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