

ST. ANNE'S COLLEGE OF ENGINEERING AND TECHNOLOGY
DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING
CCS345 ETHICS AND AI

UNIT III AI STANDARDS AND REGULATION

Model Process for Addressing Ethical Concerns During System Design - Transparency of Autonomous Systems-Data Privacy Process- Algorithmic Bias Considerations - Ontological Standard for Ethically Driven Robotics and Automation Systems

Model Process for Addressing Ethical Concerns During System Design IEEE Std 7000:

The goal of this standard is to enable organizations to **design systems** with explicit consideration of

1. individual and societal ethical values,

a) transparency,

b) sustainability,

c) privacy,

d) fairness, and

e) accountability,

2. as well as values typically considered in system engineering, such as a) efficiency b) effectiveness.

Scope

- The standard establishes processes for engineers and technologists to incorporate ethical values throughout concept exploration and development.
- It helps address ethical concerns or risks during system design by aligning innovation management processes, system design approaches, and software engineering methods.
- The standard does not provide specific guidance on designing algorithms for ethical values like fairness and privacy.

Purpose

- IEEE Std 7000 aims to design systems with ethical values like transparency, sustainability, privacy, fairness, and accountability, while balancing time and budget constraints.
- Top executives' commitment to uphold organizational values is crucial.

Applicability and constraints

IEEE Std 7000 is a standard that supports organizations in identifying stakeholder values and engaging in value-based system or service development. It is applicable within any life cycle model or set of methods for systems and software engineering. The process applies during system conception and design for organizations seeking to uncover, address, and monitor value concerns for a system intended for a given context.

The use of IEEE Std 7000 does not guarantee that the system as designed and built is ethical, as the ethicality achieved depends on the moral capabilities and choices of those who use the standard and the commitment of the organization offering the system to adhere to the recommendations made as a result of ethically aligned design.

The standard does not prescribe what is ethical and does not provide ethical guidance for individual engineers or whistle-blowers. However, adopting and implementing ethical value processes in the design and deployment of new products and services demonstrates an organization's awareness of its social responsibility and impact on stakeholder values.

Moreover, IEEE Std 7000 provides processes for organizations that assume accountability for the ethical decisions they take. This standard helps organizations in the following: —

- Understanding and anticipating value implications and consequences of their systems and taking investment decisions based on them
- Identifying ethical value requirements (EVR) and priorities for system design to be integrated into system requirements
- Choosing system design alternatives according to value priorities while avoiding or mitigating value harms or ethical pitfalls
- Keeping control of the long-term value-based sustainability of a system through ongoing supervision and information management
- Creating transparency and responsibility for the choices made and the system's resulting functionality

This standard is most applicable to organizations that are building a system for a known context or at least known typical use cases for the products, services, and systems they build.

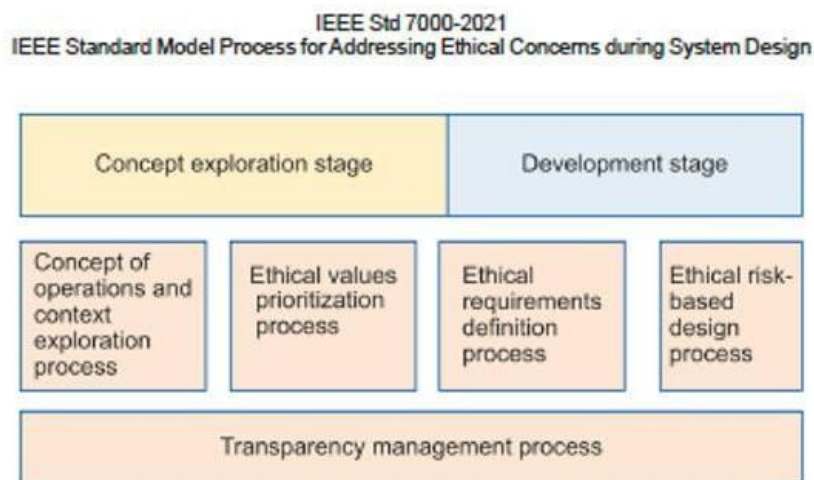


Figure 1—Relationship of processes and stages in IEEE Std 7000

Process Overview

This standard is applicable to enhance or modify existing legacy systems, such as care robots and AI chat systems. It is less useful for generic products or systems with indefinite deployment contexts. The standard outlines processes for organizations and projects to address ethical values of software-based systems during design and development, aligning with any engineering methods, life cycle model, and management style. These processes can be used for new design and improvement of existing systems' ethical attributes.

IEEE Std 7000 is a standard that provides organizations with ethical requirements and design activities for systems engineering. It enables the pragmatic application of value-based system design methodology, aligning innovation management processes, IT system design approaches, and software engineering methods. The process involves considering potential values and harms during concept exploration and development of the concept of operations (ConOps). The Ethical Values Elicitation and Prioritization Process involves stakeholders identifying potential positive and negative system consequences, stakeholder virtues, and ethical duties. These values are prioritized by top management, scrutinized for legal expectations and internationally applied ethical guidelines, resulting in a list of value priorities for the system.

These value priorities are then analyzed more systematically and conceptually as the basis for the Ethical Requirements Definition Process, which generates EVR and value-based system requirements. IEEE Std 7000 is compatible with many existing development practices, including iterative and incremental life-cycle models and agile methods.

The Ethical Risk-Based Design Process translates value-based requirements into design characteristics and determines controls that can mitigate risks to values. Controls are system requirements or organizational policies and procedures. As EVRs are instantiated in the system design, the value dispositions are validated for incorporation of the specified values.

TRANSPARENCY OF AUTONOMOUS SYSTEMS - IEEE STANDARD P 7001:

- Transparency can be defined as the **extent to which the system discloses the processes or parameters** that relate to its functioning.
- Transparency can also be considered as the property that makes it possible to **discover how and why the system** made a particular **decision** or acted the way it did.

P7001 STANDARD Scope and Structure

- The aim of P7001 is to provide a standard that sets out “**measurable, testable levels of transparency**, so that autonomous systems can be objectively assessed
- An autonomous system is defined in P7001 as “a system that has the capacity to **make decisions itself, in response to some input data or stimulus**, with a varying degree of human intervention depending on the system’s level of autonomy”.

The intended users of P7001 are

- a) specifiers,
- b) designers,
- c) manufacturers,
- d) operators and maintainers of autonomous systems.

Defining Transparency in P7001

P7001 defines transparency as “the transfer of information from an autonomous system or its designers to a stakeholder, which is honest, contains information relevant to the causes of some action, decision or behaviour and is presented at a level of abstraction and in a form meaningful to the stakeholder.

The P7001 standard acknowledges the interconnectedness of AI technology with the larger socio-technical system, emphasizing the responsibility of system designers to provide relevant information to stakeholders. The standard emphasizes honesty, ensuring only non-false or deceptive information. To make an autonomous system transparent, stakeholders should have the ability to request explanations of the system's status or decision-making principles.

Explainability is defined as the extent to which the internal state and decision-making processes are accessible to non-experts, aiming to satisfy both transparency and explainability. However, providing an explanation does not guarantee complete transparency of a system's actions.

Furthermore, P7001 is generic; it is intended to apply to all autonomous systems including robots (autonomous vehicles, assisted living robots, drones, robot toys, etc.), as well as software-only AI systems, such as medical diagnosis AIs, chatbots, loan recommendation systems, facial recognition systems, etc.

To ensure the transparency of autonomous systems to a range of stakeholders the IEEE P7001 standard address the following issues.

Transparency Is Not the Same for Everyone

P7001 defines five distinct groups of stakeholders, and AIS must be transparent to each group, in different ways and for different reasons. These stakeholders split into two groups: nonexpert end users of autonomous systems (and wider society), and experts including safety certification engineers or agencies, accident investigators, and lawyers or expert witnesses. Stakeholders are beneficiaries of the

standard, as distinct from users of the standard: designers, developers, builders and operators of autonomous systems.

Let us now look at the transparency needs of each of these five groups.

1. Transparency for End Users

Taking a care robot as an example, transparency means the user can begin to predict what the robot might do in different circumstances. A vulnerable person might feel very unsure about robots, so it is important that the robot is helpful, predictable—never does anything that frightens them—and above all safe. It should be easy to learn what the robot does and why, in different circumstances. The robot's responses would allow the user to build a mental model of how the robot will behave in different situations.

2. Transparency for the Wider Public and Bystanders

It is very important therefore that the whole of society has a basic level of understanding of how these systems work, so we can confidently share work or public spaces with them. That understanding is also needed to inform public debates—and hence policy—on which robots/AIs are acceptable, which are not, and how they should be regulated. This kind of transparency needs public engagement, for example through panel debates and science cafés, supported by high quality documentaries targeted at distribution by mass media (e.g., YouTube and TV), which present emerging robotics and AI technologies and how they work in an interesting and understandable way. Balanced science journalism—avoiding hype and sensationalism—is also needed. For this stakeholder group, P7001 defines levels of transparency starting with a requirement that follows a proposed Turing Red Flag law: “An autonomous system should be designed so that it is unlikely to be mistaken for anything besides an autonomous system, and should identify itself at the start of any interaction with another agent. Successive levels build upon this by requiring that systems provide warnings and information about data collected or recorded, since data on bystanders may well be captured.

3. Transparency for Safety Certifiers

For safety certification of an AIS, transparency is important because it exposes the system's decision-making processes for assurance and independent certification. The type and level of evidence required to satisfy a certification agency or regulator that a system is safe and fit for purpose depends on how critical the system is. An autonomous vehicle autopilot requires a much higher standard of safety certification than, say, a music recommendation AI, since a fault in the latter is unlikely to endanger life. Safe and correct behaviour can be tested by verification, and fitness for purpose tested by validation. Put simply, verification asks “is this system, right?” and validation asks “is this the right system?”. At the lowest level of transparency, certification agencies or regulators need to see evidence (i.e., documentation) showing how the designer or manufacturer of an AIS has verified and validated that system. This includes as a minimum a technical specification for the system. Higher levels of transparency may need access to source code and all materials needed (such as test metrics or benchmarks) to reproduce the verification and validation processes. For learning systems, this includes details of the composition and provenance of training data sets.

4. Transparency for Incident/Accident Investigators

Robots and other AI systems can and do act in unexpected or undesired ways. When they do it is important that we can find out why. Autonomous vehicles provide us with a topical example of why transparency for accident investigation is so important. Discovering why an accident happened through investigation requires details of the situational events leading up to and during the accident and, ideally, details of the internal decision-making process in the robot or AI prior to the accident. Established and trusted processes of air accident investigation provide an excellent

model of good practice for AIS–processes, which have without doubt contributed to the outstanding safety record of modern commercial air travel. One example of best practice is the aircraft Flight Data Recorder, or “black box”; a functionality we consider essential in autonomous systems.

5. Transparency for Lawyers and Expert Witnesses

Following an accident, lawyers or other expert witnesses who have been obliged to give evidence in an inquiry or court case or to determine insurance settlements, require transparency to inform their evidence. Both need to draw upon information available to the other stakeholder groups: safety certification agencies, accident investigators and users. They especially need to be able to interpret the findings of accident investigations. In addition, lawyers and expert witnesses may well draw upon additional information relating to the general quality management processes of the company that designed and/or manufactured the robot or AI system. Does that company, for instance, have ISO 9001 certification for its quality management systems? A higher level of transparency might require that a designer or manufacturer provides evidence that it has undertaken an ethical risk assessment of a robot.

EX : TRANSPARENCY ASSESSMENT FOR ROBOT TOY:

STAKEHOLDERS	ISSUES ADDRESSED
<i>Users</i>	• <u>User Manual</u> is provided for parents
<i>Lawyers and expert witnesses:</i>	• <u>Risk Assessment</u> is performed by the manufacturer and documented
<i>Public</i>	• <u>Interactive Online visual guide</u> shows how to safeguard from Hacking and explains the risk of addiction

To ensure the **transparency of autonomous systems to a range of stakeholders** the IEEE P7001 standard address the following issues.

STAKEHOLDERS	ISSUES ADDRESSED
<i>Users</i>	ensuring users understand what the system does and why, with the intention of building trust;
<i>Lawyers and expert witnesses:</i>	ensuring that, following an accident, these groups are able to give evidence;
<i>Public</i>	enabling the public to assess technology (<i>e.g. driverless cars</i>) (and, if appropriate, build confidence).

Data Privacy Process- IEEE STANDARD P7002

Scope

- This standard defines requirements for a systems/software engineering process for privacy-oriented considerations regarding products, services, and systems utilizing employee, customer or other external user's personal data.
- It extends across the life cycle from policy through development, quality assurance, and value realization.
- It includes a use case and data model (including metadata).
- It applies to organizations and projects that are developing and deploying products, systems, processes, and applications that involve personal information.
- By providing specific procedures, diagrams, and checklists, users of this standard will be able to perform a conformity assessment on their specific privacy practices.
- Privacy impact assessments (PIAs) are described as a tool for both identifying where privacy controls and measures are needed and for confirming they are in place.

Managing Privacy and Security Risk

Assumptions:

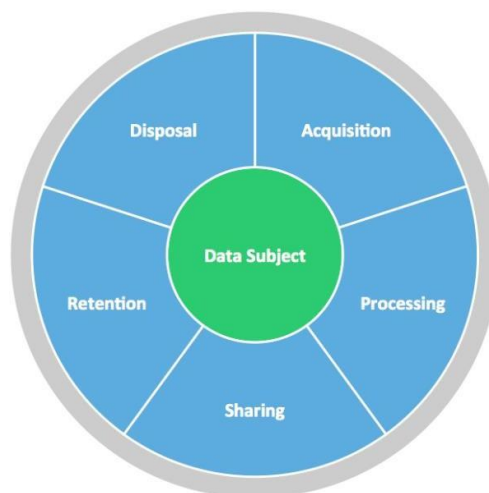
- Systems are built with flaws
- Good privacy and security today is better than perfect privacy and security tomorrow
- Good privacy and security today may not be good tomorrow
- Bad guys do NOT think rationally
- Whatever can go wrong probably already has.

The intersection of 5G, ML, and Privacy

- A bigger, faster network with more sensors in more devices communicating more information about more things. Many of those things will implicate personal data directly or by inference.
- Larger data sets from more diverse sources enabling deeper analytics about more things – including individuals' behaviour, interactions, habits, and more.
- Retrofitting for inappropriate collection, use, or disclosure of personal data will vary from impractical to impossible.

Data Privacy Life Cycle

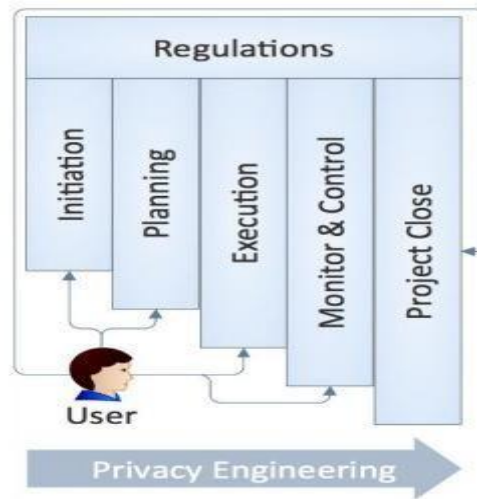
A product, service or system that processes data about an identifiable individual has to account for, or be accountable to, the data subject at every stage of the life cycle of the data.



Data Privacy Process in Projects

- 1. Initiation:** Does this project implicate personal data? Are we allowed to do what we are thinking about?
- 2. Planning:** How can we respect the user while using data about them?
- 3. Execution:** Are there data protection checkpoints and/or gates?
- 4. Monitor & Control:** Q/A tests for data protection?

5. Project Close: Release to production with data protection instrumentation (metrics and controls)



Privacy by Design

1. Proactive not reactive; preventative not remedial
2. Privacy as the default
3. Privacy embedded into design
4. Full functionality – positive-sum, not zero-sum
5. End-to-end security – full lifecycle protection
6. Visibility and transparency – keep it open
7. Respect for user privacy – keep it user-centric

Privacy Engineering

Privacy engineering is a specialty discipline of systems engineering focused on removing conditions that can create problems for people when system operations process their information.

• Predictability

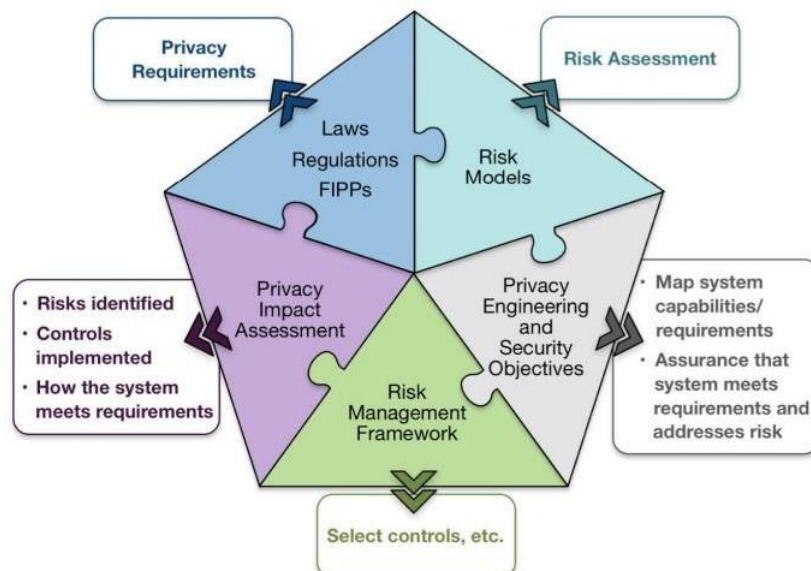
enabling reliable assumptions by individuals, owners, and operators about PII and its processing by an information system

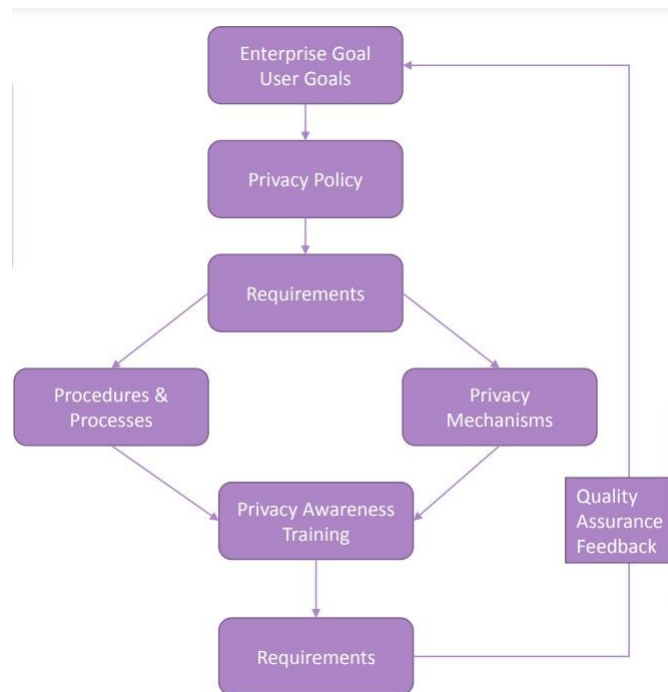
• Manageability

providing the capability for granular administration of PII including alteration, deletion, and selective disclosure

• Dissociability

enabling the processing of PII or events without association to individuals or devices beyond the operational requirement of the system





Privacy engineering development process

Examples

Functional Requirements

- The system must provide a privacy notice to users
- The system must collect an opt-in consent from users
- The system must encrypt all PII at rest and in motion
- Unattended patch management
- All non-required ports closed by default

Non-functional Requirements

- Where there are user configurable choices, the default choices should be privacy protective
- Identifying and contacting the system privacy office should be built into the UX
- Mandatory breach notification
- Unique default password per device (no “admin”/“admin”)

ALGORITHMIC BIAS CONSIDERATIONS: The IEEE P7003 standard

Any system **that will produce different results for some people** than for others is open to challenges of being biased.

Examples could include:

Security camera applications that detect theft or suspicious behaviour.

Marketing automation applications that calibrate offers, prices, or content to an individuals’ preferences and behaviour.

Amazon’s Recruiting Engine biased against women

The Amazon recruiting engine is an artificial intelligence algorithm that was created to analyze the resumes of job applicants applying to Amazon and decide which ones would be called for further interviews and selection. However, the **Amazon algorithm** turned out to be biased against women in the recruitment process. When Amazon studies the algorithm, they found that it automatically handicapped the resumes that contained words like “women” and also automatically downgraded the graduates of two all-women colleges. Therefore, Amazon finally discarded the algorithm and didn’t use it to evaluate candidates for recruitment.

SCOPE OF P7003:

The IEEE P7003 standard will provide **a framework**, which helps developers of algorithmic systems and those responsible for their deployment to identify and mitigate

- a) unintended,
- b) unjustified and/or
- c) inappropriate biases in the outcomes of the algorithmic system.

Algorithmic systems in this context refers to the **combination of algorithms, data and the output deployment process** that together determine the outcomes that affect end users.

Unjustified bias refers to differential treatment of individuals based on criteria for which **no operational justification** is given.

Inappropriate bias refers to bias that is **legally or morally unacceptable** within the social context where the system is used, e.g. algorithmic systems that produce outcomes with differential impact strongly correlated with protected characteristics (such as race, gender, etc).

Introduction

Bias is a significant ethical and societal concern within data science-based AI systems, distinct from other forms of automation. It can lead to unjust or unfair decisions that disproportionately affect certain individuals or groups. While bias can exist in classic AI systems, it is often more prevalent in machine learning applications, raising concerns about perpetuating and amplifying societal inequalities.

Manifestations of Bias

Bias can manifest in various ways throughout the design, testing, and application of AI systems. From biased training datasets to algorithmic decisions based on spurious correlations, multiple factors contribute to the emergence of bias. Moreover, biases embedded in societal practices can further exacerbate the problem, impacting everything from criminal justice to predictive policing and internet services.

Cultural and Data Set Biases

Cultural biases in data sets, such as those found in image recognition networks trained on datasets disproportionately representing certain demographics, can lead to skewed outcomes. Additionally, incomplete or poor-quality data can reinforce biases, while the lack of diversity in AI development teams further contributes to biased outcomes.

Challenges in Mitigating Bias

Addressing bias in AI systems poses significant challenges. There's a risk of sacrificing accuracy when attempting to counteract bias, and even perfectly unbiased algorithms may still discriminate indirectly. Moreover, societal disagreements on concepts of justice and fairness complicate efforts to mitigate bias effectively.

The Role of Human Interpretation

Human decision-makers play a crucial role in perpetuating or mitigating bias within AI systems. Overreliance on algorithmic recommendations and disregarding other information can exacerbate biases. However, the extent to which bias should be avoided and the trade-offs involved remain contentious issues.

Debates and Solutions

Debates around bias in AI extend beyond technical considerations to encompass broader philosophical and political discussions. Questions about the nature of justice, the role of AI in shaping societal outcomes, and the responsibilities of developers and policymakers are central to addressing bias effectively. Various approaches, from diversifying development teams to implementing affirmative action in algorithms, are proposed, reflecting differing views on justice and fairness.

Conclusion

Bias in AI is a multifaceted issue that intersects with complex societal and philosophical debates. While efforts to mitigate bias are necessary, achieving perfect neutrality may be elusive. Instead, discussions about AI ethics offer insights into broader questions of social justice and the future of human-machine interactions.

ONTOLOGICAL STANDARD FOR ETHICALLY DRIVEN ROBOTICS AND AUTOMATION SYSTEMS: IEEE P 7007 STANDARD

Ontologies are formal representations of the concepts, relations, and constraints in a domain of knowledge. They are widely used in artificial intelligence (AI) to provide a common vocabulary, structure, and reasoning for various tasks and applications.

The ontological specification reports provide methods to assess AI systems and organizations in their ethical performance regarding the key ethical principles of transparency, accountability, bias, and privacy.

EXAMPLES OF ONTOLOGIES:

ONTOLOGY CATEGORY	CONCEPTS	RELATIONSHIPS
medical ontology	Include concepts such as disease, symptom, treatment, and drug.	• "a disease can have multiple symptoms" • "a drug can be used to treat multiple diseases."
product ontology	include concepts such as product, category, attribute, and value	• "a product can belong to multiple categories" and • "an attribute can have multiple values."

ETHICAL RISKS:

1. For example, an ontology for a criminal justice AI system might include concepts such as crime, criminal, and victim. If the ontology is biased against certain groups of people, such as African Americans, then the AI system will be more likely to recommend harsher punishments for members of those groups.
2. For example, an ontology for a military AI system might include concepts such as weapon, target, and enemy. If the ontology is biased against certain groups of people, then the AI system will be more likely to identify members of those groups as enemies and recommend that they be attacked.

Scope:

The standard establishes a set of ontologies with different abstraction levels that contain concepts, definitions and axioms which are necessary to establish ethically driven methodologies for the design of Robots and Automation Systems.

Purpose:

The standard establishes a set of definitions and their relationships that will enable the development of Robotics and Automation Systems in accordance with worldwide Ethics and Moral theories, with a particular emphasis on aligning the ethics and engineering communities to understand how to pragmatically design and implement these systems in unison. These definitions allow for a precise communication among global experts of different domains that includes Robotics, Automation and Ethics

Ontological Standards

- Ontologies are information artifacts that represent consensual knowledge in an explicit and formal way.
- Very good tool standardization initiatives: formalize the consensus around a specific domain.
- Main uses: – Vocabulary disambiguation among groups of humans, robots, and other groups of agents that share the same conceptualization. – Conceptual model (e.g. in databases). – Symbolic model to support different forms of reasoning

Group Expectation

This standard should be used as guide to the design, development and operation of products and services related robots and robotic systems with respect to ethics. It should help decision makers and robot designers to address ethical issues regarding user experience, safety, data protection, data privacy and transparency”

Possible use

- a guide for teaching ethical design;
- a reference by policy makers and governments to draft AI related policies;
- a common vocabulary to enable the communication among government agencies and other professional bodies around the world;
- part of decision-making during investment in companies and technologies;
- a framework to create systems that can act ethically;

Stakeholders:

- Manufacturers
- Service and solution providers
- Equipment suppliers in the robotics and users

BENEFITS:

The use of ontologies for representing knowledge in any domain has several benefits that include the following:

- a) A formal definition of concepts of a particular domain in a language-independent representation,
- b) Tools for analyzing concepts and their relationships in searching for inconsistency, incompleteness, and redundancy

Ex: Webprotegy analysis tool

- c) Language being used in the communication process among robots from different manufacturers is standardized