneers adopting AI ethics principles in Russia[75]

The press release announces the board of Sberbank™4) has adopted ethical principles behind AI development and use across Sber™5) Group. The five adopted principles are: secure AI, explainable AI, reliable AI, responsible AI and fair AI. Each principle has been documented with explanatory examples. The principles have been designed taking into account the requirements of the National Strategy for the Development of Artificial Intelligence until 2030.

AI Alliance Russia. *Artificial Intelligence Code of Ethics*[76]

The Code applies to relationships related to the ethical aspects of the creation (design, construction, piloting), implementation and use of AI technologies at all stages that are currently not regulated by the legislation of the Russian Federation and/or by acts of technical regulation. Section 1, principles of ethics and rules of conduct, outlines the core principles and attributes associated with each: protecting the interests and rights of human beings collectively and as individuals (including human-centred and humanistic approach, respect for human autonomy and freedom of will, compliance with the law, non-discrimination , assessment of risks and humanitarian impact); the need for conscious responsibility when creating and using AI (including a risk-based approach, a responsible attitude, considering and taking precautions, no harm principle, identification of AI in communication with a human, data security, information security, voluntary certification and code compliance, control of the recursive self-improvement of AI systems); humans are always responsibility for the consequences of the application of an AI system (supervision , responsibility); AI technologies to be applied and implemented where it can benefit people (including the application of an AI system in accordance with its intended purpose, stimulating the development of AI); interests of developing AI technologies above the interests of competition (correctness of AI systems comparisons, development of competencies, collaboration of developers); the importance of maximum transparency and truthfulness in information on the level of development, capabilities and risks of AI technologies (including the credibility of information about AI and raising awareness of the ethics of AI application).

A.9 Asia Pacific

Australian Government Department of Industry, Science, Energy and Resources. *Australia's Artificial Intelligence Ethics Framework*[77]

The voluntary Artificial Intelligence (AI) Ethics Framework guides businesses and governments to responsibly design, develop and implement AI and is composed of four pillars: AI Ethics principles, Applying the principles, Testing the principles and Developing the framework. This lifecycle-like approach, while not unique to Australia, is infrequently encountered in other resources. Australia's eight AI Ethics principles are designed to ensure AI is safe, secure and reliable. They consider on: Human, societal and environmental wellbeing, Human-centred values, Fairness, Privacy protection and security, Reliability and safety, Transparency and explainability, Contestability and Accountability.

AI Singapore. *An Overview of AI Ethics and Governance*[78]

4) Sberbank is a trademark of Sberbank of Russia. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO or IEC.

5) Sber is a trademark of Sberbank of Russia. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO or IEC.

The article discusses governance through the prism of operationalising AI principles. It identifies five common themes or overarching principles of ethical AI; beneficence, non-maleficence, autonomy, justice and explicability. The article points out that the first four principles correspond with the four traditional bioethics principles, and that they are joined by a new enabling principle of explicability for AI. The document contains references to open source technology tools from IBM®[6] and Microsoft®[7] that can assist Python™[8] developers when incorporating these principles in their work.

AI Forum New Zealand. *Trustworthy AI in Aotearoa: The AI Principles*[79]

To help maintain public trust in the development and use of AI in New Zealand, the Law, Society and Ethics Working Group of the AI Forum has published a set of five guiding principles designed to provide high-level guidance for anyone involved in designing, developing and using artificial intelligence in New Zealand (AI stakeholders), with the goal of ensuring New Zealanders have access to trustworthy AI. The five guiding principles are: Fairness and justice [including (respect for) Applicable laws, Human rights, Rights of Māori, Democratic values and Principles of equity and fairness], Reliability, security and privacy, Transparency, Human oversight and accountability and Wellbeing). This publication is noteworthy for the fact this it is one of the few, if any, to explicit consider and respect the rights of Indigenous people.

Singapore's Approach to AI Governance (ISAGO)[80]

The article discusses the evolution between 2019 and 2020 of Singapore's Model AI Governance Framework from the first to the second edition. The second edition builds on the two guiding principles of the first edition: Decisions made by AI should be explainable, transparent and fair and AI systems should be human-centric. The second edition includes additional considerations (such as robustness and reproducibility) and refines the original Model Framework for greater relevance and usability. For instance, the section on customer relationship management has been expanded to include considerations on interactions and communications with a broader network of stakeholders. The second edition of the Model Framework continues to take a sector- and technology-agnostic approach that can complement sector-specific requirements and guidelines. The practical guidance concerns: Internal governance structures and measures, Determining the level of human involvement in AI-augmented decision making, Operations management and Stakeholder interaction and communication. The work of PDPC is supported by the Advisory Council on the Ethical Use of AI and data.

Singapore Digital (SG:D). *Compendium of Use Cases, Practical illustrations of the Model AI governance framework*[81]

Complementing the Singaporean Model Framework and ISAGO is a Compendium of Use Cases (Compendium) that demonstrates how local and international organizations across different sectors and sizes implemented or aligned their AI governance practices with all sections of the Model Framework. The Compendium also illustrates how the featured organizations have effectively put in place accountable AI governance practices and benefitted from the use of AI in their line of business. These real-world use cases are intended to inspire other companies to do the same and to foster responsible development and adoption of AI.

Advisory Council of the Ethical Use of AI and Data, the Infocomm Media Development Authority (IMDA) and the PDPC. *A Guide to Job Redesign in the Age of AI*[82]

Singapore's first guide that helps organizations and employees understand how existing job roles can be redesigned to harness the potential of AI, so that the value of their work is increased. It provides an industry-agnostic and practical approach to help companies manage AI's impact on employees, and for organizations that are adopting AI to prepare themselves for the digital future and provides guidance on practical steps in four areas of job redesign: Transforming jobs, Charting clear pathways between

6) IBM is a registered trademark of International Business Machines Corporation ("IBM"). This information is given for the convenience of users of this document and does not constitute an endorsement by ISO or IEC.

7) Microsoft is a registered trademark of Microsoft Corporation. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO or IEC.

8) Python is a trademark of the Python Software Foundation. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO or IEC.

jobs, Clearing barriers to digital transformation and Enabling effective communication between employers and employees. The guide is a practical example of addressing the ethical risks and societal sustainability impacts frequently highlighted in other documents in this appendix.

Singapore Computer Society. *AI Ethics and governance body of knowledge*[83]

The AI Ethics and Governance Body of Knowledge (BoK), developed by the Singapore Computer Society (SCS) and supported by the Info-communications Media Development Authority, supports businesses to hire trained AI ethics professionals to deploy AI responsibly. The BoK is a living document based on the Model AI Governance Framework and will be enhanced over time. It forms the basis of future AI ethics and governance training and certification for professionals, both in ICT and non-ICT domains, and also seeks to facilitate the development of curricula on AI governance. The Toolkit referenced was a precursor to the BoK and can be of interest to those less advanced in their AI ethics journey and provides answers to fundamental questions such as: “Why care for AI ethics?”, “What are the benefits of understanding AI ethics?” and “Who is responsible for AI ethics?”.

Annex B
(informative)

Use case studies

B.1 General

Professionals, researchers, regulators and individuals need to be aware of the ethical and societal concerns associated with AI systems and applications. ISO/IEC TR 24030 and various publicly available information provides examples of positive impacts of AI systems and applications. The case studies have been chosen for the purpose of illustrating the ethical and societal concerns or issues related to systems and applications using AI functionality.

B.2 Use cases

B.2.1 Use case No. 1

Table B.1 illustrates concerns about accountability, professional responsibility, privacy, safety and security, community involvement and development, respect for the rule of law and labour practices.

Table B.1 — AI components for vehicle platooning on public roads (self-driving vehicles)

Description	The overall concept of automated platooning is that the lead vehicle would be driven as usual by a trained (professional) driver, and the following vehicles would be driven fully automatically by the system, allowing the drivers to perform tasks other than driving their vehicles. The EU roadmap for truck platooning [EU project - expectation and non-formal skills to empower migrants and to boost local economy (ENSEMBLE)] envisions market introduction of multi-brand platooning by 2025 (use case No. 31 in Reference [15]). Several pilot projects have been carried out since about the year 2000 (use cases No. 27, 28, 29, 32 and 33, in Reference [15]). While a few AI components are already used in the pilot projects (for example lane keeping), future products are likely to incorporate AI solutions on several functional levels.	
Ethical and societal concerns	Stress or boredom for the drivers, constant monitoring, safety, system security, and reliability, risk of hacking and hijacking a long-haul freight truck poses great danger, trust over system reliability when driving next to a computer-controlled platoon. Highly unpredictable traffic environment, legislative situation, standardization. The use case illustrates concerns about accountability (see 6.2.1), professional responsibility (see 6.2.4), privacy (see 6.2.6), safety and security (see 6.2.7), community involvement and development (see 6.2.9), respect for the rule of Law (see 6.2.10), labour practices (see 6.2.13).	
References	[15]	Extract from use case No. 9.

B.2.2 Use case No. 2

Table B.2 illustrates concerns about human rights, fairness and non-discrimination, privacy, safety and security, human control of technology, and respect for the rule of law.

Table B.2 — Behavioural and sentiment analytics (security and law enforcement)

Description	Use on-premise CCTV systems and analysis to ascertain a person's emotional state and goal from their gestures, facial expression, and actions. Use these insights to deter undesired behaviours, adapt the narrative to suit the state of the person, provide dynamic content according to the person's emotional responses, or detect object theft and other criminal behaviours. However: Surveillance cameras often have low resolution and can be in poorly lit environments with a poor top-down view angle. A lot of suspicious behaviour can be indiscernible behind passers-by or large crowds. Unwanted behaviours are much less frequent than normal behaviours and can take on various forms.	
Ethical and societal concerns	Bias, security threats, privacy threats. The use case illustrates concerns about human rights (see 5.1), fairness and non-discrimination (see 6.2.2), privacy (see 6.2.6), safety and security (see 6.2.7), human control of technology (see 6.2.8), respect for the rule of law (see 6.2.10).	
References	[15]	Extract from use case No. 14.

B.2.3 Use case No. 3

Table B.3 illustrates concerns about privacy.

Table B.3 — Enhancing traffic management efficiency and infraction detection accuracy with AI technologies

Description	Big data-enabled AI technologies are applied to monitoring and managing the traffic in a large municipality in China. Multi-sourced data (traffic flow, vehicle data, pedestrian movement, etc.) is monitored, from which illegal operation of vehicles, unexpected incidents, surges of traffic, etc. are detected and analysed using ML methods. ML tasks (including training and deployment) are carried out on a platform supporting the integration of various ML frameworks, models and algorithms. The platform is based on heterogeneous computing resources. The efficiency and accuracy of infraction detection, and the effectiveness of traffic management are significantly improved, with much reduced human effort and overall solution cost.	
Ethical and societal concerns	Privacy threats. The use case illustrates concerns about privacy (see 6.2.6). Safety concern (due to the risk of hacking, disrupting traffic management systems in dangerous ways, for example setting all traffic lights at a junction to 'go' simultaneously) (see 6.2.7).	
References	[15]	Extract from use case No. 29.

B.2.4 Use case No. 4

Table B.4 illustrates concerns about accountability, fairness and non-discrimination, transparency, privacy, safety and security, respect for the rule of law and respecting international human rights.

Table B.4 — Law enforcement, administration of justice and democratic processes

Description	AI and robotics can significantly enhance law enforcement's surveillance capabilities. The International Criminal Police Organization is an inter-governmental organization with 194 member countries in 2020. During the INTERPOL-UNICRI Global Meeting on Artificial Intelligence for Law Enforcement, held in November 2020, the following topics were discussed[63]: — The potential misuses of AI; — Law enforcement use of AI, including special panels on the use of AI to combat online child sexual abuse and terrorist use of the internet and social media; — Latest AI developments for law enforcement in the private sector; — Developments in related areas such as the use of AI in the criminal justice system, AI and criminal liability and the interaction between AI and drones.	
Ethical and societal concerns	Ethical challenges posed by the adoption of AI technology to be addressed, such as privacy concerns associated with these technologies, including issues such as when and where it is permissible to use sensors, potential misuses of AI and for such AI systems. The safeguarding of democratic processes. The use case illustrates concerns about accountability (see 6.2.1), fairness and non-discrimination (see 6.2.2), transparency (see 6.2.3), privacy (see 6.2.6), safety and security (see 6.2.7) and showing respect for the rule of law (see 6.2.10) as well as respecting international human rights (see 5.1).	
References	[103]	The highlighting of the need for collaboration and useable frameworks.
	[104]	The opportunities and risks of AI and robotics for law enforcement.

B.2.5 Use case No. 5

Table B.5 illustrates concerns about accountability, fairness and non-discrimination, transparency, privacy and safety and security.

Table B.5 — Conversational AI and chatbots

Description	In an article – "Ethics and Artificial Intelligence with IBM Watson's Rob High", see Reference [105], Rob High, the CTO of IBM Watson has featured on the subject of ethical considerations in relation to chatbots. Establishing trust between machines and humans works similarly to building trust between humans. A brand can build up trust by aligning their expectations to reality, learning from their mistakes, and correcting them, listening to feedback from customers and being transparent. Hence, transparency can be a critical consideration when designing a customer service chatbot. It all comes down to the simple question of whether it is obvious that users are talking to a human or a machine? Customers can usually tell the difference between the two, and they expect that brands are honest about it. Customers hardly expect the chatbot to be perfect, but they would like to know what they can and cannot do.	
Ethical and societal concerns	Failure to appreciate that one is talking to an AI system can lead to confusion and frustration, or to a sense of being duped or misled, none of which are good for mental health, confidence, nor for interpersonal or societal trust. One of the keys to keeping AI ethical is for it to be transparent. The author states that when a person interacts with a chatbot, they need to know they are talking to a chatbot and not a live person. The use case illustrates accountability (see 6.2.1), transparency (see 6.2.3), privacy (see 6.2.6) and safety and security (see 6.2.7).	
References	[15]	Use case No. 106.
	[105]	Ethical considerations in relation to chatbots
	[106]	The publication describes that the use of language is not a "neutral" or "objective" medium. Rather, it reflects existing societal values and judgements.

B.2.6 Use case No. 6

Table B.6 illustrates concerns about accountability, fairness and non-discrimination, transparency, privacy and safety and security.

Table B.6 — Conversational AI and chatbots

Description	Using AI in mental health settings, to provide patient care or additional social interactions.	
Ethical and Societal Concerns	Mental health services meet people at their most vulnerable, so the use of conversational AI within mental health services requires careful consideration. In particular, any conversational interactions with such users need the utmost ethical and critical attention. For example, conversational AI can be harmful to users due to their limited capacity to re-create human interaction and to provide tailored treatment. This can, however, be balanced against the benefits of having more verbal interactions than is otherwise possible in a care setting, for example due to limited staffing numbers. Oversight is also required to ensure the AI system has no acquired harmful concepts or language. The use case illustrates accountability (see 6.2.1), fairness and non-discrimination (see 6.2.2), transparency (see 6.2.3), privacy (see 6.2.6) and safety and security (see 6.2.7)	
References	[15]	Use case No. 106.

B.2.7 Use case No. 7

Table B.7 illustrates concerns about fairness and non-discrimination, professional responsibility, promotion of human values, community involvement and development and respect for the rule of law.

Table B.7 — AI to understand adulteration in commonly used food items

Description	Food adulteration is becoming a menace, especially with adulterants that are either carcinogenic or harmful to body parts like the kidneys. For example, milk has been adulterated with soda, urea and detergents, and mangoes and bananas have been prematurely ripened using calcium carbide. There is no frugal way to identify these types of adulteration. An experiment of controlled adulteration was conducted and hyperspectral reflectance readings were taken. AI helped to find the patterns in the hyperspectral signature and was able to reliably classify (90 % ++) samples that were either unadulterated or adulterated.	
Ethical and societal concerns	Fair treatment. Different sources of bias, incorrect AI system use, improperly trained model, incorrect classification (->false accusations). Failure to apply such systems equally for a whole society; risking omission of correct accusations through failure to apply the technology. The use case illustrates concerns about fairness and non-discrimination (see 6.2.2), professional responsibility (see 6.2.4), promotion of human values (see 6.2.5), community involvement and development (see 6.2.9), respect for the rule of law (see 6.2.10).	
References	[15]	Extract from use case No. 19.

B.2.8 Use case No. 8

Table B.8 illustrates concerns about fairness and non-discrimination, transparency and explainability and professional responsibility.

Table B.8 — Credit scoring using KYC data (Banking and financial services)

Description	It can often be difficult to build a risk scorecard using only KYC (Know Your Customer) data, which often has noise and incompleteness issues. However, if realized, it can be used to provide an objective score to all loan applicants, even new-to-credit ones. Nonlinear classification algorithms are suitable for this purpose. Several variables are collected from the customer during the KYC process, such as the age of the customer, self-reported income, type of occupation, loan purpose, etc. All these features can be added to a non-linear risk model and their complex interactions allowed to take place.	
Ethical and societal concerns	KYC data obtained from very rural areas can be noisy and have several missing values. Appropriate pre-processing and treatment are necessary before feeding to the model algorithm. The use case illustrates concerns about fairness and non-discrimination (see 6.2.2), transparency and explainability (see 6.2.3), professional responsibility (see 6.2.4).	
References	[15]	Extract from use case No. 27.

B.2.9 Use case No. 9

Table B.9 illustrates concerns about accountability, professional responsibility, promotion of human values, privacy, safety and security, human control of technology and community involvement and development.

Table B.9 — AI situation explanation service for people with visual impairments

Description	The use case supports the daily life of people with visual impairments through AI vision technologies. This service helps these people to recognize and avoid dangerous objects on the move, and identify people, text, and objects, and acquaintances by taking into account various surrounding situations. It also supports a captioning service to help people with visual impairments understand the current situation or photos.	
Ethical and societal concerns	Deployment of such services potentially exposes the user's daily life to automatic recording and monitoring if privacy safeguards are lacking or security measures are inadequate. Failure to provide such services on an equitable level, where available, risks exacerbating wealth and social disparities. The use case illustrates concerns about accountability (see 6.2.1), professional responsibility (see 6.2.4), promotion of human values (see 6.2.5), privacy (see 6.2.6), safety and security (see 6.2.7), human control of technology (see 6.2.8), community involvement and development (see 6.2.9).	
References	[15]	Extract from use case No. 64.

B.2.10 Use case No. 10

Table B.10 illustrates concerns about promotion of human values, privacy, community involvement and development.

Table B.10 — AI solution to intelligent campus

Description	Based on big data and artificial intelligence technology, the scheme brings teaching, examination, learning and management into the integrated system of mutual cooperation, based on accompanying data acquisition and dynamic big data analysis, combined with process evaluation, to help teachers and students to realize teaching according to their aptitude and individualized learning, to help managers to supervise and assist decision-making, and to greatly promote the transformation of education, learning and management to intelligence.	
Ethical and Societal Concerns	Disclosure of privacy data for teachers and students The use case illustrates concerns about promotion of human values (see 6.2.5), privacy (see 6.2.6), community involvement and development (see 6.2.9).	
References	[15]	Extract from use case No. 85.

B.2.11 Use case No. 11

Table B.11 illustrates concerns about accountability, professional responsibility and Human Control of Technology.