studocu

1707362639601 Ethics and AI Unit 3

Ethics and AI (Anna University)

Scan to open on Studocu

Studocu is not sponsored or endorsed by any college or university
Downloaded by Lokesh S (lokesh@psgitech.ac.in)

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

3.1 MODEL PROCESS FOR ADDRESSING ETHICAL CONCERNS DURING SYSTEM DESIGN

Engineering design ethics

Engineering design ethics concerns issues that arise during the design of technological products, processes, systems, and services.

This includes issues such as safety, sustainability, user autonomy, and privacy. Ethical concern with respect to technology has often focused on the user phase. Technologies, however, take their shape during the design phase.

The engineering design process thus underlies many ethical issues in technology, even when the ethical challenge occurs in operation and use.

Engineering Design

- Engineering design ethics concerns issues that arise during the design of technological products, processes, systems, and services. This includes issues such as safety, sustainability, user autonomy, and privacy.
- Ethical concern with respect to technology has often focused on the user phase. Technologies, however, take their shape during the design phase. The engineering design process thus underlies many ethical issues in technology, even when the ethical challenge occurs in operation and use.
- The character of the engineering design process has been much debated, but for present purposes it may be described as an iterative process divided into different phases.

The following phrases are the simplest and most accepted (Pahl and Beitz 1996):

- Problem analysis and definition, including the formulation of design requirements and the planning for the design and development of the product, process, system, or service.
- Conceptual design, including the creation of alternative conceptual solutions to the design problem, and possible reformulation of the problem.
- Embodiment design, in which a choice is made between different conceptual solutions, and this solution is then worked out in structural terms.
- Detail design, leading to description that can function as a guide to the production process.

- In each phase, engineering design is a systematic process in which use is made of technical and scientific knowledge. This process aims at developing a solution that best

This document is available free of charge on studocu

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

meets the design requirements. Nevertheless, the final design solution does not simply follow from the initially formulated function because design problems are usually ill-structured. Nigel Cross (1989) has argued that proposing solutions often helps clarify the design problem, so that any problem formulation turns out to be partly solution-dependent. It is impossible to make a complete or definite list of all possible alternative solutions to a problem. It is also extremely difficult to formulate any criterion or set of criteria with which alternatives can be ordered on a scale from "good" or "satisfactory" to "bad" or "unsatisfactory," even though any given feature of the design may be assessed in terms of some given criterion such as speed or efficiency.

Problem formulation:

- Problem definition is of special importance because it establishes the framework and boundaries within which the design problem is solved.
- It can make quite a difference—including an ethical difference—from whose point of view a problem is formulated. The problem of designing an Internet search engine looks different from the perspective of a potential user concerned about privacy than from the perspective of a provider concerned about selling banner advertisements.
- The elderly or physically disabled will have different design requirements than the young or healthy.
- An important ethical question in this phase concerns what design requirements to include in the problem definition. Usually design requirements will be based on the intended use of the artifact and on the desires of a client or user.
- In addition, legal requirements and technical codes and standards play a part. The latter may address, if only implicitly, ethical issues in relation to safety or environmental concerns. Nevertheless, some ethical concerns may not have been adequately translated into design requirements.
- Engineering codes of ethics, for example, require that engineers hold "paramount the safety, health and welfare of the public," an obligation that should be translated into design requirements.
- The idea that morally relevant values should find their way into the design process has led to a number of new design approaches. An example is eco-design or sustainable design, aimed at developing sustainable products (Stitt 1999).
- Another example is value-sensitive design, an approach in information technology that accounts for values such as human well-being, human dignity, justice, welfare, and human rights throughout the design process (Friedman 1996).

Conceptual design.

- Design is a creative process, especially during the conceptual phase. In this phase the designer or design team thinks out potential solutions to a design problem.
- Although creativity is not a moral virtue in itself, it is nevertheless important for good design, even ethically. Ethical concerns about a technology may on occasion be overcome or diminished by clever design.
- One interesting example is the design of a storm surge barrier in the Eastern Scheldt estuary in the Netherlands (Van de Poel and Disco 1996). In the 1950s,

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

the government decided to dam up the Eastern Scheldt for safety reasons after a huge storm had flooded the Netherlands in 1953, killing more than 1,800 people.

- In the 1970s, the construction plan led to protests because of the ecological value of the Eastern Scheldt estuary, which would be destroyed. Many felt that the ecological value of the estuary should be taken into account.
- Eventually, a group of engineering students devised a creative solution that would meet both safety and ecological concerns: a storm surge barrier that would be closed only in cases of storm floods. Eventually this solution was accepted as a creative, although more expensive, solution to the original design problem.

Embodiment design.

- During embodiment design, one solution concept is selected and worked out. In this phase, important ethical questions pertain to the choice between different alternatives.

- One issue is tradeoffs between various ethically relevant design requirements. While some design requirements may be formulated in such terms that they can be clearly met or not —for example, that an electric apparatus should be compatible with 220V—others may be formulated in terms of goals or values that can never be fully met.

- Safety is a good example. An absolutely safe car does not exist; cars can only be more or less safe. Such criteria as safety almost always conflict with other criteria such as cost, sustainability, and comfort. This raises a question about morally acceptable tradeoffs between these different design criteria.

Detail design.

- During detail design, a design solution is further developed, including the design of a production process. Examples of ethical issues addressed at this phase are related to the choice of materials: Different materials may have different environmental impacts or impose different health risks on workers and users.
- Choices with respect to maintainability, ability to be recycled, and the disposal of artifacts may have important impacts on the environment, health, or safety.
- The design of the production process may invoke ethical issues with respect to working conditions or whether or not to produce the design, or parts of it, in low-wage countries.

Design as a Social Process

Engineering design is usually not carried out by a single individual, but by design teams embedded in larger organizations.

The design of an airplane includes hundreds of people working for several years. Organizing such design processes raises ***a number of ethical issues***.

This document is available free of charge on studocu

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

The first issue is the allocation of responsibilities. What is the best way to allocate responsibility for safety in the design process? One option would be to make someone in particular responsible.

A potential disadvantage of this solution is that others—whose design choices may be highly relevant—do not take safety into account.

Another approach might be to make safety a common responsibility, with the danger that no one in particular feels responsible for safety and that safety does not get the concern it deserves.

A second issue is decision-making. During design, many morally relevant tradeoffs have to be made. Sometimes such decisions are made explicitly, but many times they occur implicitly and gradually, evolving from earlier decisions and commitments. Such patterned decision making may lead to negative results that never would have been chosen if the actors were not immersed in the problematic decision-making pattern (Vaughan 1996). This raises ethical issues about how to organize decision making in design because different arrangements for making decisions predispose different outcomes in ethical terms (Devon and van de Poel 2004).

A third issue is what actors to include. Engineering design usually affects many people with interests and moral values other than those of the designers. One way to do right to these interests and values is to give different groups, including users and other stakeholders, a role in the design and development process itself. Different approaches have been proposed to this issue, such as participatory design in information technology development (Schuler and Namioka 1993). Constructive technology assessment likewise aims to include stakeholders in the design and development process in order to improve social learning processes at both the technical and normative levels with respect to new technologies (Schot and Rip 1997).

3.2 TRANSPARENCY IN AUTONOMOUS SYSTEM:

Transparency:

"AI transparency helps ensure that all stakeholders can clearly understand the workings of an AI system,including how it makes decisions and processes data"

AI transparency also involves being open about data handling and model limitation

Transparency Is Not the Same for Everyone

- ✓ Transparency is not a singular property of systems that would meet the needs of all stakeholders. In this regard, transparency is like any other ethical or socio-legal value (Theodorou et al., 2017).
- ✓ Clearly a naive user does not require the same level of understanding of a robot as the engineer who repairs it. By the same reasoning, a naive user may require explanations for aspects of reasoning and behaviour that would be obvious and transparent to developers and engineers

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

3.2.1 Transparency for End Users

- For users, transparency (or explainability as defined in P7001) is important because it both builds and calibrates confidence in the system, by providing a simple way for the user to understand what the system is doing and why.
- 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.
- A higher level of explainability might be the ability to respond to questions such as "Robot: what would you do if I fell down?" or "Robot: what would you do if I forget to take my medicine?" The robot's responses would allow the user to build a mental model of how the robot will behave in different situations.

3.2.2 Transparency for the Wider Public and Bystanders

- Robots and AIs are disruptive technologies likely to have significant societal impact .
- 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

3.2.3 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?".

This document is available free of charge on studocu

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

- ✧ 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.

3.2.4 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 .

3.2.5 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 or AI system using, for instance, BS 8611 Guide to the ethical design of robots and robotic systems (BSI, 2016).

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

3.2.6 System Transparency Assessment for a Robot Toy

- RoboTED is an Internet (WiFi) connected device with cloud-based speech recognition and conversational AI (chatbot) with local speech synthesis; RoboTED's eyes are functional cameras allowing the robot to recognise faces; RoboTED has touch sensors, and motorised arms and legs to provide it with limited baby-like movement and locomotion—not walking but shuffling and crawling.
- Our ethical risk assessment (ERA) exposed two physical (safety) hazards including tripping over the robot and batteries overheating. Psychological hazards include addiction to the robot by the child, deception (the child coming to believe the robot cares for them), over-trusting of the robot by the child, and over-trusting of the robot by the child's parents.
- Privacy and security hazards include weak security (allowing hackers to gain access to the robot), weak privacy of personal data especially images and voice clips, and no event data logging making any investigation of accidents all but impossible4 .
- The ERA leads to a number of recommendations for design changes. One of those is particularly relevant to the present paper: the inclusion of an event data recorder, so our outline transparency assessment, given below in Table 3, will assume this change has been made.

TABLE 3 | Outline system transparency assessment (STA) for RoboTED.

Stakeholder Group	Transparency level(s)	Evidenced by
[i] users	1, 2	A user manual is provided for parents. As well as detailing how parents can show children how best to use RoboTED, the manual explains the risks (addiction, deception and over-trusting) and how to minimise these. The manual also shows how to guard against hacking and check personal data has been deleted (level 1). An interactive online visual guide is also provided, for both parents and children (level 2)
[ii] general public	1	P7001 level 1 requires that a robot identifies itself as an autonomous system, following Walsh (2016). When powered up, or on waking from sleep mode, RoboTED announces itself as a robot
[iii] certification agencies	2	RoboTED has been certified as safe against standard EU EN 62115 (2020) *Safety of Electric Toys*, and descriptions of the system and how it has been validated are available for safety certifiers. This meets P7001 level 2
[iv] accident Investigators	2	The robot is equipped with a data logging system as outlined in **Table 2**
[v] lawyers and expert witnesses	2	P7001 level 2 requires that a system has been subjected to an ethical risk assessment, which can be made available to lawyers or expert witnesses. This is the case for RoboTED

3.2.7 System Transparency Specification for a Vacuum Cleaner Robot

- ❖ Consider now a fictional company that designs and manufactures robot vacuum cleaners for domestic use. Let us call this company nextVac.
- ❖ Let us assume that nextVac is well established in the domestic market and has a reputation both for the quality of its products and responsible approach to design and manufacture. nextVac now wishes to develop a new line of robot vacuum cleaners for use in healthcare settings: including hospitals, clinics and elder care homes and elder care homes.
- ❖ nextVac begins the design process with a scoping study in which they visit healthcare facilities and discuss cleaning needs with healthcare staff, facilities managers and cleaning

This document is available free of charge on studocu

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

contractors. Mindful of the additional safety, operational and regulatory requirements of the healthcare sector (over and above their domestic market), nextVac decides to capture the transparency needs of the new product–while also reflecting the findings of the scoping study–in a System Transparency Specification (STS),guided by IEEE P7001.

- Their intention is to follow the STS with an initial product design specification. In turn this specification will be subjected to an Ethical Risk Assessment (ERA), guided by BS8611. Depending on the findings of the ERA, the company will iterate this process until a product specification emerges that is technically feasible, tailored to customer needs, and addresses both ethical risks and transparency needs.

- The outline STS for nextVac's proposed new vacuum cleaning robot for healthcare, leads to a number of clear technical design requirements, especially for stakeholder groups [i], [ii], and [iv], alongside process requirements for groups [iii] and [v]. The STS will thus feed into and form part of the product design specification.

TABLE 4 | Outline system transparency specification (STS) for nextVac.

Stakeholder Group	Transparency level(s) Required	Rationale
[i] users	1, 2 (see **Table 1**)	A comprehensive user manual is required, covering both use and maintenance. The manual should be written in compliance with standard IEC/IEEE std 82,079 *Preparation of information for use*, as recommended by P7001 (level 1). An interactive online visual guide is also required, for both operators of the cleaning robot and facilities managers (level 2). Levels 3 and 4 are not required as the robot is not expected to need a complex human robot interface. The robot will only require a limited number of behaviours and these will be indicated by warning lights and sounds, see group [ii] below
[ii] general public	1, 2	The robot's design will ensure that its machine nature is apparent; lights and sounds will provide simple audio-visual indications of what the robot is doing at any time (level 1). The robot will provide physical cues showing the location of sensors, and publicly available information will explain what data is stored and why (see [iv] accident Investigators in this table), and that this data will not include any personal data (level 2)
[iii] certification agencies	3	The robot will be certified as safe against relevant standards, such as ISO 10218 (2011) (noting that ISO 10218 is a generic standard for the safety of industrial robots). Descriptions of the system and how it has been validated will be made available to safety certifiers (level 2). In addition, a high level model (simulation) of the robot will be developed and made available (level 3)
[iv] accident Investigators	3 (see **Table 2**)	The robot will be equipped with a data logging system, which records high level decisions (as outlined in **Table 2**). Noting that the data logging system will not record any personal data. Levels 4 and 5 are not considered essential, as the cleaning robot will only require a limited number of behaviours, nor will it learn
[v] lawyers and expert witnesses	4	nextVac already has certification of quality management (QM) to standard ISO 9001 (level 1). Ethical risk assessment (ERA) against BS8611 will be undertaken (level 2). nextVac has in place processes of ethical governance (level 3). nextVac also maintains complete audit trails for QM, ERA and ethical governance processes (level 4)

3.2.8 Security, Privacy and Transparency

Security and privacy practices are generally embedded within the fabric of autonomous systems. Security standards, especially for regulated industries such as transportation, utilities and finance, receive particular attention by system architects and auditors, but transparency within these mature frameworks tends to be addressed indirectly. To adequately consider

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

transparency for security and privacy, STA and STS statements must be tied closely to prevailing information security standards.

3.2.9 Challenges and Limitations

(1) The comparative youth of the field makes it difficult to assess what it is practical to require now in terms of transparency, let alone what might be practical within the lifetime of the standard.

(2)The heterogeneous nature of transparency is a problem. Is the simple provision of information (e.g., a log) sufficient, or must the information be in a contextualised form (e.g., an explanation) Across and within the stakeholder groups, there was discussion over whether contextualisation was desirable since it necessarily creates a system-generated interpretation of what is happening, which could introduce biases or errors in reporting.

(3)It was sometimes difficult to foresee what transparency might be wanted for, and without knowing the purpose of transparency it was hard to determine what should be required and how compliance might be measured.

3.3 DATA PRIVACY PROCESS

As technology continues to advance at an unprecedented rate, the use of artificial intelligence (AI) has become increasingly prevalent in many areas of our lives. From generative AI that can create any content using a simple prompt to smart home devices that learn our habits and preferences, AI has the potential to revolutionize the way we interact with technology.

3.3.1 Importance of privacy:

- In the digital era, personal data has become an incredibly valuable commodity. The vast amounts of data generated and shared online daily have enabled businesses, governments, and organisations to gain new insights and make better decisions. However, this data also contains sensitive information that individuals may not want to share, or organizations have used without their consent. That is where privacy comes in.
- Privacy is crucial for a variety of reasons. For one, it protects individuals from harm, such as identity theft or fraud. It also helps to maintain individual autonomy and control over personal information, which is essential for personal dignity and respect. Furthermore, privacy allows individuals to maintain their personal and professional relationships without fear of surveillance or interference.
- The importance of privacy in the digital era cannot be overstated. It is a fundamental human right that is necessary for personal autonomy, protection, and fairness. As AI continues to become more prevalent in our lives, we must remain vigilant in protecting our privacy to ensure that technology is used ethically and responsibly.

This document is available free of charge on studocu

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

3.3.2 Privacy Challenges

- AI presents a challenge to the privacy of individuals and organisations because of the complexity of the algorithms used in AI systems. As AI becomes more advanced, it can make decisions based on subtle patterns in data that are difficult for humans to discern.
- This means that individuals may not even be aware that their personal data is being used to make decisions that affect them.

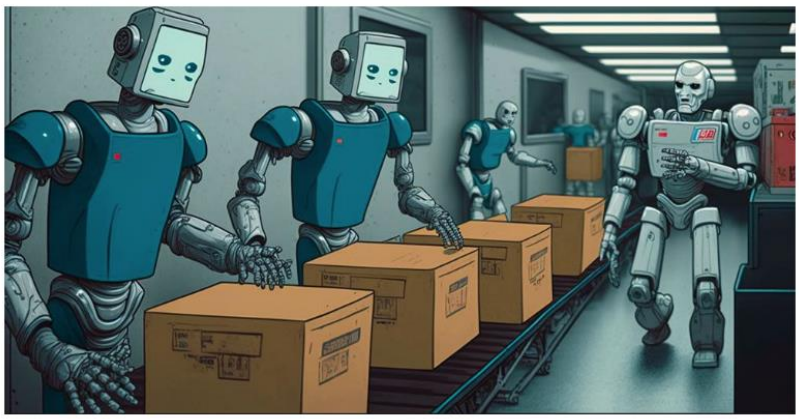
3.3.2.1The issue of violation of Privacy:

While AI technology offers many potential benefits, there are also several significant challenges posed by its use. One of the primary challenges is the potential for AI systems require vast amounts of (personal) data, and if this data falls into the wrong hands it can be used for nefarious purposes, such as identity theft or cyberbullying.

3.3.2.2 The issue of bias and discrimination:

- Another challenge posed by AI technology is the potential for bias and discrimination. AI systems are only as unbiased as the data they are trained on; if that data is biased, the resulting system will be too. This can lead to discriminatory decisions that affect individuals based on factors such as race, gender, or socioeconomic status. It is essential to ensure that AI systems are trained on diverse data and regularly audited to prevent bias.
- For example, imagine an AI system used by a hiring company to screen job applications. If the system is biased against women or people of colour, it may use data about a candidate's gender or race to unfairly exclude them from consideration. This harms the individual applicant and perpetuates systemic inequalities in the workforce.

3.3.2.3 The issue of job displacement for workers



- A third challenge posed by AI technology is the potential for job loss and economic disruption. As AI systems become more advanced, they are increasingly capable of performing tasks that were previously done by humans.

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

This can lead to job displacement, economic disruption in certain industries, and the need for individuals to retrain for new roles.

- But the issue of job loss is also connected to privacy in a number of important ways. For one thing, the economic disruption caused by AI technology can lead to increased financial insecurity for workers. This, in turn, can lead to a situation where individuals are forced to sacrifice their privacy to make ends meet.
- For example, imagine a worker has lost their job due to automation. They are struggling to pay their bills and make ends meet and are forced to turn to the gig economy to make money. In order to find work, they may be required to provide personal information to a platform, such as their location, work history, and ratings from previous clients. While this may be necessary to find work, it also raises serious concerns about privacy, as this data may be shared with third parties or used to target ads.

3.3.3.4 The issue of data abuse practices

- Finally, another significant challenge posed by AI technology is the potential for misuse by bad actors. AI can be used to create convincing fake images and videos, which can be used to spread misinformation or even manipulate public opinion. Additionally, AI can be used to create highly sophisticated phishing attacks, which can trick individuals into revealing sensitive information or clicking on malicious links.
- For example, consider a case in which an evil actor uses artificial intelligence to create a fake video showing a politician engaging in illegal or immoral behaviour. Even if the video is clearly fake, it may still be shared widely on social media, leading to serious reputational harm for the politician in question. This not only violates their privacy but also has the potential to cause real-world harm.

3.3.3 Underlying Privacy Issues in the age of AI

- In the age of AI, privacy has become an increasingly complex issue. With the vast amount of data being collected and analysed by companies and governments, individuals' private information is at greater risk than ever before.
- Some of these issues include invasive surveillance, which can erode individual autonomy and exacerbate power imbalances, and unauthorised data collection, which can compromise sensitive personal information and leave individuals vulnerable to cyber attacks. These problems are often compounded by the power of BigTech companies, which have vast amounts of data at their disposal and significant influence over how that data is collected, analysed and used.

3.3.4Data collection and use by AI technologies:

- One of the most significant impacts of AI technology is the way it collects and uses data. AI systems are designed to learn and improve through the analysis of vast amounts of data.
- As a result, the amount of personal data collected by AI systems continues to grow, raising concerns about privacy and data protection.

This document is available free of charge on studocu

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

- We only have to look at the various generative AI tools, such as ChatGPT, Stable Diffusion or any of the other tools currently being developed, to see how our data (articles, images, videos, etc.) are being used, often without our consent.

3.3.5 The use of AI in Surveillence

- One of the most controversial uses of AI technology is in the area of surveillence. AI-based surveillance systems have the potential to revolutionise law enforcement and security, but they also pose significant risks to privacy and civil liberties.
- AI-based surveillance systems use algorithms to analyse vast amounts of data from a range of sources, including cameras, social media, and other online sources. This allows law enforcement and security agencies to monitor individuals and predict criminal activity before it occurs.
- Recently, The European Union (EU) Parliament has taken a significant step towards protecting individual privacy in the age of AI. A majority of the EU Parliament is now in favour of a proposal to ban the use of AI surveillance in public spaces.

3.3.6 Real life examples:

CASE 1. Google's Location Tracking

Due to privacy concerns, Google's location-tracking practices have come under intense scrutiny in recent years. The company tracks the location of its users, even when they have not given explicit permission for their location to be shared. This revelation came to light in 2018 when an Associated Press investigation found that Google services continued to store location data, even when users turned off location tracking. This was a clear breach of user trust and privacy, and Google faced significant backlash from users and privacy advocates.

Since 2018, Google has changed its location tracking policies and improved transparency regarding how it collects and uses location data. However, concerns remain regarding the extent of data collected, how it is used, and who has access to it. As one of the world's largest tech companies, Google's actions have far-reaching implications for individuals and society at large.

One of the biggest issues with Google's location tracking practices is the potential for the misuse of personal data. Location data is incredibly sensitive, and if it falls into the wrong hands, it can be used to track individuals' movements, monitor their behaviour, and even be used for criminal activities. The implications of location data being leaked or hacked can be dire, and it is essential for companies like Google to ensure that they have robust security measures in place to protect user data. Also, there is the issue of third-party access to user data, which can be used for advertising purposes or even sold to other companies for profit.

CASE 2. AI-Powered Recommendations: My Personal Experience with Google's Suggestion Engine

An example of privacy concerns in the age of AI is the invasive nature of Big Tech companies. I recently shared a personal experience I had about watching a show on Amazon Prime on Apple TV. Two days after finishing the show, I received news recommendations related to the show on a Google app on an iPhone, while I never watched that show on my

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

iPhone. An alarming practice and it begs the question: does Google have full access to all of our apps and activities?

As someone who has been working with big data for over a decade, I know it is technically possible, but it is concerning that it is allowed. For this level of personalised recommendation to be made, Google would need to access information from other apps on the iPad (even with my privacy settings preventing this practice) or eavesdropping on my conversations using the microphone of my iPhone or iPad and connect it to the my Google account. Both are not allowed and are a massive breach of privacy.

The example of Google's suggestive algorithm highlights the significant privacy concerns in the age of AI. The fact that Google is able to make personalised recommendations based on seemingly unrelated activities raises questions about the company's access to our private data. While this level of personalisation is technically possible, it is important to consider the ethical implications of such practices. As we continue relying more on AI and big data, it is critical to ensure privacy is respected and protected. It is vital that companies and policymakers take the necessary steps to establish clear guidelines and regulations to ensure that AI technology is developed and used in a way that upholds fundamental human rights and values.

CASE 3. The Use of AI in Hiring and Recruitment

The use of AI in hiring and recruitment has become increasingly popular in recent years. Companies are turning to AI-powered tools to screen and select job candidates, citing benefits such as increased efficiency and objectivity. However, these tools can also raise significant concerns about fairness and bias. One notable example is the case of Amazon's AI-powered recruiting tool, which was found to discriminate against women because the system was trained on resumes from mostly male candidates.

This highlights the potential for AI to perpetuate existing biases and discrimination, and the need for careful consideration and testing of these tools to ensure they are not inadvertently perpetuating unfair practices. As the use of AI in hiring and recruitment continues to grow, it is crucial that we prioritise transparency and accountability to prevent discrimination and ensure fairness in the workplace.

3.4 ALGORITHMIC BIAS CONSIDERATIONS

- The IEEE P7003 Standard for Algorithmic Bias Considerations is one of eleven IEEE ethics related standards .
- Which are currently under development as part of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.
- The purpose of the IEEE P7003 standard is to provide individuals or organizations creating algorithmic systems with development framework to avoid unintended, unjustified and inappropriately differential outcomes for users.
- The IEEE Standards Association (IEEE SA) launched the IEEE Global Initiative on Ethics for Autonomous and Intelligence Systems in April 2016.

This document is available free of charge on studocu

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

- Early 2018 the main pillars of the Global Initiative are:
  - a public discussion document "Ethically Aligned Design: A vision for Prioritizing human Well-being with Autonomous and Intelligent Systems", on establishing ethical and social implementations for intelligent and autonomous systems and technology aligned with values and ethical principles that prioritize human well-being in a given cultural context;
  - a set of eleven working groups to create the IEEE P70xx series ethics standards, and associated certification programs, for Intelligent and Autonomous systems.
- The IEEE P70xx series of ethics standards aims to translate the principles of Ethically Aligned Design document into actionable guidelines that can be used as practical industry standards.
- The eleven IEEE P70xx standards that are currently under development are:
  - IEEE P7000: Model Process for Addressing Ethical Concerns During System Design
  - IEEE P7001: Transparency of Autonomous Systems
  - IEEE P7002: Data Privacy Process
  - IEEE P7003: Algorithmic Bias Considerations
  - IEEE P7004: Standard on Child and Student Data Governance
  - IEEE P7005: Standard on Employer Data Governance
  - IEEE P7006: Standard on Personal Data AI Agent Working Group
  - IEEE P7007: Ontological Standard for Ethically Driven Robotics and Automation Systems
  - IEEE P7008: Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems
  - IEEE P7009: Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems
  - IEEE P7010: Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems
- IEEE P7003 is aimed to be used by people/organizations who are developing and/or deploying automated decision (support) systems (which may or may not involve AI/machine learning) that are part of products/services that affect people.
- Typical examples would include anything related to personalization or individual assessment, including any system that performs a filtering function by selecting to prioritize

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

the ease with which people will find some items over others (e.g. search engines or recommendation systems).

- 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 individual's preferences and behaviour.
- The requirements specification provided by the IEEE P7003 standard will allow creators
    - to communicate to users, and
    - regulatory authorities,
    - that up-to-date best practices were used in the design
    - testing and evaluation of the algorithm to attempt to avoid unintended, unjustified and inappropriate differential impact on users.

Example, an online retailer developing a new product recommendation system might use the IEEE P7003 standard as follows:

- Early in the development cycle, after outlining the intended functions of the new system IEEE P7003 guides the developer through a process of considering the likely customer groups, in order to identify if there are subgroups that will need special consideration (e.g. people with visual impairments).
- In the next phase of the development, the developer is establishing a testing dataset to validate if the system is performing as desired.
- Referencing P7003 the developer is reminded of certain methods for checking if all customer groups are sufficiently represented in the testing data to avoid reduced quality of service for certain customer groups

- Throughout the development process IEEE P7003 challenges the developer to think explicitly about the criteria that are being used for the recommendation process and the rationale, i.e. justification, for why these criteria are relevant and why they are appropriate (legally and socially).
- This process of analysis will help the business to be aware of the context for which this recommendation system can confidently be used, and which uses would require additional testing (e.g. age ranges of customers, types of products).
- The IEEE P7003 standard will provide a framework, which helps developers of algorithmic systems and those responsible for their deployment to identify and mitigate unintended, unjustified and/or inappropriate biases in the outcomes of the algorithmic system.

This document is available free of charge on studocu

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

- 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.
- The standard will describe specific methodologies that allow users of the standard to assert how they worked to address and eliminate issues of unintended, unjustified and inappropriate bias in the creation of their algorithmic system. This will help to design systems that are more easily auditable by external parties (such as regulatory bodies).

Elements include:

- a set of guidelines for what to do when designing or using such algorithmic systems following a principled methodology (process), engaging with stakeholders (people), determining and justifying the objectives of using the algorithm (purpose), and validating the principles that are actually embedded in the algorithmic system (product);
- a practical guideline for developers to identify when they should step back to evaluate possible bias issues in their systems, and pointing to methods they can use to do this;
- benchmarking procedures and criteria for the selection of validation data sets for bias quality control;
- methods for establishing and communicating the application boundaries for which the system has been designed and validated, to guard against unintended consequences arising from out-of-bound application of algorithms;
- methods for user expectation management to mitigate bias due to incorrect interpretation of systems outputs by users (e.g. correlation vs. causation), such as specific action points/guidelines on what to do if in doubt about how to interpret the algorithm outputs;

3.4.1 Structure

Standard document will consist of three main section categories:

1. Foundational sections covering issues related to the fundamentals

   of understanding algorithmic bias;

2. Algorithmic system design and implementation orientated sections addressing actionable recommendations for identifying and mitigating algorithmic bias;

3. Use cases providing examples of systems where the use of the

   P7003 standard could provide clear benefits.

Foundational sections

- Foundational sections are currently envisioned to include sections on ‘Taxonomy of Bias’, ‘Legal frameworks related to Bias’, ‘Psychology of Bias’ and ‘Cultural context of Bias’.

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

- Even though the presence of these foundational sections may appear unusual for an industry standard, we believe that they play an important part in an 'ethics' standard such as IEEE P7003.
- The foundational sections provide a framework of understanding that should allow the designers of algorithmic systems to go beyond a mechanistic 'tick-box' compliance exercise towards a deeper engagement with the underlying ethical issues of algorithmic bias.

System Design and Implementation sections

- The 'algorithmic system design and implementation' orientated sections are currently envisaged to include sections on 'Algorithmic system design stages', 'Person categorizations and identifying of affected groups', 'Representativeness and balance of testing/training/validation data', 'System outcomes evaluation','Evaluation of algorithmic processing', Assessment of resilience against external biasing manipulation', 'Assessment of scope limits for safe system usage' and 'Transparent documentation', though it is anticipated that further sections will be added as work progresses.
- The intent of these sections is to provide a clear framework of guidance including challenge questions to help designers identify unintended bias issues that would go unnoticed unless specifically looked for. A possible comparison would be the way in which explicit questioning of everyday behavior is required in order to identify and mitigate unconscious bias in management practices.
- Solutions to identified causes of algorithmic bias will likely primarily take the form of listing classes of solution methods, with links to relevant work being published at venues such as FairWare, FAT*, KDD and similar publications, in order to reflect the context dependent nature of optimal solutions and the dynamic development in the research on improved methods.

Use Cases

- The Use Cases form an annex to the IEEE P7003 standard document listing a number of illustrative examples of algorithmic systems that resulted in nintended bias, or that highlight specific types of concerns about bias that could be ddressed by following the framework provided by IEEE P7003.
- The inclusion of the Use Cases, and their standardized presentation format, were proposed by a working group participant with experience of industry engagement with standards.
- They form an important element for 'making the case' for using ethics standards within a corporate context.

Some examples of the use cases that have been gathered so far include:

- "Tay the Nazi chatbot", an example of deliberate system behavior corruption through biased manipulation of inputs by an external 'adversary';

This document is available free of charge on studocu

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

- "The use of facial expression recognition to support diagnostic assessment for patient prioritization", an example of a sensitive application context where

differences in operational capability of the system for different population groups can easily result in reputation damaging claims of unjustified bias;

- "Beauty contest judging algorithm that appeared biased to favor lighter skin tones", an example of bias in the training data resulting in biased outcomes that undermined the credibility of the statement purpose of the algorithm (to produce objective beauty contest judgements);

3.4.2 Methodology

- Methodologically, the content of the P70xx standards are developed by the working group members through an open deliberation process in which each participant is encourage to suggest content or amendments for the standard document.
- In order to reflect the broad socio-technical nature of the AI ethics issues addressed by the P70xx standards, the working group members are drawn from a broad range of stakeholders including civil-society organizations, industry and a wide range of academic disciplines.
- Participation in the working groups is on an individual basis.
- Even through the participants are affiliated with particular stakeholder organizations, all voices in the standard development process are treated as equals
- .With the exception of the working group chair and vice-chair, IEEE membership is not required and does not change the status of the participant within the working group.
- For the P7003 Standard for Algorithmic Bias Considerations the working group currently consists of 78 participants identifying as having expertise in: Computer Science (18), Engineering (8), Law (6), Business/Entrepreneurship (6), Policy (6), Humanities (4), Social Sciences (3), Arts (2) and Natural Sciences (1).
- Once the IEEE P7003 draft document is completed and approved by the IEEE P7003 working group, it will be submitted for balloting approval to the IEEE-SA.
- The IEEE-SA will send out an invitation-to-ballot to all IEEE-SA members who have expressed an in interest in the subject, i.e. Algorithmic Bias.
- If the draft receives at least 75% approval, the draft is submitted to the IEEE-SA Standards Board Review Committee, which checks that the proposed standard is compliant with the IEEE-SA Standards Board Bylaws and Operations Manual.
- The Standards Board then votes to approve the standard, which requires a simple majority.
- At that point, about 2.5 to 3 years after the proposal for Number in brackets indicate number of participants who identified as having this expertise as part of an informal internal survey.

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

- Many participants chose not to respond while some chose to indicate multiple expertise. developing the standard was first submitted, the standard is published for use.

3.4.3 Conclusion

- As part of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems a series of eleven ethics standards are under development, designated IEEE P7000 through IEEE P7010.
- The IEEE P7003 Standard for Algorithmic Bias Considerations aims to provide an actionable framework for improving fairness of algorithmic decision-making systems that are increasingly being developed and deployed by industry, government and other organizations.
- The IEEE P7003 standard is currently transitioning from an initial exploratory phase into a consolidation and specification phase.
- Participation in the IEEE P7003 working group is open to all who are interested in contributing towards reducing and mitigating unintended, unjustified and societally unacceptable bias in algorithmic decisions.

3.5 ONTOLOGICAL STANDARD FOR ETHICALLY DRIVEN ROBOTICS AND AUTOMATION SYSTEMS

- In the rapidly evolving fields of artificial intelligence (AI) and robotics, the elaboration of ethical concerns, considerations, and requirements helps illustrate the nature of technology's reach and impact on society where there is a legal void.
- Thus, establishing ethics in AI and robotics is fundamental to identifying their potential risks and benefits , especially in our widespread wrecked world.
- Ethical considerations help to create a much-desired relationship between technology and human values and address the impacts a technology can have, thereby addressing issues of trust, safety, security, data privacy, and algorithmic bias.
- The need for an ethical framework is urgent because of the increasing adoption and use of autonomous and intelligent systems (A/ISs) in many domains, such as health care, education, finance, and insurance services.
- In 2016, IEEE established its Global Initiative on Ethics of Autonomous and Intelligent Systems with the aim of ensuring that every stakeholder involved in the design, development, and management of A/ISs is educated trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity.
- The IEEE Robotics and Automation Society (RAS)/Standards Association (SA) 7007 Ontologies for Ethically Driven Robotics and Automation Systems Working Group (IEEE 7007 WG) was established in 2017.

This document is available free of charge on studocu

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

- During the past four years, this group has been working to create an ontological standard to enable the development of ethically driven robotics and automation systems.
- This standard was scrutinized by the global community in 2021, and it was officially approved by the IEEE SA on 24 September 2021.
- Due to the relevance of this standard, the IEEE 7007 WG has been selected as a recipient of the IEEE SA Emerging Technology Award "for developing an innovative ontological standard on the ethics of artificial intelligence"

3.5.1 Regulatory Frameworks

- There are various international regulatory initiatives in the area of emerging technologies with an impact on AI and robotics
- Current international regulatory requirements are contained in a combination of nonlegally binding ethical standards, frameworks, and guidelines as well as legally binding instruments
- The IEEE Ethics Certification Program for Autonomous and Intelligent Systems is a world first in setting standards for the ethical certification of products, services, and systems deploying AI and robotics in the public and private sectors.
- Certification is essential to guarantee that these technologies operate as expected when they are interacting with human and nonhuman agents.
- Different from these frameworks, the standard developed by the IEEE 7007 WG has a formal and ontological representation that can be used not only as a foundation to elaborate public policies but also to create computational systems.
- In fact, IEEE Standard 7007 is the first global ontological standard that contains the concepts, definitions, and axioms that are necessary to establish ethical methodologies for the design, development, and deployment of AI and robotics.

3.5.2 IEEE 7007 WG

- The IEEE 7007 WG is under the umbrella of the IEEE SA P7000 series devoted to ethics in A/IS.
- In this scope, several WGs were formed—15 to date— to deliver a broad range of standards and/or recommended practices. Among the goals of the IEEE 7007 WG are to
  - Establish 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
  - Align the ethics and engineering communities to understand how to pragmatically design and implement these systems in unison
  - Develop a precise communication framework among global experts of different domains, including robotics, automation, and ethics.

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

- To attain these goals, the IEEE 7007 WG developed a set of ontologies for representing the domain in a more precise way.
- As a result, IEEE Standard 7007 contains a set of ontologies that represents norms and ethical principles (NEP), data privacy and protection (DPP), transparency and accountability, and ethical violation management (EVM)
- The 2development of this standard was a complex process requiring a dedicated lifecycle.
- For this purpose, the IEEE 7007 WG developed an agile, collaborative, and iterative methodology called the robotic standard development lifecycle.
- The usefulness of ontologies in standardization is two fold.
- On the one hand, standardization processes are set to produce a body of knowledge that reflects a consensual view of practitioners around a topic, defining, among other aspects, a standard knowledge structure in a domain, including common concepts, relationships, and attributes.
- Ontologies and their methods provide a formal approach to that aspect of the standardization process, which is expected to produce a sounder standard.
- On the other hand, the ontologies themselves, as formal artifacts, can be seen as products of the standardization process that can be used directly in data processing and automatic reasoning.
- As an example, one can cite IEEE 1872-2015, which set forth to establish clear definitions for common terms in robotics and automation.

3.5.3 IEEE 7007 ONTOLOGICAL STANDARD FOR ETHICALLY DRIVEN ROBOTICS AND AUTOMATION SYSTEMS

Top-Level Ontology:

- As a core ontology, the ethically driven robotics and autonomous systems (ERAS) ontology represents a mid level set of formalization and commitments that are platform independent and intended to fit between an upper top level or foundational ontology and lower-domain and application-specific ontologies.
- While some potential users of the standard may intend to align the ERAS core formalizations with existing top-level ontologies specific to their application domain, other user communities will only require a minimal top level set of conceptualizations to complete the formalization of the concepts, terms, and commitments axiomatized in the ERAS ontology.
- For that purpose, the four ERAS subdomain ontologies are augmented with axioms sufficient to complete the definitions and commitments expressed in the core ERAS models.

This document is available free of charge on studocu

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

- These axioms are expressed formally using the Common Logic Interchange Format (CLIF) .
- The ERAS top level ontology (ERAS-TLO) formalizations define a minimal set of terms deemed relevant to the characterization of ethically oriented agents and autonomous systems.
- It is not intended to be applicable as a TLO in other contexts.

NEP Ontology:

- The NEP ontology subdomain formalizes the terminology and ontological commitments associated with ethical theories and principles that characterize the norms of expected behaviors for norm-oriented agents and autonomous systems.
- This includes axioms for concepts, such as norms, ethical theory, situation plan repertoire, agent plans, plan actions, and agent actions as well as the corresponding relationships, such as "selects plans from," "subscribes to," "satisfies," and "constrains plans for."
- Figure 1 depicts a brief and partial view of a subset of the NEP terms with a Unified Modeling Language (UML) class diagram.

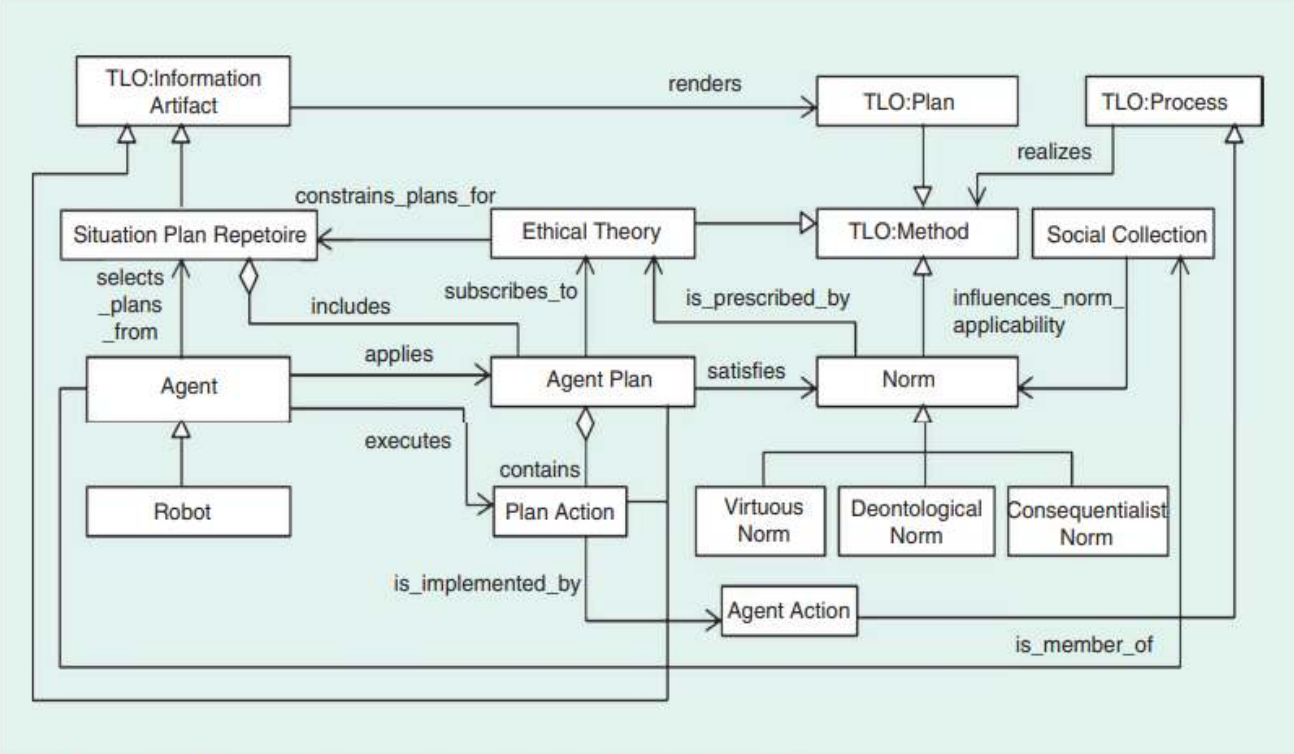


Figure 1. A partial UML model of the ERAS NEP ontology. UML: Unified Modeling Language.

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

DPP Ontology:

- The DPP ontology represents concepts and relationships among the diverse agents, entities, and organizations that may be involved at different stages in data gathering, processing, transfer,retention, and storage and in which autonomous systems may be deployed.
- Thus, the DPP ontology represents concepts like the natural person, caregiver, data protection authority, controller, and authorized accessor as well as the different types and processing of personal data (e.g., health data, economic data, and social data) and corresponding data process access.
- DPP principles, like privacy by design, data protection by design, data protection by default, and human rights by design, were also included in the standard.
- It is crucial to represent this domain formally because of the relevance of the existing regulations worldwide about DPP.
- In addition, evaluating the impact of driven robotics and automation systems on personal data and, hence, on the processing of personal information is essential to the regulation of A/IS.
- As stated in the standard, "Data privacy is a highly complex and increasingly regulated area of law, in which the regulatory regime is rapidly evolving.
- No standard can provide unconditional consistency with all applicable laws and regulations, which continue to change rapidly in this area, and may also vary at the local, state and regional level.
- Users of this Standard are responsible for keeping apprised of such laws and regulations."

Transparency and Accountability Ontology:

- The transparency and accountability ontology subdomain formalizes the vocabulary and ontological commitments relevant for terms capable of expressing the concepts and relationships necessary to enable ethical autonomous systems with capabilities that provide informative explanations for plans and associated actions.
- Ethically aware agents require the ability to be transparent in their interactions with other agents.
- An agent qualifies as an autonomous transparent agent if it is enabled with an always-available mechanism capable of reporting its behavior, intentions, perceptions, goals, and constraints in a manner that permits authorized users and collaborating agents to understand its past and expected future behaviors.

EVM:

- The EVM ontology subdomain presents axioms to formalize the terminology associated with capabilities to detect, assess, and manage ethical and legal norm violations occurring within or generated by autonomous system behavior.

This document is available free of charge on studocu

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

- This includes concepts such as norm violation, norm violation incident, responsibility ascription, ascription justification, grounds for ascription, agent accountability, event causation, liability sanction, and ethical behavior monitor.
- Figure 2 presents a partial view of the EVM concepts and relationships in a UML class diagram.
- Agent system components or other agents providing an ethical behavior monitoring service may detect and record norm violations using norm violation incident information artifacts.
- A norm violation elicits a responsibility ascription process as a social interaction process to identify those responsible for the violation.
- A responsibility ascription process that results in the ascription of responsibility to one or more agents is justified by an ascription justification information artifact.
- This category represents the collection of facts formulated and asserted by an authoritative agent or agency to ascribe responsibilities for ethical or legal norm violations.
- It is composed of constituent grounds for ascription information artifacts.
- Ethical violation as well as transparency and accountability ontologies identify accountability and legal responsibility as important real-world concepts impacting AI and robotics.
- Legal responsibility and its manifestations in terms of culpability as well as civil and criminal liability , have influenced the content of the standard.
- The parameters between accountability and responsibility are also reflected with use of terminology that conveys a spectrum of potential agents who may be held responsible (e.g., partial or distributed responsibility).
- An important observation here is that the EVM core axioms restrict autonomous system agent responsibility ascription to a set of specific system ethical norm violations and when human agents are involved in the collective distributed responsibility chain.
- Autonomous systems cannot be ascribed any responsibility for legal norm violations.
- An autonomous system acting as a single agent cannot be ascribed responsibility for any type of norm violation.
- Distributed responsibility is applicable only when the autonomous system is a member of a human-directed team and when an action by the system caused a norm violation

Downloaded by Lokesh S (lokesh@psgitech.ac.in)

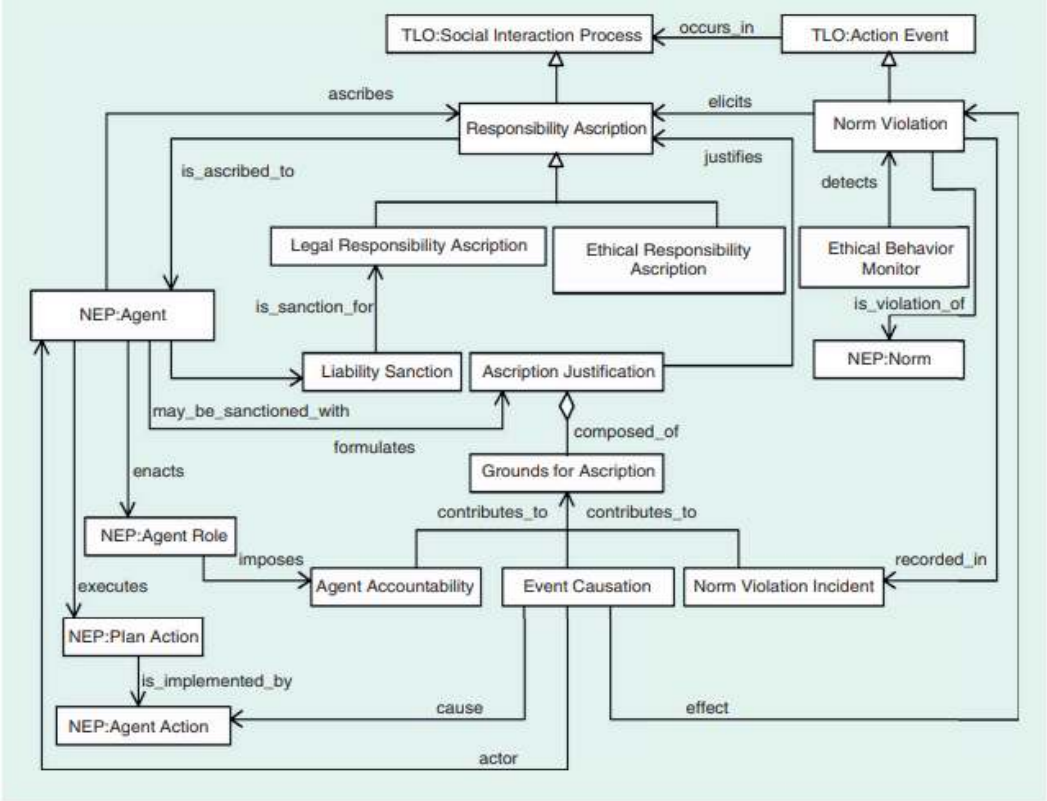


Figure 2. A partial UML model of the ERAS EVM ontology.

3.5.4 Conclusions

- IEEE Standard 7007 is the first global ontological standard elaborated to establish ethical methodologies for the design, development, and deployment of A/IS.
- It contains a set of ontologies that represents, explicitly and formally, core concepts that are relevant to dealing with NEP, transparency and accountability, EVM, and DPP.
- It is expected that this work has a significant impact worldwide in being used to teach ethical design; for both human and institutional capacity building in the domain of the ethics of AI; to create computational ethically aligned systems; to create a taxonomy to support the elaboration of public policies; and to strengthen digital cooperation across nations applied together with the other members of the IEEE P7000 family.

This document is available free of charge on studocu

Downloaded by Lokesh S (lokesh@psgitech.ac.in)