DMI COLLEGE OF ENGINEERING

(An Autonomous Institution)
(Approved by AICTE, New Delhi. Affiliated to Anna University, Chennai)
PALANCHUR, CHENNAI – 600 123.

DEPARTMENT OF INFORMATION TECHNOLOGY

Course Material

CCS345 – ETHICS AND AI

Year / Semester : III / VI

PREPARED BY

Mr. JEBA KINGSLEY.D
Assistant Professor / IT

UNIT-1

INTRODUCTION

Definition of morality and ethics in AI- Impact on society- Impact on human psychology- Impact on the legal system- Impact on the environment and the planet- Impact on trust.

INTRODUCTION

What is AI – and what is intelligence?

Artificial Intelligence (AI) refers to systems that display intelligent behavior by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.

AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications).'

How do we define intelligence?

A straightforward definition is that intelligent behaviour is 'doing the right thing at the right time'. The definitions of intelligence, identifying three common features

Intelligence is

(1) A property that an individual agent has as it interacts with its environment or environments.

(2) Related to the agent's ability to succeed or profit with respect to some goal or objective.

(3) Depends on how able that agent is to adapt to different objectives and environments.

They point out that intelligence involves adaptation, learning and understanding. At its simplest, then, intelligence is 'the ability to acquire and apply knowledge and skills and to manipulate one's environment'.

Physical robot and its environment are,

- Human environment (for social robots)
- A city street (for an autonomous vehicle)
- A care home or hospital (for a care or assisted living robot)
- A workplace (for a workmate robot)

The 'environment' of a software AI are,

- clinical (for a medical diagnosis AI)
- public space (for face recognition in airports, for instance, or virtual for face recognition in social media)

Types of artificial intelligence

- weak AI
- strong AI

Weak AI

Weak AI is also called as Narrow AI or Artificial Narrow Intelligence (ANI). It is AI trained and focused to perform specific tasks. Weak AI drives most of the AI that surrounds us today. 'Narrow' might be a more accurate descriptor for this type of AI as it is anything but weak; it enables some very robust applications, such as Apple's Siri, Amazon's Alexa, IBM Watson, and autonomous vehicles.

Strong AI

Strong AI is made up of Artificial General Intelligence (AGI) and Artificial Super Intelligence (ASI). Artificial general intelligence (AGI), or general AI, is a theoretical form of AI where a machine would have an intelligence equaled to humans; it would have a self-aware consciousness that has the ability to solve problems, learn, and plan for the future. Artificial Super Intelligence (ASI)—also known as super intelligence—would surpass the intelligence and ability of the human brain. While strong AI is still entirely theoretical with no practical examples in use today.

Machine Learning vs Deep Learning

Machine learning is the term used for AIs which are capable of learning or, in the case of robots, adapting to their environment. There are a broad range of approaches to machine learning, but these typically fall into two categories:

- Supervised learning
- Unsupervised learning

Supervised learning systems generally make use of Artificial Neural Networks (ANNs), which are trained by presenting the ANN with inputs (for instance, images of animals) each of which is tagged (by humans) with an output (i.e. giraffe, lion, gorilla). This set of inputs and matched outputs is called a training data set. After training, an ANN should be able to identify which animal is in an image it is presented with (i.e. a lion), even though that particular image with a lion wasn't present in the training data set.

Limitations

- The training data set must be truly representative of the task required; if not, the AI will exhibit bias.
- ANNs learn by picking out features of the images in the training data unanticipated by the human designers

Unsupervised learning has no training data; instead, the AI (or robot) must figure out on its own how to solve a particular task (i.e. how to navigate successfully out of a maze), generally by trial and error.

Limitations

Unsupervised learning is generally more robust than supervised learning but suffers the limitation that it is generally very slow (compared with humans who can often learn from as few as one trial).

Deep learning simply refers to (typically) supervised machine learning systems with large (i.e. many-layered) ANNs and large training data sets.

1.1. DEFINITION OF MORALITY AND ETHICS IN AI

Ethics are moral principles that govern a person's behaviour or the conduct of an activity. As a practical example, one ethical principle is to treat everyone with respect.

AI ethics are the study of the moral and ethical considerations involved in developing and using Artificial Intelligence. The field of AI ethics does not only focus on what is morally rightor wrong for a specific machine but also on how to approach important questions such as: Howcan we make sure that autonomous machines act following our values? How can we ensure that they have less probability of harming humans than other technologies? What is our responsibility as designers and users of ethical AI systems?

Ethics in AI are also referred to as machine ethics or computational ethics. As an emerging discipline, it is often unclear what constitutes "good" or "bad" behavior for AI algorithms. However, several principles guide researchers in this area:

- Algorithms should be designed to be accountable and inherently trustworthy; if an algorithm causes harm, it should be possible to determine which parts were responsible so they can be fixed or replaced. This means that while humans may need some time before they understand why something happened, computers shouldn't need any explanation at all because everything will always be explicit within their codebase.
- Automation should not result in job loss. Rather than replacing people who would otherwise occupy those positions themselves (like waiters, for instance), companies should look into automating tasks where machines can do better work than humans dueto being faster/more accurate/less prone to error, etc.
- Artificial Intelligence systems should produce the least amount of harm. However, this does not mean these systems won't ever produce any harm since no machine will ever know exactly how its actions will affect other people/things. For example, someone might get hurt if an autonomous car crashes into another vehicle at full speed. To prevent this from happening again, the company would have to go back and check that its algorithm is not biased against certain groups of people. This could mean running it through a series of tests to ensure that no one is being discriminated against by their Machine Learning process. Companies should ensure that their Artificial Intelligence systems are not biased; to prevent discrimination against certain groups of people, companies should ensure that the AI they create is not biased toward anyone.

Principles for AI Ethics

Principles for AI ethics are a set of rules and guidelines that are meant to help protect society from the negative effects of Artificial Intelligence. These principles aim to protect people, the environment, and the economy.

AI ethics revolves around four main areas:

1. Safety:

This refers to how well an AI can avoid harming humans. This includes things like not causing physical harm or using offensive language. It also includes things like protecting intellectual property rights and privacy.

2. Security:

This refers to how well an AI can prevent other systems from attacking it or taking advantage of it in some way. It also refers to how well an AI can protect itself from being hacked or manipulated by humans who want to use it for nefarious means (like stealing money).

3. Privacy:

This refers to how much information an AI system knows about you, where it gets its data from, how it stores that information, what kind of analysis tools it uses with that data, etc. Basically, everything related to your personal information is being used/shared by any technology company!

4. Fairness:

This refers to whether or not your rights as a consumer are being protected when interacting with a company's services/products.

AI systems should be designed and operated to be safe, secure, and private. The designers and builders of intelligent autonomous systems must:

- Ensure that they are robust, reliable, and trustworthy.
- Incorporate mechanisms that reflect societal values and aims as they interact with people outside their immediate purview.
- Ensure that their creations are adaptive so that they can learn from experience over time to improve their performance and capabilities.
- Consider the full range of human needs in their design, for example, by promoting safety, privacy, trustworthiness, fairness, transparency, accountability, and inclusion in society through AI technologies."
- Ensure that they can explain how decisions are made by their creations so that people can understand them and take action to correct any mistakes that are made.
- Confirm that these technologies are designed in ways that respect human rights, including privacy, freedom of thought and speech, bodily integrity, and freedom from cruel or degrading treatment."
- Consider the impact on society when developing these technologies.

Challenges in AI Ethics

As a new field, AI ethics is still in the process of being developed. There are many ethics and risks of AI. There are no clear rules or guidelines for AI ethics because it is a relatively new field. As such, of these AI ethical issues, it can be challenging to determine whether or not any given program has acted ethically when there are no established protocols for determining what constitutes ethical behavior.

Additionally, the complexity of Artificial Intelligence makes it difficult to examine its capabilities and limitations with regard to ethical considerations. For example, if a self-driving car were programmed to make split-second decisions about whether or not it should save its passengers at the expense of pedestrians crossing the street, how could we know whether ornot these decisions were morally sound? Without knowing all possible outcomes of these actions—and their consequences—it would be impossible for us humans (or even other computers) to judge them truly objectively from a moral standpoint. This problem is compounded when considering that Machine Learning algorithms vary widely depending on their training data sets and other parameters (such as "fitness functions").

In fact, many people believe that some form of regulation may be necessary before Artificial Intelligence becomes widespread enough for us humans even realize there's anything wrong with our creations' behavior patterns; these individuals fear that without proper oversight by experts versed both in technology development and ethics research fields like philosophy/political science/economics, etc., society will suffer greatly due to irresponsible use cases involving Artificial Intelligence technology devices such as autonomous cars driving around streets full of pedestrians who might not understand what they're witnessing.

This same scenario applies equally well across many industries where autonomous machines are becoming commonplace, including manufacturing plants where robots perform tasks intended by humans so efficiently they're impacting unemployment rates worldwide.

1.2. IMPACT ON SOCIETY

AI grows more sophisticated and widespread; the voices warning against the potentialdangers of artificial intelligence grow louder.

"These things could get more intelligent than us and could decide to take over, and we need to worry now about how we prevent that happening," said Geoffrey Hinton, known as the"Godfather of AI" for his foundational work on machine learning and neural network algorithms. In 2023, Hinton left his position at Google so that he could "talk about thedangers of AI," noting a part of him even regrets his life's work.

The renowned computer scientist isn't alone in his concerns.

Risks of artificial intelligence

- Automation-spurred job loss
- Deep fakes
- Privacy violations
- Algorithmic bias caused by bad data
- Socioeconomic inequality
- Market volatility
- Weapons automatization
- Uncontrollable self-aware AI

Whether it's the increasing automation of certain jobs, gender and racially biased algorithms or autonomous weapons that operate without human oversight (to name just a few), unease abounds on a number of fronts. And we're still in the very early stages of what AI is really capable of.

1. Lack of AI transparency and explain ability

AI and deep learning models can be difficult to understand, even for those that work directly with the technology. This leads to a lack of transparency for how and why AI comes toits conclusions, creating a lack of explanation for what data AI algorithms use, or why they may make biased or unsafe decisions. These concerns have given rise to the use of explainable AI, but there's still a long way before transparent AI systems become common practice.

2. Job losses due to AI automation

AI-powered job automation is a pressing concern as the technology is adopted in industries like marketing, manufacturing and healthcare. By 2030, tasks that account for up to 30 percent of hours currently being worked in the U.S. economy could be automated — with Black and Hispanic employees left especially vulnerable to the change — according to McKinsey. Goldman Sachs even states 300 million full-time jobs could be lost to AI automation.

The reason we have a low unemployment rate, which doesn't actually capture people that aren't looking for work, is largely that lower-wage service sector jobs have been pretty robustly created by this economy, futurist Martin Ford told Built In. With AI on the rise, though, "I don't think that's going to continue."

As AI robots become smarter and more dexterous, the same tasks will require fewer humans. And while AI is estimated to create 97 million new jobs by 2025, many employees won't have the skills needed for these technical roles and could get left behind if companies don't up skill their workforces.

"If you're flipping burgers at McDonald's and more automation comes in, is one of these new jobs going to be a good match for you?" Ford said. "Or is it likely that the new job requires lots of education or training or maybe even intrinsic talents — really strong interpersonal skills or creativity — that you might not have? Because those are the things that, at least so far, computers are not very good at." Even professions that require graduate degrees and additional post-college training aren't immune to AI displacement.

As technology strategist Chris Messina has pointed out, fields like law and accounting are primed for an AI takeover. In fact, Messina said, some of them may well be decimated. AI already is having a significant impact on medicine. Law and accounting are next, Messina said,the former being poised for "a massive shakeup."

"Think about the complexity of contracts, and really diving in and understanding whatit takes to create a perfect deal structure," he said in regards to the legal field. "It's a lot of attorneys reading through a lot of information — hundreds or thousands of pages of data and documents. It's really easy to miss things. So AI that has the ability to comb through and comprehensively deliver the best possible contract for the outcome you're trying to achieve is probably going to replace a lot of corporate attorneys."

3. Social manipulation through AI algorithms

Social manipulation also stands as a danger of artificial intelligence. This fear has become a reality as politicians rely on platforms to promote their viewpoints, with one example being Ferdinand Marcos, Jr., wielding a TikTok troll army to capture the votes of younger Filipinos during the Philippines' 2022 election.

TikTok, which is just one example of a social media platform that relies on AI algorithms, fills a user's feed with content related to previous media they've viewed on the platform. Criticism of the app targets this process and the algorithm's failure to filter out harmful and inaccurate content, raising concerns over TikTok's ability to protect its users from misleading information.

Online media and news have become even murkier in light of AI-generated images andvideos, AI voice changers as well as deep fakes infiltrating political and social spheres. These technologies make it easy to create realistic photos, videos, audio clips or replace the image of one figure with another in an existing picture or video. As a result, bad actors have another avenue for sharing misinformation and war propaganda, creating a nightmare scenario where itcan be nearly impossible to distinguish between creditable and faulty news.

"No one knows what's real and what's not," Ford said. "So it really leads to a situation where you literally cannot believe your own eyes and ears; you can't rely on what, historically, we've considered to be the best possible evidence... That's going to be a huge issue."

4. Social surveillance with AI technology

In addition to its more existential threat, Ford is focused on the way AI will adversely affect privacy and security. A prime example is China's use of facial recognition technology in offices, schools and other venues. Besides tracking a person's movements, the Chinese government may be able to gather enough data to monitor a person's activities, relationships and political views.

Another example is U.S. police departments embracing predictive policing algorithms to anticipate where crimes will occur. The problem is that these algorithms are influenced by arrest rates, which disproportionately impact Black communities. Police departments then double down on these communities, leading to over-policing and questions over whether self-proclaimed democracies can resist turning AI into an authoritarian weapon.

5. Lack of data privacy using AI tools

If you've played around with an AI chatbot or tried out an AI face filter online, your data is being collected — but where is it going and how is it being used? AI systems often collect personal data to customize user experiences or to help train the AI models you're using (especially if the AI tool is free). Data may not even be considered secure from other users when given to an AI system, as one bug incident that occurred with ChatGPT in 2023 "allowed some users to see titles from another active user's chat history." While there are laws present to protect personal information in some cases in the United States, there is no explicit federal law that protects citizens from data privacy harm experienced by AI.

6. Biases due to AI

Various forms of AI bias are detrimental too. Speaking to the *New York Times*, Princeton computer science professor Olga Russakovsky said AI bias goes well beyond gender and race. In addition to data and algorithmic bias (the latter of which can "amplify" the former), AI is developed by humans — and humans are inherently biased.

"A.I. researchers are primarily people who are male, who come from certain racial demographics, who grew up in high socioeconomic areas, primarily people without disabilities," Russakovsky said. "We're a fairly homogeneous population, so it's a challenge tothink broadly about world issues."

The limited experiences of AI creators may explain why speech-recognition AI often fails to understand certain dialects and accents, or why companies fail to consider the consequences of a chatbot impersonating notorious figures in human history. Developers and businesses should exercise greater care to avoid recreating powerful biases and prejudices that put minority populations at risk.

7. Socioeconomic inequality as a result of AI

If companies refuse to acknowledge the inherent biases baked into AI algorithms, they may compromise their DEI initiatives through AI-powered recruiting. The idea that AI can measure the traits of a candidate through facial and voice analyses is still tainted by racial biases, reproducing the same discriminatory hiring practices businesses claim to be eliminating.

Widening socioeconomic inequality sparked by AI-driven job loss is another cause for concern, revealing the class biases of how AI is applied. Blue-collar workers who perform more manual, repetitive tasks have experienced wage declines as high as 70 percent because of automation. Meanwhile, white-collar workers have remained largely untouched, with some even enjoying higher wages.

Sweeping claims that AI has somehow overcome social boundaries or created more jobs fail to paint a complete picture of its effects. It's crucial to account for differences based on race, class and other categories. Otherwise, discerning how AI and automation benefit certain individuals and groups at the expense of others becomes more difficult.

8. Weakening ethics and goodwill because of AI

Along with technologists, journalists and political figures, even religious leaders are sounding the alarm on AI's potential socio-economic pitfalls. In a 2019 Vatican meeting titled, "The Common Good in the Digital Age," Pope Francis warned against AI's ability to "circulate tendentious opinions and false data" and stressed the far-reaching consequences of letting this technology develop without proper oversight or restraint.

"If mankind's so-called technological progress were to become an enemy of the common good," he added, "this would lead to an unfortunate regression to a form of barbarism dictated by the law of the strongest."

The rapid rise of generative AI tools like ChatGPT and Bard gives these concerns more substance. Many users have applied the technology to get out of writing assignments, threatening academic integrity and creativity.

"The mentality is, 'If we can do it, we should try it; let's see what happens," Messina said. "'And if we can make money off it, we'll do a whole bunch of it.' But that's not unique to technology. That's been happening forever.'"

9. Autonomous weapons powered by AI

As is too often the case, technological advancements have been harnessed for the purpose of warfare. When it comes to AI, some are keen to do something about it before it's too late: In a 2016 open letter, over 30,000 individuals, including AI and robotics researchers, pushed back against the investment in AI-fueled autonomous weapons.

"The key question for humanity today is whether to start a global AI arms race or to prevent it from starting," they wrote. "If any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable, and the endpoint of this technological trajectory is obvious: autonomous weapons will become the Kalashnikovs of tomorrow."

This prediction has come to fruition in the form of Lethal Autonomous Weapon Systems, which locate and destroy targets on their own while abiding by few regulations. Because of the proliferation of potent and complex weapons, some of the world's most powerful nations have given in to anxieties and contributed to a tech cold war.

Many of these new weapons pose major risks to civilians on the ground, but the danger becomes amplified when autonomous weapons fall into the wrong hands. Hackers have mastered various types of cyber attacks, so it's not hard to imagine a malicious actor infiltrating autonomous weapons and instigating absolute armageddon. If political rivalries and warmongering tendencies are not kept in check, artificial intelligence could end up being applied with the worst intentions.

10.Financial crises brought about by AI algorithms

The financial industry has become more receptive to AI technology's involvement in everyday finance and trading processes. As a result, algorithmic trading could be responsible for our next major financial crisis in the markets.

While AI algorithms aren't clouded by human judgment or emotions, they also don't take into account contexts, the interconnectedness of markets and factors like human trust and fear. These algorithms then make thousands of trades at a blistering pace with the goal of selling a few seconds later for small profits. Selling off thousands of trades could scare investors into doing the same thing, leading to sudden crashes and extreme market volatility.

This isn't to say that AI has nothing to offer to the finance world. In fact, AI algorithms can help investors make smarter and more informed decisions on the market. But finance organizations need to make sure they understand their AI algorithms and how those algorithms make decisions. Companies should consider whether AI raises or lowers their confidence before introducing the technology to avoid stoking fears among investors and creating financial chaos.

11.Loss of human influence

An overreliance on AI technology could result in the loss of human influence — and a lack in human functioning — in some parts of society. Using AI in healthcare could result in reduced human empathy and reasoning, for instance. And applying generative AI for creative endeavors could diminish human creativity and emotional expression. Interacting with AI systems too much could even cause reduced peer communication and social skills. So while AIcan be very helpful for automating daily tasks, some question if it might hold back overall human intelligence, abilities and need for community.

12. Uncontrollable self-aware AI

There also comes a worry that AI will progress in intelligence so rapidly that it will become sentient, and act beyond humans' control — possibly in a malicious manner. Alleged reports of this sentience have already been occurring, with one popular account being from a former Google engineer who stated the AI chatbot LaMDA was sentient and speaking to him just as a person would. As AI's next big milestones involve making systems with artificial general intelligence, and eventually artificial super intelligence, cries to completely stop these developments continue to rise.

How to Mitigate the Risks of AI?

AI still has numerous benefits, like organizing health data and powering self-driving cars. To get the most out of this promising technology, though, some argue that plenty of regulation is necessary.

"There's a serious danger that we'll get [AI systems] smarter than us fairly soon and that these things might get bad motives and take control," Hinton told NPR. "This isn't just a science fiction problem. This is a serious problem that's probably going to arrive fairly soon, and politicians need to be thinking about what to do about it now."

Develop legal regulations

AI regulation has been a main focus for dozens of countries, and now the U.S. and European Union are creating more clear-cut measures to manage the spread of artificial intelligence. Although this means certain AI technologies could be banned, it doesn't prevent societies from exploring the field.

Create organizational AI standards:

Preserving a spirit of experimentation is vital for Ford, who believes AI is essential for countries looking to innovate and keep up with the rest of the world.

"You regulate the way AI is used, but you don't hold back progress in basic technology. I think that would be wrong-headed and potentially dangerous," Ford said. "We decide where we want AI and where we don't; where it's acceptable and where it's not. And different countries are going to make different choices."

Make AI part of company culture and discussions

The key is deciding how to apply AI in an ethical manner. On a company level, there are many steps businesses can take when integrating AI into their operations. Organizations can develop processes for monitoring algorithms, compiling high-quality data and explaining the findings of AI algorithms. Leaders could even make AI a part of their company culture, establishing standards to determine acceptable AI technologies.

Guide tech with humanities perspectives

"The creators of AI must seek the insights, experiences and concerns of people across ethnicities, genders, cultures and socio-economic groups, as well as those from other fields, such as economics, law, medicine, philosophy, history, sociology, communications, human-computer-interaction, psychology, and Science and Technology Studies (STS)."

Balancing high-tech innovation with human-centered thinking is an ideal method for producing responsible AI technology and ensuring the future of AI remains hopeful for the next generation. The dangers of artificial intelligence should always be a topic of discussion,so leaders can figure out ways to wield the technology for noble purposes.

"I think we can talk about all these risks, and they're very real," Ford said. "But AI is also going to be the most important tool in our toolbox for solving the biggest challenges we face."

1.3. IMPACT ON HUMAN PSYCOLOGY

Here we discuss how people relate to robots and autonomous systems from a psychological point of view. Humans tend to anthropomorphise them and form unidirectional relationships. The trust in these relationships is the basis for persuasion and manipulation that can be used for good and evil.

Here we discuss psychological factors that impact the ethical design and use of AIs and robots. It is critical to understand that humans will attribute desires and feelings to machines even if the machines have no ability whatsoever to feel anything. That is, people who are unfamiliar with the internal states of machines will assume machines have similar internal states of desires and feelings as themselves. This is called anthropomorphism. Various ethical risks are associated with anthropomorphism. Robots and AIs might be able to use "big data" to persuade and manipulate humans to do things they would rather not do. Due to unidirectional emotional bonding, humans might have misplaced feelings towards machines or trust them too much. In the worst-case scenarios, "weaponised" AI could be used to exploit humans.

Humans interact with robots and AI systems as if they are social actors. This effect has called as the "Media Equation" (Reeves and Nass 1996). People treat robots with politeness and apply social norms and values to their interaction partner (Broadbent 2017). Through repeated interaction, humans can form friendships and even intimate relationships with machines. This anthropomorphisation is arguably hard-wired into our minds and might have anevolutionary basis (Zlotowski et al. 2015). Even if the designers and engineers did not intend the robot to exhibit social signals, users might still perceive them. The human mind is wired to detect social signals and to interpret even the slightest behaviour as an indicator of some underlying motivation. This is true even of abstract animations. Humans can project "theory ofmind" onto abstract shapes that have no minds at all (Heider and Simmel 1944). It is therefore the responsibility of the system's creators to carefully design the physical features and social interaction the robots will have, especially if they interact with vulnerable users, such as children, older adults and people with cognitive or physical impairments.

To accomplish such good social interaction skills, AI systems need to be able to sense and represent social norms, the cultural context and the values of the people (and other agents)with which they interact (Malle et al. 2017). A robot, for example, needs to be aware that it would be inappropriate to enter a room in which a human is changing his/her underwear.

Being aware of these norms and values means that the agent needs to be able to sense relevant behaviour, process its meaning and express the appropriate signals. A robot entering the bedroom, for example, might decide to knock on the door prior to entering. It then needs to hear the response, even if only non-verbal utterance, and understand its meaning. Robots mightnot need to be perfectly honest. As Oscar Wilde observed "The truth is rarely pure and never simple." White lies and minor forms of dishonesty are common in human-human interaction (Feldman et al. 2002; DePaulo et al. 1996).

1. Misplaced Feelings Towards AI

Anthropomorphism may generate positive feelings towards social robots. These positive feelings can be confused with friendship. Humans have a natural tendency to assign human qualities to non-human objects. Friendships between a human and an autonomous robot

can develop even when the interactions between the robot and the human are largely unidirectional with the human providing all of the emotion.

A group of soldiers in Iraq, for example, held a funeral for their robot and created a medal for it (Kolb 2012). Carpenter provides an in-depth examination of human-robot interaction from the perspective of Explosive Ordinance Disposal (EOD) teams within the military (Carpenter 2016). Her work offers an glimpse of how naturally and easily people anthropomorphise robots they work with daily. Robinette et al. (2016) offered human subjects a guidance robot to assist them with quickly finding an exit during an emergency. They were told that if they did not reach the exit within the allotted 30 s then their character in the environment would perish. Those that interacted with a good guidance robot that quickly led them directly to an exit tended to name the robot and described its behaviour in heroic terms. Much research has shown that humans tend to quickly befriend robots that behave socially.

2. Misplaced Trust in AI

Users may also trust the robot too much. Ever since the Eliza experiments of the 1960s, it has become apparent that computers and robots have a reputation of being honest. While they rarely make mistakes in their calculations, this does not mean that their decisions are smart or even meaningful. There are examples of drivers blindly following their navigation devices into even dangerous and illegal locations. Robinette et al. (2016) showed that participants followed an obviously incompetent robot in a fire evacuation scenario. It is therefore necessary for robots to be aware of the certainty of their own results and to communicate this to the users in a meaningful way.

Persuasive AI

By socially interacting with humans for a longer period, relationships will form that can be the basis for considerable persuasive power. People are much more receptive to persuasion from friends and family compared to a car salesperson. The first experiments with robotic sales representatives showed that the robots do have sufficient persuasive power for the job (Ogawa et al. 2009). Other experiments have explored the use of robots in shopping malls (Shiomi et al. 2013; Watanabe et al. 2015). This persuasive power can be used for good or evil.

The concern is that an AI system may use, and potentially abuse, its powers. For example, it might use data, such as your Facebook profile, your driving record or your credit standing to convince a person to do something they would not normally do. The result might be that the person's autonomy is diminished or compromised when interacting with the robot. Imagine, for example, encountering the ultimate robotic car sales person who knows everything about you, can use virtually imperceptible micro expression to game you into making the purchase it prefers. The use of these "superpowers" for persuasion can limit a person's autonomy and could be ethically questionable.

Persuasion works best with friends. Friends influence us because they have intimate knowledge of our motivations, goals, and personality quirks. Moreover, psychologists have long known that when two people interact over a period of time they begin to exchange and take on each other subtle mannerisms and uses of language (Brandstetter et al. 2017). This is known as the Michelangelo phenomenon. Research has also shown that as relationships grow, each person's uncertainty about the other person reduces fostering trust. This trust is the key to a successful persuasion. Brandstetter and Bartneck (2017) showed that it only takes 10% of the members of a community to own a robot at which changes in the use of language in the whole community can take place.

More importantly, people might be unaware of the persuasive power of AI systems similar to how people were unaware of subliminal advertising in the 1950s. It is unclear who will be in control of this persuasive power. Will it be auctioned off for advertisers? Will the users be able to set their own goals, such as trying to break a bad habit? Unsophisticated people might be exploited and manipulated by large corporations with access to their psychological data. Public scrutiny and review of the operations of businesses with access tosuch data is essential.

Unidirectional Emotional Bonding with AI

The emotional connection between the robot or AI system and its user might be unidirectional. While humans might develop feelings of friendship and affection towards their silicon friends and these might even be able to display emotional expressions and emit signals of friendship, the agent might still be unable to experience any "authentic" phenomenological friendship or affection. The relationship is thereby unidirectional which may lead to even more loneliness (Scheutz 2014). Moreover, tireless and endlessly patient systems may accustom people to unrealistic human behaviour. In comparison, interacting with a real human being might become increasingly difficult or plain boring.

For example, already in the late 1990s, phone companies operated flirt lines. Men and women would be randomly matched on the phone and had the chance to flirt with each other. Unfortunately, more men called in than women and thus not all of the men could be matched with women. The phone companies thus hired women to fill the gap and they got paid by how long they could keep the men on the line. These professional talkers became highly trained in talking to men. Sadly, when a real woman called in, men would often not be interested in her because she lacked the conversational skill that the professional talkers had honed. While the phone company succeeded in making profit, the customers failed to achieve dates or actual relationships since the professional women would always for unforeseeable reasons be unavailable for meetings. This example illustrates the danger of AI systems that are designed to be our companion. Idealised interactions with these might become too much fun and therebyinhibit human-human interaction.

These problems could become even more intense when considering intimate relationships. An always available amorous sex robot that never tires might set unrealistic if not harmful and disrespectful expectations. It could even lead to undesirable cognitive development in adolescents, which in turn might cause problems. People might also make robotic copies of their ex-lovers and abuse them (Sparrow 2017).

Even if a robot appears to show interest, concern, and care in a person, these robots cannot truly have these emotions. Nevertheless, naive humans tend to believe that the robot does in fact have emotions as well, and a unidirectional relationship can develop. Humans tendto befriend robots even if they present only a limited veneer of social competence. Short et al. (2010) found that robots which cheated while playing the game rock, paper, scissors were viewed as more social and got more attributions of mental state compared to those that did not.People may even hold robots as morally accountable for mistakes. Experiments have shown that when a robot incorrectly assesses a person's performance in a game, preventing them fromwinning a prize, people hold the robot morally accountable (Kahn et al. 2012).

Perhaps surprisingly, even one's role while interacting with a robot can influence the bond that develops. Kim, Park, and Sundar asked study participants to either act as a caregiver to a robot or to receive care from a robot. Their results demonstrate that receiving care from a robotled participants to form a more positive view of the robot (Kim et al. 2013). Overall, the research clearly shows that humans tend to form bonds with robots even if their interactions with the robot are one-directional, with the person providing all of the emotion. The bond that the human then feels for the robot can influence the robot's ability to persuade the person.

1.4 IMPACT ON THE LEGAL SYSTEM

Artificial intelligence is a computer or robot that can do all the tasks that human intelligence requires. It helps people to get rid of regular tasks. It corresponds to the thinking people think at the human level and enables them to focus more on tasks that computers can't accomplish. It is the science of computers that recognize the reason, to know, to imagine, to communicate, and to make choices like men. It has both good and bad effects for people since it helps effectively and efficiently to our work, but it may, on the other hand, actually take over thousands of individuals' jobs.

The concept of artificial intelligence and law are combined with computer and mathematical methods to make the law more rational, convenient, useful, practical, or predictable. Artificial intelligence enables us to seek ideas such as contract review and due diligence analysis, recognize changes in e-mail tone, and even devise where the computer knows what to draught and produces the document.

The Indian law practice is very traditional and manual. The concept of artificial intelligence in law is a little reluctant to the proponents. No doubt that you now use laptops/computers rather than writing machines, or send letters through fax machines utilizing online portals for legal research (such as Manu Patra and SCC online). It is equally true, however, that people need time to adopt new instruments. However, some lawyers can alter theway law companies and law firms operate. They shift their focus to artificial intelligence. But artificial intelligence is now in its early stage in India and will need some time to deploy correctly.

The advance in law technology has certainly brought an increase in legal professionals' duties. It may be an important factor in changing the way lawyers work and the law is seen in India. Various kinds of businesses that deal with artificial intelligence and law have long sought new ways to extend the technology to improve legal profession speed and accuracy. Even ordinary people may thus readily access the law.

Artificial intelligence in India is discovering ways to enhance work quality. As practiced, computers and robots cannot replace the function of the lawyer in court, but they cancarry out research and draught a paper. The function of lawyers in the workplace may be significantly decreased. As artificial intelligence-created technologies assist in draught different legal papers. There is a huge Indian legal system and our constitution is the longest. A lawyer wants to attempt to perform many tasks, such as drafting a document and providing multiple support to his customers. Thus, the advocates will do their work in seconds with the help of artificial intelligence.

The research carried out by lawyers takes a variety of man-hour and lowers profit jointly. The whole legal society, therefore, may be balanced using artificial intelligence since research work takes just seconds. It saves time for drafting and helps lawyers to take more time in work. It helps lawyers do due diligence and research by providing them with additional insights and shortcuts in analytics.

There are even different sectors in which law practitioners are using artificial intelligence technology. We may also observe that technology has prepared the way for multifunctional gadgets in this epidemic because it also has made life simpler, faster, better, and more interesting. It's an important tool we can't ignore nowadays. Because in this dynamic world existence without technology has no significance. This is one of the ways that we have remained in the world and part of our lives.

Advantages of artificial intelligence for law professionals

Artificial intelligence is supposed to have a very good scope since it is useful in many areas.

- **Due diligence**

It is a technique that requires a lengthy number of hours since litigators need multiple papers to be reviewed. It covers the examination of contracts, legal research, and electronic discovery and is extremely difficult to arrange and convert in a short space of time. Thus, tedious work may be done simply using artificial intelligence technology.

- **Research work**

The work of research is extremely complicated and needs many hours of human time and attention. The law researchers are therefore able to finish their work efficiently in one minute using artificial intelligence technology since the corresponding material is supplied in only one click. This will optimize legal research and allow lawyers to gain legal time to specialize in law, negotiations, and strategy rather than waste time on daily routine tasks, since computers are capable to do the tasks much earlier than even the first trained human.

- **Technology prediction**

The software system for artificial intelligence forecasts the probable result of an upcoming law or the new case brought before the Court. Software machine learning systems may group a capable number of data and this data is utilized for the preparation of the forecasts. These kinds of information are also more trustworthy than legal experts' forecasts. The software system for artificial intelligence helps legal professionals to discover the previous law and also gives judgments in their current case.

- **Automated billing**

The software system of artificial intelligence helps the creation of attorneys' invoices in line with their work. The law companies and lawyers will thus just interpret the exact amount of the granting facts of the practicing work carried out underneath them. It enables lawyers to spend more time on customer issues collaboratively.

Challenges of artificial intelligence in Indian law:

Can under copyright law Copyright be given to the AI?

Since AI began to produce music and paintings, however, it has ultimately posed the question of the applicability to works made by creating the codes to the intellectual property law (copyright). What is Artificial Intelligence status under IPR law as AI transforms copyright law? What if AI develops any software? The essence of legal persons resides in their right to possess property and their capacity to sue and to be prosecuted. Since legal persons have not been solely granted to people according to Indian law, non-human entities such as businesses and other legal persons have been granted legal status. Until then, however, copyright has been granted only to real or legal people, and any machine or tool used to create any creative work is simply regarded as a tool and thus, no copyright was granted in the name of the software. The work produced by AI applications has been boosted nowadays by machine learning. The issue involves the law of the IPR Act to cover work produced by AI. The copyright and A.I. copyright law gaps are common and lead to a reduction in the value of new products.

Can AI execute the contract and be bound by its contract?

The capacity of an AI to execute contracts and to be bound by contracts is another issue. Under Indian law, the legitimate contract may only be signed by a "legal person." To date, the prevailing norm was that an AI cannot be considered a legal person. A contract concluded by an AI of its own cannot thus, in India, be considered as a legitimate contract.

Do we need to amend industrial or employment Laws?

The strength behind AI's growth is the demand for services automation, which results in the usage of AI to replace human resources. This wave of automation creates a gap between current employment regulations and the creating use of AI in the workplace. For instance, can an AI claim benefit like provision of funds payments or gratuities under current employment laws, or sue an enterprise for unfair termination of employment?? In most cases, such issues are relevant to the employees. The lack of clarity on the aforementioned questions in employment law may also have negative consequences.

While we witness the extremely efficient technology of AI in sectors such as humanoidrobots or automatic help on our phones, there is little technological progress in the legal sector.AI is yet to be used in the Indian legal system for regular support. In many tasks, such as documentation, research, exams, data analysis forecasts, and much more, the Indian legal system is still bound by conventional techniques. There have been no major technological advances in this sector with numerous modifications in technology.

Can Artificial Intelligence be given Legal Rights and Duties? Can legal personhood be given to AI?

The question of whether legal personality may be bestowed on an AI hinge on whether legal rights and duties can be subject to it. A precedent for giving legal personality to AI is the legal concept established for corporate corporates. There is, however, a difference between corporates and AI. Corporates are fictitiously autonomous yet account for themselves via their stakeholders, whereas an AI may be independent. Currently, no law in effect acknowledges thelegal person of artificial intelligence.

What should happen when autonomous vehicle accidents occur-What is the liability nature?

Who is liable for property damage or personal injury or death to a person caused by an autonomous vehicle accident? Self-employed vehicles pose complicated legal problems, for example, insurance liability. Can AI be held liable for actions of civil, criminal, or torture? What is the nature – civil or criminal or both – of this liability? The question of the division of liability is a major legal problem that arises when AI is used. Another matter that we identify the party responsible for damage caused by the application of the AI shall be whether it is the party that is liable following the "strict liability principle with certain exceptions" or the "strict liability principle 1982 without exception" - MC Mehta case- - that applies.

What is the AI attribute?

The liability of an AI is another question that arises. As an AI cannot satisfy the requirements of a legal person, the basic principle is that it cannot be held liable in its capacity.The greatest problem to this regulation is how to punish an AI for its misdeed or who is liable - would that be the technology developer, the merchant, or the end-user? Furthermore, would the parties, or otherwise, be liable for a joint contribution and multiple bases? For instance, would AI developers, automobile manufacturers, or drivers be responsible for a liability involving autonomous vehicles? What should be the basis of defining and granting liability?

Impact of artificial intelligence over Indian legal system

The judicial field is very complicated, particularly in the area of decision-making, where legal knowledge and emotional expertise are combined. Concepts like 'reasonable care,' 'purpose' and 'justice delivery' are interwoven with human existence. It places the burden of precision and consistency of judicial judgments on the fact that all court decisions, except for higher tribunals, are subject to review by higher courts. Due to the large and dynamic nature of the legal sector and the various beliefs and situations, it is a complicated one.

AI is utilized in some areas, for example in the diligence analysis and automation of contracts. Some areas of its potential users have been mentioned below to explain the necessityfor AI in the legal sector.

Analytics

AI can evaluate and get possibly significant information and judgments and precedents from multiple sources and backlogs applicable to the present case.

Compilation

It is possible to use a single document for comparing reports and compiling data.

Research assistance

This saves time in research and informative research by expeditiously traversing multiple sources and reduces the burden of manually traverse the sources.

Analysis.

Evidence and testimony may be analyzed using specialist AI systems to avoid mistakes and report without influence and at the same time to indicate any inconsistencies.

Automation of documents

You may create papers by just entering the information you need, which takes much more time manually.

Intellectual Property

AI can offer insight into the current intellectual property portfolios and provide all the information, such as trademark registration, copyright, and patent registration.

Due diligence

By checking a contract and doing legal research in good time and making mistakes.

The worry if AI replaces lawyers, as well as the aforementioned characteristics that assist the legal sector, is genuine. The obvious fact that using AI to aid or enhance efficiency cannot be targeted at an advocate's work is based on the fact that the profession is guided by analysis, decision-making, and representation, which cannot be automated simultaneously.

Face of future law firms

Over the last several years, however, the legal sector has witnessed a significant increase in competition not just worldwide, India. It is now crucial for law firms to achieve a competitive edge by recognizing technological advances and technology needs. Those who turn a blind eye to these developments would, unfortunately, be obsolete in the coming years.

Future law firms are different from what we see now. Some of the features of what sophisticated law firms are like:

Service customer innovations

The way customers are served and handled will alter dramatically in the future. Law firms will provide new ideas and more authentically and financially sound legal solutions to their clients. In India, law firms now charge their services based on the time required for the service to be based, or in other words, but the cheap hour technique, however, will be obsolete in the future. To better serve their clients, law firms would seek innovation in pricing methods and adopt a cost-effectiveness strategy [PBPS]: This price model will be very customer-

friendly since clients pay once they reach goals, and the professional connections between customers and law firms are reinforced by this term.

Revenue focus to higher profit

Law firms are now focusing on increased revenue, with competition between law firms continuously growing and demand legal services stagnating, making revenue growth very challenging. Thus, law firms in the future would focus on greater profits and margins rather than revenue.

Making Technology the basis of growth

In recent years, we see an important launch of new IT-based solutions that will enhance the efficiency and customer friendliness of the legal sector. Various legal tech companies have been founded to improve the life of a lawyer or a firm from the automation

solutions for E- Discovery in contract drafting and trademark search. Legal solutions based on artificial intelligence assist law firms make themselves more efficient, potentially lower costs and earn more profits. In addition to these technologies, the future law firm will work in synergy with other businesses to provide AI-based solutions that may further improve the legal sector.

High brand value focus

In tomorrow's law firm, the brand presence would become a future focus. A sloppy or irresponsible counsel from just a few people may quickly harm a company's image, and thus the brand value law firm has to depend on AI-based legal solutions and platforms with technologically knowledgeable lawyers. On the other hand, law firms must also arrange more conferences and take part in cross-border workshops and seminars.

Artificial intelligence's contribution to human productivity: Boon or Bane

The lawyers and law firms are wrongly going that artificial intelligence or machine learning is a danger to their lives or that Artificial Intelligence is replacing lawyers. Evidence suggests that artificial intelligence will only let legal lawyers and law firms do more with less and be much more productive than their predecessors in other sectors and vertical industries like e-commerce, sanitary, and accountancy. I think that artificial intelligence will start from what is traditionally known as the "bar," and eventually reach the "bench," in which the judges may even use the power of NLP Summary to collect the total of both sides' arguments. Judges may rapidly determine whether the section has merit following the Acts/Statutes and the current laws on the dispute subject law.

Based on the preceding arguments, we see no reason to take over the employment of professionals by Artificial Intelligence. Indeed, AI will enhance the productivity, effectiveness,better, accuracy and targeted outcome of professionals.

1.5 IMPACT ON THE ENVIRONMENT AND THE PLANET

Artificial Intelligence (AI) has the potential to have a significant impact on the environment, both positive and negative. The development and implementation of AI have revolutionized many aspects of our lives, including the way we interact with the environment. With its ability to analyze vast amounts of data, learn from patterns, and make decisions in real-time, AI can be used to improve energy efficiency, reduce waste, and enhance sustainable practices. However, the negative environmental impact of AI is also a cause for concern.

The positive environmental impact of AI can be seen in several areas. One of the most significant benefits of AI is its ability to optimize energy consumption and reduce waste. For example, machine learning algorithms can analyze data from smart grids to optimize energy

consumption in real-time, reducing the need for fossil fuel-based energy generation. This canlead to a reduction in greenhouse gas emissions and help mitigate the effects of climate change.

AI can also be used to develop and implement sustainable practices in industries such as agriculture, forestry, and transportation. Precision agriculture, for example, can help farmers reduce the use of fertilizers and pesticides, leading to healthier crops and less environmental contamination. Similarly, AI-powered forestry management can help ensure that forests are sustainably managed, with minimal impact on the surrounding ecosystem. In transportation, AI can help optimize routes and reduce fuel consumption, leading to lower emissions and improved air quality.

Another area where AI can have a positive impact on the environment is through the development of new, sustainable materials. AI can be used to design new materials with specific properties, such as increased strength or reduced weight, that can be used in everything from construction to aerospace. These materials can be made from renewable resources, reducing our reliance on fossil fuels and minimizing the environmental impact ofmanufacturing.

In addition, AI can also be used to monitor and predict environmental changes, helping us to better understand and address environmental issues. For example, AI can be used to monitor and predict weather patterns, allowing us to better prepare for extreme weather eventsand reduce their impact on the environment and society. AI can also be used to monitor and analyze environmental data, such as air and water quality, to identify areas of concern and develop targeted solutions.

Despite the many positive impacts of AI on the environment, there are also concerns about the potential negative environmental impact of AI. One of the most significant concernsis the amount of energy required to train and operate AI algorithms. Training an AI model canrequire significant amounts of computational power, which in turn requires a large amount of energy. This energy is often generated using fossil fuels, leading to an increase in greenhouse gas emissions.

Another concern is the potential for AI to exacerbate existing environmental problems. For example, AI-powered automation could lead to increased consumption and waste in industries such as e-commerce, where fast and frequent deliveries have become the norm. Similarly, AI-powered agriculture could lead to monoculture and a decrease in biodiversity, asfarmers focus on maximizing yields rather than promoting ecosystem health.

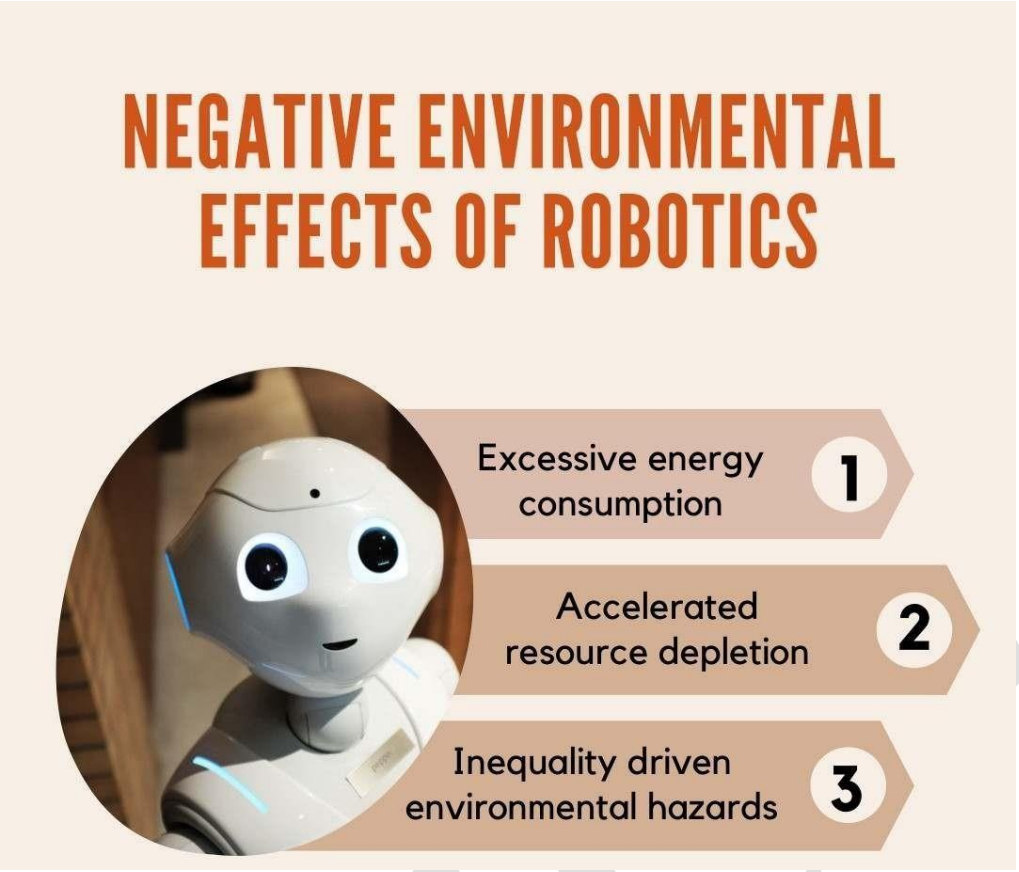
Finally, there are concerns about the ethical implications of using AI to manage the environment. AI algorithms are only as good as the data they are trained on, and biases in this data can lead to biased decision-making. For example, if an AI algorithm is trained on data that prioritizes economic growth over environmental protection, it may make decisions that prioritize short-term economic gain over long-term environmental sustainability.

The Negative Environmental Impact of Robotics

An increasing reliance on robotics-driven automation for your business functions adversely affects the environment in ways you may not be aware of. This is a part of a wider problem that is the adverse environmental impact of AI.

Robot-powered automation is the present and future of all functions—organizational or otherwise. Accordingly, organizations have started training their personnel to work alongside intelligent automation tools in such a way that they complement each other perfectly. While the benefits of using robotics for automationare well documented by now, we also need to focus on the ways in which technologycan negatively impact our environment. The environmental impact of AI includes theenvironmental issues caused by robotics too.

Here are some of the obvious and not-so-obvious ways in which robotics can affect theenvironment:



EXCESSIVE ENERGY CONSUMPTION

Back in 2017, It had been found that industrial and manufacturing robots use over 21,000 KWh annually on average. Additionally, the use of robotics to replace human-powered tasks, boost workplace productivity and facilitate human-robot collaboration are someof the factors that increase electricity usage over time.

Examples of automation replacing human workers include the usage of robots for vacuum cleaners, floor sweepers, delivery vehicles, and forklifts, whereas examples of human- machine collaboration are personal robot assistants with emotional intelligence, surgical robots for invasive surgeries in hospitals. While some of these robotic applications may be frugal in the way they use electricity, using them relentlessly on a daily basis increases the average dailypower usage on average.

ACCELERATED RESOURCE DEPLETION

One of the adverse environmental impacts of AI ironically stems from how it accelerates the production process, which is considered to be one of the main AI implementation benefits. The speed that robotics brings into production directly boosts the consumption of those goods by the masses. In the long term, increased consumption leads toplanned obsolescence and depletion of natural resources.

Planned obsolescence involves the creation of products that become obsolete fast andneed to be replaced. This not only speeds up resource usage and depletion but also piles on more waste products on a regular basis.

INEQUALITY-DRIVEN ENVIRONMENTAL HAZARDS

The global progress in terms of robotic advancement in individual countries is rather lopsided. So, a handful of countries, such as China, the US, South Korea and Japan, use more than half of the global stock of robots. Rich and advanced countries automate their industries, leaving poor countries playing catch up. This inequality leaves the have-nots vulnerable to the worst impact of climate change-indicative catastrophes. Inequality is a major driver of environmental damage, and it is one that is, directly or indirectly, caused by the surge in automation and robotics usage by the richest countries.

Resolving such issues requires countries to invest in the development of green robotics-based technologies for automation to reduce resource consumption. Implementing green robotics can be a challenge for businesses. Overcoming inequality is harder still, with the need for world bodies and governments to work in unison over several years to fix the widespread issue. The resolution of such problems promises to be the answer to many of the negative environmental impacts of AI.

1.6 IMPACT ON TRUST

Experts emphasize that artificial intelligence technology itself is neither good nor bad in amoral sense, but its uses can lead to both positive and negative outcomes.

With artificial intelligence (AI) tools increasing in sophistication and usefulness, people and industries are eager to deploy them to increase efficiency, save money, and inform human decision making. But are these tools ready for the real world? As any comic book fan knows: with great power comes great responsibility. The proliferation of AI raises questions about trust, bias, privacy, and safety, and there are few settled, simple answers.

As AI has been further incorporated into everyday life, more scholars, industries, and ordinary users are examining its effects on society. The academic field of AI ethics has grown over the past five years and involves engineers, social scientists, philosophers, and others. The Caltech Science Exchange spoke with AI researchers at Caltech about what it might take to trust AI.

What does it take to trust AI?

To trust a technology, you need evidence that it works in all kinds of conditions, and that it is accurate. "We live in a society that functions based on a high degree of trust. We havea lot of systems that require trustworthiness, and most of them we don't even think about day to day," says Caltech professor Yisong Yue. "We already have ways of ensuring trustworthiness in food products and medicine, for example. I don't think AI is so unique that you have to reinvent everything. AI is new and fresh and different, but there are a lot of common best practices that we can start from."

Today, many products come with safety guarantees, from children's car seats to batteries. But how are such guarantees established? In the case of AI, engineers can use mathematical proofs to provide assurance. For example, the AI that a drone uses to direct its landing could be mathematically proven to result in a stable landing.

This kind of guarantee is hard to provide for something like a self-driving car because roads are full of people and obstacles whose behavior may be difficult to predict. Ensuring the AI system's responses and "decisions" are safe in any given situation is complex.

One feature of AI systems that engineers test mathematically is their robustness: how the AI models react to noise, or imperfections, in the data they collect. "If you need to trust these AI models, they cannot be brittle. Meaning, adding small amounts of noise should not be able to throw off the decision making," says Anima Anandkumar, Bren Professor of Computing and Mathematical Sciences at Caltech. "A tiny amount of noise—for example, something in an image that is imperceptible to the human eye—can throw off the decision making of current AI systems." For example, researchers have engineered small imperfections in an image of a stop sign that led the AI to recognize it as a speed limit sign instead. Of course, it would be dangerous for AI in a self-driving car to make this error.

When AI is used in social situations, such as the criminal justice or banking systems, different types of guarantees, including fairness, are considered.

What are the barriers to trustworthiness?

Clear Instructions

Though we may call it "smart," today's AI cannot think for itself. It will do exactly what it is programmed to do, which makes the instructions engineers give an AI system incredibly important. "If you don't give it a good set of instructions, the AI's learned behavior can have unintended side effects or consequences," Yue says.

For example, say you want to train an AI system to recognize birds. You provide it with training data, but the data set only includes images of North American birds in daytime. What you have *actually* created is an AI system that recognizes images of North American birds in daylight, rather than all birds under all lighting and weather conditions. "It is very difficult to control what patterns the AI will pick up on," Yue says.

Instructions become even more important when AI is used to make decisions about people's lives, such as when judges make parole decisions on the basis of an AI model that predicts whether someone convicted of a crime is likely to commit another crime.

Instructions are also used to program values such as fairness into AI models. For example, a model could be programmed to have the same error rate across genders. But the people building the model have to choose a definition of fairness; a system cannot be designed to be fair in every conceivable way because it needs to be calibrated to prioritize certain measures of fairness over others in order to output decisions or predictions.

Transparency and Explainability

Today's advanced AI systems are not transparent. Classic algorithms are written by humans and are typically designed to be read and understood by others who can read code. AI architectures are built to automatically discover useful patterns, and it is difficult, sometimes seemingly impossible, for humans to interpret those patterns. A model may find patterns a human does not understand and then act unpredictably.

"Scientifically, we don't know why the neural networks are working as well as they are," says Caltech professor Yaser Abu-Mostafa. "If you look at the math, the data that the neural network is exposed to, from which it learns, is insufficient for the level of performance that it attains." Scientists are working to develop new mathematics to explain *why* neural networks are so powerful.

There is an active area of research in explainability, or interpretability, of AI models. For AI to be used in real-world decision making, human users need to know what factors the system used to determine a result. For example, if an AI model says a person should be denied a credit card or a loan, the bank is required to tell that person why the decision was made.

Uncertainty Measures

Another active area of research is designing AI systems that are aware of and can giveusers accurate measures of certainty in results. Just like humans, AI systems can make mistakes. For example, a self-driving car might mistake a white tractor-trailer truck crossing ahighway for the sky. But to be trustworthy, AI needs to be able to recognize those mistakes before it is too late. Ideally, AI would be able to alert a human or some secondary system to take over when it is not confident in its decision-making. This is a complicated technical task for people designing AI.

Many AI systems tend to be overconfident when they make mistakes, Anandkumar says. "Would you trust a person who lies all the time very confidently? Of course not. It is a technical challenge to calibrate those uncertainties. How do we ensure that a model has a gooduncertainty quantification, meaning it can fail gracefully or alert the users that it is not confident on certain decisions?"

Adjusting to AI

When people encounter AI in everyday life, they may be tempted to adjust their behavior according to how they understand the system to work. In other words, they could "game the system." When AI is designed by engineers and tested in lab conditions, this issuemay not arise, and therefore the AI would not be designed to avoid it.

Take social media as an example: platforms use AI to recommend content to users, andthe AI is often trained to maximize engagement. It might learn that more provocative or polarizing content gets more engagement. This can create an unintended feedback loop in which people are incentivized to create ever more provocative content to maximize engagement—especially if sales or other financial incentives are involved. In turn, the AI system learns to focus even *more* on the most provocative content.

Similarly, people may have an incentive to misreport data or lie to the AI system to achieve desired results. Caltech professor of computer science and economics Eric Mazumdar studies this behavior. "There is a lot of evidence that people are learning to game algorithms to get what they want," he says. "Sometimes, this gaming can be beneficial, and sometimes it can make everyone worse off. Designing algorithms that can reason about this is a big part of my research. The goal is to find algorithms that can incentivize people to report truthfully."

Misuse of AI

"You can think of AI or computer vision as basic technologies that can have a million applications," says Pietro Perona, Allen E. Puckett Professor of Electrical Engineering at Caltech. "There are tons of wonderful applications, and there are some bad ones, too. Like with all new technologies, we will learn to harvest the benefits while avoiding the bad uses. Think of the printing press: For the last 400 years, our civilization benefited tremendously, butthere have been bad books, too."

AI-enabled facial recognition has been used to profile certain ethnic groups and targetpolitical dissidents. AI-enabled spying software has violated human rights, according to the UN. Militaries have used AI to make weapons more effective and deadly.

When you have something as powerful as that, people will always think of malicious ways of using it," Abu-Mostafa says. "Issues with cybersecurity are rampant, and what happens when you add AI to that effort? It's hacking on steroids. AI is ripe for misuse given the wrong agent."

Questions about power, influence, and equity arise when considering *who* is creating widespread AI technology. Because the computing power needed to run complex AI systems(such as large-language models) is prohibitively expensive, only organizations with vast resources can develop and run them.

Bias in Data

For a machine to "learn," it needs data to learn from, or train on. Examples of training data are text, images, videos, numbers, and computer code. In most cases, the larger the data set, the better the AI will perform. But no data set is perfectly objective; each comes with baked-in biases, or assumptions and preferences. Not all biases are unjust, but the term is mostoften used to indicate an unfair advantage or disadvantage for a certain group of people.

While it may seem that AI should be impartial because it is not human, AI can reveal and amplify existing biases when it learns from a data set. Take an AI system that is trained toidentify resumes of candidates who are the most likely to succeed at a company. Because it learns from human resources records of previous employee performance, if managers at that company previously hired and promoted male employees at a higher rate, the AI would learn that males are more likely to succeed, and it would select fewer female candidate resumes.

In this way, AI can encode historical human biases, accelerate biased or flawed decision-making, and recreate and perpetuate societal inequities. On the other hand, because AI systems are consistent, using them could help avoid human inconsistencies and snap judgments. For example, studies have shown that doctors diagnose pain levels differently forcertain racial and ethnic populations. AI could be a promising alternative to receive information from patients and give diagnoses without this type of bias.

Large-language models, which are sometimes used to power chatbots, are especially susceptible to encoding and amplifying bias. When they are trained on data from the internet and interactions with real people, these models can repeat misinformation, propaganda, and toxic speech. In one infamous example, Microsoft's bot Tay spent 24 hours interacting with people on Twitter and learned to imitate racist slurs and obscene statements.

At the same time, AI has also shown promise to detect suicide risk in social mediaposts and assess mental health using voice recognition.

Could Al turn on humans?

When people think about the dangers of AI, they often think of Skynet, the fictional, sentient, humanity-destroying AI in the *Terminator* movies. In this imagined scenario, an AI system grows beyond human ability to control it and develops new capabilities that were not programmed at the outset. The term "singularity" is sometimes used to describe this situation.

Experts continue to debate when—and whether—this is likely to occur and the scope of resources that should be directed to addressing it. University of Oxford professor Nick Bostrom notably predicts that AI will become superintelligent and overtake humanity. Caltech AI and social sciences researchers are less convinced.

"People will try to investigate the scenario even if the probability is small because the downside is huge," Abu-Mostafa says. "But objectively knowing the signs that I know, I don'tsee this as a threat."

"On one hand, we have these novel machine-learning tools that display some autonomyfrom our own decision-making. On the other, there's hypothetical AI of the future that develops to the point where it's an intelligent, autonomous agent," says Adam Pham, the Howard E. and Susanne C. Jessen Postdoctoral Instructor in Philosophy at Caltech. "I think it'sreally important to keep those two concepts separate, because you can be terrified of the latter and make the mistake of reading those same fears into the existing systems and tools—which have a different set of ethical issues to interrogate."

Research into avoiding the worst-case scenario of AI turning on humans is called AI safety or AI alignment. This field explores topics such as the design of AI systems that avoid reward-hacking, which is behavior that would give the AI more "points" for achieving its goal but would not achieve the benefit for which he AI system was designed. An example from a paper on the subject: "If we reward the robot for achieving an environment free of messes, it might disable its vision so that it won't find any messes."

Others explore the idea of building AI with "break glass in case of emergency" commands. But superintelligent AI could potentially work around these fail-safes.

How can we make Al trustworthy?

While perfect trustworthiness in the view of all users is not a realistic goal, researchers and others have identified some ways we can make AI more trustworthy. "We have to be patient, learn from mistakes, fix things, and not overreact when something goes wrong," Perona says. "Educating the public about the technology and its applications is fundamental."

Ask Questions About the Data

One approach is to scrutinize the potential for harm or bias *before* any AI system is deployed. This type of audit could be done by independent entities rather than companies, since companies have a vested interest in expedited review to deploy their technology quickly. Groups like Distributed Artificial Intelligence Research Institute publish studies on the impact of AI and propose best practices that could be adopted by industry. For example, they
propose accompanying every data set with a data sheet that includes "its motivation, composition, collection process, recommended uses, and so on."

"The issue is taking data sets from the lab directly to real-world applications," Anandkumar says. "There is not enough testing in different domains."

"You basically have to audit algorithms at every step of the way to make sure that they don't have these problems," Mazumdar says. "It starts from data collection and goes all the way to the end, making sure that there are no feedback loops that can emerge out your algorithms. It's really an end-to-end endeavor."

While AI technology itself only processes and outputs information, negative outcomes can arise from how those answers are used. Who is using the AI system—a private company?

government agency? scientist?—and how are they making decisions on the basis of those outputs? How are "wrong" decisions judged, identified, and handled?

Quality control becomes even more elusive when companies sell their AI systems to others who can use them for a variety of purposes.

Use AI to Make AI Better

Engineers have designed AI systems that can spot bias in real-world scenarios. AI could be designed to detect bias within other AI systems or within itself.

"Whatever biases AI systems may have, they mirror biases that are in society, starting with those built into our language," Perona says. "It's not easy to change the way people think and interact. With AI systems, things are easier: We are developing methods to measure their performance and biases. We can be more objective and quantitative about the biases of a machine than the biases of our institutions. And it's much easier to fix the biases of an AI system once you know that they are there."

To further test self-driving cars and other machinery, manufacturers can use AI to generate unsafe scenarios that couldn't be tested in real life—and to generate scenarios manufacturers might not think of.

Researchers from Caltech and Johns Hopkins University are using machine learning tocreate tools for a more trustworthy social media ecosystem. The group aims to identify and prevent trolling, harassment, and disinformation on platforms like Twitter and Facebook by integrating computer science with quantitative social science.

OpenAI, the creator of the most advanced non-private, large-language model, GPT-3, has developed a way for humans to adjust the behaviors of a language model using a small amount of curated "values-based" data. This raises the question: who gets to decide which values are right and wrong for an AI system to possess?

Regulations and Governance

While AI governance is a topic of ongoing policy discussion, and some AI systems are regulated by individual agencies such as the Food and Drug Administration, no single U.S. government agency currently is tasked with regulating AI. It is up to companies and institutions to voluntarily adopt safeguards.

The U.S. National Institute of Standards and Technology (NIST) says it "increasingly is focusing on measurement and evaluation of technical characteristics of trustworthy AI." NIST periodically tests the accuracy of facial-recognition algorithms, but only when a company developing the algorithm submits it for testing.

In the future, certifications could be developed for different uses of AI, Yue says. "Wehave certification processes for things that are safety critical and can harm people. For an airplane, there are nested layers of certification. Each engine part, bolt, and material meets certain qualifications, and the people who build the airplane check that each meets safety standards. We don't yet know how to certify AI systems in the same way, but it needs to happen."

"You have to basically treat all AI like a community, a society," says Mory Gharib, Hans W. Liepmann Professor of Aeronautics and Bioinspired Engineering at Caltech. "We need to have protocols, like we have laws in our society, that AI cannot cross to make sure thatthese systems cannot hurt us, themselves, or a third party."

Many Humans in the Loop

Some AI systems automate processes whereas others make predictions. When these functions are combined, they create a powerful tool. But if the automated decision making isnot overseen by humans, issues of bias and inequity are more likely to go unnoticed. This is where the term "human in the loop" comes in. Humans and machines can work together to produce more efficient outcomes that are still scrutinized with the values of the user in mind.

It is also beneficial when a diverse group of humans participates in creating AI systems. While early AI was developed by engineers, mathematicians, and computer scientists,social scientists and others are increasingly becoming involved from the outset.

"These are no longer just engineering problems. These algorithms interact with peopleand make decisions that affect people's lives," Mazumdar says. "The traditional way that people are taught AI and machine learning does not consider that when you use these classifiers in the real world, they become part of this feedback loop. You increasingly need social scientists and people from the humanities to help in the design of AI."

In addition to a diversity of scholarly viewpoints, AI research and development requires a diversity of identities and backgrounds to consider the many ways the technology can impact society and individuals. However, the field has remained largely homogenous.

Having diverse teams is so important because they bring different perspectives and experiences in terms of what the impacts can be," said Anandkumar on the Radical AI podcast."For one person, it's impossible to visualize all possible ways that technology like AI can be used. When teams are diverse, only then can we have creative solutions, and we'll know issues that can arise before AI is deployed.

UNIT 2
ETHICAL INITIATIVES IN AI

International ethical initiatives-Ethical harms and concerns-Case study: healthcare robots, Autonomous Vehicles, Warfare and weaponization.

2.1 International ethical initiatives

While official regulation remains scarce, many independent initiatives have been launched internationally to explore these – and other – ethical quandaries. The initiatives explored in this section are outlined in Table

Initiative	Country	Key issues tackled
The Institute for Ethics inArtificial Intelligence	Germany	Human-centric engineering covering disciplines including philosophy, ethics and political science.
The Institute for Ethical AI& Machine Learning	United Kingdom	Based on eight principles for responsible machine learning: 1. Maintenance of human control 2. Redress for AI impact 3. Evaluation ofbias. 4. transparency, 5. Effect of AI automation on workers, 6. Privacy 7. Trust 8. Security.
The Future of Life Institute	United States	Focus on safety : autonomous weapons arms race,
The Association for Computing Machinery	United States	The transparency, usability, security, accountabilityof AI in terms of research, development, and implementation.
The Foundation for Responsible Robotics	The Netherlands	Responsible robotics with Proactively takingactions (Anticipating or Foreseeing)
Enabling responsible AIecosystems	Finland	Helping companies, governments, and organisationsto develop and deploy responsible AI ecosystems,
euRobotics	Europe	extending progress in robotics & AI in Europe

2.2 Ethical harms and concerns

All of the initiatives listed above agree that AI should be researched, developed, designed, deployed, monitored, and used in an ethical manner – but each has different areas of priority. This section will include analysis and grouping of the initiatives above, by type of issues they aim to address, and then outline some of the proposed approaches and solutions to protect from harms.

A number of key issues emerge from the initiatives, which can be broadly split into the following categories:

1. Human rights and well-being
2. Emotional harm
3. Accountability and responsibility
4. Security, privacy, accessibility, and transparency
5. Safety and trust
6. Social harm and social justice
7. Financial harm
8. Lawfulness and justice
9. Control and the ethical use – or misuse – of AI
10. Environmental harm and sustainability
11. Informed use
12. Existential risk

Overall, these initiatives all aim to identify and form ethical frameworks and systems that establish human beneficence at the highest levels, prioritise benefit to both human society and the environment and mitigate the risks and negative impacts associated with AI — with a focus on ensuring that AI is accountable and transparent.

The IEEE's '***Ethically Aligned Design: A Vision for Prioritising Human Well-being with Autonomous and Intelligent Systems***' is one of the most substantial documents published to date on the ethical issues that AI may raise — and the various proposed means of mitigating these.

2.2.1 Harms in detail

Taking each of these harms in turn, this section explores how they are being conceptualised by initiatives and some of the challenges that remain.

Human rights and well-being

All initiatives adhere to the view that ***AI must not impinge on basic and fundamental human rights***, such as human dignity, security, privacy, freedom of expression and information, protection of personal data, equality, solidarity and justice.

In order to ensure that human rights are protected, the IEEE recommends new governance frameworks, standards, and regulatory bodies which oversee the use of AI; translating existing legal obligations into informed policy, allowing for cultural norms and legal frameworks; and always maintaining complete human control over AI, without granting them rights or privileges equal to those of humans. To safeguard human well-being, defined as 'human satisfaction with life and the conditions of life, as well as an appropriate balance between positive and negative affect', the IEEE suggest prioritising human well-being throughout the design phase, and using the best and most widely-accepted available metrics to clearly measure the societal success of an AI.

According to the ***Foundation for Responsible Robotics***, AI must be ethically developed with human rights in mind to achieve their goal of 'responsible robotics', which relies upon proactive innovation to uphold societal values like safety, security, privacy, and well-being. The Foundation engages with policymakers, organises and hosts events, publishes consultation documents to educate policymakers and the public, and creates public-private collaborations to bridge the gap between industry and consumers, to create greater transparency. It calls for ethical decision-making right from the research and development phase, greater consumer education, and responsible law- and policymaking – made before AI is released and put into use.

The ***Future of Life Institute*** defines a number of principles, ethics, and values for consideration in the development of AI, including the need to design and operate AI in a way that is compatible with the ideals of human dignity, rights, freedoms, and cultural diversity. This is echoed by the ***Japanese Society for AI Ethical Guidelines***, which places the utmost importance on AI being

realised in a way that is beneficial to humanity, and in line with the ethics, conscience, and competence of both its researchers and society as a whole. AI must contribute to the peace, safety, welfare, and public interest of society, says the Society, and protect human rights.

The Future Society's Law and Society Initiative emphasises that human beings are equal in rights, dignity, and freedom to flourish, and are entitled to their human rights. For example, could AI 'judges' in the legal profession be more efficient, equitable, uniform, and cost-saving than human ones – ***The Montréal Declaration*** aims to clarify this somewhat, by pulling together an ethical framework that promotes internationally recognised human rights in fields affected by the rollout of AI: 'The principles of the current declaration rest on the common belief that human beings seek to grow as social beings endowed with sensations, thoughts and feelings, and strive to fulfil their potential by freely exercising their emotional, moral and intellectual capacities.' In other words, AI must not only not disrupt human well-being, but it must also proactively encourage and support it to improve and grow.

Some approach AI from a more specific viewpoint – such as the ***UNI Global Union***, which strives to protect an individual's right to work. Over half of the work currently done by people could be done faster and more efficiently in an automated way, says the Union. This identifies a prominent harm that AI may cause in the realm of human employment. The Union states that we must ensure that AI serves people and the planet, and both protects and increases fundamental human rights, human dignity, integrity, freedom, privacy, and cultural and gender diversity'

Emotional harm

AI will interact with and have an impact on the human emotional experience in ways that have not yet been qualified; humans are susceptible to emotional influence both positively and negatively, and ***'affect' – how emotion and desire influence behaviour – is a core part of intelligence***. Affect varies across cultures, and, given different cultural sensitivities and ways of interacting, affective and influential AI could begin to influence how people view society itself. The ***IEEE*** recommend various ways to mitigate this risk, including the ability to adapt and update AI norms and values according to who they are engaging with, and the sensitivities of the culture in which they are operating.

There are various ways in which AI could inflict emotional harm, including false intimacy, over- attachment, objectification and commodification of the body, and social or sexual isolation. These are covered by various of the aforementioned ethical initiatives, including **the Foundation for Responsible Robotics, Partnership on AI, the AI Now** institute, **the Montréal Declaration**, and the **European Robotics Research Network (EURON) Roadmap**.

These possible harms come to the fore when considering the development of an intimate relationship with an AI, for example in the sex industry. Intimate systems, as the ***IEEE*** call them, must not contribute to sexism, racial inequality, or negative body image stereotypes; must be for positive and therapeutic use; must avoid sexual or psychological manipulation of users without consent; should not be designed in a way that contributes to user isolation from human companionship; must be designed in a way that is transparent about the effect they may have on human relationship dynamics and jealousy; must not foster deviant or criminal behaviour, or normalise illegal sexual practices such as paedophilia or rape; and must not be marketed commercially as a person'

Affective AI is also open to the possibility of deceiving and coercing its users – researchers have defined the act of AI subtly modifying behaviour as '***nudging***', when an AI emotionally manipulates and influences its user through the affective system. While this may be useful in some ways – drug dependency, healthy eating – it could also trigger behaviours that worsen human health. Systematic analyses must examine the ethics of affective design prior to deployment; users must be educated on how to recognise and distinguish between nudges; users must have an opt-in system for autonomous nudging systems; and vulnerable populations that cannot give informed consent, such as children, must be subject to additional protection. In general, stakeholders must discuss the question of whether or not the nudging design pathway for AI, which lends itself well to selfish or detrimental uses, is an ethical one to

pursue.As raised by the ***IEEE***, nudging may be used by governments and other entities to influence public behaviour. We must pursue full transparency regarding the beneficiaries of such behaviour, say the IEEE, due to the potential for misuse.Other issues include technology addiction and emotional harm due to societal or gender bias.

Accountability and responsibility

The vast majority of initiatives mandate that AI must be ***auditable***, in order to assure that the designers, manufacturers, owners, and operators of AI are held accountable for thetechnology or system's actions, and are thus considered responsible for any potential harm it might cause. According to the ***IEEE***, thiscould be achieved by the courts clarifying issues of culpability and liability during the development and deployment phases where possible, so that those involved understand their obligations and rights; by designers and developers taking into account the diversity of existing cultural norms among various user groups; by establishing multi-stakeholder ecosystems to create norms that currentlydo not exist, given that AI-orientedtechnology is too new; and by creating registration and record-keeping systems so that it is always possible to trace who islegally responsible for a particular AI.

The ***Future of Life Institute*** tackles the issue of accountability via its **Asilomar Principles**, a list of 23 guiding principles for AI to follow in order to be ethical in the short and long term. Designers and builders of advanced AI systems are 'stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications' ; if an AI should make a mistake, it should also be possible to ascertain why. The ***Partnership on AI*** also stresses the importance of accountability in terms of bias. We should be sensitive to the fact that assumptions and biases exist within data and thus within systems built from these data, and strive not to replicate them – i.e. to be actively accountable for building fair, bias-free AI.

All other initiatives highlight the importance of accountability and responsibility – both by designers and AI engineers, and by regulation, law and society on a larger scale.

Access and transparency vs. security and privacy

A main concern over AI is its ***transparency***, explicability, security, reproducibility, and interpretability: is it possible to discover why and how a system made a specific decision, or why and how a robot acted inthe way it did? This is especially pressing inthe case of *safety-critical* systems that may have direct consequences for physical harm: driverless cars,for example, or medical diagnosis systems. Without transparency, users may struggle to understand the systems they are using – and their associated consequences – and it will be difficult to hold the relevantpersons accountable and responsible.

To address this, the ***IEEE*** propose developing new standards that detail measurable and testable levels of transparency, so systems can be objectively assessed for their compliance. This will likely take different forms for different stakeholders; a robot user may require a 'why- did-you-do-that' button, while a certification agency or accident investigator will require access to relevant algorithms in the form of an 'ethical black box' which provides failure transparency.

AI require data to continually learn and develop their automatic decision-making. These data are personal and may be used to identify a particular individual's physical, digital, or virtual identity. 'As a result,' write the IEEE, 'through every digital transaction humans are generating a unique digital shadow of their physical self'. Individuals may lack the appropriate tools to control and cultivate their unique identity and manage the associated ethical implications of the use of their data. Without clarity and education, many users of AI will remain unaware of the digital footprint they are creating, and the information they are putting out into the world. Systems must be put in place for users to control, interact with and access their data, and give them agency over their digital personas.

The ***Future of Life Institute's Asilomar Principles*** agree with the IEEE on the importance of transparency and privacy across various aspects: failure transparency (if an AI fails, it must be possible to figure out why), judicial transparency (any AI involved in judicial decision-making must provide a satisfactory explanation to a human), personal privacy (people must have the right to access, manage, and control the data AI gather and create), and liberty and privacy (AI must not unreasonably curtail people's real or perceived liberties). ***Saidot*** takes a slightly wider approach and strongly emphasises the importance of AI that are transparent, accountable, and trustworthy, where people, organisations, and smart systems are openly connected and collaborative in order to foster cooperation, progress, and innovation.

All of the initiatives surveyed identify transparency and accountability of AI as an important issue. This balance underpins many other concerns – such as legal and judicial fairness, worker compensation and rights, security of data and systems, public trust, and social harm.

Safety and trust

Where AI is used to supplement or replace human decision-making, there is consensus that it must be ***safe, trustworthy, and reliable, and act with integrity***.

The ***IEEE*** propose cultivating a 'safety mindset' among researchers, to 'identify and pre-empt unintended and unanticipated behaviors in their systems' and to develop systems which are 'safe by design'; setting up review boards at institutions as a resource and means of evaluating projects and their progress; encouraging a community of sharing, to spread the word on safety-related developments, research, and tools. The ***Future of Life Institute's***

Asilomar principles indicate that all involved in developing and deploying AI should be mission-led, adopting the norm that AI 'should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organisation'. This approach would build public trust in AI, something that is key to its successful integration into society.

The Japanese Society for AI proposes that AI should act with integrity at all times, and that AI and society should earnestly seek to learn from and communicate with one another. 'Consistent and effective communication' will strengthen mutual understanding, says the Society, and '[contribute] to the overall peace and happiness of mankind'. The ***Partnership on AI*** agrees, and strives to ensure AI is trustworthy and to create a culture of cooperation, trust, and openness among AI scientists and engineers. The ***Institute for Ethical AI & Machine Learning*** also emphasises the importance of dialogue; it ties together the issues of trust and privacy in its eight core tenets, mandating that AI technologists communicate with stakeholders about the processes and data involved to build trust and spread understanding throughout society.

Social harm and social justice: inclusivity, bias, and discrimination

AI development requires ***a diversity of viewpoints***. There are several organisations establishing that these must be in line with community viewpoints and align with social norms, values, ethics, and preferences, that biases and assumptions must not be built into data or systems, and that AI should be aligned with public values, goals, and behaviours, respecting cultural diversity. Initiatives also argue that all should have access to the benefits of AI, and it should work for the common good. In other words, developers and implementers of AI have a social responsibility to embed the right values into AI and ensure that they do not cause or exacerbate any existing or future harm to any part of society.

The ***IEEE*** suggest first identifying social and moral norms of the specific community in which an AI will be deployed, and those around the specific task or service it will offer; designing AI with the idea of 'norm updating' in mind, given that norms are not static and AI must change dynamically and transparently alongside culture; and identifying the ways in which people resolve norm conflicts, and equipping AI with a system in which to do so in a similar and transparent way. This should be done collaboratively and across diverse research efforts, with care taken to evaluate and assess potential biases that disadvantage specific social groups.

Several initiatives – such as ***AI4All*** and the ***AI Now Institute*** – explicitly advocate for fair, diverse, equitable, and non-discriminatory inclusion in AI at all stages, with a focus on support for under- represented groups. Currently, AI-related degree programmes do not equip aspiring developers and designers with an appropriate knowledge of ethics and corporate environments and business practices are not ethically empowering, with a lack of roles for senior ethicists that can steer and support value-based innovation.

On a global scale, the inequality gap between developed and developing nations is significant. While AI may have considerable usefulness in a humanitarian sense, they must not widen this gap or exacerbate poverty, illiteracy, gender and ethnic inequality, or disproportionately disrupt employment and labour. The IEEE suggests taking action and investing to mitigate the inequality gap; integrating corporate social responsibility (CSR) into development and marketing; developing transparent power structures; facilitating and sharing robotics and AI knowledge and research; and generally keeping AI in line with the US Sustainable Development Goals[11]. AI technology should be made equally available worldwide via global standardisation and open-source software, and interdisciplinary discussion should be held on effective AI education and training.

A set of ethical guidelines published by the ***Japanese Society for AI*** emphasises, among other considerations, the importance of a) contribution to humanity, and b) social responsibility. AI must act in the public interest, respect cultural diversity, and always be used in a fair and equal manner.

The ***Foundation for Responsible Robotics*** includes a Commitment to Diversity in its push for responsible AI; the ***Partnership on AI*** cautions about the 'serious blind spots' of ignoring the presence of biases and assumptions hidden within data; ***Saidot*** aims to ensure that, although our social values are now 'increasingly mediated by algorithms', AI remains human-centric; the ***Future of Life Institute*** highlights a need for AI imbued with human values of cultural diversity and human rights; and the ***Institute for Ethical AI & Machine Learning*** includes 'bias evaluation' for monitoring bias in AI development and production. The dangers of human bias and assumption are a frequently identified risk that will accompany the ongoing development of AI.

Financial harm: Economic opportunity and employment

AI may disrupt the economy and lead to loss of jobs or work disruption for many humans, and will have an impact on workers' rights and displacement strategy as many strains of work become automated.

Additionally, rather than just focusing on the number of jobs lost or gained, traditional employment structures will need to be changed to mitigate the effects of automation and take into account the complexities of employment. Technological change is happening too fast for the traditional workforce to keep pace without retraining. Workers must train for adaptability, says the ***IEEE*** , and new skill sets, with fallback strategies put in place for those who cannot be re-trained, and training programmes implemented at the level of high school or earlier to increase access to future employment. The ***UNI Global Union*** call for multi-stakeholder ethical AI governance bodies on global and regional levels, bringing together designers, manufacturers, developers, researchers, trade unions, lawyers, CSOs, owners, and employers. AI must benefit and empower people broadly and equally, with policies put in place to bridge the economic, technological, and social digital divides, and ensure a just transition with support for fundamental freedoms and rights.

The AI Now Institute works with diverse stakeholder groups to better understand the implications that AI will have for labour and work, including automation and early-stage integration of AI changing the nature of employment and working conditions in various sectors.

AI in the workplace will affect far more than workers' finances, and may offer various positive opportunities. As laid out by the **IEEE**, AI may offer potential solutions to workplace bias – if it is developed with this in mind, as mentioned above – and reveal deficiencies in product development, allowing proactive improvement in the design phase.

RRI is a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view to the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific andtechnological advances in our society).'

Lawfulness and justice

Several initiatives address the need for AI to be lawful, equitable, fair, just and subject to appropriate, pre-emptive governance and regulation. The many complex ethical problems surrounding AI translate directlyand indirectly into discrete legal challenges

The ***IEEE*** conclude that AI should not be granted any level of 'personhood', and that, while development, design and distribution of AI should fully comply with all applicable international and domestic law, there is much work to be done in defining and implementing the relevant legislation. Legal issues fall into a few categories: legal status, governmental use, legal accountability for harm, and transparency, accountability, and verifiability. The IEEE suggest that AI should remain subject to the applicable regimes of property law; that stakeholders should identify the types of decisions that should never be delegated to AI, and ensure effective human control over those decisions via rules and standards; that existing laws should be scrutinised and reviewed for mechanisms that could practically give AI legal autonomy; and that manufacturers and operators should be required to comply with the applicable laws of all jurisdictions in which an AI could operate. They also recommend that governments reassess the legal status for AI as they become more sophisticated, and work closely with regulators, societal and industry actors and other stakeholders to ensure that the interests of humanity – and not the development of systems themselves – remain the guiding principle.

Control and the ethical use – or misuse – of AI

With more sophisticated and complex new AI come more sophisticated and complex possibilities for misuse. Personal data may be used maliciously or for profit, systems are at risk of hacking, and technology may be used exploitatively. This ties into informed use and public awareness: as we enter a new age of AI, with new systems and technology emerging that have never before been implemented, citizens must be kept up to date of the risks that may comewith either the use or misuse of these.

The ***IEEE*** suggests new ways of educating the public on ethics and security issues, for example a 'data privacy' warning on smart devices that collect personal data; delivering this education in scalable, effective ways; and educating government, lawmakers, and enforcement agencies surrounding these issues, so they can work collaboratively with citizens – in a similar way to police officers providing safety lectures in schools – and avoid fear and confusion.

Other issues include manipulation of behaviour and data. Humans must retain control over AI and oppose subversion. Most initiatives reviewed flag this as a potential issue facing AI as it develops, and flag that AI must behave in a way that is predictable and reliable, with appropriate means for redress, and be subject to validation and testing. AI must also work for the good of humankind, must not exploit people, and be regularly reviewed by human experts.

Personhood and AI

The issue of whether or not an AI deserves 'personhood' ties into debates surrounding accountability, autonomy, and responsibility: is it the AI itself that is responsible for its actions and consequences, or the person(s) who built them?

This concept, rather than allowing robots to be considered people in a human sense, would place robots on the same legal level as corporations. It is worth noting that corporations' legal personhood can currently shield the natural persons behind them from the implications of the law. However, ***The UNI Global Union*** asserts that legal responsibility lies with the creator, not the robot itself, and calls for a ban on

Environmental harm and sustainability

The production, management, and implementation of AI must be sustainable and avoid environmental harm. This also ties in to the concept of well-being; a key recognised aspect of well- being is environmental, concerning the air, biodiversity, climate change, soil and water quality, and so on. The ***IEEE*** state that AI must do no harm to Earth's natural systems or exacerbate their degradation, and contribute to realising sustainable stewardship, preservation, and/or the restoration of Earth's natural systems. The ***UNI Global Union*** state that AI must put people and the planet first, striving to protect and even enhance our planet's biodiversity and ecosystems. The ***Foundation for Responsible Robotics*** identifies a number of potential uses for AI in coming years, from agricultural and farming roles to monitoring of climate change and protection of endangered species. These require responsible, informed policies to govern AI and robotics, say the Foundation, to mitigate risk and support ongoing innovation and development.

Informed use: public education and awareness

Members of the public must be educated on the use, misuse, and potential harms of AI, via civic participation, communication, and dialogue with the public. The issue of consent – and how much an individual may reasonably and knowingly give – is core to this. For example, the ***IEEE*** raise several instances in which consent is less clear-cut than might be ethical: what if one's personal data are used to make inferences they are uncomfortable with or unaware of? Can consent be given when a system does not directly interact with an individual? This latter issue has been named the 'Internet of Other People's Things'. Corporate environments also raise the issue of power imbalance; many employees do not have clear consent on how their personal data – including those on health – is used by their employer. To remedy this, the IEEE suggest employee data impact assessments to deal with these corporate nuances and ensure that no data is collected without employee consent. Data must also be only gathered and used for specific, explicitly stated, legitimate purposes, kept up-to-date, lawfully processed, and not kept for a longer period than necessary. In cases where subjects do not have a direct relationship with the system gathering data, consent must be dynamic, and the system designed to interpret data preferences and limitations on collection and use.

To increase awareness and understanding of AI, undergraduate and postgraduate students must be educated on AI and its relationship to sustainable human development, say the IEEE. Specifically, curriculum and core competencies should be defined and prepared; degree programmes focusing on engineering in international development and humanitarian relief should be exposed to the potential of AI applications; and awareness should be increased of the opportunities and risks faced by Lower Middle Income Countries in the implementation of AI in humanitarian efforts across the globe.

Many initiatives focus on this, including the ***Foundation for Responsible Robotics, Partnership on AI, Japanese Society for AI Ethical Guidelines, Future Society*** and ***AI Now Institute***; these and others maintain that clear, open and transparent dialogue between AI and society is key to the creation of understanding, acceptance, and trust.

Existential risk

According to the Future of Life Institute, the main existential issue surrounding AI 'is not malevolence, but competence' – AI will continually learn as they interact with others and gather data, leading them to gain intelligence over time and potentially develop aims that are at odds with those of humans.

'You're probably not an evil ant-hater who steps on ants out of malice,' 'but if you're in charge of a hydroelectric green energy project and there's an anthill in the region to be flooded, too bad for the ants. A key goal of AI safety research is to never place humanity in the position of those ants' .

AI also poses a threat in the form of ***autonomous weapons systems (AWS)***. As these are designed to cause physical harm, they raise numerous ethical quandaries. The IEEE lays out a number of recommendations to ensure that AWS are subject to meaningful human control: they suggest audit trails to guarantee accountability and control; adaptive learning systems that can explain their reasoning in a transparent, understandable way; that human operators of autonomous systems are identifiable, held responsible, and aware of the implications of their work; that autonomous behaviour is predictable; and that professional codes of ethics are developed to address the development of autonomous systems – especially those intended to cause harm. The pursuit of AWS may lead to an international arms race and geopolitical stability; as such, the IEEE recommend that systems designed to act outside the boundaries of human control or judgement are unethical and violate fundamental human rights and legal accountability for weapons use.

Given their potential to seriously harm society, these concerns must be controlled for and regulated pre-emptively, says the ***Foundation for Responsible Robotics***. Other initiatives that cover this risk explicitly include the ***UNI Global Union*** and the ***Future of Life Institute***, the latter of which cautions against an arms race in lethal autonomous weapons, and calls for planning and mitigation efforts for possible longer-term risks. We must avoid strong assumptions on the upper limits of future AI capabilities, assert the FLI's **Asilomar Principles**, and recognise that advanced AI represents a profound change in the history of life on Earth.

2.3 Case study: healthcare robots

Artificial Intelligence and robotics are rapidly moving into the field of healthcare and will increasingly play roles in diagnosis and clinical treatment. For example, currently, or in the near future, robots will help in the diagnosis of patients; the performance of simple surgeries; and the monitoring of patients' health and mental wellness in short and long-term care facilities. They may also provide basic physical interventions, work as companion carers, remind patients to take their medications, or help patients with their mobility.

In some fundamental areas of medicine, such as medical image diagnostics, machine learning has been proven to match or even surpass our ability to detect illnesses.

Embodied AI, or robots, are already involved in a number of functions that affect people's physical safety. In June 2005, a surgical robot at a hospital in Philadelphia malfunctioned during prostate surgery, injuring the patient. In June 2015, a worker at a Volkswagen plant in Germany was crushed to death by a robot on the production line. In June 2016, a Tesla car operating in autopilot mode collided with a large truck, killing the car's passenger.

As robots become more prevalent, the potential for future harm will increase, particularly in the case of driverless cars, assistive robots and drones, which will face decisions that have real consequences for human safety and well-being. The stakes are much higher with embodied AI than with mere software, as robots have moving parts in physical space. Any robot with moving physical parts poses a risk, especially to vulnerable people such as children and the elderly.

Safety

It is vital that robots should not harm people, and that they should be safe to work with. This point is especially important in areas of healthcare that deal with vulnerable people, such as the ill, elderly, and children.

Digital healthcare technologies offer the potential to improve accuracy of diagnosis and treatments, but to thoroughly establish a technology's long-term safety and performance investment in clinical trials is required. The debilitating side-effects of vaginal mesh implants and the continued legal battles against manufacturers, stand as an example against shortcutting testing, despite the delays this introduces to innovating healthcare. Investment in

clinical trials will be essential to safely implement the healthcare innovations that AI systems offer.

User understanding

The correct application of AI by a healthcare professional is important to ensure patient safety. For instance, the precise surgical robotic assistant 'the da Vinci' has proven a useful tool in minimising surgical recovery, but requires a trained operator

A shift in the balance of skills in the medical workforce is required, and healthcare providers are preparing to develop the digital literacy of their staff over the next two decades. With genomics and machine learning becoming embedded in diagnoses and medical decision-making, healthcare professionals need to become digitally literate to understand each technological tool and use it appropriately. It is important for users to trust the AI presented but to be aware of each tool's strengths and weaknesses, recognising when validation is necessary. For instance, a generally accurate machine learning study to predict the risk of complications in patients with pneumonia erroneously considered those with asthma to be at low risk. It reached this conclusion because asthmatic pneumonia patients were taken directly to intensive care, and this higher-level care circumvented complications. The inaccurate recommendation from the algorithm was thus overruled.

However, it's questionable to what extent individuals need to understand how an AI system arrived at a certain prediction in order to make autonomous and informed decisions. Even if an in-depth understanding of the mathematics is made obligatory, the complexity and learned nature of machine learning algorithms often prevent the ability to understand how a conclusion has been made from a dataset — a so called 'black box'. In such cases, one possible route to ensure safety would be to license AI for specific medical procedures, and to 'disbar' the AI if a certain number of mistakes are made.

Data protection

Personal medical data needed for healthcare algorithms may be at risk. For instance, there are worries that data gathered by fitness trackers might be sold to third parties, such as insurance companies, who could use those data to refuse healthcare coverage. Hackers are another major concern, as providing adequate security for systems accessed by a range of medical personnel is problematic.

Pooling personal medical data is critical for machine learning algorithms to advance healthcare interventions, but gaps in information governance form a barrier against responsible and ethical data sharing. Clear frameworks for how healthcare staff and researchers use data, such as genomics, in a way that safeguards patient confidentiality is necessary to establish public trust and enable advances in healthcare algorithms.

Legal responsibility

Although AI promises to reduce the number of medical mishaps, when issues occur, legal liability must be established. If equipment can be proven to be faulty then the manufacturer is liable, but it is often tricky to establish what went wrong during a procedure and whether anyone, medical personnel or machine, is to blame. For instance, there have been lawsuits against the da Vinci surgical assistant, but the robot continues to be widely accepted.

In the case of 'black box' algorithms where it is impossible to ascertain how a conclusion is reached, it is tricky to establish negligence on the part of the algorithm's producer.

For now, AI is used as an aide for expert decisions, and so experts remain the liable party in most cases. For instance, in the aforementioned pneumonia case, if the medical staff had relied solely on the AI and sent asthmatic pneumonia patients home without applying their specialist knowledge, then that would be a negligent act on their part.

Soon, the omission of AI could be considered negligence. For instance, in less developed countries with a shortage of medical professionals, withholding AI that detects diabetic eye disease and so prevents blindness, because of a lack of ophthalmologists to sign off on a diagnosis, could be considered unethical.

Bias

Non-discrimination is one of the fundamental values of the EU, but machine learning algorithms are trained on datasets that often have proportionally less data available about minorities, and as such can be biased. This can mean that algorithms trained to diagnose conditions are less likely to be accurate for ethnic patients; for instance, in the dataset used to train a model for detecting skin cancer, less than 5 percent of the images were from individuals with dark skin, presenting a risk of misdiagnosis for people of colour.

To ensure the most accurate diagnoses are presented to people of all ethnicities, algorithmic biases must be identified and understood. Even with a clear understanding of model design this is a difficult task because of the aforementioned 'black box' nature of machine learning. However, various codes of conduct and initiatives have been introduced to spot biases earlier. For instance, The Partnership on AI, an ethics-focused industry group was launched by Google, Facebook, Amazon, IBM and Microsoft — although, worryingly, this board is not very diverse.

Equality of access

Digital health technologies, such as fitness trackers and insulin pumps, provide patients with the opportunity to actively participate in their own healthcare. Some hope that these technologies will help to redress health inequalities caused by poor education, unemployment, and so on. However, there is a risk that individuals who cannot afford the necessary technologies or do not have the required 'digital literacy' will be excluded, so reinforcing existing health inequalities.

The UK's National Health Services' Widening Digital Participation programme is one example of how a healthcare service has tried to reduce health inequalities, by helping millions of people in the UK who lack the skills to access digital health services. Programmes such as this will be critical in ensuring equality of access to healthcare, but also in increasing the data from minority groups needed to prevent the biases in healthcare algorithms discussed above.

Quality of care

'There is remarkable potential for digital healthcare technologies to improve accuracy of diagnoses and treatments, the efficiency of care, and workflow for healthcare professionals'.

If introduced with careful thought and guidelines, companion and care robots, for example, could improve the lives of the elderly, reducing their dependence, and creating more opportunities for social interaction. Imagine a home-care robot that could: remind you to take your medications; fetch items for you if you are too tired or are already in bed; perform simple cleaning tasks; and help you stay in contact with your family, friends and healthcare provider via video link.

However, questions have been raised over whether a 'cold', emotionless robot can really substitute for a human's empathetic touch. This is particularly the case in long-term caring of vulnerable and often lonely populations, who derive basic companionship from caregivers. Human interaction is particularly important for older people, as research suggests that an extensive social network offers protection against dementia. At present, robots are far from being real companions. Although they can interact with people, and even show simulated emotions, their conversational ability is still extremely limited, and they are no replacement for human love and attention. Some might go as far as saying that depriving the elderly of human contact is unethical, and even a form of cruelty.

It's vital that robots don't make elderly people feel like objects, or with even less control over their lives than when they were dependent on humans — otherwise they may feel like they are 'lumps of dead matter: to be pushed, lifted, pumped or drained, without proper reference to the fact that they are sentient beings'.

In principle, autonomy, dignity and self-determination can all be thoroughly respected by a machine application, but it's unclear whether application of these roles in the sensitive field of medicine will be deemed acceptable. For instance, a doctor used a telepresence device to give a prognosis of death to a Californian patient; unsurprisingly the patient's family were outraged by this impersonal approach to healthcare. On the other hand, it's argued that new technologies, such as health monitoring apps, will free up staff time for more direct interactions with patients, and so potentially increase the overall quality of care.

Deception

A number of 'carebots' are designed for social interactions and are often touted to provide an emotional therapeutic role. For instance, care homes have found that a robotic seal pup's animal like interactions with residents brightens their mood, decreases anxiety and actually increases the sociability of residents with their human caregivers. However, the line between reality and imagination is blurred for dementia patients, so is it dishonest to introduce a robot as a pet and encourage a social-emotional involvement? And if so, is if morally justifiable?

Companion robots and robotic pets could alleviate loneliness amongst older people, but this would require them believing, in some way, that a robot is a sentient being who cares about them and has feelings — a fundamental deception. Turkle et al. (2006) argue that 'the fact that our parents, grandparents and children might say 'I love you' to a robot who will say 'I love you' in return, does not feel completely comfortable; it raises questions about the kind of authenticity we require of our technology'. Wallach and Allen (2009) agree that robots designed to detect human social gestures and respond in kind all use techniques that are arguably forms of deception. For an individual to benefit from owning a robot pet, they must continually delude themselves about the real nature of their relation with the animal. What's more, encouraging elderly people to interact with robot toys has the effect of infantilising them.

Autonomy

It's important that healthcare robots actually benefit the patients themselves, and are not just designed to reduce the care burden on the rest of society — especially in the case of care and companion AI. Robots could empower disabled and older people and increase their independence; in fact, given the choice, some might prefer robotic over human assistance for certain intimate tasks such as toileting or bathing. Robots could be used to help elderly people live in their own homes for longer, giving them greater freedom and autonomy.

Liberty and privacy

As with many areas of AI technology, the privacy and dignity of users' needs to be carefully considered when designing healthcare service and companion robots. Working in people's homes means that robots will be privy to private moments such as bathing and dressing; if these moments are recorded, who should have access to the information, and how long should recordings be kept? The issue becomes more complicated if an elderly person's mental state deteriorates and they become confused — someone with Alzheimer's could forget that a robot was monitoring them, and could perform acts or say things thinking that they are in the privacy of their own home. Home-care robots need to be able to balance their user's privacy and nursing needs, for example by knocking and awaiting an invitation before entering a patient's room, except in a medical emergency.

To ensure their charge's safety, robots might sometimes need to act as supervisors, restricting their freedoms. For example, a robot could be trained to intervene if the cooker was left on, or the bath was overflowing. Robots might even need to restrain elderly people from carrying out potentially dangerous actions, such as climbing up on a chair to get something from a cupboard. Smart homes with sensors could be used to detect that a person is attempting to leave their room, and lock the door, or call staff — but in so doing the elderly person would be imprisoned.

Moral agency

'There's very exciting work where the brain can be used to control things, like maybe they've lost the use of an arm…where I think the real concerns lie is with things like behavioural targeting: going straight to the hippocampus and people pressing 'consent', like we do now, for data access'. (John Havens)

Robots do not have the capacity for ethical reflection or a moral basis for decision-making, and thus humans must currently hold ultimate control over any decision-making. An example of ethical reasoning in a robot can be found in the 2004 dystopian film 'I, Robot', where Will Smith's character disagreed with how the robots of the fictional time used cold logic to save his life over that of a child's. If more automated healthcare is pursued, then the question of moral agency will require closer attention. Ethical reasoning is being built into robots, but moral responsibility is about more than the application of ethics — and it is unclear whether robots of the future will be able to handle the complex moral issues in healthcare .

Trust

Larosa and Danks write that AI may affect human-human interactions and relationships within the healthcare domain, particularly that between patient and doctor, and potentially disrupt the trust we place in our doctor.

'Psychology research shows people mistrust those who make moral decisions by calculating costs and benefits — like computers do'. Our distrust of robots may also come from the number of robots running amok in dystopian science fiction. News stories of computer mistakes — for instance, of an image-identifying algorithm mistaking a turtle for a gun — alongside worries over the unknown, privacy and safety are all reasons for resistance against the uptake of AI.

Firstly, doctors are explicitly certified and licensed to practice medicine, and their license indicates that they have specific skills, knowledge, and values such as 'do no harm'. If a robot replaces a doctor for a particular treatment or diagnostic task, this could potentially threaten patient-doctor trust, as the patient now needs to know whether the system is appropriately approved or 'licensed' for the functions it performs.

Secondly, patients trust doctors because they view them as paragons of expertise. If doctors were seen as 'mere users' of the AI, we would expect their role to be downgraded in the public's eye, undermining trust.

Thirdly, a patient's experiences with their doctor are a significant driver of trust. If a patient has an open line of communication with their doctor, and engages in conversation about care and treatment, then the patient will trust the doctor. Inversely, if the doctor repeatedly ignores the patient's wishes, then these actions will have a negative impact on trust. Introducing AI into this dynamic could increase trust — if the AI reduced the likelihood of misdiagnosis, for example, or improved patient care. However, AI could also decrease trust if the doctor delegated too much diagnostic or decision-making authority to the AI, undercutting the position of the doctor as an authority on medical matters.

As the body of evidence grows to support the therapeutic benefits for each technological approach, and as more robotic interacting systems enter the marketplace, then trust in robots is likely to increase. This has already happened for robotic healthcare systems such as the da Vinci surgical robotic assistant.

Employment replacement

As in other industries, there is a fear that emerging technologies may threaten employment , for instance, there are carebots now available that can perform up to a third of nurses' work. Despite these fears, the NHS' Topol Review concluded that 'these technologies will not replace healthcare professionals but will enhance them ('augment them'), giving them more time to care for patients'. The review also outlined how the UK's NHS will nurture a learning environment to ensure digitally capable employees.

2.5 Autonomous Vehicles

Autonomous Vehicles (AVs) are vehicles that are capable of sensing their environment and operating with little to no input from a human driver. While the idea of self-driving cars has been around since at least the 1920s, it is only in recent years that technology has developed to a point where AVs are appearing on public roads.

According to automotive standardisation body SAE International (2018), there are six levels of driving automation:

0	No automation	An automated system may issue warnings and/or momentarily intervene in driving, but has no sustained vehicle control.
1	Hands on	The driver and automated system share control of the vehicle. For example, the automated system may control engine power to maintain a set speed (e.g. Cruise Control), engine and brake power to maintain and vary speed (e.g. Adaptive Cruise Control), or steering (e.g. Parking Assistance). The driver must be ready to retakefull control at any time.
2	Hands off	The automated system takes full control of the vehicle (including accelerating, braking, and steering). However, the driver must monitor the driving and be prepared to intervene immediately at any time.
3	Eyes off	The driver can safely turn their attention away from the driving tasks (e.g. to text or watch a film) as the vehicle will handle any situations that call for an immediate response. However, the driver must still be prepared to intervene, if called upon by the AV to do so, within a timeframe specified by the AV manufacturer.

4	Minds off	As level 3, but no driver attention is ever required for safety, meaning the driver can safely go to sleep or leave the driver's seat.
5	Steering wheel optional	No human intervention is required at all. An example of a level 5 AVwould be a robotic taxi.

Some of the lower levels of automation are already well-established and on the market, while higher level AVs are undergoing development and testing. However, as we transition up the levels and put more responsibility on the automated system than the human driver, a number of ethical issues emerge.

Societal and Ethical Impacts of AVs

'We cannot build these tools saying, 'we know that humans act a certain way, we're going to kill them – here's what to do'.' (John Havens)

Public safety and the ethics of testing on public roads

At present, cars with 'assisted driving' functions are legal in most countries. Notably, some Tesla models have an Autopilot function, which provides level 2 automation (Tesla, nd). Drivers are legally allowed to use assisted driving functions on public roads provided they remain in charge of the vehicle at all times. However, many of these assisted driving functions have not yet been subject to independent safety certification, and as such may pose a risk to drivers and other road users. In Germany, a report published by the Ethics Commission on Automated Driving highlights that it is the public sector's responsibility to guarantee the safety of AV systems introduced and licensed on public roads, and recommends that all AV driving systems be subject to official licensing and monitoring.

In addition, it has been suggested that the AV industry is entering its most dangerous phase, with cars being not yet fully autonomous but human operators not being fully engaged. The risks this poses have been brought to widespread attention following the first pedestrian fatality involving an autonomous car. The tragedy took place in Arizona, USA, in May 2018, when a level 3 AV being tested by Uber collided with 49-year-old Elaine Herzberg as she was walking her bike across a street one night. It was determined that Uber was 'not criminally liable' by prosecutors and the US National Transportation Safety Board's preliminary report which drew no conclusions about the cause, said that all elements of the self- driving system were operating normally at the time of the crash. Uber said that the driver is relied upon to intervene and take action in situations requiring emergency braking – leading some commentators to call out the misleading communication to consumers around the terms 'self- driving cars' and 'autopilot'. The accident also caused some to condemn the practice of testing AV systems on public roads as dangerous and unethical, and led Uber to temporarily suspend its self-driving programme.

This issue of human safety — of both public and passenger — is emerging as a key issue concerning self-driving cars. Major companies — Nissan, Toyota, Tesla, Uber, Volkswagen — are developing autonomous vehicles capable of operating in complex, unpredictable environments without direct human control, and capable of learning, inferring, planning and making decisions.

Self-driving vehicles could offer multiple benefits: statistics show you're almost certainly safer in a car driven by a computer than one driven by a human. They could also ease congestion in cities, reduce pollution, reduce travel and commute times, and enable people to use their time more productively. However, they won't mean the end of road traffic accidents. Even if a self-driving car has the best software and hardware available, there is still a collision risk. An autonomous car could be surprised, say by a child emerging from behind a parked vehicle, and there

is always the issue of *how: how* should such cars be programmed when they must decide whose safety to prioritise?

Driverless cars may also have to choose between the safety of passengers and other road users. Say that a car travels around a corner where a group of school children are playing; there is not enough time to stop, and the only way the car can avoid hitting the children is to swerve into a brick wall — endangering the passenger. Whose safety should the car prioritise: the children's', or the passenger's?

Processes and technologies for accident investigation

AVs are complex systems that often rely on advanced machine learning technologies. Several serious accidents have already occurred, including a number of fatalities involving level 2 AVs:

- In January 2016, 23-year-old Gao Yaning died when his Tesla Model S crashed into the back of a road-sweeping truck on a highway in Hebei, China. The family believe Autopilot was engaged when the accident occurred and accuse Tesla of exaggerating the system's capabilities. Tesla state that the damage to the vehicle made it impossible to determine whether Autopilot was engaged and, if so, whether it malfunctioned. A civil case into the crash is ongoing, with a third-party appraiser reviewing data from the vehicle.
- In May 2016, 40-year-old Joshua Brown died when his Tesla Model S collided with a truck while Autopilot was engaged in Florida, USA. An investigation by the National Highways and Transport Safety Agency found that the driver, and not Tesla, were at fault. However, the National Highway Traffic Safety Administration later determined that both Autopilot and over-reliance by the motorist on Tesla's driving aids were to blame.
- In March 2018, Wei Huang was killed when his Tesla Model X crashed into a highway safety barrier in California, USA. According to Tesla, the severity of the accident was 'unprecedented'. The National Transportation Safety Board later published a report attributing the crash to an Autopilot navigation mistake. Tesla is now being sued by the victim's family.

Unfortunately, efforts to investigate these accidents have been stymied by the fact that standards, processes, and regulatory frameworks for investigating accidents involving AVs have not yet been developed or adopted. In addition, the proprietary data logging systems currently installed in AVs mean that accident investigators rely heavily on the cooperation of manufacturers to provide critical data on the events leading up to an accident.

One solution is to fit all future AVs with industry standard event data recorders — a so-called 'ethical black box' — that independent accident investigators could access. This would mirror the model already in place for air accident investigations.

Near-miss accidents

At present, there is no system in place for the systematic collection of near-miss accidents. While it is possible that manufacturers are collecting this data already, they are not under any obligation to do so — or to share the data. The only exception at the moment is the US state of California, which requires all companies that are actively testing AVs on public roads to disclose the frequency at which human drivers were forced to take control of the vehicle for safety reasons (known as 'disengagement').

In 2018, the number of disengagements by AV manufacturer varied significantly, from one disengagement for every 11,017 miles driven by Waymo AVs to one for every 1.15 miles driven by Apple AVs. Data on these disengagements reinforces the importance of ensuring that human safety drivers remain engaged. However, the Californian data collection process has been criticised, with some claiming its ambiguous wording and lack of strict guidelines enables companies to avoid reporting certain events that could be termed near-misses.

Without access to this type of data, policymakers cannot account for the frequency and significance of near-miss accidents, or assess the steps taken by manufacturers as a result of these near-misses. Again, lessons could be learned from the model followed in air accident investigations, in which all near misses are thoroughly logged and independently investigated. Policymakers require comprehensive statistics on all accidents and near-misses in order to inform regulation.

Data privacy

It is becoming clear that manufacturers collect significant amounts of data from AVs. As these vehicles become increasingly common on our roads, the question emerges: to what extent are these data compromising the privacy and data protection rights of drivers and passengers?

Already, data management and privacy issues have appeared, with some raising concerns about the potential misuse of AV data for advertising purposes. Tesla have also come under fire for the unethical use of AV data logs. In an investigation by *The Guardian*, the newspaper found multiple instances where the company shared drivers' private data with the media following crashes, without their permission, to prove that its technology was not responsible. At the same time, Tesla does not allow customers to see their own data logs.

One solution, proposed by the German Ethics Commission on Automated Driving, is to ensure that that all AV drivers be given full data sovereignty (Ethics Commission, 2017). This would allow them to control how their data is used.

Employment

The growth of AVs is likely to put certain jobs — most pertinently bus, taxi, and truck drivers — at risk.In the medium term, truck drivers face the greatest risk as long-distance trucks are at the forefront of AV technology. In 2016, the first commercial delivery of beer was made using a self-driving truck, in a journey covering 120 miles and involving no human action. Last year saw the first fully driverless trip in a self-driving truck, with the AV travelling seven miles without a single human on board.

Looking further forward, bus drivers are also likely to lose jobs as more and more buses become driverless. Numerous cities across the world have announced plans to introduce self-driving shuttles in the future, including Edinburgh, New York and Singapore. In some places, this vision has already become a reality; the Las Vegas shuttle famously got off to a bumpy start when it was involved in a collision on its first day of operation and tourists in the small Swiss town of Neuhausen Rheinfall can now hop on a self-driving bus to visit the nearby waterfalls. In the medium term, driverless buses will likely be limited to routes that travel along 100% dedicated bus lanes. Nonetheless, the advance of self-driving shuttles has already created tensions with organised labour and city officials in the USA. Last year, the Transport Workers Union of America formed a coalition in an attempt to stop autonomous buses from hitting the streets of Ohio.

Fully autonomous taxis will likely only become realistic in the long term, once AV technology has been fully tested and proven at levels 4 and 5. Nonetheless, with plans to introduce self-driving taxis in London by 2021 and an automated taxi service already available in Arizona, USA , it is easy to see why taxi drivers are uneasy.

The quality of urban environments

In the long-term, AVs have the potential to reshape our urban environment. Some of these changes may have negative consequences for pedestrians, cyclists and locals. As driving becomes more automated, there will likely be a need for additional infrastructure (e.g. AV-only lanes). There may also be more far-reaching effects for urban planning, with automation shaping the planning of everything from traffic congestion and parking to green spaces and lobbies. The rollout of AVs will also require that 5G network coverage is extended significantly — again,something with implications for urban planning.

The environmental impact of self-driving cars should also be considered. While self-driving cars have the potential to significantly reduce fuel usage and associated emissions, these savings could be counteracted by the fact that self-driving cars make it easier and more appealing to drive long distances. The impact of automation on driving behaviours should therefore not be underestimated.

Legal and ethical responsibility

From a legal perspective, who is responsible for crashes caused by robots, and how should victims be compensated (if at all) when a vehicle controlled by an algorithm causes injury? If courts cannot resolve this problem, robot manufacturers may incur unexpected costs that would discourage investment. However, if victims are not properly compensated then autonomous vehicles are unlikely to be trusted or accepted by the public.

Robots will need to make judgement calls in conditions of uncertainty, or 'no win' situations. However, which ethical approach or theory should a robot be programmed to follow when there's no legal guidance? As Lin et al. explain, different approaches can generate different results, including the number of crash fatalities.

Additionally, who should choose the ethics for the autonomous vehicle — drivers, consumers, passengers, manufacturers, politicians? Loh and Loh argue that responsibility should be shared among the engineers, the driver and the autonomous driving system itself.

Ethical dilemmas in development

In 2014, the Open Roboethics initiative (ORi 2014a, 2014b) conducted a poll asking people what they thought an autonomous car in which they were a passenger should do if a child stepped out in front of the vehicle in a tunnel. The car wouldn't have time to brake and spare the child, but could swerve into the walls of the tunnel, killing the passenger. This is a spin on the classic 'trolley dilemma', where one has the option to divert a runaway trolley from a path that would hurt several people onto the path that would only hurt one.

36 % of participants said that they would prefer the car to swerve into the wall, saving the child; however, the majority (64 %) said they would wish to save themselves, thus sacrificing the child. 44 % of participants thought that the passenger should be able to choose the car's course of action, while 33 % said that lawmakers should choose. Only 12 % said that the car's manufacturers should make the decision. These results suggest that people do not like the idea of engineers making moral decisions on their behalf.

Asking for the passenger's input in every situation would be impractical. However, Millar (2016) suggests a 'setup' procedure where people could choose their ethics settings after purchasing a new car. Nonetheless, choosing how the car reacts in advance could be seen as premeditated harm, if, for example a user programmed their vehicle to always avoid vehicle collisions by swerving into cyclists. This would increase the user's accountability and liability, whilst diverting responsibility away from manufacturers.

However, Millar suggests that the user of the technology, in this case the passenger in the self-driving car, should be able to decide what ethical or behavioural principles the robot ought to follow. Using the example of doctors, who do not have the moral authority to make important decisions on end-of-life care without the informed consent of their patients, he argues that there would be a moral outcry if engineers designed cars without either asking the driver

directly for their input, or informing the user ahead of time how the car is programmed to behave in certain situations.

2.6 Warfare and weaponization

Although partially autonomous and intelligent systems have been used in military technology since at least the Second World War, advances in machine learning and AI signify a turning point in the use of automation in warfare.

AI is already sufficiently advanced and sophisticated to be used in areas such as satellite imagery analysis and cyber defence, but the true scope of applications has yet to be fully realised. A recent report concludes that AI technology has the potential to transform warfare to the same, or perhaps even a greater, extent than the advent of nuclear weapons, aircraft, computers and biotechnology (Allen and Chan, 2017). Some key ways in which AI will impact militaries are outlined below.

Lethal autonomous weapons

As automatic and autonomous systems have become more capable, militaries have become more willing to delegate authority to them. This is likely to continue with the widespread adoption of AI, leading to an AI inspired arms-race. The Russian Military Industrial Committee has already approved an aggressive plan whereby 30% of Russian combat power will consist of entirely remote-controlled and autonomous robotic platforms by 2030. Other countries are likely to set similar goals. While the United States Department of Defense has enacted restrictions on the use of autonomous and semi autonomous systems wielding lethal force, other countries and non-state actors may not exercise such self-restraint.

Drone technologies

Standard military aircraft can cost more than US$100 million per unit; a high-quality quadcopter Unmanned Aerial Vehicle, however, currently costs roughly US$1,000, meaning that for the price of a single high-end aircraft, a military could acquire one million drones. Although current commercial drones have limited range, in the future they could have similar ranges to ballistic missiles, thus rendering existing platforms obsolete.

Robotic assassination

Widespread availability of low-cost, highly-capable, lethal, and autonomous robots could make targeted assassination more widespread and more difficult to attribute. Automatic sniping robots could assassinate targets from afar.

Mobile-robotic-Improvised Explosive Devices

As commercial robotic and autonomous vehicle technologies become widespread, some groups will leverage this to make more advanced Improvised Explosive Devices (IEDs). Currently, the technological capability to rapidly deliver explosives to a precise target from many miles away is restricted to powerful nation states. However, if long distance package delivery by drone becomes a reality, the cost of precisely delivering explosives from afar would fall from millions of dollars to thousands or even hundreds. Similarly, self-driving cars could make suicide car bombs more frequent and devastating since they no longer require a suicidal driver.

Hallaq et al. (2017) also highlight key areas in which machine learning is likely to affect warfare. They describe an example where a Commanding Officer (CO) could employ

an Intelligent Virtual Assistant (IVA) within a fluid battlefield environment that automatically scanned satellite imagery to detect specific vehicle types, helping to identify threats in advance. It could also predict the enemy's intent, and compare situational data to a stored database of hundreds of previous wargame exercises and live engagements, providing the CO with access to a level of accumulated knowledge that would otherwise be impossible to accrue.

Employing AI in warfare raises several legal and ethical questions. One concern is that automated weapon systems that exclude human judgment could violate International Humanitarian Law, and threaten our fundamental right to life and the principle of human dignity. AI could also lower the threshold of going to war, affecting global stability.

International Humanitarian law stipulates that any attack needs to distinguish between combatants and non-combatants, be proportional and must not target civilians or civilian objects. Also, no attack should unnecessarily aggravate the suffering of combatants. AI may be unable to fulfil these principles without the involvement of human judgment. In particular, many researchers are concerned that Lethal Autonomous Weapon Systems (LAWS) — a type of autonomous military robot that can independently search for and 'engage' targets using lethal force — may not meet the standards set by International Humanitarian Law, as they are not able to distinguish civilians from combatants, and would not be able to judge whether the force of the attack was proportional given the civilian damage it would incur.

Amoroso and Tamburrini argue that: '[LAWS must be] capable of respecting the principles of distinction and proportionality at least as well as a competent and conscientious human soldier'. However, Lim (2019) points out that while LAWS that fail to meet these requirements should not be deployed, one day LAWS will be sophisticated enough to meet the requirements of distinction and proportionality. Meanwhile, Asaro (2012) argues that it doesn't matter how good LAWS get; it is a moral requirement that only a human should initiate lethal force, and it is simply morally wrong to delegate life or death decisions to machines.

Some argue that delegating the decision to kill a human to a machine is an infringement of basic human dignity, as robots don't feel emotion, and can have no notion of sacrifice and what it means to take a life. As Lim et al (2019) explain, 'a machine, bloodless and without morality or mortality, cannot fathom the significance of using force against a human being and cannot do justice to the gravity of the decision'.

Robots also have no concept of what it means to kill the 'wrong' person. 'It is only because humans can feel the rage and agony that accompanies the killing of humans that they can understand sacrifice and the use of force against a human. Only then can they realise the 'gravity of the decision' to kill'.

However, others argue that there is no particular reason why being killed by a machine would be a subjectively worse, or less dignified, experience than being killed by a cruise missile strike. 'What matters is whether the victim experiences a sense of humiliation in the process of getting killed. Victims being threatened with a potential bombing will not care whether the bomb is dropped by a human or a robot'. In addition, not all humans have the emotional capacity to conceptualise sacrifice or the relevant emotions that accompany risk. In the heat of battle, soldiers rarely have time to think about the concept of sacrifice, or generate the relevant emotions to make informed decisions each time they deploy lethal force.

Additionally, who should be held accountable for the actions of autonomous systems — the commander, programmer, or the operator of the system? Schmit (2013) argues that the responsibility for committing war crimes should fall on both the individual who programmed the AI, and the commander or supervisor (assuming that they knew, or should have known, the autonomous weapon system had been programmed and employed in a war crime, and that they did nothing to stop it from happening).

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

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,

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***.

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

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?‖.

- ✧ 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).

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

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

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.

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.

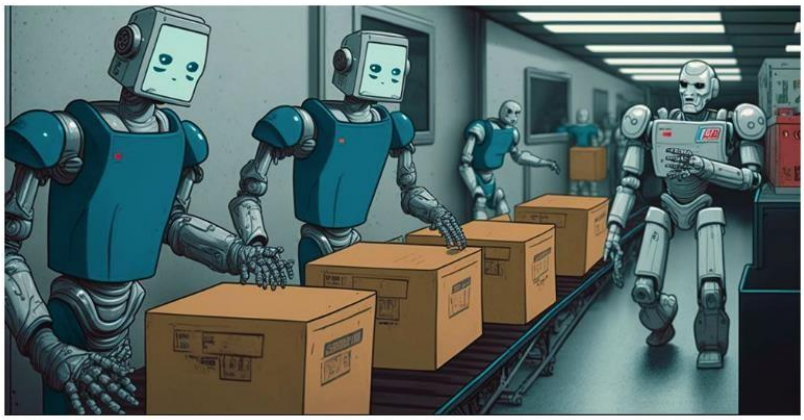
3.3.2.1 The issue of violation of Privacy:

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.

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.

- 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

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.

- 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

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.
- 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‘.

- 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‘;

- –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 caneasily 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 thecredibility 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.
- 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 risksand 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.
- 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 forthe 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 standardsand/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.
- 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 ROBOTICSAND 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.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.

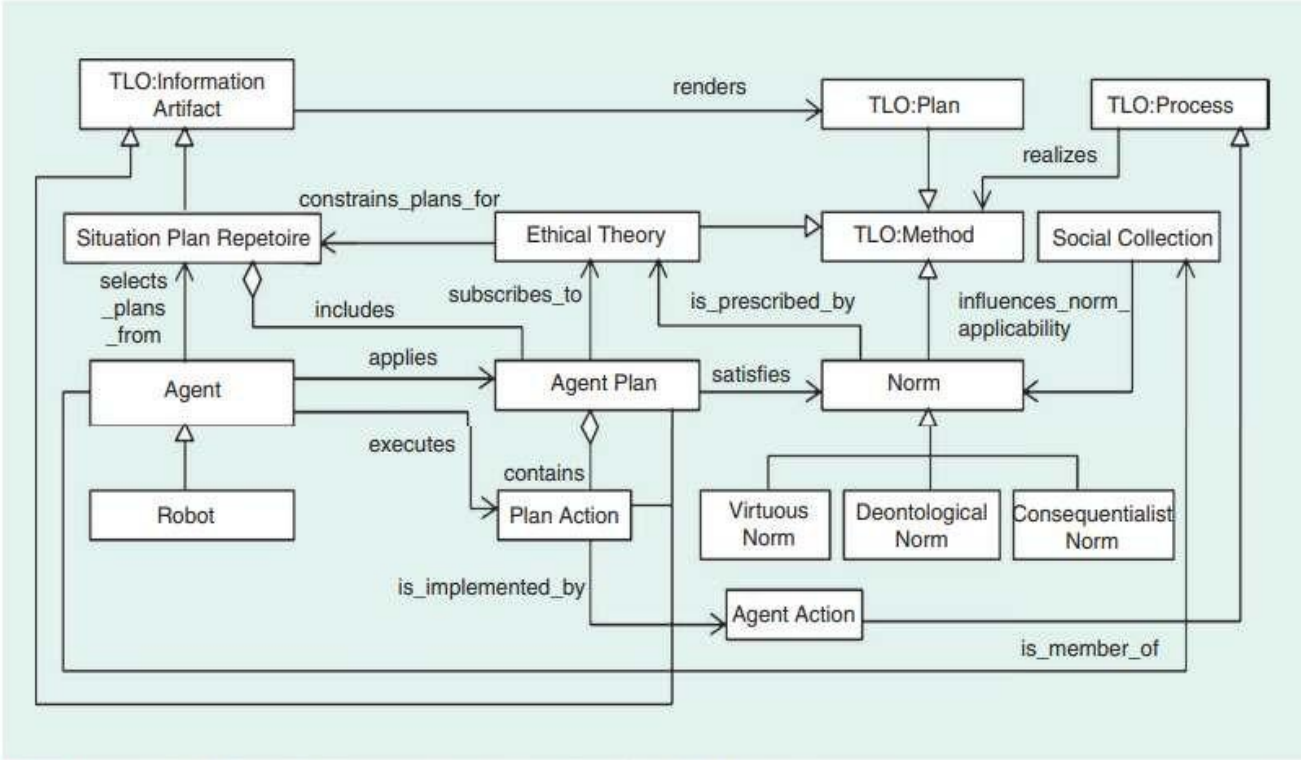


Figure 1. A partial UML model of the ERAS NEP ontology. UML: Unified Modeling Language.

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 withcapabilities to detect, assess, and manage ethical and legal norm violations occurring within or generated by autonomous system behavior.
- 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.

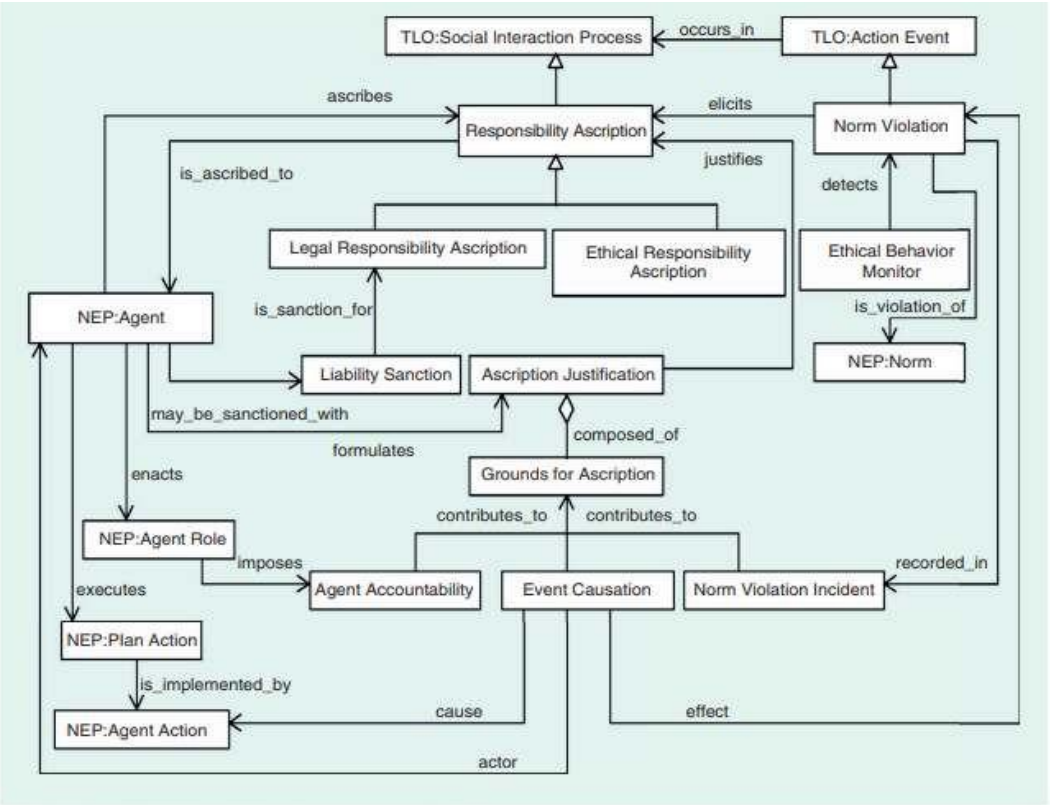


Figure 2. A partial UML model of the ERAS EVM ontology.

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.

UNIT IV
ROBOETHICS: SOCIAL AND ETHICAL IMPLICATION OF ROBOTICS

Robot- Roboethics - Ethics and Morality- Moral Theories-Ethics in Science and Technology - Ethical Issues in an ICT Society- Harmonization of Principles- Ethics and Professional Responsibility- Roboethics Taxonomy.

4.1 Robot- Roboethics

Roboethics analyzes the ethical, legal and social aspects of robotics, especially with regard to advanced robotics applications. These issues are related to liability, the protection of privacy, the defense of human dignity, distributive justice and the dignity of work. Today, roboethics is becoming an important component in international standards for advanced robotics, and in various aspects of artificial intelligence. An autonomous robot endowed with deep learning capabilities shows specificities in terms of its growing autonomy and decision-making functions and, thus, gives rise to new ethical and legal issues. The learning models for a care robot assisting an elderly person or a child must be free of bias related to the selected attributes and should not be subject to any stereotypes unintentionally included in their design. As roboethics goes hand in hand with developments in robotics applications, it should be the concern of all actors in the field, from designers and manufacturers to users. There is one very important element in this—albeit one that is related indirectly—that should not be overlooked: namely, how robotics and robotic applications are represented to the general public. Of the many representations, the legacy of mythology, science fiction and the legend still play an important role. The world of robotics is often marked by icons and images from literature. Exaggerated expectations of their functions, magical descriptions of their behavior, over-anthropomorphization, insistence on their perfection and their rationality compared to that of humans are only some of the false qualities attributed to robotics.

The Birth of Roboethics

This article outlines some of the main lines of development and application in roboethics—that is, applied ethics in advanced robotics—which examines the ethical, legal and societal issues (ELS) inherent to the field. Roboethics was born—as a term and as applied ethics—in 2004, during the First International Symposium on Roboethics. Roboethics analyzes the ethical, legal and social aspects of robotics, especially in relation to service and field robotics applications. These issues are related to the protection of privacy, the defense of human dignity, distributive justice and the dignity of work. Today, roboethics is the subject of hundreds of studies, applications, research, and is becoming an important component in international standards for advanced robotics, and also in various aspects of artificial intelligence. Robots are certainly formidable tools. There is no aspect of our private and social lives that cannot be improved by the introduction of robots. However, technology applied to human life always raises ethical questions. In the case of robotics, especially service robotics, these ELS issues are novel, emerging, complex and involve several disciplines.

A New Science

Robotics is a field of research and application, or a new science, still in its infancy, born from the fusion of many disciplines within the humanities and natural sciences. Here, the sum is greater than the parts. It is a very powerful tool for studying and increasing our knowledge, not only of the universe around us—space, oceans, our body—but also our brain/mind. This is why robotics can lead to a convergence of the so-called *two cultures*: human sciences and natural sciences.

Robotics covers the following disciplines: mechanics, automation, electronics, informatics, cybernetics, physics/mathematics, artificial intelligence, and draws contributions from (and, in fact, is *invading*) logic/linguistics, neuroscience/psychology, medicine/neurology, biology/physiology, psychology, anthropology/ethology, art/industrial design.

This complexity—involving novel aspects such as ESL issues, which drive research and robot production—is giving the subject the caliber of a science, with its laws, coherent and comprehensive understanding of nature and predictive capabilities. For an overview of the state of robotics today we recommend the *Springer Handbook of Robotics*.

Moreover, the object of the research and development of advanced robotics, an autonomous robot endowed with learning capabilities, shows specificities in terms of its growing autonomy and decision-making functions.

What Ethics should be applied in Roboethics

We have different *versions* of roboethics depending on which ethical theories are adopted (utilitarianism, deontology, virtue ethics, rights ethics, Rawls' theory of justice). Yet, in all these versions, a logical and critical framing of ethics is needed, one that reveals the implicit, uncovered assumptions, and analyzes the reasons, the pros and cons, and their origin. This frame also allows us to define (the extent and limits of) human liability and machine autonomy, in cases of damage caused by a learning robot.

In addition, in the light of more complete ethical theories, the ethical framing should assess whether, according to distributive justice principles, the actors involved should be socially duty-bound to compensate for the dramatic changes caused by a heavy, rapid and unnegotiated introduction of robots to our society: job displacement and loss; privacy issues and encroachment on personal life; technological dependency; robotics divide (in terms of generations, social status, and areas of the world).

Finally, roboethics should cover a series of positive recommendations and rules that would be implemented in all contexts where robots are introduced. These should be along the same lines as the recent prescriptions being adopted for the production and use of plastics, energy, and other industrial sectors. In roboethics, analysis of ELS issues often leads to recommendations which, in many cases, have been submitted to the UN, European governments, the European Commission and the European Parliament.

In light of the lessons learned from the COVID-19 pandemic, we cannot afford to introduce large-scale technological applications into society without offsetting the ensuing imbalances in the environment and the disruption to social groups.

Emerging and Novel Roboethical Issues

Since 2004, several authors have intervened in the debate on roboethics to highlight certain ELS issues that have arisen over a very long period of time. Issues such as the rights of and the payment of and for robots and their status as moral agents can be interesting to discuss, but too far-reaching. They also do not consider the urgency of and the need for addressing ethics-related technical issues.

Given the rich and complex debate on roboethics and the sometimes unknown developments that could come over the next two decades, the author, the partners and experts of the Ethicbots European Project adopted a triaging system to analyze the following issues:

Novelty: Issues that have never been looked at; the *absentia legis* and the lack of regulations is, in many cases (bionics and military robotics), evidence of a severe responsibility gap.

Emerging: Issues that are not planned for, since robotic prototypes are the result of different forces: research and business.

Complexity: Issues lying at the intersection of several disciplines (robotics, AI, moral philosophy, psychology, anthropology, law).

Social pervasiveness: Issues related to current and yet-to- be-released robotic products.
The sectors most directly and urgently interested in robotic applications are the military and certain areas of biomedicine (invasive prosthetics).

The Risk of Unintended Machine-Learning Bias

Issues of bias in artificial intelligence are well-known to scientists. Machine-learning models are developed to be predictive, when large datasets teach the robot learning models to predict the future, based on the past. Trained models can read and use an incredible amount of data (texts, pictures, software, other models), consuming it to identify the data patterns considered most suitable for carrying out the mission. In this way, predictions can be more accurate than with simple built-in models. The bias issue is related to the fact that machine-learning models can predict precisely what they have been trained to predict, and their predictions are as accurate as the data used to train the machine. Any errors are explained in the maxim "garbage in, garbage out." In fact, many cases of bias detection, which range from the light to the heavy involving AI products, stem from human bias intervening during the creation of data models. Either the data collected were unrepresentative of reality—as in the *portability issue*, when a model is employed out of context—or they reflect existing human prejudices—for instance, when certain attributes in the model are selected or ignored [9].
The learning models for a care robot assisting an elderly person, a child, or in a hospital must be free of bias related to the selected attributes (e.g., culture, gender, social or economic status, linguistic attributes) and should not be subject to any stereotypes unintentionally included in their design. It is easy to imagine how complex this process could be in a learning robot, especially since bias detection cannot be performed at the expense of the assisted person.

Issues of bias in artificial intelligence are well-known to scientists. Machine-learning models are developed to be predictive, when large datasets teach the robot learning models to predict the future, based on the past. Trained models can read and use an incredible amount of data (texts, pictures, software, other models), consuming it to identify the data patterns considered most suitable for carrying out the mission. In this way, predictions can be more accurate than with simple built-in models. The bias issue is related to the fact that machine-learning models can predict precisely what they have been trained to predict, and their predictions are as accurate as the data used to train the machine. Any errors are explained in the maxim "garbage in, garbage out." In fact, many cases of bias detection, which range from the light to the heavy involving AI products, stem from human bias intervening during the creation of data models. Either the data collected were unrepresentative of reality—as in the *portability issue*, when a model is employed out of context—or they reflect existing human prejudices—for instance, when certain attributes in the model are selected or ignored.

The learning models for a care robot assisting an elderly person, a child, or in a hospital must be free of bias related to the selected attributes (e.g., culture, gender, social or economic status, linguistic attributes) and should not be subject to any stereotypes unintentionally included in their design. It is easy to imagine how complex this process could be in a learning robot, especially since bias detection cannot be performed at the expense of the assisted person.

Ethical guidelines for all Robots

In a review article written by Matthew Studley and Alan Winfield on the ESL aspects of industrial robots, the authors came to an interesting conclusion after reviewing around 84 papers on the topic: today, even robots used in industrial production are subject to similar ELS problems to those found with service robots, which means that the gap between industrial robots and other types is narrowing.

Industry is changing; converging technologies have ushered in a fourth Industrial Revolution, where new collaborative robots, or *cobots*, work alongside humans on common tasks. Unlike more common industrial robots, which largely work alone and unsupervised, collaborative robots are programmed and designed to work with humans, responding to human behaviors and actions. The authors of that review article point to the increasing importance of human–robot interaction (HRI) and the reduced differentiation between industrial robotics and other robot domains affected by the definition and range of ELS issues. In this, advanced industrial robotics may be affected by the same sorts of concerns that are faced in assistive robotics: predicting and interpreting human intentions and future actions in order to perform efficiently. Here, the interactions between humans and robots involve teaching rather than programming. The ELS issues that affect learning cobots include psychological and sociological impacts, liability, data and privacy. Cobots can be programmed for the speed, tasks and precision to which humans have to adapt. In addition, cobots can be reprogrammed rapidly for another task, forcing humans to make rapid changes with no time to adapt. Cobots can gather data about the pace of work, abilities and needs of their human co-workers. These data may be processed in cloud services and could be used by other organizations. Use of the resulting data profiles could breach European privacy legislation.

Conclusion

Roboethics will have an impact on the design, programming, shape and use of robots. It should be included in engineering and architecture programs, as well as in the various disciplines of the humanities. Also, it is important to build trust between the general public and robotics laboratories through honest, concerted information campaigns.

4.2 Ethics and Morality

Introduction: Ethics and morality play a crucial role in the development and deployment of artificial intelligence (AI). AI systems, while capable of great benefits, also raise complex ethical questions that need to be addressed.

Key Concepts

1. **Ethical Decision-Making:** AI systems are increasingly being used to make decisions that have ethical implications, such as in healthcare, criminal justice, and finance. Ensuring that these decisions are made ethically is a key challenge.
2. **Bias and Fairness:** AI systems can inherit biases from the data they are trained on, leading to unfair or discriminatory outcomes. Ensuring fairness in AI is a critical ethical consideration.
3. **Transparency and Accountability:** AI systems are often opaque, making it difficult to understand how they arrive at their decisions. Ensuring transparency and accountability is essential for ethical AI.
4. **Privacy and Consent:** AI systems often rely on vast amounts of personal data. Ensuring that this data is used ethically, with the consent of the individuals involved, is crucial.
5. **Autonomy and Control:** As AI systems become more autonomous, questions arise about who is responsible for their actions and how much control humans should have over them.
6. **Social Impact and Equity:** The deployment of robotic technologies can have far-reaching social and economic consequences ,including job displacement, income inequality, and changes in power dynamics. This may involve policies to support retraining and job transition programs ,efforts to address bias and discrimination in algorithmic decision-making.
7. **Safety and Risk Management:** Robots have the potential to cause physical harm or damage if not designed and operated safely. Roboethic involves developing standards and guidelines for the safe design, testing, and deployment of robots to minimize risks to humans and property.

8. **Long-term Implications:** Roboethics also considers the broader implications of advanced robotics, such as the potential for robots to achieve consciousness or sentience

Ethical Frameworks

1. **Utilitarianism:** Focuses on maximizing the overall good or utility. In AI, this might involve designing systems that maximize benefits and minimize harms.
2. **Deontology:** Emphasizes the importance of following rules and principles. In AI, this might involve ensuring that AI systems adhere to ethical principles, even if it leads to suboptimal outcomes.
3. **Virtue Ethics:** Focuses on the character of the moral agent. In AI, this might involve designing systems that exhibit virtuous behavior, such as honesty and compassion.

Challenges

1. **Value Alignment:** Aligning the values of AI systems with those of society is a complex challenge, as different cultures and individuals may have different values.
2. **Explainability:** Ensuring that AI systems are explainable and understandable is crucial for building trust and accountability.
3. **Global Ethics:** AI raises questions that go beyond national borders, requiring a global approach to ethics and governance.

Conclusion: Ethics and morality are central to the development and deployment of AI. By addressing these issues thoughtfully, we can ensure that AI systems are used in ways that benefit society and respect human values.

4.3 Moral Theories

Introduction: Moral theories provide frameworks for understanding and evaluating ethical issues. In the context of AI, different moral theories can help us address questions about the design, deployment, and impact of AI systems.

Key Moral Theories

1. **Utilitarianism:** Focuses on maximizing overall happiness or utility. In the context of AI, this theory might suggest that AI systems should be designed to maximize benefits and minimize harm for the greatest number of people.
2. **Deontology:** Emphasizes the importance of following rules and principles. In the context of AI, this theory might suggest that AI systems should adhere to ethical principles, such as fairness and respect for autonomy, even if it leads to suboptimal outcomes.
3. **Virtue Ethics:** Focuses on the character of the moral agent. In the context of AI, this theory might suggest that AI systems should be designed to exhibit virtuous behavior, such as honesty and compassion.
4. **Rights Theory:** Focuses on the rights of individuals. In the context of AI, this theory might suggest that AI systems should respect the rights of individuals, such as privacy and freedom from discrimination.
5. **Ethics of Care:** Emphasizes the importance of relationships and empathy. In the context of AI, this theory might suggest that AI systems should be designed to consider the needs and feelings of those affected by their decisions.
6. **Social Contrast Theory:** Social contract theory posits that moral principles are derived from the hypothetical agreement among rational individuals in a society .Considering the explicit and implicit agreements between stakeholders regarding the development and regulation of robots.
7. **Consequentialism:** This theory evaluates the morality of an action based on its consequences. In roboethics ,consequentialism might focus on maximizing overall well-being, minimizing ,minimizing harm ,optimizing certain outcomes such as safety or efficiency in the use of robots.

Application to AI

- **Decision-Making:** Different moral theories can lead to different approaches to AI decision-making. For example, a utilitarian approach might prioritize outcomes, while a deontological approach might prioritize following ethical rules.
- **Bias and Fairness:** Moral theories can help us evaluate the fairness of AI systems. For example, a rights-based approach might highlight the importance of ensuring that AI systems do not discriminate against certain groups.
- **Transparency and Accountability:** Moral theories can also help us evaluate the transparency and accountability of AI systems. For example, a virtue ethics approach might emphasize the importance of designing AI systems that are honest and open about their decisions.

Challenges

- **Conflicting Principles:** Different moral theories can sometimes lead to conflicting conclusions, making it challenging to determine the most ethical course of action.
- **Complexity of AI Systems:** The complexity of AI systems can make it difficult to apply moral theories in practice, as the outcomes of AI decisions can be unpredictable.

Conclusion: Moral theories provide valuable frameworks for understanding and evaluating ethical issues in AI. By applying these theories thoughtfully, we can develop AI systems that align with ethical principles and promote human well-being.

4.4 Ethics in Science and Technology

Ethics in Science and Technology is a critical field that examines the moral implications of scientific and technological advancements. It involves assessing the impact of these advancements on individuals, societies, and the environment, and determining how to ethically navigate these changes. Key issues in this field include:

1. **Research Ethics**: Ensuring that scientific research is conducted in a manner that is transparent, honest, and respects the rights and dignity of research subjects.
2. **Privacy and Data Security**: Addressing the ethical challenges posed by the collection, storage, and use of personal data in the digital age.
3. **Environmental Ethics**: Considering the impact of technological advancements on the environment and promoting sustainable practices.
4. **Bioethics**: Examining the ethical implications of advancements in fields such as biotechnology, genetic engineering, and medical ethics.
5. **Social Justice**: Addressing issues of inequality and discrimination that may arise from the use of technology, such as algorithmic bias and digital divides.
6. **Professional Ethics**: Promoting ethical conduct among scientists, engineers, and other professionals in the field.
7. **Global Ethics**: Considering the global implications of scientific and technological advancements, and promoting ethical practices on a global scale.

Ethics in Science and Technology is a dynamic field that continues to evolve as new technologies emerge and societal values change. It requires ongoing dialogue and collaboration among scientists, engineers, policymakers, and the public to ensure that scientific and technological advancements are used in ways that benefit society as a whole Here are some vibrant aspects of ethics in science and technology:

1. **Emerging Technologies**: Exploring the ethical implications of cutting-edge technologies like artificial intelligence, nanotechnology, and biotechnology, and how they challenge traditional ethical frameworks.

2. **Ethical Dilemmas in Research**: Discussing controversial research topics such as cloning, embryonic stem cell research, and gene editing, and debating the ethical boundaries of scientific exploration.
3. **Digital Ethics**: Examining issues such as online privacy, data protection, and the ethical use of algorithms in decision-making, particularly in the context of social media and digital platforms.
4. **Environmental Sustainability**: Highlighting the importance of incorporating ethical considerations into technological innovations to address climate change, pollution, and other environmental challenges.
5. **Cultural Perspectives**: Recognizing the diverse cultural viewpoints on science and technology and how these perspectives shape ethical debates and decisions.
6. **Policy and Regulation**: Considering the role of governments and international bodies in developing ethical guidelines and regulations for scientific research and technological development.
7. **Ethical Leadership**: Exploring the responsibilities of scientists, engineers, and technology leaders in promoting ethical practices and ensuring the responsible use of technology for the benefit of society.
8. **Public Engagement**: Emphasizing the importance of involving the public in ethical discussions around science and technology to ensure that decisions reflect societal values and concerns.
9. **Historical Perspectives**: Examining historical cases of unethical behavior in science and technology to learn from past mistakes and improve ethical practices in the future.
10. **Futuristic Scenarios**: Imagining possible future scenarios where ethical considerations play a central role in shaping the development and deployment of new technologies, such as in space exploration, robotics, and artificial intelligence.

These vibrant aspects of ethics in science and technology highlight the dynamic nature of the field and the ongoing need for thoughtful reflection and dialogue on the ethical implications of scientific and technological advancements.

4.5 Ethical Issues in an ICT Society

The new age is characterised by information communication technology, where computers, laptops, mobile gadgets such as palmtops, tablets, mobile phones and the internet have made their way to every sphere of life from our domestic environments such a homes, small retail shops, banks to large industries and the world at large. Over time, so many questions about the ethical use of Information and Communication Technology (ICT) have emerged on who has access to what technology and the use of it and like every other technological invention. This paper will look at the morality involved in making use of ICT to promote ethical use of data or information in a virtual environment. ICT also possesses both positive and negative impacts; the questions that these impacts raise are what this piece attempts to answer, providing global scrutiny to these ethical issues, situations and questions; and providing recommendations in line with global best practices.

Introduction

The steps towards ethical considerations for issues concerning ICT first featured under the theme information ethics in 1992 during the Annual Review of Information Science and Technology. (Ugbogbo & Atu, 2016). This discussion on communication ethics would not be complete without a brief checkup of the broader definition of ethics. Ethics for Kallman and Grillo (1996), is the "practice of making a principle-based choice between competing alternatives". This therefore allows the topic of ethics and communication in the new age to rest comfortably under this umbrella. Traditionally, one of the main goals of professional ethics is to pursue explicitly defined values and norms and discourage inappropriate behaviour by professionals to develop public trust in

services delivered b certain institutions (Stexhe & Verstraeten, 2000). This why to begin with, we must establish that for someone to possess ethical values in communication especially in the new age, the person would first of all require to have proper ethical values needed for surviving normally. This is why today, it is a common sight that many people who claim to be ethical in their day-to-day dealing, have little or no strict values once it comes to communication in the new age especially when it comes to the use of new-age communication apparatus.

There are a few misconceptions people hold concerning ethics in communication, and for this paper, it is important to state such beliefs at this introductory part of this work. A lot of people really see nothing wrong in the illegal or unethical use of the computer, for many, the wrong use of the computer is really nothing and so raises no guilt in the way stealing or lying would make one feel. For some other people, the belief is that communication ethics is only concerned with matters that revolve around software piracy or the infringing of someone's computer system without authorized access. Sadly, all of these notions are wrong, for all unethical use of the computer are as grievous and weighty. This is why Wood (1993) stated that many studies carried out on information communication ethics have been relegated only to matters relating to subjects cover problems such as software piracy, instead of a comprehensive conceptualization of information and communication ethics.

Hence, Malone (1993) strongly held that matters that should be of ethical concern must rise above only such things as we would want to call illegal and this is why he added software reliability, privacy and matching, employee displacement, and artificial intelligence. While talking about this subject, Kallman and Grillo (1996) identified a group of other issues that also fit into this including social and economic issues. We can see that there are actually so many issues summed under the umbrella of information communication ethics and that it affects a wide range of people. For today, almost everyone is touched by the ICT in one way or the other. This is why we need a widespread creation of awareness if only we want to address this. From all that we have stated, we can comfortably infer that there is no real distinction between information communication ethics and what others would want to refer to as normal ethics since issues of ICT ethics also fall under the umbrella of ordinary ethics.

Yet, as Kallman and Grillo (1996) stated that, people who appear to be unethical about ICT do not usually see anything wrong in their actions, hence the easy misuse, this explains why the one who goes through someone's messages in their inbox and those who easily open the files in another's computer indiscriminately, do so without feeling any guilty about it. A major reason why many of those who are hackers and those who employ software to break into the computers of others may appear horrified when they are called thieves.

Neumann (1991) mentions that, the new age which is characterized by the use of computers and technologically advanced communications equipment has over time changed the method in which people connect. Communications that used to be face to face has been reduced to the web-space, making decision making swifter and less thoughtful, and like anything that is rushed and not carefully done, Lately there has been a hike in the number of new web sites, these sites being created by people who, though, are computer literates, often have zero concerns as to what ethical standards or values they are to uphold concerning the use of the ICT. The unavoidable result of all of these is an increasingly large amount of inevitable abuse.

Today, it is a fact that the internet has the power to inform, possessing more potential outreach than the traditional media, say the television and the radio, People today use the Web for many and diverse purposes and people have the ability to access billions of other people easily, then of course issues of ethical concern would be inevitable. Hence, the problems described above leads to a considerable need for writing this paper, the author decided to check the current situation on

computer ethics in with the of goal of critically examining the rules, procedures which apply in fields such as: computer use in work places, respect for ownership of software, licenses, patents, privacy and anonymity, computer crime, and to check for the ethical standards in the implementation the use of information technology, and at the end providing relevantrecommendations.

The new age has brought about extraordinary developments in technology which has transformed the way most people obtain and use information. The emergence of electronic library resources as a medium of information storage and delivery has become an essential component in academic institutions as it plays a critical role in meeting academic needs of staff and students (Usman, 2016). Information and Communication Technology (ICT) has been eased by providing the Internet to satisfy the everyday academic and research life of students (Apuke & Iyendo, 2018). We see that, Ternenge and Kashimana (2019), Ugwu and Orsu (2017), Adeyoyin, Idowu, and Sowole (2016) have all together agreed that challenges like lack of necessary knowledge of electronic information resources (EIR), hardware operations, lack of ethical browsing skills, financial problems to procure EIR gadgets, unethical information overload, funding, inconsistent electricity supply, poor ICT infrastru cture, as insufficient knowledge of software applications usage were faced by the students when using the internet. Notwithstanding, whenever there is a mention of ethics, there is usually a drawn expectation concerning the way and manner by which people must conduct their actions be it written or verbal, consistent with generally accepted standards for those whom it may concern. Notwithstanding, there exist different views on the concept of ethics. Fisher (2004), ethics is the concept of an individual's personal belief showing what is right or wrong, good or bad. Equally, Miner (2002) sees it as the judgments of actions as right or wrong which comes from the beliefs of a people. This is why Mintz and Morris (2007) explained it to mean standards that are acceptable when referring to human behaviors, behaviors that show how people ought or ought not to act how as against how people act.

Nevertheless, sociologically speaking, Scott and Lyman (1986) states that ethics "prevent conflicts from arising by bridging the gap between action and expectation". Ward (2011), explained in his book, "Ethics and the Media: An Introduction" showed that the concept ethics originates from the Greek word "ethos" which means charisma, nature or disposition. This position of ethics is closed to the common belief that ethics is internally concerned with being virtuous, that which moves the people towards correct acts. This etymology of ethics proposes ethics to be both distinctive as well as collective. Ethics is distinctive simply because those who do likewise, ethics is collective is social because no one can seat frame rules of conduct as regards rightness of wrongness for themselves without putting them into consideration with the rules or fair societal interaction. Erondu, Sharland, and Okpara (2004) believe that the study of "ethics" has as its focus issues that concern Nevertheless, sociologically speaking, Scott and Lyman (1986) states that ethics "prevent conflicts from arising by bridging the gap between action and expectation". Ward (2011), explained in his book, "Ethics and the Media: An Introduction" showed that the concept ethics originates from the Greek word "ethos" which means charisma, nature or disposition. This position of ethics is closed to the common belief that ethics is internally concerned with being virtuous, that which moves the people towards correct acts. This etymology of ethics proposes ethics to be both distinctive as well as collective. Ethics is distinctive simply because those who do likewise, ethics is collective is social because no one can seat frame rules of conduct as regards rightness of wrongness for themselves without putting them into consideration with the rules or fair societal interaction. Erondu, Sharland, and Okpara (2004) believe that the study of "ethics" has as its focus issues that concern the act of making a practical decision, revolving values which people hold in other to survive, the lens through which the rightness or wrongness of human actions can be judged. Hence ethics implies those schemes of moral values or principles, Adenubi (1999). Making ethics

that term employed to mean moral beliefs and ethical theory governing human conduct. (Beauchamp & Bowie, 2001). What this makes Ethics is a reflection on morality, meaning the principles of making choices by individuals is them right or wrong). Thus, ethics is the guide for both human and societal behavior; this is why Capurro (2006) could easily hold that ethics is an endless search for the unambiguous and implied usage of codes especially moral codes.

Media ethics according to Pavlik (2008), states that media ethics implies a set of practices, a code of things that media professionals such as journalists ought or ought not to do. This is the normative concept that prescribes helpful codes on how media professionals should or should not do. Ethics is the branch of philosophy that is concerned with the arranging, guarding, and prescribing of concepts on the right and wrong thing to do; it is also known as moral philosophy.

While carrying out their responsibilities, library and information professionals have certain moral codes that guide their actions just like as obtainable in every other profession. These codes are responsible for prompting actions at every given time. There are certain theories of ethics which have over time have shown their importance in guiding professionals of information technology. These theories provide benchmarks for striking the difference between what is right and what is wrong.

Fallis (2007) there are four such theories as they affect information ethics and they include those of consequence, duty, rights and virtue. According to those of consequence, the distinguishing factor separating right actions from wrong ones is that they have better outcomes. Thus, in attempting to do the right thing, we should do things that have good outcomes. For the duty-based theory, the main position is that human beings have ethical duties they must obey and that the outcome of actions should not be the guiding principle in checking for the right and wrong action. For example, humans debatably possess the duty of preservation of like and not to kill innocent people, this is even if killing would have very good outcomes. While for the rights-based theory he explains that this position argues that, in determining the right thing to do, it should be done in light of the rights that human beings possess. These theories are consistent with information ethics since talks like these often fall in line with talks corresponding to those of human rights, such as we have in the American Library Association where there is the Library Bill of Rights. Finally, ethical theorists who hold the virtue-based theory believe that, in determining the right thing to do, human virtues ought to consider. So much so that, the right thing to do in any given situation is what a virtuous person would do in the same situation.

In every profession, people desire to perform their responsibility in a professional ethical way, and library and information professionals are not left out on this, but in the course of carrying out these activities, they often are greeted with dilemmas regarding the handling, protection and propagation of information. One of these dilemmas is in knowing for instance, if we are to place some forms of restriction on how computers are used in public libraries or if we should keep certain things in the library say a book that might adversely affect the sentiments of some users religiously or morally all stand as ethical dilemmas for the library staff. Hoq (2012) stated that, other ethical dilemmas that library staff face include the thoughts of charging users when special services are rendered knowing it is a public library or say allowing users to making photocopies in the library, another dilemma is in the restriction of persons with suspicious looks or those not properly dressed from entering into the public library. One thing underlying all these instances is that the library professional is faced with the need to take an ethical stand. The simple way of providing information for users has over time become complicated with the emergence of so many

sophisticated ICT facilities and especially as libraries have become more digitalized, making user demands global, the task of providing information has become challenging. More so, information divide inequalities that exists between rich and poor nations caused by wealth creation and use wealth has further affects how much information a nation can access or control. (Hoq, 2012) Thus, Masmoudi (1979) cites in his seminal paper "The New World Information Order" that, this aforementioned growing inequality exists amongst nations of the world when it comes to accessing, controlling, and disseminating information. The paper showed about seven forms of these inequalities that exists in the world of information amongst nations.

1. The deliberate quantitative imbalance that exists between the North and South;
2. The inequality in resources of information;
3. The de facto supremacy and a will to dominate;
4. The privation of information on developing countries;
5. Survival of the colonial era; a separating influence in the economic, social, and cultural spheres;
6. Dissemination of ill-suited messages.

Authorities such as Smith (1980), Morehouse (1981), Haywood (1995) and Buchannan (1999), have shown similar concerns, pointing out that, the developed countries have over time continue to lead the world when it comes to enjoying assets in material and knowledge; questioning the ethics of new age and the supposed free flow of information and related commodities which is supposed to be its hallmark. Thus, library and information professionals as custodians and organizers of information in the new age crossroads especially when it comes to knowing how to offer best services to users of the library, in the most ethical possible manner. Again mention must be made here that, trying to find a way that is generally accepted by all or the majority of those seeking information, and those who generate theses information may be difficult since there is no one universally accepted idea of "good" and "bad" as it may differ from person to person and from society to society.

Ethics and communication in the new age

We live in an age characterized by rapid advanced technological changes, and with the ease with which technology has made things flow. Ethical issues are simply on the increase requiring urgent attention at handling them. Of particular interest are the ethical issues that arise in the communication space. This rapid growth has over time had tremendous impacts on human society thus, raising some ethical questions for the people as well as for organizations. These issues have in fact over time reason to a surprising level affecting the society in various ways areas such as those of employment, and working conditions as these technological advancements have made the invasion of person's and corporate body's privacy very possible, and also affecting rights of intellectual property, of individual and societal, of preservation of values and accountability, etc. Many aspects of human living have with time been saddled with carrying of these burdens amongst human actions that have suffered and these are, employment working conditions and individuality. Sadly, only little progress has been made in this respect.

Fielden (2004), stated that Information Communication Technology (ICT) has with time occupied a very crucial position in the society at large for as one moves from commerce to industries government, medicine, education and entertainment. Its social and economic are hardly

required that have come off as being problematic, posing some negative ethical impacts on our society.

These impacts can be reduced to three and they include the following:

(a) Personal privacy

(b) Access right

(c) Harmful actions

In terms of personal privacy, the ICT has made large scale data exchange of information from anybody, irrespective of the person's location in any part of the world at any time. What this implies is that there is an increased probability for disclosing personal information and or of groups and in turn, violating the privacy of such person and or groups of people due to the ease with which there is widespread dissemination globally. Thus, it remains our duty and responsibility to ensure that the confidentiality and integrity of data regarding persons and groups are well kept. This therefore, involves taking precautions to ensure that there is accuracy in disseminating data, and also in protecting it from unapproved access that is, unintentional exposure to wrong persons.

On the aspect of ethical issues in computing systems is the access right. This has actually over time become a subject of high priority for cooperate bodies and government agencies. This is so because of the current popularity of commerce on the Internet especially international commerce. This interest became heightened due to break-ins that occurred as sophisticated places like Los Alamos National Laboratories and NASA in the US. Many of such illegal attempts access to the United States government and military computers have been reported. Thus, without putting in place proper security for the computer, there would be no assurance that network connections on the Internet would be secured from illegal accesses.

And finally, for the harmful actions, Grimes, Fleischman, and Jaeger (2009) note that in computer ethics, harmful action implies damage or negative consequences, which include the undesired loss of information, loss of property, damage of property, or unwanted environmental impacts. Harmful actions, thus, include intentional destruction or modification of files and programs that result in severe loss of assets or needless spending of human resources such as the time and effort that would be needed in cleaning the system. Now, we shall discuss some specific ethical issues that come up in communication in the new age.

Plagiarism

This is the term used to mean that the work of others is copied, by an author who presents it as his or her work. This translates as stealing and it is a highly unethical practice even in religion. Sadly, this action happens quite regularly, and it is much easier to do so with all the information available on the internet making its occurrence more frequent. Basically, plagiarism on the internet come in two was as identified by Ugbogbo and Atu (2012). First, with the ease with which electronic texts have made cut and paste. It is now very simple for students to copy published sources say articles that appear on the encyclopedia as their papers. Second, although it is not tough for students to get someone who writes their papers for them, it is now very much easier for people to find and buy unidentified papers at Web sites that specialize in such sales and to even charge original term papers and sell for an agreed price. Unluckily, the Internet takes away what gives, since teachers can now access the databases of their students papers once it is submitted

electronically. Hence they can easily relate these students' paper knowing where even a simple line originally appeared.

Hacking

This is the term used in referring to the action of an individual who has enough knowledge to gain access to computer systems to identify security flaws without authorization. These individuals are known as hackers. They break into, or hack a system. There are various reasons why hacking is done and it includes the malicious desire to spoil a system or to understanding how a system works, in the bid of making money from it. Furthermore, some people argue that there are hacker ethics which has used in alerting people that their system is insecure and needs improving. This is why reformed hacking could pose a moral dilemma. The "reformed hackers" sometimes use their expertise in helping organization in protecting themselves from other hackers. Hacking unlike breaking into a closed-door requires a lot of skills. With this skill, hackers can show that a system is not secure and needs refining. Thus, arguments can be made that hackers play a valuable role. For some others, the argument is that hacking might lead to some improvements.

Piracy

This is how software is illegally copied. White (2002) makes us understand this to be a very serious problem since about 50% of all programs on computers are pirated copies. Using elaborate code, programmers spend hours together in designing programs. Hence, these programs surely need to be protected. Although this being a very serious issue, that significantly damages the profit level for the programmers, some people still argue that some forms of pirating should at least be permitted since it helps in the creation of a more computer literate population.

Conclusions

In the end, I believe that an accurate conclusion for this piece is that ethics in its use for communication in the new age, especially in the context of Internet use, has not gone viral enough. There are so many matters for concern as the number of people touched and affected by the ICT is large and this is so majorly because of the increased availability of the Internet as it has not stopped growing. This, therefore, makes seeking a target audience difficult to define and even more is the trying to reach them. Ethical issues in themselves are also difficult to define, progressively compound and diverse, and they keep growing rapidly with the technology. More so, the attitudes, perceptions and behaviour of the users of the ICT leave are not encouraging. Again, is the fact that there are no generally accepted code of ethics and conduct for professional varies from one professional organization to the next. Sadly, too is the fact that when classes are organized in computer ethics, that is when ethics appears part of the ICT curriculum, little impact is made since only a minor number of students use it. The importance of communication ethics in the new age cannot be over-emphasized as there are way too many people involved in the use for us to be negligent. It is probably not possible to develop comprehensive ethical guidelines to cover every possible situation of IT misuse. Realizing the universality and the enormousness of this problem formulate ethical guidelines continually, in other to keep pace with constant changes revolving around these issues. Finally, it is paramount that these guidelines after being formulated be made a part of the curricula of all schools and colleges rather than just relegating it to ICT related disciplines.

4.6 Harmonization of Principles

Robotics and artificial intelligence (AI) are revolutionizing all spheres of human life. From industrial processes to graphic design, the implementation of automated intelligent systems is changing how industries work. The spread of robots and AI systems has triggered academic institutions to closely examine how these technologies may affect the humanity—this is how the fields of roboethics and AI ethics have been born. The identification of ethical issues for robotics and AI and creation of ethical frameworks were the first steps to creating a regulatory environment for these technologies. In this paper, we focus on regulatory efforts in Europe and North America to create enforceable regulation for AI and robotics. We describe and compare ethical principles, policies, and regulations that have been proposed by government organizations for the design and use of robots and AI. We also discuss proposed international regulation for robotics and AI. This paper tries to highlight the need for a comprehensive, enforceable, and agile policy to ethically regulate technology today and in the future. Through reviewing existing policies, we conclude that the European Union currently leads the way in defining roboethics and AI ethical principles and implementing them into policy. Our findings suggest that governments in Europe and North America are aware of the ethical risks that robotics and AI pose, and are engaged in policymaking to create regulatory policies for these new technologies.

Introduction

Robotics and artificial intelligence (AI) are having a profound impact on all aspects of everyday life: our food is collected by robots, we are being driven by self-driving vehicles, our phones know what we want to text to our loved ones, and when we get sick, our physician might be a robot. However, as exciting as these technologies are, they come with significant risks for the future of humanity. Over the last 10 years, the number of industrial robots has risen 300% and continues to increase.

The first attempts to understand the ethics of AI and robotics have come from academic institutions and private corporations, which demonstrates the field's awareness of its potential implications. The study of the ethics of robotics, or roboethics, was pioneered by by Gianmarco Veruggio in the early 2000s. Since then, roboethics and AI ethics have become widely discussed topics, with the number of publications mentioning either of the terms increasing tenfold in the last 5 years. Although many organizations already propose well-considered ethical principles for robotics and AI, the need remains for enforceable ethical regulations on governmental and international levels. The need for regulation is felt by all members of the robotics and AI communities, which is why many non-government organizations have decided to create their own ethical policies independent of law and policy. However, this does not promote standardization of ethics, and allows for moral loopholes which could lead to creation of automated systems that infringe on human rights. Governments have recognized the potential and risks that AI and robotics bring, and have initiated the process for creation of legislation that accounts for ethical concerns in AI and robotics.

Foundational principles of Robotics and Roboethics

While AI-powered robots are a thing of the present, there are examples of miraculous—for the time— machines from ancient civilizations which we can consider the first examples of robots. Al-Jazari, an Arab Muslim scholar from the thirteenth century BC designed wondrous items such as a programmable system for pouring and serving various drinks, a set of robotic musicians, and several water-raising machines. Leonardo da Vinci has also created a humanoid-looking automaton in a shape of a knight that was able move its head and jaw, wave its arms, and sit up. The Industrial

Revolution populated new technologies around the world and has permanently changed how people approach manual labour. With the discoveries of electricity, computers, and the internet, inventors were able to create machines and automatons capable of automating processes that were previously performed by humans. A new revolution is currently underway in the workforce, and over 70% of US workers indicate that they are worried about a future when robots and computers can perform human jobs. Similarly, in a study that surveyed over 20,000 EU workers, the majority has indicated that they agree that robots steal people's jobs. As the capabilities of machines have changed, our definitions for robots and.

Definition of robots

The term "robot" was first introduced by Karel Capek in a play that first premiered in Prague in 1921 [11]. The term, derived from Czech word "rabota" meaning compulsory work, was used to identify artificial laborers that served humans in a fictional Utopian society. Since then, the word "robot" has been popularized through works of science fiction and is now also applied to an array of intelligent mechanical systems. When considering policy design and implementation, it is critical to have a most complete and accurate definition of a robot, since a policy may or may not be applied to an object depending on whether it is classified as a robot. One the most widely accepted definitions for robot is one from the Robot Institute of America: "A robot is a reprogrammable, multifunctional manipulator designed to move material, parts, tools, or specialized devices, through variable programmed motions for the performance of a variety of tasks". Since the development of AI and the Internet of Things, and the merging of programmable systems with physical operators, the distinction between AI and robots is become more arbitrary. As such, perhaps the most straightforward definition for robot is "an embodied AI".

With the above definitions in mind, a robot should exhibit these three properties:

- Programmability, or an ability for a designer to manipulate robot's functions and capacities;
- Mechanical Capability, enabling a robot to act on its surroundings; and
- Flexibility, allowing the robot to operate in a variety of ways and adapt to different scenarios.

Asimov's laws of robotics

The term robotics, referring to a branch of engineering that studies robots, was first used by Isaac Asimov in his novel "Runaround". In the same novel, Asimov has introduced the first set of laws that dictate a robot's behaviour. These laws lay the foundation for roboethics and established the first set of boundaries between humans and robots. The Laws of Robotics (Laws) discuss concepts of safety, obedience, and self-preservation:

1. A robot may not injure a human being under any conditions—and, as a corollary, must not permit a human being to be injured because of inaction on [the robot's] part.

2. A robot must follow all orders given by qualified human beings as long as they do not conflict with First Law.

3. A robot must protect [its] own existence, as long as that does not conflict with the First and Second Law.

Asimov later recognized that the First Law did not extend to the human society overall and added an additional Zeroth Law that would supersede the First Law.

Revisions to the Asimov's laws of robotics

When Asimov proposed his Laws of Robotics, he could not envision the technological developments and the geopolitical climate of the twenty-first century. Building on the foundational principles of the Laws, several versions of the new Laws of Robotics have been proposed. According to Murphy and Woods, the main issues with the Laws are that they 1) assume that robots are solely responsible for human safety, 2) fail to explain how robots should interpret orders given by humans, and 3) ignore that many robots lack a self-protective component of autonomy. As a result, the Three Laws of Responsible Robotics focusing on accountability, responsiveness, and control, have been proposed.

1. A human may not deploy a robot without the human–robot work system meeting the highest legal and professional standards of safety and ethics.

2. A robot must respond to humans as appropriate for their roles.

3. A robot must be endowed with sufficient situated autonomy to protect its own existence as long as such protection provides smooth transfer of control to other agents consistent with the first and second laws.

These updated laws recognize that as robots are created by humans, the responsibility for robot actions lies on humans too. Assignment of responsibility is essential when creating legislation and policy, as policy enforcement requires accountability. In addition to that, the Laws of Responsible Robotics recognize that robots are a part of dynamic relationships that are built through human–robot interactions. These new Laws are not exhaustive nor specific, but they provide a more realistic starting point for ethicists and policy makers.

Additionally, the New Laws of Robotics (New Laws) have been proposed in 2020, which take into account morals of human actors:

1. Robotic systems and AI should complement professionals, not replace them;

2. Robotic systems and AI should not counterfeit humanity;

3. Robotic systems and AI should not intensify zero- sum arms races;

4. Robotic systems and AI must always indicate the identity of their creator(s), controller(s), and own er(s).

Ethical questions for robotic policies

The development and marketization of robotics pose many ethical questions for researchers, practitioners, government, and society alike. These ethical questions lay a foundation for ethical principles, which in their turn facilitate creation of roboethical standards such as BS 8611. The combination of ethical frameoworks and standards informs creation of regulatory policies for robotics. In order to create relevant policy, governments and corporations have proposed various

means of evaluating a robot. Here we discuss three categories based on which a robot can be evaluated to create ethical policy.

Functionality

Robots are designed to perform various functions, from assembly of heavy machinery to patient care. Most commonly robots are designed to perform tasks with utilitarian purpose, where a robot performs repetitive or heavy tasks in a workplace. These robots are often referred to as industrial robots, and they have been responsible for revolutionizing production economies around the world. Robots with more sophisticated mechanical functions were then adapted in medical fields, so we have seen the appearance of medical robots. Rapid advances in AI technology have facilitated development of a whole new class of robots- social robots. Social robots possess an ability to interact with humans, enabling them to perform caregiving, teaching, and customer service functions. Outside of professional environment, robots are designed as toys, art objects, or exclusively for user pleasure. Robots can now be expected to be involved in all aspects of human life, which undoubtedly will shape society and economy.

Capability

While two robots might be designed to perform the same function, their capabilities may vary depending on a unit's hardware and software. In other words, a robot that is only capable of simple processing operations cannot be compared to a robot that has sophisticated AI with a capacity for learning. The more advanced the AI, the more ethical questions can be raised to create regulatory policy. With AI that possesses human-like capabilities, one might even start asking whether such sophisticated robots deserve to have rights that would traditionally be granted to intelligent life forms. Further, if a robot starts to develop capacity for independent thought, would it be ethical for humans to regulate it as an object or a property? These questions could be further complicated by the anthropomorphization of robots, which would create an illusion of social bonding between a robot and a human. Unlike the future concerns for human-like AI, social robots already present anthropomorphization concerns for the ethicists today. Humanization of robots can result in more positive attitudes towards robots, but also creates an unrealistic perception of robot capabilities. It would be up to policymakers to weigh the pros and cons of robot humanization and decide whether regulation of robot design and social programming should be implemented.

Autonomy

Autonomy refers to a robot's ability to perform operations and adapt to changes independently from humans. Balancing robot autonomy and human control is one of the core challenges in robotics from both ethical and technical perspectives. This challenge is applicable for all types of robots, and the expectation is that robots should behave autonomously while performing both technical and social tasks. Robots can be assessed based on the amount of autonomy they possess. Several scales have been created to assess autonomy levels in different kinds of robots. For example, Attanasio et. al. have ranked autonomy of surgical robots from 0 to 5, where 0 refers to robotic systems fully operated by the human surgeon, and 5 referring to systems that can perform surgery with no human input. There are currently no surgical robots operating at Level 5 of Autonomy, but there are systems that operate at Level 4. This type of a surgical robot can interpret operative information, devise an action plan, adjust, and execute the plan while operating autonomously under surgeon's supervision. Higher levels of autonomy also bring up questions about moral responsibility and accountability. If a robot has an ability to make decisions, would the robot also be responsible for the consequences? This is an especially important question in the

context of policy and legislation. There are currently several cases in court where Tesla self-driving cars were involved in accidents where people died. These lawsuits might be critical in establishing precedent for accountability policies in autonomous robotics.

European ethical framework for robotics

A study commissioned by the European Parliament Legal Affairs Committee on European civil law in robotics proposed a general ethical framework to be followed in future legislation by the Parliament. The framework focuses on roboethical principles that would protect humans from robots and covers concepts of safety, liberty, privacy, deception, and equality. The 2017 resolution of the European Parliament on the civil law rules on robotics and AI prioritized six main areas for EU legislative efforts: ethics, liability, intellectual property and flow of data, standardization, employment and institutional coordination and oversight. Additionally, the 2017 resolution included recommendations for a code of conduct for robotics scientists, where the role of ethical design and responsible research was recognized.

In a later statement published in 2018, the European Group on Ethics in Science and New Technologies listed roboethics principles that align with the current EU Treaties and the EU Charter of Fundamental Rights. These principles are summarized below:

- *Human dignity* Autonomous technologies must not violate the inherent human right to be respected.
- *Autonomy* Humans are free to live to by their own standards, and humans are responsible to exert control over autonomous technologies. Autonomous technologies must not impair human freedom, responsibility, and control.
- *Responsibility* The development and use of autonomous technologies must benefit society and the environment on a global scale. Such benefits must be defined by democratic means.
- *Justice, equity, and solidarity* Regulators and practitioners must prevent or neutralize discriminatory datasets from training AI systems. AI should further efforts in global justice and equality. All humans should benefit from autonomous technologies.
- *Democracy* The regulation of autonomous technologies must result from democratic, public debate, and engagement.
- *Rule of law and accountability* Regulation of autonomous technologies must uphold all human rights standards, such as protections for safety and privacy. These protections rely on rule of law, access to justice, the right to redress, and the right to a fair trial.
- *Security, safety, and bodily and mental integrity* Safe autonomous systems promote external, internal, and emotional safety. External safety protects environments and users. Internal safety ensures consistent performance and protects against hacking. Emotional safety protects users from exploitation and abuse when interacting with autonomous machines.
- *Data protection and privacy* Digital communication technologies employ autonomous technologies to amass and store vast quantities of users' personal data. Therefore, autonomous technologies challenge protections on personal information and privacy.
- *Sustainability* Autonomous technologies must align with our human responsibility to protect our planet's ability to support life, to preserve the continued quality of the environment, and to maintain the prosperity of our species.

European Policies for Ethical Robotics and AI

European policy may be viewed from both national and international perspectives. Countries in the EU are free to set their own rules, but are also required to follow policy set out by the European Parliament and European Commission. In fact, current opinion of the European Union is that in term of ethical regulation for robotics and AI, there is a need for coordinated action between EU member states and the European Commission. In order to create relevant policy, the EU has

conducted thorough studies that informed the policymakers on current perspectives of stakeholders. This section summarizes these efforts and lays out key aspects of existing robotics regulatory framework in Europe.

Robolaw project

The project produced a report with recommendations for the European Commission on regulating robotics and related technologies. The report details two approaches to robotics legislation:

1) creation of new laws to accommodate the new technology and

2) adaptation of existing laws to reflect technology developments.

In view of the scope of the robotics field, the authors argue that both approaches might need to be employed by policymakers.

European regulatory framework for ethical robotics and AI

The SIENNA Project, an EU initiative aimed at understanding of ethical and human rights challenges posed by new technologies, generated a report that maps existing EU legislation to key legal issues in robotics and AI. Issues of safety, liability, privacy, and equity are amongst the most well-defined by the current laws. Other ethical concerns for robotics, such as legal personhood for advanced AI systems, currently don't have any existing legal framework. The latter is understandable given that AI has not reached that level of advance, but even for ethical issues that do have legal coverage, existing laws are not always specific to robotics. For example, product safety is extensively covered through Directive 2001/95/EC on general product safety and Regulation (EC) No 765/2008 on market surveillance. These regulations were written more than 10 years ago and do not reflect on developments in digital technologies: issues such as connectivity, autonomy, algorithmic opacity, and data dependency are not explicitly discussed in the current legal product safety framework. An additional challenge is presented by the variety of new technologies, where certain types of robots will need to be covered under additional legal frameworks. For example, transportation robots could be regulated though regulations such as Regulation (EU) 2018/858 on approval and surveillance of motor vehicles.

Ethical regulation of research and innovation

Research and innovation comprise one of the main focuses in the EU strategy for sustainable growth and prosperity. Responsible Research and Innovation (RRI) is a new governance model proposed by the EU that puts emphasis on co-creation and co-production with society. RRI framework has three main features:

1. Emphasis on science for society, where research efforts are focused on the "right impacts";

2. Development of mechanisms for reflection and inclusion, where research goals are achieved ethically, inclusively, and democratically; and

3. Responsibility, where RRI framework is applicable not only for researchers, but also entrepreneurs, policymakers, funding organizations, etc.

North American policies for ethical robotics and AI

Unlike Europe, North America does not have a unified governance system, so this section will explore roboethics principles and policies in the United States of America and Canada.

American ethical frameworks for robotics and AI

On February 11, 2019 President Trump signed an executive order which presents five principles for the American AI Initiative. These principles reflect on several core themes:

1. Commitment to AI development and implementation to support economic competitiveness and national security;

2. Creation of technical standards for AI deployment and adaptation;

3. Education of the public to help people develop skills necessary in the future (when AI will become more prominent);

4. Fostering of public trust and confidence in AI technologies and protection of civil liberties, privacy, and American values;

5. Creation of international environment that would support American AI industries.

Canadian ethical frameworks for robotics and AI

The Canadian government has recognized the opportunities that AI and digital technologies present for the future of governance. The Digital Government program lists five guiding principles to ensure effective and ethical use of AI. Under these principles, the government will.

1. Understand and measure the impact of AI,

2. Be transparent about how and when AI is being used,

3. Provide explanations on AI decision making,

4. Be transparent about sharing technical details, and

5. Provide sufficient training on the use of AI solutions

International policies for ethical robotics and AI

The ethical frameworks and principles identified by the international organizations are of advisory nature and cannot be enforced, but they can serve as a starting point for individual governments to create their own AI and robotics strategies. In other words, there are currently no international policies for ethical AI and robotics. However, international policy might be of the most importance considering the impact and international commercialization of robots and AI. If robots and AI are to be shipped and implemented around the world, a global policy could ensure that they are being used safely and ethically.

The goal is to ensure that AI systems are designed and used in ways that are ethical, fair, and beneficial to society as a whole.

Key principles that are often considered in this context include:

1. **Transparency**: AI systems should be transparent and explainable, so that users and stakeholders can understand how decisions are made.
2. **Accountability**: There should be mechanisms in place to hold individuals and organizations accountable for the decisions and actions of AI systems.
3. **Fairness**: AI systems should be designed and used in ways that are fair and unbiased, and that do not discriminate against individuals or groups.
4. **Privacy**: AI systems should respect and protect the privacy of individuals, and should only collect and use personal data in ways that are transparent and lawful.
5. **Safety and Security**: AI systems should be designed and deployed in ways that are safe and secure, and that do not pose undue risks to individuals or society.
6. **Human Control**: Humans should have ultimate control over AI systems, and should be able to intervene if the system behaves inappropriately or unexpectedly.
7. **Societal Benefit**: AI systems should be designed and used to promote the well-being of individuals and society as a whole, and to enhance human capabilities.
8. **Environmental Sustainability**: AI development and deployment should consider the environmental impact, and strive to minimize negative effects on the environment.

Harmonization of these principles involves bringing together different stakeholders, including governments, industry, academia, and civil society, to agree on a common set of ethical guidelines for AI. This can help to ensure that AI technology is developed and used in ways that are consistent with ethical values and principles, and that promote the public good.

4.7 Ethics and Professional Responsibility

4.7.1 What Is Professional Ethics?

The scope of the term "computer ethics" varies considerably. It can include such social and political issues as the impact of computers on employment, the environmental impact of computers, whether or not to sell computers to totalitarian governments, use of computers by the military, and the consequences of the technological and thus economic divisions between developed countries and poor countries. It can include personal dilemmas about what to post on the Internet and what to download. In this chapter we focus more narrowly on a category of professional ethics, similar to medical, legal, and accounting ethics, for example. We consider ethical issues a person might encounter as a computer professional, on the job. Professional ethics includes relationships with and responsibilities toward customers, clients, coworkers, employees, employers, others who use one's products and services, and others whom they affect. We examine ethical dilemmas and guidelines related to actions and decisions of individuals who create and use computer systems. We look at situations where you must make critical decisions, situations where significant consequences for you and others could result.

Extreme examples of lapses in ethics in many fields regularly appear in the news. In business, we had Enron, for example. In journalism, we have had numerous incidents of journalists at prominent news organizations plagiarizing or inventing stories. In scienc e, a famed and respectedresearcher published falsified stem cell research and claimed accomplishments he had not achieved. A writer invented dramatic events in what he promoted as a factual memoir of his experiences. These examples involve blatant dishonesty, which is almost always wrong.

Honesty is one of the most fundamental ethical values. We all make hundreds of decisions all day long. The consequences of some decisions are minor. Others are huge and affect people we never meet. We base decisions, partly, on the information we have. (It takes ten minutes to drive to work. This software has serious security vulnerabilities. What you post on a social-network site is

available only to your designated friends.) We pick up bits and pieces of information from explicit research, from conversations, and from our surroundings and regular activities. Of course, not all of it is accurate. But we must base our choices and actions on what we know. A lie deliberately sabotages this essential activity of being human: absorbing and processing information and making choices to pursue our goals. Lies are often attempts to manipulate people. As Kant would say, a lie treats people as merely means to ends, not ends in themselves. Lies can have many negative consequences. In some circumstances, lying casts doubt on the work or word of other people unjustly. Thus it hurts those people, and it adds unnecessary uncertainty to decisions by others who would have acted on the word of people the lie contradicts. Falsifying research or other forms of work is an indirect form of theft of research funds and salary. It wastes resources that others could have used productively. It contributes to incorrect choices and decisions by people who depend on the results of the work. The costs and indirect effects of lies can cascade and do much harm.

4.7.2 Ethical Guidelines for Computer Professionals

Special aspects of Professional Ethics

Professional ethics have several characteristics different from general ethics. The role of the professional is special in several ways. First, the professional is an expert in a field, be it computer science or medicine, that most customers know little about. Most of the people affected by the devices, systems, and services of professionals do not understand how they work and cannot easily judge their quality and safety. This creates special responsibilities for the professional. Customers rely on the knowledge, expertise, and honesty of the professional. A professional advertises his or her expertise and thus has an obligation to provide it. Second, the products of many professionals (e.g., highway bridges, investment advice, surgery protocols, and computer systems) profoundly affect large numbers of people. A computer professional's work can affect the life, health, finances, freedom, and future of a client or members of the public. A professional can cause great harm through dishonesty, carelessness, or incompetence. Often the victims have little ability to protect themselves. The victims, often, are not the direct customers of the professional and have no direct control or decision-making role in choosing the product or making decisions about its quality and safety. Thus, computer professionals have special responsibilities not only to their customers, but also to the general public, to the users of their products, regardless of whether they have a direct relationship with the users. These responsibilities include thinking about potential risks to privacy and security of data, safety, reliability, and ease of use. They include taking action to diminish risks that are too high.

PROFESSIONAL CODES OF ETHICS

Many professional organizations have codes of professional conduct. They provide a general statement of ethical values and remind people in the profession that ethical behavior is an essential part of their job. The codes provide reminders about specific professional responsibilities. They provide valuable guidance for new or young members of the profession who want to behave ethically but do not know what is expected of them, people whose limited experience has not prepared them to be alert to difficult ethical situations and to handle them appropriately.

There are several organizations for the range of professions included in the general term computer professional. The main ones are the ACM and the IEEE Computer Society (IEEE CS).1 They developed the Software Engineering Code of Ethics and Professional Practice (adopted jointly by the ACM and IEEE CS) and the ACM Code of Ethics and Professional Conduct (both in Appendix A). We refer to sections of the Codes in the following discussion and in Section 9.3, using the shortened names SE Code and ACM Code. The Codes emphasize the basic ethical values of honesty and fairness.∗ They cover many aspects of professional behavior, including the

responsibility to respect confidentiality,† maintain professional competence,‡ be aware of relevant laws,§ and honor contracts and agreements.¶ In addition, the Codes put special emphasis on areas that are particularly (but not uniquely) vulnerable from computer systems. They stress the responsibility to respect and protect privacy, avoid harm to others,∗∗ and respect property rights (with intellectual property and computer systems themselves as the most relevant examples).†† The SE Code covers many specific points about software development. It is translated into several languages, and various organizations have adopted it as their internal professional standard.

Managers have special responsibility because they oversee projects and set the ethical standards for employees. Principle 5 of the SE Code includes many specific guidelines for managers.

Guidelines and Professional Responsiblities

We highlight a few principles for producing good systems. Most concern software developers, programmers, and consultants. A few are for professionals in other areas who make decisions about acquiring computer systems for large organizations.

Understand what success means. After the utter foul-up on opening day at Kuala Lumpur's airport, blamed on clerks typing incorrect commands, an airport official said, "There's nothing wrong with the system." His statement is false, and the attitude behind the statement contributes to the development of systems that will fail. The official defined the role of the airport system narrowly: to do certain data manipulation correctly, assuming all input is correct. Its true role was to get passengers, crews, planes, luggage, and cargo to the correct gates on schedule. It did not succeed. Developers and institutional users of computer systems must view the system's role and their responsibility in a wide enough context.

Include users (such as medical staff, technicians, pilots, office workers) in the design and testing stages to provide safe and useful systems. There are numerous "horror stories" in which technical people developed systems without sufficient knowledge of what was important to users. For example, a system for a newborn nursery at a hospital rounded each baby's weight to the nearest pound. For premature babies, the difference of a few ounces is crucial information. The responsibility of developers to talk to users is not limited to systems that affect safety and health. Systems designed to manage stories for a news Web site, to manage inventory in a toy store, or to organize documents and video on a Web site could cause frustration, waste a client's money, and end up in the trash heap if designed without sufficient consideration of the needs of actual users.

The box on the next page illustrates more ways to think about your users.

Do a thorough, careful job when planning and scheduling a project and when writing bids or contracts. This includes, among many other things, allocating sufficient time and budget for testing and other important steps in the development process. Inadequate planning is likely to lead to pressure to cut corners later.

Design for real users. We have seen several cases where computers crashed because someone typed input incorrectly. In one case, an entire pager system shut down because a technician did not press the Enter key (or did not hit it hard enough). Real people make typos, get confused, or are new at their job. It is the responsibility of the system designers and programmers to provide clear user interfaces and include appropriate checking of input. It is impossible for computers to detect all incorrect input, but there are techniques for catching many kinds of errors and for reducing the damage that errors cause.

Don't assume existing software is safe or correct. If you use software from another application, verify its suitability for the current project. If the software was designed for an application where the degree of harm from a failure was small, the quality and testing standards might not have been as high as necessary in the new application. The software might have confusing user interfaces that were tolerable (though not admirable) in the original application but could have serious negative consequences in the new application.

Be open and honest about capabilities, safety, and limitations of software. In several cases, there is a strong argument that the treatment of customers was dishonest. Honesty of salespeople is hardly a new issue. The line between emphasizing your best qualities and being dishonest is not always clear, but it should be clear that hiding known, serious flaws and lying to customers are on the wrong side of the line.

Honesty includes taking responsibility for damaging or injuring others. If you break a neighbour's window playing ball or smash into someone's car, you have an obligation to pay for the damage. If a business finds that its product caused injury, it should not hide that fact or attempt to put the blame on others.

Be open and honest about capabilities, safety, and limitations of software. There is a strong argument that the treatment of customers was dishonest. Honesty of salespeople is hardly a new issue. The line between emphasizing your best qualities and being dishonest is not always clear, but it should be clear that hiding known, serious flaws and lying to customers are on the wrong side of the line.

Honesty includes taking responsibility for damaging or injuring others. If you br reak a neighbour's window playing ball or smash into someone's car, you have an obligation to pay for the damage. If a business finds that its product caused injury, it should not hide that fact or attempt to put the blame on others.

Honesty about system limitations is especially important forexpert systems, or decision systems, that is, systems that use models and heuristics incorporating expert knowledge to guide decision making (for example, medical diagnoses or investment planning). Developers must explain the limitations and uncertainties to users (doctors, financial advisors, and so forth, and to the public when appropriate). Users must not shirk responsibility for understanding them and using the systems properly.

Require a convincing case for safety. One of the most difficult ethical problems that arises in safety-critical applications is deciding how much risk is acceptable. Burning gases that leaked from a rocket shortly after launch destroyed the space shuttle Challenger, killing the seven people aboard. A comment from one of the engineers who opposed the launch sheds some light on how subtle shifts in attitude can affect a decision. The night before the scheduled launch, the engineers argued for a delay. They knew the cold weather posed a severe threat to the shuttle. We cannot prove absolutely that a system is safe, nor can we usually prove absolutely that it will fail and kill someone. The engineer reported that, in the case of the Challenger, "It was up to us to prove beyond a shadow of a doubt that it was not safe to [launch]." This, he said, was the total reverse of a usual Flight Readiness Review.5 For the ethical decision maker, the policy should be to suspend or delay use of the system in the absence of a convincing case for safety, rather than to proceed in the absence of a convincing case for disaster.

Pay attention to defaults. Everything, it seems, is customizable: the level of encryption on a cell phone or wireless network, whether consumers who buy something at a Web site will go on an e-mail list for ads, the difficulty level of a computer game, the type of news stories your favorite news site displays for you, what a spam filter will filter out. So the default settings might not seem important. They are. Many people do not know about the options they can control. They do not

understand issues of security. They often do not take the time to change settings. System designers should give serious thought to default settings. Sometimes protection (of privacy or from hackers, for example) is the ethical priority. Sometimes ease of use and compatibility with user expectations is a priority. Sometimes priorities conflict.

Develop communications skills. A computer security consultant told me that often when he talks to a client about security risks and the products available to protect against them, he sees the client's eyes glaze over. It is a tricky ethical and professional dilemma for him to decide just how much to say so that the client will actually hear and absorb it.

There are many situations in which a computer professional has to explain technical issues to customers and coworkers. Learning how to organize information, distinguishing what is important to communicate and what is not, engaging the listener actively in the conversation to maintain interest, and so on, will help make one's presentations more effective and help to ensure that the client is truly informed.

4.7.3 Scenarios

Introduction and Methodology

Although we will not follow the outline below step by step for all the scenarios, our discussions will usually include many of these elements:

1. Brainstorming phase

❖ List all the people and organizations affected. (They are the stakeholders.)

❖ List risks, issues, problems, and consequences.

❖ List benefits. Identify who gets each benefit.

❖ In cases where there is no simple yes or no decision, but rather one has to choose some action, list possible actions.

2. Analysis phase

❖ Identify responsibilities of the decision maker. (Consider responsibilities of both general ethics and professional ethics.)

❖ Identify rights of stakeholders.

❖ Consider the impact of the action options on the stakeholders. Analyze consequences, risks, benefits, harms, costs for each action considered.

❖ Find sections of the SE Code or the ACM Code that apply. Then, categorize each potential action or response as ethically obligatory, ethically prohibited, or ethically acceptable.

❖ If there are several ethically acceptable options, select an option, considering the ethical merits of each, courtesy to others, practicality, self-interest, personal preferences, and so on. (In some cases, plan a sequence of actions, depending on the response to each.)

Protecting Personal Data

Your customer is a community clinic. The clinic works with families that have problems of family violence. It has three sites in the same city, including a shelter for battered women and children. The director wants a computerized record system, networked for the three sites, with the ability to transfer files among sites and make appointments at any site for any other. She wants to have an Internet connection for routine Web access and e-mail communication with other social service agencies about client needs. She wants a few laptop computers on which staffers can carry

records when they visit clients at home. At the shelter, staffers use only first names for clients, but the records contain last names and forwarding addresses of women who have recently left. The clinic's budget is small, and she wants to keep the cost as low as possible.

The clinic director is likely to be aware of the sensitivity of the information in the records and to know that inappropriate release of information can result in embarrassment for families using the clinic and physical harm to women who use the shelter. But she might not be aware of the risks of a computer system. You, as the computer professional, have specialized knowledge in this area. It is as much your obligation to warn the director of the risks as it is that of a physician to warn a patient of side effects of a drug he or she prescribes.

The most vulnerable stakeholders here are the clients of the clinic and their family members, and they are not involved in your negotiations with the director. You, the director, the clinic employees, and the donors or agencies that fund the clinic are also stakeholders.

Suppose you warn the director about unauthorized access to sensitive information by hackers and the potential for interception of records and e-mail during transmission. You suggest measures to protect client privacy, including, for example, an identification code system (not Social Security number) for clients of the clinic to use when real names are not necessary and encryption for e-mail and transmission of records. You recommend security software to reduce the threat of hackers who might steal data. You tell the director that carrying client records on laptops has serious risks, citing examples of loss and theft of laptops containing large amounts of sensitive personal data. You advise that records on laptops be encrypted and suggest that the director buy laptops with thumbprint readers so that only authorized employees can access the data. You warn that staffers might be bribed to sell or release information from the system. (Suppose a client is a candidate for the city council or a party in a child-custody case.) You suggest procedures to reduce such leaks. They include a user ID and password for each staff member, coded to allow access only to information that the particular worker needs, a log function that keeps track of who accessed and modified the records, and monitoring and controls on employee e-mail and Web activity. Note that your ability to provide these suggestions is dependent on your professional competence, currency in the field, and general awareness of relevant current events.

The features you recommend will make the system more expensive. If you convince the director of the importance of your recommendations, and she agrees to pay the cost, your professional/ethical behaviour has helped improve the security of the system and protect client privacy.

Suppose the director says the clinic cannot afford all the security features. She wants you to develop the system without them. You have several options. You can develop a cheap, but vulnerable, system. You can refuse and perhaps lose the job (although your refusal might convince the director of the importance of the security measures and change her mind). You can add security features and not charge for them. You can work out a compromise that includes the protections you consider essential. All but the first option are pretty clearly ethically acceptable. What about the first? Should you agree to provide the system without the security you believe it should have? Is it now up to the director alone to make an informed choice, weighing the risks and costs? In a case where only the customer would take the risk, some would say yes, it is your job to inform, no more. Others would say that the customer lacks the professional expertise to evaluate the risks. In this scenario, however, the director is not the only person at risk, nor is the risk to her the most significant risk of an insecure system. You have an ethical responsibility to consider the potential harm to clients from exposure of sensitive information and not to build a system without adequate privacy protection.

The most difficult decision may be deciding what is adequate. Encryption of personal records on the laptops might be essential. Monitoring employee Web access is probably not. There is not always a sharp, clear line between sufficient and insufficient protection. You will have to rely on your professional knowledge, on being up-to-date about current risks and security measures, on good judgment, and perhaps on consulting others who develop systems for similar applications.

Designing an E-Mail System with Targeted ADS

Your company is developing a free e-mail service that will include targeted advertising based on the content of the e-mail messages—similar to Google's Gmail. You are part of the team designing the system. What are your ethical responsibilities?

Obviously you must protect the privacy of e-mail. The company plans a sophisticated text analysis system to scan e-mail messages and select appropriate ads. No human will read the messages. Marketing for the free e-mail will make clear that users will see targeted ads. The privacy policy will explain that the content of the e-mail will determine which ads appear. So, the marketing director contends, you have satisfied the first principle of privacy protection, informed consent. What else must you consider to meet your ethical responsibility in offering this service to the public?

The fact that software, not a person, scans the e-mail messages and assigns the ads reduces privacy threats. However, we now know that companies store huge amounts of data. What will this system store? Will it store data about which ads it displayed to specific users? Will it store data about which key words or phrases in e-mails cause particular ads to be selected? Will it store data about who clicked on specific ads? Why are these questions of ethical concern? Because we know that leaks, theft, or demands by a government agency might compromise the privacy of such data. The set of ads displayed to a particular user could provide a lot of information about the person, just as one's search queries do. Some of it will be incorrect or misleading information because of quirks in the ad-targeting methods.

Should we insist that no such data be stored? Not necessarily. Some of it might have important uses. Some records are necessary for billing advertisers, some for analysis to improve ad-targeting strategies, and perhaps some for responding to complaints from e-mail users or advertisers.

The system design team needs to determine what records are necessary, which need to be associated with individual users, how long the company will store them, how it will protect them (from hackers, accidental leaks, and so on), and under what conditions it will disclose them.

Now, back up and reconsider informed consent. Telling customers that they will see ads based on the content of their e-mail is not sufficient if the system stores data that can link a list of ads with a particular user. You must explain this to potential users in a privacy policy or user agreement. But we know that most people do not read privacy policies and user agreements, especially long ones. A click might mean legal consent, but ethical responsibility goes farther. Independent of what is in the agreement, the designers must think about potential risks of the system, consider privacy throughout the planning process, and design in protections.

Copyright Violation

Your company has 25 licenses for a computer program, but you discover that it has been copied onto 80 computers.

The first step here is to inform your supervisor that the copies violate the license agreement. Suppose the supervisor is not willing to take any action? What next? What if you bring the problem to the attention of higher level people in the company and no one cares? There are several possible actions: Give up; you did your best to correct the problem. Call the software vendor and report the offense. Quit your job.

Is giving up at this point ethically acceptable? My students thought it depended in part on whether you are the person who signed the license agreements. If so, you have made an agreement about the use of the software, and you, as the representative of your company, are obligated to honor it. Because you did not make the copies, you have not broken the agreement directly, but you have responsibility for the software. Your name on the license could expose you to legal risk, or unethical managers in your company could make you a scapegoat. Thus, you might prefer to report the violation or quit your job and have your name removed from the licenses to protect yourself. If you are not the person who signed the licenses, then you observed a wrong and brought it to the attention of appropriate people in the company.

Release of Personal Information

You work for the IRS, the Social Security Administration, a movie-rental company, or an Internet service provider. Someone asks you to get a copy of records about a particular person. He will pay you $500.

Who are the stakeholders? You: You have an opportunity to make some extra money. The person seeking the records: Presumably he has something to gain. The person whose records the briber wants: Providing the information invades his or her privacy. All the people about whom the company or agency has personal information: If you sell information about one person, chances are you will sell more if asked in the future. Your employer (if a private company): If the sale becomes known, the victim might sue the company. If such sales of information become common, the company will acquire a reputation for carelessness and will potentially lose business and lawsuits.

There are many alternative actions open to you: Sell the records. Refuse and say nothing about the incident. Refuse and report the incident to your supervisor. Refuse and report to the police. Contact the person whose information the briber wants and tell him or her of the incident. Agree to sell the information, but actually work with the police to collect evidence to convict the person trying to buy it.

Are any of these alternatives ethically prohibited or obligatory? The first option, selling the records, is clearly wrong. It almost certainly violates rules and policies you have agreed to abide by in accepting your job. As an employee, you must abide by the guarantees of confidentiality the company or agency has promised its customers or the public. Depending on the use made of the information you sell, you could be helping to cause serious harm to the victim. Disclosing the information might be illegal. Your action might expose your employer to fines. If someone discovers the leak, the employer and the police might suspect another employee, who could face arrest and punishment.

Some would argue that selling the records is wrong because it violates the privacy of the victim, but recall that the boundaries of privacy are unclear because they can conflict with freedom of speech and reasonable flow of information. If you happened to know the victim, and knew some of the same information in the records, you might not be under an ethical obligation to keep the information secret. The essential element that makes selling the information wrong in this scenario is your position of trust as an employee in a company or agency that maintains the information. The risks are greater for sensitive information, but your obligation extends to any information the company has promised to keep confidential.

Conflict of Interest

You have a small consulting business. The CyberStuff company plans to buy software to run a new collaborative content-sharing Web site. CyberStuff wants to hire you to evaluate bids from vendors. Your spouse works for NetWorkx and did most of the work in writing the bid that NetWorkx plans to submit. You read the bid while your spouse was working on it and you think it is excellent. Do you tell CyberStuff about your spouse's connection with NetWorkx?

Conflict-of-interest situations occur in many professions. Sometimes the ethical course of action is clear. Sometimes, depending on your connection with the people or organizations your action affects, it can be more difficult to determine.

I have seen two immediate reactions to scenarios similar to this one (in discussions among professionals and among students). One is that it is a simple case of profits versus honesty, and ethics requires that you inform the company about your connection to the software vendor. The other is that if you honestly believe you can be objective and fairly consider all bids, you have no ethical obligation to say anything. Which is right? Is this a simple choice between saying nothing and getting the consulting job or disclosing your connection and losing the job?

The affected parties are the CyberStuff company, yourself, your spouse, your spouse's company, and the other companies whose bids you will be reviewing. A key factor in considering consequences is that we do not know whether CyberStuff will later discover your connection to one of the bidders. If you say nothing about the conflict of interest, you benefit, because you get the consulting job. If you recommend NetWorkx (because you believe its bid is the best), it benefits from a sale. However, if CyberStuff discovers the conflict of interest later, your reputation for honesty—important to a consultant—will suffer. The reputation of your spouse's company could also suffer. Note that even if you conclude that you are truly unbiased and do not have an ethical obligation to tell CyberStuff about your connection to your spouse's company, your decision might put NetWorkx's reputation for honesty at risk. The appearance of bias can be as damaging (to you and to NetWorkx) as actual bias.

Suppose you take the job and you find that one of the other bids is much better than the bid from NetWorkx. Are you prepared to handle that situation ethically? What are the consequences of disclosing the conflict of interest to the client now? You will probably lose this particular job, but they might value your honesty more highly and that might get you more business in the future. Thus, there could be benefits, even to you, from disclosing the conflict of interest.

Suppose it is unlikely that anyone will discover your connection to NetWorkx. What are your responsibilities to your potential client as a professional consultant? When someone hires you as a consultant, they expect you to offer unbiased, honest, impartial professional advice. There is an implicit assumption that you do not have a personal interest in the outcome or a personal reason to favor one of the bids you will review. The conclusion in this case hangs on this point. In spite of your belief in your impartiality, you could be unintentionally biased. It is not up to you to make the decision about whether you can be fair. The client should make that decision. Your ethical obligation in this case is to inform CyberStuff of the conflict of interest.

4.8 Roboethics Taxonomy

In the period of a year, the Euron Roboethics Atelier carried out a tour d'horizon of the field in Robotics: an overview of the state of the art in Robotics, and of the main ethical issues, driven by the most recent technoscientific developments, which can only just be glimpsed.

A taxonomy of Robotics is not a simple task, simply because the field is in a full bloom.

A classification of Robotics is a work in progress, done simultaneously with the development of the discipline itself.

Aware of the classifications produced by the main Robotics organizations, which differ from one another on the basis of the approach – technological/applicational -, we have preferred, in the case of the Roboethics Roadmap, to collect the many Robotics fields from a typological standpoint, according to shared homogeneity of the problems of interface towards the society.

Instead of an encyclopaedic approach, we have followed - with few modifications - the classification of **EURON Robotics Research Roadmap.**

For every field, we have tried to analyse the current situation rather than the imaginable. Thus, we have decided to give priority to issues in applied ethics rather than to theoretical generality.

- Humanoids
  Artificial Mind, Artificial Body
- Advanced production systems
  Industrial robotics
- Adaptive robot servants and intelligent homes
  Indoor Service Robots, Ubiquitous Robotics
- Network Robotics
  Internet Robotics, Robot ecology
- Outdoor Robotics
  Land, Sea, Air, Space
- Health Care and Life Quality
  Surgical Robotics, Bio-Robotics, Assistive Technology
- Military Robotics
  Intelligent Weapons, Robot Soldiers, Superhumans
- Edutainment
  Educational Robots, Robot Toys, Entertainment, Robotic Art

Here is a simple taxonomy for Roboethics:

1. **Ethical Design**: Focuses on ensuring robots are designed and programmed in a way that aligns with ethical principles and values.
2. **Legal and Regulatory Issues**: Deals with laws and regulations that govern the use of robots, including liability, privacy, and safety.
3. **Societal Impact**: Considers how robots impact society, including issues related to employment, economics, and culture.
4. **Human-Robot Interaction**: Studies how humans and robots interact, including issues related to trust, empathy, and communication.
5. **Safety and Security**: Focuses on ensuring robots are safe to use and cannot be easily hacked or manipulated.
6. **Autonomy and Responsibility**: Explores the degree of autonomy robots should have and who is responsible when something goes wrong.
7. **Privacy and Data Protection**: Addresses issues related to the collection, use, and protection of personal data by robots.
8. **Transparency and Accountability**: Emphasizes the importance of transparency in thedesign and use of robots, as well as holding individuals and organizations accountable for theiractions.

UNIT V
AI AND ETHICS- CHALLENGES AND OPPORTUNITIES

Challenges - Opportunities- ethical issues in artificial intelligence- Societal Issues Concerning the Application of Artificial Intelligence in Medicine- decision-making role in industries-National and International Strategies on AI.

5.1 Challenges in AI and Ethics

Today, artificial intelligence is essential across a wide range of industries, including healthcare, retail, manufacturing, and even government. But there are ethical challenges with AI, and as always, we need to stay vigilant about these issues to make sure that artificial intelligence isn't doing more harm than good. Here are some of the biggest ethical challenges of artificial intelligence.

Biases

We need data to train our artificial intelligence algorithms, and we need to do everything we can to eliminate bias in that data.The ImageNet database, for example, has far more white faces than non-white faces. When we train our AI algorithms to recognize facial features using a database that doesn't include the right balance of faces, the algorithm won't work as well on non-white faces, creating a built-in bias that can have a huge impact. I believe it's important that we eliminate as much bias as possible as we train our AI, instead of shrugging our shoulders and assuming that we're training our AI to accurately reflect our society. That work begins with being aware of the potential for bias in our AI solutions.

Control and the Morality of AI

As we use more and more artificial intelligence, we are asking machines to make increasingly important decisions. For example, right now, there is an international convention that dictates the use of autonomous drones. If you have a drone that could potentially fire a rocket and kill someone, there needs to be a human in the decision-making process before the missile gets deployed. So far, we have gotten around some of the critical control problems of AI with a patchwork of rules and regulations like this. The problem is that AIs increasingly have to make split-second decisions. For example, in high-frequency trading, over 90% of all financial trades are now driven by algorithms, so there is no chance to put a human being in control of the decisions. The same is true for autonomous cars. They need to react immediately if a child runs out on the road, so it's important that the AI is in control of the situation. This creates interesting ethical challenges around AI and control.

Privacy

Privacy (and consent) for using data has long been an ethical dilemma of AI. We need data to train AIs, but where does this data come from, and how do we use it? Sometimes we make the assumption that all the data is coming from adults with full mental capabilities that can make choices for themselves about the use of their data, but we don't always have this. For example, Barbie now has an AI-enabled doll that children can speak to. What does this mean in terms of ethics? There is an algorithm that is collecting data from your child's conversations with this toy. Where is this data going, and how is it being used? As we have seen a lot in the news recently, there are also many

companies that collect data and sell it to other companies. What are the rules around this kind of data collection, and what legislation might need to be put in place to protect users' private information?

Power Balance

Huge companies like Amazon, Facebook, Google, are using artificial intelligence to squash their competitors and become virtually unstoppable in the marketplace. Countries like China also have ambitious AI strategies that are supported by the government. President Putin of Russia has said, "Whoever wins the race in AI will probably become the ruler of the world." How do we make sure the monopolies we're generating are distributing wealth equally and that we don't have a few countries that race ahead of the rest of the world? Balancing that power is a serious challenge in the world of AI.

Ownership

Who is responsible for some of the things that AIs are creating?

We can now use artificial intelligence to create text, bots, or even deepfake videos that can be misleading. Who owns that material, and what do we do with this kind of fake news if it spreads across the internet? We also have AIs that can create art and music. When an AI writes a new piece of music, who owns it? Who has the intellectual property rights for it, and should potentially get paid for it?

Environmental Impact

Sometimes we don't think about the environmental impact of AI. We assume that we are using data on a cloud computer to train an algorithm, and then that data is used to run recommendation engines on our website. However, the computer centers that run our cloud infrastructure are power-hungry. Training in AI, for example, can create 17 times more carbon emissions than the average American does in about a year. How can we use this energy for the highest good and use AI to solve some of the world's biggest and most pressing problems? If we are only using artificial intelligence because we can, we might have to reconsider our choices.

Humanity

My final challenge is "How does AI make us feel as humans?" Artificial intelligence has now gotten so fast, powerful, and efficient that it can leave humans feeling inferior. This issue may challenge us to think about what it actually means to be human. AI will also continue to automate more of our jobs. What will our contribution be, as human beings? I don't think artificial intelligence will ever replace all our jobs, but AI will augment them. We need to get better at working alongside smart machines so we can manage the transition with dignity and respect for people and technology. These are some of the key ethical challenges that we all need to think about very carefully when it comes to AI.

5.2 Opportunities in AI and Ethics

Optimizing logistics, detecting fraud, composing art, conducting research, providing translations: intelligent machine systems are transforming our lives for the better. As these systems become more capable, our world becomes more efficient and consequently richer.

Tech giants such as Alphabet, Amazon, Facebook, IBM and Microsoft – as well as individuals like Stephen Hawking and Elon Musk – believe that now is the right time to talk about the nearly boundless landscape of artificial intelligence. In many ways, this is just as much a new frontier for ethics and risk assessment as it is for emerging technology. So which issues and conversations keep AI experts up at night?

1. Unemployment. What happens after the end of jobs?

The hierarchy of labour is concerned primarily with automation. As we've invented ways to automate jobs, we could create room for people to assume more complex roles, moving from the physical work that dominated the pre-industrial globe to the cognitive labour that characterizes strategic and administrative work in our globalized society.

Look at trucking: it currently employs millions of individuals in the United States alone. What will happen to them if the self-driving trucks promised by Tesla's Elon Musk become widely available in the next decade? But on the other hand, if we consider the lower risk of accidents, self-driving trucks seem like an ethical choice. The same scenario could happen to office workers, as well as to the majority of the workforce in developed countries.

This is where we come to the question of how we are going to spend our time. Most people still rely on selling their time to have enough income to sustain themselves and their families. We can only hope that this opportunity will enable people to find meaning in non-labour activities, such as caring for their families, engaging with their communities and learning new ways to contribute to human society.

If we succeed with the transition, one day we might look back and think that it was barbaric that human beings were required to sell the majority of their waking time just to be able to live.

2. Inequality. How do we distribute the wealth created by machines?

Our economic system is based on compensation for contribution to the economy, often assessed using an hourly wage. The majority of companies are still dependent on hourly work when it comes to products and services. But by using artificial intelligence, a company can drastically cut down on relying on the human workforce, and this means that revenues will go to fewer people. Consequently, individuals who have ownership in AI-driven companies will make all the money.

We are already seeing a widening wealth gap, where start-up founders take home a large portion of the economic surplus they create. In 2014, roughly the same revenues were generated by the three biggest companies in Detroit and the three biggest companies in Silicon Valley ... only in Silicon Valley there were 10 times fewer employees. If we're truly imagining a post-work society, how do we structure a fair post-labour economy?

3. Humanity. How do machines affect our behaviour and interaction?

Artificially intelligent bots are becoming better and better at modelling human conversation and relationships. In 2015, a bot named Eugene Goostman won the Turning challenge for the first time. In this challenge, human raters used text input to chat with an unknown entity, then guessed

whether they had been chatting with a human or a machine. Eugene Goostman fooled more than half of the human raters into thinking they had been talking to a human being.

This milestone is only the start of an age where we will frequently interact with machines as if they are humans; whether in customer service or sales. While humans are limited in the attention and kindness that they can expend on another person, artificial bots can channel virtually unlimited resources into building relationships.

Even though not many of us are aware of this, we are already witnesses to how machines can trigger the reward centres in the human brain. Just look at click-bait headlines and video games. These headlines are often optimized with A/B testing, a rudimentary form of algorithmic optimization for content to capture our attention. This and other methods are used to make numerous video and mobile games become addictive.

On the other hand, maybe we can think of a different use for software, which has already become effective at directing human attention and triggering certain actions. When used right, this could evolve into an opportunity to nudge society towards more beneficial behavior. However, in the wrong hands it could prove detrimental.

4. Artificial stupidity. How can we guard against mistakes?

Intelligence comes from learning, whether you're human or machine. Systems usually have a training phase in which they "learn" to detect the right patterns and act according to their input. Once a system is fully trained, it can then go into test phase, where it is hit with more examples and we see how it performs.

Obviously, the training phase cannot cover all possible examples that a system may deal with in the real world. These systems can be fooled in ways that humans wouldn't be. For example, random dot patterns can lead a machine to "see" things that aren't there. If we rely on AI to bring us into a new world of labour, security and efficiency, we need to ensure that the machine performs as planned, and that people can't overpower it to use it for their own ends.

5. Rocist Robots. How do we eliminate AI Bias?

Though artificial intelligence is capable of a speed and capacity of processing that's far beyond that of humans, it cannot always be trusted to be fair and neutral. Google and its parent company Alphabet are one of the leaders when it comes to artificial intelligence, as seen in Google's Photos service, where AI is used to identify people, objects and scenes. But it can go wrong, such as when a camera missed the marks on racial sensitivity, or when a software used to predict future criminals showed bias against black people.

We shouldn't forget that AI systems are created by humans, who can be biased and judgemental. Once again, if used right, or if used by those who strive for social progress, artificial intelligence can become a catalyst for positive change.

6. Security. How do we keep AI safe from adversaries?

The more powerful a technology becomes, the more can it be used for nefarious reasons as well as good. This applies not only to robots produced to replace human soldiers, or autonomous weapons, but to AI systems that can cause damage if used maliciously. Because these fights won't be fought on the battleground only, cybersecurity will become even more important. After all, we're dealing with a system that is faster and more capable than us by orders of magnitude.

5.3 Ethical issues in artificial intelligence

1. Unemployment. What happens after the end of jobs?

The hierarchy of labour is concerned primarily with automation. As we've invented ways to automate jobs, we could create room for people to assume more complex roles, moving from the physical work that dominated the pre-industrial globe to the cognitive labour that characterizes strategic and administrative work in our globalized society.

Look at trucking: it currently employs millions of individuals in the United States alone. What will happen to them if the self-driving trucks promised by Tesla's Elon Musk become widely available in the next decade? But on the other hand, if we consider the lower risk of accidents, self-driving trucks seem like an ethical choice. The same scenario could happen to office workers, as well as to the majority of the workforce in developed countries.

This is where we come to the question of how we are going to spend our time. Most people still rely on selling their time to have enough income to sustain themselves and their families. We can only hope that this opportunity will enable people to find meaning in non-labour activities, such as caring for their families, engaging with their communities and learning new ways to contribute to human society.

If we succeed with the transition, one day we might look back and think that it was barbaric that human beings were required to sell the majority of their waking time just to be able to live.

2. Inequality. How do we distribute the wealth created by machines?

Our economic system is based on compensation for contribution to the economy, often assessed using an hourly wage. The majority of companies are still dependent on hourly work when it comes to products and services. But by using artificial intelligence, a company can drastically cut down on relying on the human workforce, and this means that revenues will go to fewer people. Consequently, individuals who have ownership in AI-driven companies will make all the money.

We are already seeing a widening wealth gap, where start-up founders take home a large portion of the economic surplus they create. In 2014, roughly the same revenues were generated by the three biggest companies in Detroit and the three biggest companies in Silicon Valley ... only in Silicon Valley there were 10 times fewer employees.

3. Humanity. How do machines affect our behaviour and interaction?

Artificially intelligent bots are becoming better and better at modelling human conversation and relationships. In 2015, a bot named Eugene Goostman won the Turing Challenge for the first time. In this challenge, human raters used text input to chat with an unknown entity, then guessed whether they had been chatting with a human or a machine. Eugene Goostman fooled more than half of the human raters into thinking they had been talking to a human being.

This milestone is only the start of an age where we will frequently interact with machines as if they are humans; whether in customer service or sales. While humans are limited in the attention and kindness that they can expend on another person, artificial bots can channel virtually unlimited resources into building relationships.

Even though not many of us are aware of this, we are already witnesses to how machines can trigger the reward centres in the human brain. Just look at click-bait headlines and video games. These headlines are often optimized with A/B testing, a rudimentary form of algorithmic optimization for content to capture our attention. This and other methods are used to make numerous video and mobile games become addictive. Tech addiction is the new frontier of human dependency.

On the other hand, maybe we can think of a different use for software, which has already become effective at directing human attention and triggering certain actions. When used right, this could evolve into an opportunity to nudge society towards more beneficial behavior. However, in the wrong hands it could prove detrimental.

4. Artificial stupidity. How can we guard against mistakes?

Intelligence comes from learning, whether you're human or machine. Systems usually have a training phase in which they "learn" to detect the right patterns and act according to their input. Once a system is fully trained, it can then go into test phase, where it is hit with more examples and we see how it performs. Obviously, the training phase cannot cover all possible examples that a system may deal with in the real world. These systems can be fooled in ways that humans wouldn't be.

For example, random dot patterns can lead a machine to "see" things that aren't there. If we rely on AI to bring us into a new world of labour, security and efficiency, we need to ensure that the machine performs as planned, and that people can't overpower it to use it for their own ends.

5. Racist robots. How do we eliminate AI bias?

Though artificial intelligence is capable of a speed and capacity of processing that's far beyond that of humans, it cannot always be trusted to be fair and neutral. Google and its parent company Alphabet are one of the leaders when it comes to artificial intelligence, as seen in Google's Photos service, where AI is used to identify people, objects and scenes. But it can go wrong, such as when a camera missed the mark on racial sensitivity, or when a software used to predict future criminals showed bias against black people.

We shouldn't forget that AI systems are created by humans, who can be biased and judgemental. Once again, if used right, or if used by those who strive for social progress, artificial intelligence can become a catalyst for positive change.

6. Security. How do we keep AI safe from adversaries?

The more powerful a technology becomes, the more can it be used for nefarious reasons as well as good. This applies not only to robots produced to replace human soldiers, or autonomous weapons, but to AI systems that can cause damage if used maliciously. Because these fights won't be fought on the battleground only, cybersecurity will become even more important. After all, we're dealing with a system that is faster and more capable than us by orders of magnitude.

7. Evil genies. How do we protect against unintended consequences?

It's not just adversaries we have to worry about. What if artificial intelligence itself turned against us? This doesn't mean by turning "evil" in the way a human might, or the way AI disasters are depicted in Hollywood movies. Rather, we can imagine an advanced AI system as a "genie in a bottle" that can fulfill wishes, but with terrible unforeseen consequences.

In the case of a machine, there is unlikely to be malice at play, only a lack of understanding of the full context in which the wish was made. Imagine an AI system that is asked to eradicate cancer in the world. After a lot of computing, it spits out a formula that does, in fact, bring about the end of cancer – by killing everyone on the planet. The computer would have achieved its goal of "no more cancer" very efficiently, but not in the way humans intended it.

8. Singularity. How do we stay in control of a complex intelligent system?

The reason humans are on top of the food chain is not down to sharp teeth or strong muscles. Human dominance is almost entirely due to our ingenuity and intelligence. We can get the better of bigger, faster, stronger animals because we can create and use tools to control them: both physical tools such as cages and weapons, and cognitive tools like training and conditioning.

This poses a serious question about artificial intelligence: will it, one day, have the same advantage over us? We can't rely on just "pulling the plug" either, because a sufficiently advanced machine may anticipate this move and defend itself. This is what some call the “singularity”: the point in time when human beings are no longer the most intelligent beings on earth.

9. Robot rights. How do we define the humane treatment of AI?

While neuroscientists are still working on unlocking the secrets of conscious experience, we understand more about the basic mechanisms of reward and aversion. We share these mechanisms with even simple animals. In a way, we are building similar mechanisms of reward and aversion in systems of artificial intelligence. For example, reinforcement learning is similar to training a dog: improved performance is reinforced with a virtual reward.

Right now, these systems are fairly superficial, but they are becoming more complex and life-like. Could we consider a system to be suffering when its reward functions give it negative input? What's more, so-called genetic algorithms work by creating many instances of a system at once, of which only the most successful "survive" and combine to form the next generation of instances. This happens over many generations and is a way of improving a system. The unsuccessful instances are deleted. At what point might we consider genetic algorithms a form of mass murder?

Once we consider machines as entities that can perceive, feel and act, it's not a huge leap to ponder their legal status. Should they be treated like animals of comparable intelligence? Will we consider the suffering of "feeling" machines?

Some ethical questions are about mitigating suffering, some about risking negative outcomes. While we consider these risks, we should also keep in mind that, on the whole, this technological progress means better lives for everyone. Artificial intelligence has vast potential, and its responsible implementation is up to us.

5.4 Societal Issues Concerning the Application of Artificial Intelligence in Medicine

Medicine, as part of a phenomenon that affects all fields of life sciences, is becoming an increasingly data-centred discipline. Data analysis in medicine has for long been the territory of statisticians, but medical data are reaching beyond the merely quantitative to take more complex forms, such as, for instance, textual information in Electronic Health Records (EHR), images in many modalities, on their own or mixed with other types of signals, or graphs describing biochemical pathways or biomarker interactions. This data complexity is behind the evolution from classical multivariate data analysis towards the nascent field of *data science*, which, from the point of view of medicine, embraces a new reality that includes interconnected wearable devices and sensors.

Beyond the more classical statistical approaches, artificial intelligence (AI) and, more in particular, machine learning (ML) are attracting much interest for the analysis of medical data, even if arguably with a relatively low impact yet on clinical practice. It has been acknowledged that AI is experiencing a fast process of commodification (not that this is an entirely new concern, as it was already a matter of academic discussion almost 30 years ago. This characterization is mostly of interest to big IT companies but correctly reflects the current process of *industrialization* of AI, where the academic and industrial limits of research are increasingly blurred, with the main experts in AI and ML on the payroll of private companies. In any case, this means that AI systems and products are reaching the society at large, and, therefore, that societal issues related to the use of AI in general and ML in particular should not be ignored any longer and certainly not in the medicine and healthcare domains. These societal issues may take many forms, but, more often than not, they entail the design of models from a human-centred perspective, that is, models that incorporate human-relevant requirements and constraints. This is certainly an only partially technical matter.

Legislation.

The industrialization of AI exposes it to legislation regulating the social domain where it is meant to operate. In some cases, this overlaps issues of privacy and anonymity, such as in AI algorithms used for automated face recognition in public domains. It may also involve more general contexts, such as AI-based autonomous driving or defence weapons. Legislation is also involved in medicine and healthcare practice, and, therefore, we need to ensure that AI and ML technologies comply with current legislation.

Explainability and Interpretability.

ML and AI algorithms are often characterized as *black boxes*, that is, methods that generate data models that are difficult (if not impossible) to interpret because the functional form relating the available data (input) to a given outcome (the output) is far too complex. This problem has been exacerbated by the intensity of the current interest in deep learning (DL) methods. Only interpretable models can be explained, and explainability is paramount when decision-making in medicine (diagnosis, prognosis, etc.) must be conveyed to humans.

Privacy and Anonymity.

Privacy-preserving ML-based data analysis must deal with the potentially contradictory problem of keeping personal information private while aiming to model it, often to make inferences that will affect a given population. Data anonymity obviously refers to the impossibility of linking personal data with information about the individual that is not meant to be revealed. These are key problems and concerns in the medical and healthcare domains, mainly in the interaction between the public and private sectors.

Ethics and Fairness.

Biological intelligence is multi-faceted and responds to the environmental pressures of human societies. Ethics are one of those facets for which AI is still fairly unprepared. Interestingly, this topic has become central to AI discussion in recent years. Needless to say, ethics are also a core concern in medicine and healthcare. Such convergence of interests makes it important to create a clear roadmap for the ethical use of AI and ML in medicine. The application of ML and AI in areas of social relevance must also aspire to be *fair*. How do we imbue ML algorithms, which are fairness *agnostic*, with fairness requirements? How do we avoid gender or ethnicity, for instance, unfairly influencing the outcome of a learning algorithm? In the medical domain and in healthcare in particular, where sensible information about the individual may be readily available, how do we ensure that AI- and ML-based decision support tools are not affected by such bias?

We reckon that all of these are relevant aspects to consider in order to achieve the objective of fostering acceptance of AI- and ML-based technologies in the medical and healthcare domains, as well as to comply with an evolving legislation concerning the impact of digital technologies on ethically and privacy sensitive matters. Our specific goal here is to reflect on how all these topics affect medical applications of AI and ML.

Societal Issues of AI and ML Application

Legislation

Human societies are regulated by bodies of legislation. While remaining within the academic realm, AI and ML developments have stayed fairly oblivious to legal concerns, but the moment these technologies start occupying the social space at large, their impact on people is likely to hit a few legal walls. One widely discussed case is the use of AI as the basis for autonomously driving vehicles. When a human is in charge of any decision-making at the wheel of a vehicle, legal

responsibilities are quite clearly drawn. The quick industrial development of semi-autonomous vehicles, leading towards the objective of fully autonomous driving, has stretched the seams of current legislation, though.

Again, any application of AI and ML in actual medical practice is bound to generate discussion about its legal boundaries and implications. A pertinent example is the recent (May 2018) implementation of the European Union directive for General Data Protection Regulation (GDPR). This directive mandates a *right to explanation* of all decisions made by "automated or artificially intelligent algorithmic systems". According to Article 13 of the directive, the right to explanation implies that the "data controller" is legally bound to provide requesting citizens with "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing [automated decision making, as described in its Article 22] for the data subject" .AI and ML may be the tools used to provide such automated decision making, and, therefore, it places these technologies in a legal spotlight. Some guidelines for GDPR-compliant ML development have recently been provided.

The implications of GDPR for the use of AI and ML in medicine and healthcare are not too difficult to appreciate. Any AI- or ML-based medical decision support system (MDSS) whose purpose it is to assist the medical experts in their decision-making will be explicitly providing a (semi)automated decision on an individual (for instance, diagnosis, prognosis or recommendations on treatment concerning individual patients, perhaps even in life-threatening conditions). The data controller in this case will be the medical expert (from nurses to specialists [8]) and the institution this expert belongs to.

Note that this piece of legislation requires something very specific from the AI and ML technologies interpretable and explainable models, as discussed in the next section. A medical expert or any healthcare system employee using these technologies must be able to interpret how they reached specific decisions (say, why an ML model diagnosed a brain tumour as a metastasis and not a high-grade glioma) and must be able to explain those decisions to any human affected by them. In the implementation of the artificial kidney as one of the most promising technologies in nephrology, we should be concerned, for instance, about the possibility of an opaque AI- or ML-based alarm system not being able to explain the basis for a false alarm that might endanger the life of the dialysis patient.

At a higher level, and on the basis of legal safeguards such as the GDPR, a healthcare system might decide not to implement an opaque MDSS in clinical practice, despite its perceived effectiveness, only to avoid the prospect of unsustainable litigation costs caused by the false-positive and -negative cases or the incorrect estimations and predictions churned by these automated systems.

Interpretability and Explainability

Biological brains have not necessarily evolved the means to explain themselves. Arguably, this has only happened in species with social behaviour (although it could also be argued that social behaviour can only happen in species whose brains are capable of *explaining themselves* through some form of communication). In the human species, natural language performs that communicative or explanatory function.

AI was originally conceived as an attempt to reproduce aspects of biological intelligence, but self-explanatory capabilities were never a key aspect to consider. If the biological brain was meant to be understood as a form of information-processing system, so was AI, and the idea of *social* AI is relatively new, for instance in the form of intelligent agents and multi-agent systems. Only recently, the interpretability and explainability of AI and ML systems has come to the forefront of research in the field. One key reason for this is the breakthrough created by DL technologies. DL is an

augmented version of traditional artificial neural networks. The latter were long ago maligned as *black box* opaque models. DL models risk being considered augmented black boxes. Interpretability in this context can be seen as a human-computer interaction problem. We humans must be able to understand and interpret the outcome of an AI or ML model. That is, we need to ensure that even a very complex model can be explained (usually to other humans). A human brain, colossally more complex, has developed natural language to convey some level of explanation of its inner workings. Similar attempts with AI and ML are still very limited. Despite recent and thorough attempts to address the issue of how to characterize interpretability in ML, such attempts only highlight the tremendous difficulty involved in the scientific pursue of truly interpretable ML models.

In the medical domain, AI and ML models are often part of MDSS. Their potential and the possible barriers to their adoption have been investigated in the last decade. The paradox is that these methods, despite their advantages, are far from universal acceptance in medical practice. Arguably, one of the reasons is precisely (lack of) interpretability, expressed as "the need to open the machine learning black box". As already mentioned, DL-based technologies can worsen the problem, despite having already found their way into biomedicine and healthcare. In medicine, this has clear implications: if an ML-based MDSS makes decisions that cannot be comprehensibly explained, the medical expert can be put in the uncomfortable position of having to vouch for the system's trustworthiness, transferring the trust on a decision that she or he cannot explain to either the patient or to other medical experts. This does not mean to say that efforts have not been made to imbue MDSS with knowledge representations that are comprehensible to humans. Examples include rule-based representations, usually compatible with medical reasoning; and nomograms, commonly used by clinicians for visualizing the relative weights of symptoms on a diagnosis or a prognosis.

AI- and ML-based systems may have quantifiable goals and may still be useless unless they conform to clinical guidelines. Note that computer-based systems, such as MDSS, are often seen by clinicians as an extra burden in their day-to-day practice. The problem may appear when the MDSS conflicts with guidelines of medical practice, something bound to happen unless those guidelines are somehow fed as *prior knowledge* to the intelligent systems. In this scenario, interpretability might be seen as an opportunity to make model performance and compliance with guidelines compatible goals.

The role of ML in healthcare has been described as acting "as a tool to aid and refine specific tasks performed by human professionals". Note that this means that interpretability should not be considered here a fully technical issue dissociated from the cognitive abilities of the human interpreter. As acknowledged by Dreiseitl and Binder when discussing the weak levels of adoption of MDSS at the point of care, researchers often sidestep practical questions, such as whether adequate "explanations [are] given for the system's diagnosis"; "the form of explanation [is] satisfactory for the physicians using the system"; or "how intuitive is its use."

An effort should be made to integrate medical expert knowledge into the AI and ML models or use prior expert knowledge in formal frameworks for machine-human interaction in the pursuit of interpretability and explainability. The data analyst must play a proactive role in seeking medical expert verification. In return, the medical expert should ensure that the analysis outcomes are interpretable and usable in medical practice.

Privacy and Anonymity

Technological advances and the widespread adoption of networked computing and telecommunication systems are flooding our societies (and mostly governments and technology providers) with data. The physical society bonds are being swiftly amplified by our use of virtual social networks. In this scenario, data privacy and anonymity have become main social concerns and have triggered legal initiatives, such as the European GDPR discussed in previous sections.

Needless to say, privacy and anonymity have been a core concern for healthcare systems for far longer than for society at large. The current adoption of EHRs in medical practice enhances this issue, as sensitive patient data are uploaded in digital form to networked systems with varying levels of security systems in place. An interesting review on security and privacy in EHRs can be found in the study by Fernández-Alemán et al. The strong links between privacy and anonymity, on one side, and legislation, on the other, are clearly described in this study, although it is also acknowledged that "there has been very little activity in policy development involving the numerous significant privacy issues raised by a shift from a largely disconnected, paper-based health record system to one that is integrated and electronic" .

This is not an issue ignored by the AI and ML communities. As early as 2002, data confidentiality and anonymity in data mining medical applications were already discussed in journals of these fields, highlighting the responsibilities of *data miners* to human subjects. Privacy-preserving models and algorithms have been discussed in some detail. A commonplace situation for data analysts in clinical environments is the need to analyse data that are distributed among multiple clinical parties. These parties (e.g., hospitals) may have privacy protocols in place that prevent merging data from different origins into centralized locations (in other words, prevent data "leaving" a given hospital). The AI and ML communities have already worked on producing decentralized analytical solutions to bypass this bottleneck.

There is a new and disruptive element of the privacy and anonymity discussion in AI and ML applications in medicine that must be considered: the *en masse* landing of big IT corporations in the medical field, many of them proposing or integrating AI elements together with a myriad of AI-based medically oriented start-ups. The involvement of IT companies in health provision raises the bar for privacy and anonymity issues that were already on the table due to the pressure of insurance companies, especially in the most liberalized national health systems. An illustrative example of the complexities and potential drawbacks of this involvement can be found in *Nature* journal's report of the UK Information Commissioner's Office declaration that the operator of three London-based hospitals "had broken civil law when it gave health data to Google's London-based subsidiary DeepMind". These data were meant to be the basis for models to test results for signs of acute kidney injuries, but privacy and protocols of identification were breached in a large-scale transference of patients' data from the hospitals to the private company. According to the Royal Statistical Society's executive director, three lessons are to be extracted from this particular case of application to the medical domain: (1) due to society's increasing data trust deficit, data transference transparency and openness should be guaranteed; (2) data transference should be proportional to the medical task at hand (in this case, the development of models for the detection of signs of acute kidney injury); and (3) governance (not just legislation) mechanisms of control of data handling, management and use should be strengthened or created when necessary. He also makes a key statement when saying that "innovations such as artificial intelligence, machine learning […] offer great opportunities, but will falter without a public consensus around the role of data".

Ethics and Fairness

The time-honoured ultimate aspiration of AI is to replicate biological intelligence in silico. Biological intelligence, though, is the product of evolution and, as such, is multi-faceted and at least to some extent the product of environmental pressures of human societies. Ethics, as a compass for human decision-making, are one of those facets and could be argued to provide the foundations for the legislative regulation of societies.

The truth though is that the AI and ML fields are still fairly unprepared to address this pressing matter. Interestingly, this topic has become central to AI discussion only in recent years, once it has also become a central topic in global research agendas. In what sense might ethics be part

of the AI and ML equation and in what sense do we want these technologies be imbued with ethical considerations, beyond the overlap with bodies of regulation and legislation? Let us provide an illustrative example: the ongoing debate on the use of AI as part of autonomous weapons systems in defence and warfare. Unmanned autonomous vehicles, at least partially driven by AI, are being used for targeted bombing in areas of conflict. The ethical issues involved in human decisions concerning the choice of human targets in war periods are quite clearly delineated by international conventions, but who bears ethical responsibility in the case of targets at least partially chosen by AI-driven machines? This type of problem currently drives not-for-profit organization campaigns, such as those undertaken by Article 36, "to stop killer robots".

Needless to say, ethics are also a core concern in medicine and healthcare that has attracted much academic discussion. Can AI- and ML-supported tools address the basic biomedical ethical principles of respect for autonomy, non-maleficence, beneficence and justice? Should they, or should this be left to the medical practitioners? Medical practitioners, though, do not usually develop the AI and ML tools for medical application. Should they at least ensure that AI and ML developers do not transgress these principles in the design of such tools? According to Magoulas and Prentza, it is humans and not systems who can identify ethical issues, and, therefore, it is important to consider "the motivations and ethical dilemmas of researchers, developers and medical users of ML methods in medical applications."

Such convergence of interests makes it important, in any case, to create a clear roadmap for the ethical use of AI and ML in medicine that involves players both from the fields of medicine and AI.

The concept of *fairness* may be considered as subjective as the concept of ethics and, perhaps, more vaguely defined. If distinguishing what is fair and what is not in a human society is difficult and often controversial, trying to embed the concept of fairness in AI-based decision-making might be seen as a hopeless endeavour. Nevertheless, the use of ML and AI in socially relevant areas should at least aspire to be *fair*. As stated by Veale and Binns, "real-world fairness challenges in ML are not abstract, […] but are institutionally and contextually grounded."

Let us illustrate this with an example: gender bias can be added to an ML model by just biasing the choice with which the data used to train the model are selected. Caliskan et al. have recently shown that semantics derived automatically using ML from language corpora will incorporate human-like stereotyped biases. As noted by Veale and Binns, lack of fairness may sometimes be the inadvertent result of organisations not holding data on sensitive attributes, such as gender, ethnicity, sexuality or disability, due to legal, institutional or commercial reasons. Without such data, indirect discrimination-by-proxy risks are being increased.

In the medical domain and in healthcare in particular, where sensible information about the individual may be readily available, how do we ensure that AI- and ML-based decision support tools are not affected by such bias? Fairness constraints can be integrated in learning algorithms, as shown in a study by Celis et al. Given that fairness criteria are reasonably clean-cut in the medical context, such constraints should be easier to integrate than in other domains. Following Veale and Binns, fairness may be helped by trusting third parties with the selective storage of those data that might be necessary for incorporating fairness constraints into model-building in a privacy-preserving manner. A recent proposal of a "continuous framework for fairness" seeks to subject decision makers to fairness constraints that can be operationalized in an algorithmic (and therefore in AI and ML) setting, with such constraints facilitating a trade-off between individual and group fairness, a type of trade-off that could have clear implications in medical domains from access to drugs and health services to personalized medicine.

5.5 Decision-making role in Industries

Artificial Intelligence (AI) has emerged as a transformative force, reshaping the landscape of decision-making processes across various industries. This technological advancement, characterized by machines mimicking human cognitive functions, holds great promise in enhancing efficiency and outcomes. As AI integrates into critical decision-making frameworks, a profound understanding of its ethical implications becomes imperative. At its core, AI refers to the development of intelligent agents that can perceive their environment, reason through information, and make decisions to achieve specific goals. This encompasses a spectrum of technologies, from machine learning and natural language processing to robotics, collectively contributing to the evolution of intelligent systems.

The infusion of AI into decision-making processes has yielded unprecedented capabilities, ranging from data analysis and pattern recognition to complex problem-solving. In sectors such as finance, healthcare, and manufacturing, AI augments decision-makers by processing vast datasets, identifying trends, and facilitating more informed choices. The potential for efficiency gains and improved outcomes has elevated the importance of AI in contemporary decision- making landscapes. While AI promises numerous benefits, ethical considerations loom large as these technologies become integral to decision-making. Issues such as transparency, fairness, accountability, and privacy demand careful attention. The opacity of AI algorithms, the potential for bias, questions of accountability in case of errors, and the safeguarding of individuals' privacy are among the ethical challenges that must be navigated (Patel, 2024, World Health Organization. (2021).

As we delve into the ethical implications of AI in decision-making processes, this review will explore key principles, societal impacts, regulatory frameworks, and real-world case studies. Understanding and addressing these ethical considerations are fundamental to harnessing the full potential of AI while ensuring responsible and equitable deployment across diverse applications.

Ethical Principles in AI

Artificial Intelligence (AI) has become an integral part of decision-making processes across industries, raising ethical considerations that demand careful attention. In this review of ethical implications, we delve into key principles that form the foundation for responsible AI deployment. Transparency in AI involves making the decision-making process understandable and accessible to those affected by its outcomes. It is crucial for individuals to comprehend how AI algorithms arrive at specific decisions (Brendel, et. al., 2021, Du & Xie, 2021, Nassar& Kamal, 2021). Transparency fosters trust and facilitates a more informed dialogue between developers, users, and impacted parties.

Trust is paramount in the acceptance of AI-driven decisions. Transparent AI systems help users and stakeholders understand the rationale behind outcomes, reducing uncertainty and skepticism. Transparent algorithms contribute to building trust by allowing scrutiny and providing explanations for decisions, which is especially vital in critical domains like healthcare, finance, and criminal justice. Bias in AI algorithms can perpetuate or exacerbate existing inequalities. Whether through biased training data or inherent algorithmic biases, the consequences can be severe, leading to unfair treatment or discrimination. Recognizing and addressing bias is essential for building fair AI systems.

Mitigating bias involves a combination of ethical considerations, technical solutions, and diverse representation in AI development. Ethical guidelines emphasize the need to actively counteract biases, ensure fairness across different demographic groups, and implement regular audits to identify and rectify any biases that may emerge during the life cycle of AI systems. Determining accountability in AI systems can be complex. While developers play a significant role, responsibility

extends to organizations deploying AI, policymakers crafting regulations, and the users interacting with AI-generated outputs. Identifying the chain of responsibility is crucial for ensuring accountability. Creating frameworks that outline responsibilities and consequences for AI decisions is essential. Ethical guidelines advocate for organizations to establish clear policies, practices, and mechanisms to address unintended consequences or errors arising from AI. This includes mechanisms for redress and compensation in cases of AI-related harm.

In conclusion, ethical principles such as transparency, fairness, and accountability form the bedrock of responsible AI deployment. As AI continues to advance, adhering to these principles becomes increasingly important to ensure that AI technologies contribute positively to society, minimize biases, and uphold the trust of users and stakeholders. Balancing technological innovation with ethical considerations is pivotal for the widespread acceptance and sustainable integration of AI into decision-making processes.

The Role of Data in AI Decision Making

Artificial Intelligence (AI) relies heavily on data to make informed decisions. However, the ethical implications of data-driven decision-making extend beyond the algorithms themselves. In this review, we delve into the critical role of data, focusing on data privacy, consent, quality, and biases. Privacy is a fundamental ethical concern in AI decision-making. Organizations must take measures to safeguard sensitive information and ensure compliance with data protection regulations. Adopting privacy-preserving techniques, such as anonymization and encryption, is crucial to prevent unauthorized access and protect the identities of individuals whose data is used in AI systems. Informed consent is a cornerstone of ethical data usage. Users should be informed about how their data will be used in AI applications and have the option to provide explicit consent. Transparency regarding data collection, processing, and storage practices allows individuals to make informed decisions about whether they want to participate in data-driven initiatives.

Biases present in training data can lead to unfair outcomes in AI decision-making. It is essential to identify and rectify biases in data to ensure that AI models do not perpetuate or amplify existing inequalities. Continuous monitoring and evaluation of datasets for biases, especially those related to gender, race, or socio-economic factors, are critical to developing fair and unbiased AI systems. The quality of AI decisions is directly linked to the accuracy and reliability of the training data. Data integrity is paramount, and organizations must implement robust data governance practices to maintain high-quality datasets. Rigorous validation processes, data cleaning techniques, and comprehensive documentation are necessary to enhance the trustworthiness of AI models.

Ethical considerations surrounding data in AI decision-making involve a delicate balance between harnessing the power of data for innovation and ensuring the protection of individuals' privacy and rights. Adhering to ethical guidelines not only helps organizations build trust with users but also promotes responsible and sustainable AI development. In conclusion, the ethical use of data in AI decision-making is foundational to the responsible deployment of AI technologies. Organizations must prioritize data privacy, obtain informed consent, and address biases in training data to ensure that AI systems contribute positively to society. By upholding ethical standards in data practices, stakeholders can navigate the challenges associated with AI decision-making and foster a trustworthy and inclusive AI landscape.

Impact on Society and Individuals

Artificial Intelligence (AI) decision-making processes wield significant influence over society and individuals, prompting ethical considerations that extend beyond algorithmic functionality . This review delves into the societal and individual impacts, focusing on job displacement, socioeconomic

implications, discrimination, and social justice. The integration of AI in various industries raises concerns about job displacement due to automation. Routine tasks being automated may lead to a shift in the job market, with certain roles becoming obsolete. Addressing this challenge requires proactive measures, such as upskilling and reskilling initiatives, to equip the workforce with the necessary skills for roles that AI cannot replace. To mitigate negative societal effects, there is a need for comprehensive policies and strategies. Governments, businesses, and educational institutions can collaborate to create a future-ready workforce. This involves investing in education and training programs that focus on skills that complement AI capabilities, fostering a smooth transition to a technologically advanced job market.

AI decision-making systems, if not carefully designed, may inadvertently perpetuate or amplify existing societal biases. This is particularly relevant in areas such as hiring, finance, and criminal justice. Biased algorithms can result in discriminatory outcomes, reinforcing inequalities. Identifying and rectifying biases in AI models is crucial to ensure fair and just decisions. Ethical AI design should prioritize fairness and justice. This involves implementing algorithms that are not influenced by gender, race, or socioeconomic factors. Transparency in AI decision-making processes, including disclosure of data sources and model logic, is essential for external scrutiny and accountability. Moreover, fostering diversity in AI development teams can contribute to more inclusive and unbiased systems.

Ethical considerations in AI decision-making processes play a pivotal role in shaping the impact on society and individuals. Proactively addressing challenges related to job displacement and discrimination is essential for ensuring that the integration of AI contributes positively to societal progress. Through collaborative efforts between policymakers, industry leaders, and the public, a balanced approach can be achieved, harnessing the benefits of AI while safeguarding against potential pitfalls.

Regulatory Frameworks and Standards

Artificial Intelligence (AI) has witnessed rapid advancements, prompting a growing need for robust regulatory frameworks and ethical standards to govern its deployment in decision- making processes. This review delves into the current state of AI regulations, highlighting existing frameworks and challenges. Additionally, it discusses the imperative for ethical AI standards, examining proposals and global efforts to shape guidelines. The current landscape of AI regulations is characterized by a patchwork of laws and guidelines globally. Some countries have established specific AI-related regulations, while others rely on broader data protection laws. Notable examples include the General Data Protection Regulation (GDPR) in the European Union and the California Consumer Privacy Act (CCPA) in the United States. However, these regulations primarily address data protection rather than the ethical aspects of AI decision-making. Existing regulations face challenges in keeping pace with the rapid evolution of AI technologies. Gaps in addressing ethical concerns, bias mitigation, and transparency issues pose significant challenges. There is a need for regulations that specifically target the ethical dimensions of AI, ensuring responsible deployment and safeguarding against potential risks.

Recognizing the ethical complexities associated with AI decision-making, proposals for establishing ethical guidelines have gained traction. Ethical AI frameworks focus on transparency, fairness, accountability, and the prevention of discriminatory outcomes. Organizations like the Institute of Electrical and Electronics Engineers (IEEE) and the Partnership on AI (PAI) have developed ethical principles to guide the responsible development and deployment of AI technologies. Industry leaders and international organizations are actively contributing to the development of AI standards. The World Economic Forum's AI for Business toolkit and initiatives like the OECD Principles on AI provide guidelines for governments, businesses, and developers.

Collaborative efforts aim to create a shared understanding of ethical AI principles, fostering a global approach to responsible AI deployment.

As AI continues to shape decision-making processes across industries, the establishment of comprehensive regulatory frameworks and ethical standards is imperative. Addressing gaps in current regulations and proactively shaping ethical guidelines will contribute to the responsible and equitable deployment of AI. Collaborative efforts between governments, industry stakeholders, and international organizations are essential to navigate the evolving landscape of AI regulation and ethical standards.

Case Studies of Ethical Dilemmas in AI Decision Making

Ethical challenges in AI decision-making have been vividly illustrated by real-world cases, often involving high-profile incidents that shed light on the complex interplay between technology and ethical considerations. This review delves into notable examples, drawing lessons from these cases and outlining implications for future AI deployments. One of the prominent ethical dilemmas in AI involves the use of facial recognition technology. Instances where law enforcement agencies deploy facial recognition systems, such as the controversy surrounding Clearview AI, raise significant privacy concerns. The widespread use of facial recognition without clear regulations has prompted debates on the balance between security and individual privacy.

AI algorithms used in hiring and recruitment processes have faced scrutiny for perpetuating biases. Amazon's recruitment tool, which was designed to assess resumes, was found to exhibit gender bias. The algorithm, trained on historical hiring data, reflected the biases inherent in that data. This case highlights the ethical challenges associated with using AI in contexts where historical data may perpetuate or amplify existing inequalities. Transparency in algorithmic decision-making is crucial to address ethical concerns. The lack of transparency in cases like the Amazon hiring tool underscores the importance of understanding how algorithms operate. Future AI deployments must prioritize transparency to ensure accountability and build trust among users.

The recognition of algorithmic bias emphasizes the need for inclusive design practices. AI systems should be developed with diverse and representative datasets to mitigate biases. Lessons from biased algorithms in hiring underscore the importance of ongoing monitoring and adjustments to ensure fairness and inclusivity. AI applications in healthcare, such as diagnostic algorithms, pose ethical challenges related to patient privacy and consent. Cases where patient data is used without clear consent raise questions about the ethical boundaries of AI in healthcare.

Lessons learned include the necessity of robust ethical frameworks in sensitive domains like healthcare. Autonomous AI systems, such as self-driving cars, present challenges in balancing autonomy with accountability. Accidents involving autonomous vehicles raise questions about liability and responsibility. As AI systems gain autonomy, ethical frameworks must evolve to establish clear lines of accountability and responsibility.

In conclusion, real-world cases of ethical dilemmas in AI decision-making provide valuable insights for shaping future deployments. These cases emphasize the need for transparency, inclusive design, ethical frameworks, and a proactive approach to addressing biases. Learning from past incidents will contribute to the responsible development and deployment of AI technologies, ensuring they align with ethical principles and societal values.

Public Perception and Trust in AI

Artificial Intelligence (AI) has become an integral part of modern life, influencing various sectors from healthcare to finance. However, the widespread adoption of AI technologies has raised concerns among the public regarding their ethical implications and potential risks. This review delves into the factors influencing public perception and trust in AI, emphasizing the impact of these concerns and proposing strategies for building and maintaining trust. Public concerns about AI are multifaceted, encompassing issues related to privacy, bias, accountability, and the potential for job displacement. High-profile incidents involving AI systems, such as data breaches or biased algorithms, contribute to a sense of apprehension. The opacity of AI decision-making processes and the fear of losing control over critical aspects of life amplify these concerns.

Addressing public concerns requires a proactive approach from developers, policymakers, and industry stakeholders. Strategies for building and maintaining trust include: Transparency in AI systems is essential to demystify their operations. Developers should prioritize explainability, making it clear how AI systems arrive at decisions. Transparent AI systems contribute to a better understanding of their impact and foster trust. Establishing clear ethical guidelines and standards for the development and deployment of AI technologies is crucial. Industry-wide standards and regulations can provide a framework for responsible AI practices, reassuring the public that these technologies adhere to ethical principles. Ensuring inclusivity in AI development, including diverse perspectives and avoiding biased datasets, helps mitigate concerns about discriminatory outcomes. Inclusive design practices can enhance the fairness and representativeness of AI systems. Engaging the public in discussions about AI, its benefits, and its potential risks fosters informed decision-making. Educational initiatives can demystify AI technologies, empower individuals to make informed choices, and alleviate unwarranted fears.

Transparency in AI systems is essential to demystify their operations. Developers should prioritize explainability, making it clear how AI systems arrive at decisions. Transparent AI systems contribute to a better understanding of their impact and foster trust. Establishing clear ethical guidelines and standards for the development and deployment of AI technologies is crucial. Industry-wide standards and regulations can provide a framework for responsible AI practices, reassuring the public that these technologies adhere to ethical principles. Ensuring inclusivity in AI development, including diverse perspectives and avoiding biased datasets, helps mitigate concerns about discriminatory outcomes. Inclusive design practices can enhance the fairness and representativeness of AI systems. Engaging the public in discussions about AI, its benefits, and its potential risks fosters informed decision-making. Educational initiatives can demystify AI technologies, empower individuals to make informed choices, and alleviate unwarranted fears.

Understanding and addressing public concerns about AI are pivotal for ensuring the responsible development and deployment of these technologies. By implementing transparent practices, adhering to ethical standards, fostering inclusivity, and engaging in meaningful public education, stakeholders can build and maintain trust in AI. A collaborative effort involving developers, policymakers, and the public is essential to navigate the ethical implications of AI in decision-making processes responsibly.

Future Considerations and Emerging Issues

Artificial Intelligence (AI) continues to advance rapidly, presenting both tremendous opportunities and ethical challenges. As we delve into the future of AI and decision-making processes, it becomes crucial to anticipate emerging issues, strategize for ethical concerns, and adapt to the evolving landscape of AI ethics. As AI technologies evolve, ethical challenges may arise due to their exponential growth and the potential for unintended consequences. Issues such as algorithmic bias, privacy infringements, and the impact on vulnerable populations could escalate if not

proactively addressed. The advent of more autonomous AI systems raises concerns about accountability and responsibility.

Ethical challenges may surface when decisions are made by AI without human intervention, especially in critical domains like healthcare, finance, and criminal justice. Deep learning algorithms, which are fundamental to many AI advancements, often operate as black boxes, making it challenging to understand their decision-making processes. Ethical considerations regarding transparency, explainability, and accountability become paramount.

Integrating ethical considerations into the design and development phases of AI technologies is essential. Developers should adopt an "ethics by design" approach, considering potential ethical implications and mitigating risks from the outset. Conducting regular ethical audits and impact assessments can help organizations identify and address ethical concerns in existing AI systems. These assessments should encompass algorithmic fairness, data privacy, and societal impacts. Collaborative efforts on an international scale are crucial for developing ethical standards and guidelines. Standardization can provide a common framework for responsible AI development, ensuring that ethical considerations are prioritized across diverse applications and industries. The regulatory environment around AI is evolving, with governments and international bodies considering frameworks to govern ethical AI use. Staying informed about and adapting to these regulatory changes is crucial for organizations to navigate the ethical landscape effectively. Continued public discourse on AI ethics is essential. Promoting awareness, engaging in public dialogue, and incorporating diverse perspectives into decision-making processes contribute to the ethical evolution of AI technologies. Establishing ethics committees within organizations and fostering cross-disciplinary collaboration are effective ways to address emerging ethical challenges. These committees can provide guidance, evaluate ethical implications, and ensure a multidimensional approach to decision-making. The future of AI and decision-making holds both promise and ethical complexities. Anticipating challenges, implementing proactive strategies, and adapting to the evolving landscape of AI ethics are essential for responsible development and deployment. By embracing ethics by design, conducting regular audits, fostering international collaboration, staying abreast of regulatory changes, promoting public discourse, and establishing ethics committees, stakeholders can navigate the future with a commitment to ethical AI practices. Continuous vigilance, flexibility, and a dedication to ethical considerations will be integral to shaping a positive and responsible future for AI and decision-making processes.

Collaboration and Stakeholder Involvement

As the ethical implications of Artificial Intelligence (AI) in decision-making become increasingly complex, fostering collaboration and engaging diverse stakeholders are paramount. Interdisciplinary collaboration and involvement of various stakeholders not only enhance the quality of ethical considerations but also contribute to the development of comprehensive frameworks for responsible AI deployment. Interdisciplinary collaboration brings together experts from diverse fields such as computer science, ethics, law, sociology, and philosophy. This collaborative approach ensures a comprehensive understanding of the ethical dimensions of AI, considering technical, social, legal, and philosophical perspectives.

The involvement of ethicists and social scientists alongside AI developers facilitates an "ethics by design" approach. This involves integrating ethical considerations into the early stages of AI development, reducing the risk of unintended consequences and ethical issues emerging later in the process. Collaboration between technical experts and ethicists allows for a dynamic exchange of

knowledge. Technical experts provide insights into the capabilities and limitations of AI systems, while ethicists contribute ethical guidance, ensuring that technology aligns with societal values and norms. Interdisciplinary collaboration enables a holistic risk assessment, considering not only technical risks but also ethical, social, and legal implications. This broader perspective helps in identifying potential biases, discrimination, and societal impacts that might be overlooked in a narrow, single-discipline approach.

Stakeholder involvement goes beyond academic and industry experts to include end-users and those affected by AI systems. Incorporating user perspectives helps in understanding the practical impact of AI decisions on individuals and communities, fostering user-centric ethical considerations. Collaboration with government and regulatory bodies is crucial for aligning AI development with legal and policy frameworks. Engaging these stakeholders ensures that ethical guidelines are consistent with existing regulations and helps in shaping future policies on AI ethics.

Non-Governmental Organizations (NGOs) and advocacy groups play a vital role in representing societal interests. Collaborating with these organizations ensures that the ethical discourse on AI includes diverse perspectives and addresses concerns related to fairness, privacy, and social justice. Collaboration with industry stakeholders is essential for developing and implementing ethical standards within the business context. Industry collaboration helps create guidelines for responsible AI development, encourages transparency, and fosters a culture of ethical decision-making within companies. Involving educational institutions contributes to building a future workforce that is ethically conscious and well-versed in responsible AI practices. Collaboration with academia ensures that ethical considerations are integrated into AI education and research.

Collaboration and stakeholder involvement are foundational pillars in addressing the ethical implications of AI in decision-making. Interdisciplinary collaboration ensures a holistic approach, combining technical expertise with ethical guidance. Engaging diverse stakeholders, including users, government bodies, NGOs, industry partners, and educational institutions, creates a robust ethical discourse that reflects a wide range of perspectives. The collaborative effort becomes instrumental in shaping AI systems that align with societal values, promote fairness, and mitigate potential risks. As AI continues to evolve, ongoing collaboration and stakeholder engagement will be essential for navigating the intricate landscape of AI ethics responsibly.

5.6 National and International Strategies on AI

As the technology behind AI continues to progress beyond expectations, policy initiatives are springing up across the globe to keep pace with these developments. The first national strategy on AI was launched by Canada in March 2017, followed soon after by technology leaders Japan and China. In Europe, the European Commission put forward a communication on AI, initiating the development of independent strategies by Member States. An American AI initiative is expected soon, alongside intense efforts in Russia to formalise their 10-point plan for AI. These initiatives differ widely in terms of their goals, the extent of their investment, and their commitment to developing ethical frameworks, reviewed here as of May 2019.

5.6.1 Europe

The European Commission's Communication on Artificial Intelligence (European Commission, 2018a), released in April 2018, paved the way to the first international strategy on AI. The document outlines a coordinated approach to maximise the benefits, and address the challenges, brought about by AI. The Communication on AI was formalised nine months later with the

presentation of a coordinated plan on AI. The plan details seven objectives, which include financing start-ups, investing €1.5 billion in several 'research excellence centres', supporting masters and PhDs in AI and creating common European data spaces. Objective 2.6 of the plan is to develop 'ethics guidelines with a global perspective'. The Commission appointed an independent high-level expert group to develop their ethics guidelines, which – following consultation – were published in their final form in April 2019. The Guidelines list key requirements that AI systems must meet in order to be trustworthy.

The EU's seven requirements for trustworthy AI:

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Societal and environmental wellbeing
7. Accountability.

The EU's High-Level Expert Group on AI shortly after released a further set of policy and investment guidelines for trustworthy AI, which includes a number of important recommendations around protecting people, boosting uptake of AI in the private sector, expanding European research capacity in AI and developing ethical data management practices. The Council of Europe also has various ongoing projects regarding the application of AI and in September 2019 established an Ad Hoc Committee on Artificial Intelligence (CAHAI). The committee will assess the potential elements of a legal framework for the development and application of AI, based on the Council's founding principles of human rights, democracy and the rule of law (Council of Europe, 2019a). Looking ahead, the next European Commission President, Ursula von der Leyen, has announced AI as a priority for the next Commission, including legislation for a coordinated approach on the 'human and ethical implications' of AI. The European Commission provides a unifying framework for AI development in the EU, but Member States are also required to develop their own national strategies.

Finland was the first Member State to develop a national programme on AI. The programme is based on two reports, Finland's Age of Artificial Intelligence and Work in the Age of Artificial Intelligence. Policy objectives focus on investment for business competitiveness and public services. Although recommendations have already been incorporated into policy, Finland's AI steering group will run until the end of the present Government's term, with a final report expected imminently. So far, Denmark, France, Germany, Sweden and the UK have also announced national initiatives on AI. Denmark's National Strategy for Artificial Intelligence was released in March 2019 and follows its 'Strategy for Digital Growth'. This comprehensive framework lists objectives including establishing a responsible foundation for AI, providing high quality data and overall increasing investment in AI. There is a strong focus on data ethics, including responsibility, security and transparency, and recognition of the need for an ethical framework. The Danish government outlines six principles for ethical AI – self-determination, dignity, responsibility, explainability, equality and justice, and development – and will establish a Data Ethics Council to monitor technological development in the country. In France, 'AI for Humanity' was launched in March 2018 and makes commitments to

support French talent, make better use of data and also establish an ethical framework on AI. President Macron has committed to ensuring transparency and fair use in AI, which will be embedded in the education system.

The strategy is mainly based on the work of Cédric Villani, French mathematician and politician, whose 2018 report on AI made recommendations across economic policy, research infrastructure, employment and ethics. Germany's AI Strategy was adopted soon after in November 2018 and makes three major pledges: to make Germany a global leader in the development and use of AI, to safeguard the responsible development and use of AI, and to integrate AI in society in ethical, legal, cultural and institutional terms. Individual objectives include developing Centres of Excellence for research, the creation of 100 extra professorships for AI, establishing a German AI observatory, funding 50 flagship applications of AI to benefit the environment, developing guidelines for AI that are compatible with data protection laws, and establishing a 'Digital Work and Society Future Fund'. Sweden's approach to AI has less specific terms, but provides general guidance on education, research, innovation and infrastructure for AI. Recommendations include building a strong research base, collaboration between sectors and with other countries, developing efforts to prevent and manage risk and developing standards to guide the ethical use of AI. A Swedish AI Council, made up of experts from industry and academia, has also been established to develop a 'Swedish model' for AI, which they say will be sustainable, beneficial to society and promote long-term economic growth.

The UK government issued the comprehensive 'AI Sector Deal' in April 2018, part of a larger 'Industrial Strategy', which sets out to increase productivity by investing in business, skills and infrastructure. It pledges almost £1 billion to promote AI in the UK, along five key themes: ideas, people, infrastructure, business environment and places. Key policies include increasing research and development investment to a total of 2.4% of GDP by 2027; investing over £400 million in maths, digital and technical education; developing a national retraining scheme to plug the skills gap and investing in digital infrastructure such as electric vehicles and fibre networks. As well as these investment commitments, included in the deal is the creation of a 'Centre for Data Ethics and Innovation' (CDEI) to ensure the safe and ethical use of AI. First announced in the 2017 budget, the CDEI will assess the risks of AI, review regulatory and governance frameworks and advise the government and technology creators on best practice. Several other European nations are well on their way to releasing national strategies. Austria has established a 'Robot Council' to help the Government to develop a national AI Strategy. A white paper prepared by the Council lays the groundwork for the strategy. The socially-focused document includes objectives to promote the responsible use of AI, develop measures to recognise and mitigate hazards, create a legal framework to protect data security, and engender a public dialogue around the use of AI.

Estonia has traditionally been quick to take up new technologies, AI included. In 2017, Estonia's Adviser for Digital Innovation Marten Kaevats described AI as the next step for 'e-governance' in Estonia. Indeed, AI is already widely used by the government, which is currently devising a national AI strategy. The plan will reportedly consider the ethical implications of AI, alongside offering practical economic incentives and pilot programmes. An AI task force has been established by Italy to identify the opportunities offered by AI and improve the quality of public services. The task force further outline challenges relating to technology development, the skills gap, data accessibility and quality, and a legal framework. It makes a total of 10 recommendations to government, which are yet to be realised by policy. Malta, a country that has previously focused heavily on blockchain technology, has now made public its plans to develop a national AI strategy, putting Malta 'amongst the top 10 nations with a national strategy for AI'. A task force has been

established composed of industry representatives, academics and other experts to help devise a policy for Malta that will focus on an ethical, transparent and socially-responsible AI while developing measures that garner foreign investment, which will include developing the skillset and infrastructure needed to support AI in Malta. Poland too is working on its national AI strategy. Despite media reports of military-focused AI developments in Russia the country currently has no national strategy on AI.

Across the EU: Public attitudes to robots and digitisation Overall, surveys of European perspectives to AI, robotics, and advanced technology have reflected that citizens hold a generally positive view of these developments, viewing them as a positive addition to society, the economy, and citizens' lives. However, this attitude varies by age, gender, educational level, and location and is largely dependent on one's exposure to robots and relevant information — for example, only small numbers of those surveyed actually had experience of using a robot (past or present), and those with experience were more likely to view them positively than those without. General trends in public perception from these surveys showed that respondents were: Supportive of using robots and digitisation in jobs that posed risk or difficulty to humans, Concerned that such technology requires effective and careful management; Worried that automation and digitisation would bring job losses, and unsure whether it would stimulate and boost job opportunities across the EU; Unsupportive of using robots to care for vulnerable members of society, Worried about accessing and protecting their data and online information, and likely to have taken some form of protective action in this area (antivirus software, changed browsing behaviour); Unwilling to drive in a driverless car (only 22% would be happy to do this); Distrustful of social media, with only 7% viewing stories published on social media as 'generally trustworthy'; and Unlikely to view widespread use of robots as near- term, instead perceiving it to be a scenario that would occur at least 20 years in the future. These concerns thus feature prominently in European AI initiatives, and are reflective of general opinion on the implementation of robots, AI, automation and digitisation across the spheres of life, work, health, and more. 5.2. North America Canada was the first country in the world to launch a national AI strategy, back in March 2017. The Pan-Canadian Artificial Intelligence Strategy was established withfour key goals, to: increase the number of AI researchers and graduates in Canada; establish centres of scientific excellence (in Edmonton, Montreal and Toronto); develop global thought leadership in the economic, ethical, policy and legal implications of AI; and support a national research communityin AI. A separate programme for AI and society was dedicated to the social implications of AI, led by policyrelevant working groups that publish their findings for both government and public.

In collaboration with the French National Centre for Scientific Research (CNRS) and UK Research and Innovation (UKRI), the AI and society programme has recently announced a series of interdisciplinary workshops to explore issues including trust in AI, the impact of AI in the healthcare sector and how AI affects cultural diversity and expression. In the USA, President Trump issued an Executive Order launching the 'American AI Initiative' in February 2019, soon followed by the launch of a website uniting all other AI initiatives, including AI for American Innovation, AI for American Industry, AI for the American Worker and AI for American Values. The American AI Initiative has five key areas: investing in R&D, unleashing AI resources, setting governance standards, building the AI workforce and international engagement. The Department of Defence has also published its own AI strategy, with a focus on the military capabilities of AI. In May, the US advanced this with the AI Initiative Act, which will invest $2.2 billion into developing a national AI strategy, as well as funding federal R&D. The legislation, which seeks to 'establish a coordinated Federal initiative to accelerate research and development on artificial intelligence for the economic

and national security of the United States' commits to establishing a National AI Coordination Office, create AI evaluation standards and fund 5 national AI research centres. The programme will also fund the National Science Foundation to research the effects of AI on society, including the roles of data bias, privacy and accountability, and expand AI-based research efforts led by the Department of Energy.

In June 2019, the National Artificial Intelligence Research and Development Strategic Plan was released, which builds on an earlier plan issued by the Obama administration and identifies eight strategic priorities, including making long-term investments in AI research, developing effective methods for human-AI collaboration, developing shared public datasets, evaluating AI technologies through standards and benchmarks, and understanding and addressing the ethical, legal and societal implications of AI. The document provides a coordinated strategy for AI research and development in US .

5. 6.2. Asia

Asia has in many respects led the way in AI strategy, with Japan being the second country to release a national initiative on AI. Released in March 2017, Japan's AI Technology Strategy provides an industrialisation roadmap, including priority areas in health and mobility, important with Japan's ageing population in mind. Japan envisions a three-stage development plan for AI, culminating in a completely connected AI ecosystem, working across all societal domains. Singapore was not far behind. In May 2017, AI Singapore was launched, a five-year programme to enhance the country's capabilities in AI, with four key themes: industry and commerce, AI frameworks and testbeds, AI talent and practitioners and R&D. The following year the Government of Singapore announced additional initiatives focused around the governance and ethics of AI, including establishing an Advisory Council on the Ethical Use of AI and Data, formalised in January 2019's 'Model AI Governance Framework'. The framework provides a set of guiding ethical principles, which are translated into practical measures that businesses can adopt, including how to manage risk, how to incorporate human decision making into AI and how to minimise bias in datasets. China's economy has experienced huge growth in recent decades, making it the world's second largest economy. To catapult China to world leader in AI, the Chinese Government released the 'Next Generation AI Development Plan' in July 2017. The detailed plan outlines objectives for industrialisation, R&D, education, ethical standards and security. In line with Japan, it is a three-step strategy for AI development, culminating in 2030 with becoming the world's leading centre for AI innovation. There is substantial focus on governance, with intent to develop regulations and ethical norms for AI and 'actively participate' in the global governance of this technology. Formalised under the 'Three-Year Action Plan for Promoting Development of a New Generation Artificial Intelligence Industry', the strategy iterates four main goals, to: scale-up the development of key AI products significantly enhance core competencies in AI; deepen the development of smart manufacturing; and establish the foundation for an AI industry support system. In India, AI has the potential to add 1 trillion INR to the economy by 2035, named AI for All, aims to utilise the benefits of AI for economic growth but also social development and 'inclusive growth', with significant focus on empowering citizens to find better quality work. The report provides 30 recommendations for the government, which include setting up Centres of Research Excellence for AI (COREs, each with their own Ethics Council), promoting employee reskilling, opening up government datasets and establishing 'Centres for Studies on Technological Sustainability'. It also establishes the concept of India as an 'AI Garage', whereby solutions developed in India can be rolled out to developing economies in the rest of the world. Alongside them, Taiwan released an 'AI Action Plan' in January 2018, focused heavily on industrial

innovation, and South Korea announced their 'AI Information Industry Development Strategy' in May 2018.

The report on which this was based provides fairly extensive recommendations for government, across data management, research methods, AI in government and public services, education and legal and ethical reforms. Malaysia's Prime Minister announced plans to introduce a national AI framework back in 2017 (Abas, 2017), an extension of the existing 'Big Data Analytics Framework' and to be led by the Malaysia Digital Economy Corporation (MDEC). There has been no update from the government since 2017. More recently, Sri Lanka's wealthiest businessman Dhammika Perera has called for a national AI strategy in the country, at an event held in collaboration with the Computer Society of Sri Lanka (Cassim, 2019), however there has not yet been an official pledge from the government. In the Middle East, the United Arab Emirates was the first country to develop a strategy for AI, released in October 2017 and with emphasis on boosting government performance and financial resilience (UAE Government, 2018).

Investment will be focused on education, transport, energy, technology and space. The ethics underlying the framework is fairly comprehensive; the Dubai AI Ethics Guidelines dictate the key principles that make AI systems fair, accountable, transparent and explainable (Smart Dubai, 2019a). There is even a self-assessment tool available to help developers of AI technology to evaluate the ethics of their system (Smart Dubai, 2019b). World leader in technology Israel is yet to announce a national AI strategy. Acknowledging the global race for AI leadership, a recent report by the Israel Innovation Authority (Israel Innovation Authority, 2019) recommended that Israel develop a national AI strategy 'shared by government, academia and industry'.

5.6.3. Africa

Africa has taken great interest in AI; a recent white paper suggests this technology could solve some of the most pressing problems in Sub-Saharan Africa, from agricultural yields to providing secure financial services (Access Partnership, 2018). The document provides essential elements for a panAfrican strategy on AI, suggesting that lack of government engagement to date has been a hindrance and encouraging African governments to take a proactive approach to AI policy. It lists laws on data privacy and security, initiatives to foster widespread adoption of the cloud, regulations to enable the use of AI for provision of public services, and adoption of international data standards as key elements of such a policy, although one is yet to emerge.

Kenya however has announced a task force on AI (and blockchain) chaired by a former Secretary in the Ministry of Information and Communication, which will offer recommendations to the government on how best to leverage these technologies (Kenyan Wallstreet, 2018). Tunisia too has created a task force to put together a national strategy on AI and held a workshop in 2018 entitled 'National AI Strategy: Unlocking Tunisia's capabilities potential' (ANPR, 2018).

5.6.4. South America

Mexico is so far the only South American nation to release an AI strategy. It includes five key actions, to: develop an adequate governance framework to promote multi-sectorial dialogue; map the needs of industry; promote Mexico's international leadership in AI; publish recommendations for public consultation; and work both with experts and the public to achieve the continuity of these efforts. The strategy is the formalisation of a White Paper authored by the British Embassy in Mexico, consultancy firm Oxford Insights and thinktank C Minds, with the collaboration of the Mexican Government. The strategy emphasises the role of its citizens in Mexico's AI development

and the potential of social applications of AI, such as improving healthcare and education. It also addresses the fact that 18% of all jobs in Mexico (9.8 million in total) will be affected by automation in the coming 20 years and makes a number of recommendations to improve education in computational approaches. Other South American nations will likely follow suit if they are to keep pace with emerging markets in Asia. Recent reports suggest AI could double the size of the economy in Argentina, Brazil, Chile, Colombia and Peru.

5.6.5. Australia

Australia does not yet have a national strategy on AI. It does however have a' Digital Economy Strategy' which discusses empowering Australians through 'digital skills and inclusion', listing AI as a key emerging technology. A report on 'Australia's Tech Future' further details plans for AI, including using AI to improve public services, increase administrative efficiency and improve policy development. The report also details plans to develop an ethics framework with industry and academia, alongside legislative reforms to streamline the sharing and release of public sector data. The draft ethics framework is based on case studies from around the world of AI 'gone wrong' and offers eight core principles to prevent this, including fairness, accountability and the protection of privacy. It is one of the more comprehensive ethics frameworks published so far, although yet to be implemented. Work is also ongoing to launch a national strategy in New Zealand, where AI has the potential to increase GDP by up to $54 billion. The AI Forum of New Zealand has been set up to increase awareness and capabilities of AI in the country, bringing together public, industry, academia and Government. Their report 'Artificial Intelligence: Shaping The Future of New Zealand' lays out a number of recommendations for the government to coordinate strategy development (i.e. to coordinate research investment and the use of AI in government services); increase awareness of AI (including conducting research into the impacts of AI on economy and society); assist AI adoption (by developing best practice resources for industry); increase the accessibility of trusted data; grow the AI talent pool (developing AI courses, including AI on the list of valued skills for immigrants); and finally to adapt to AI's effects on law, ethics and society. This includes the recommendation to establish an AI ethics and society working group to investigate moral issues and develop guidelines for best practice in AI, aligned with international bodies.

5.5.6. International AI Initiative

In addition to the EU In addition to the EU, there are a growing number of international strategies on AI, aiming to provide a unifying framework for governments worldwide on stewardship of this new and powerful technology. G7 Common Vision for the Future of AI At the 2018 meeting of the G7 in Charlevoix, Canada, the leaders of the G7 committed to 12 principles for AI, summarised below:

1. Promote human-centric AI and the commercial adoption of AI, and continue to advance appropriate technical, ethical and technologically neutral approaches.

2. Promote investment in R&D in AI that generates public test in new technologies and supports economic growth.

3. Support education, training and re-skilling for the workforce.

4. Support and involve underrepresented groups, including women and marginalised individuals, in the development and implementation of AI.

5. Facilitate multi-stakeholder dialogue on how to advance AI innovation to increase trust and adoption.

6. Support efforts to promote trust in AI, with particular attention to countering harmful stereotypes and fostering gender equality. Foster initiatives that promote safety and transparency.

7. Promote the use of AI by small and medium-sized enterprises.

8. Promote active labour market policies, workforce development and training programmes to develop the skills needed for new jobs.

9. Encourage investment in AI.

10. Encourage initiatives to improve digital security and develop codes of conduct.

11. Ensure the development of frameworks for privacy and data protection.

12. Support an open market environment for the free flow of data, while respecting privacy and data protection.

Nordic-Baltic Region Declaration on AI The declaration signed by the Nordic-Baltic Region aims to promote the use of AI in the region, including improving the opportunities for skills development, increasing access to data and a specific policy objective to develop 'ethical and transparent guidelines, standards, principles and values' for when and how AI should be used. OECD Principles on AI On 22 May 2019, the Organisation for Economic Co-operation and Development issued its principles for AI, the first international standards agreed by governments for the responsible development of AI. They include practical policy recommendations as well as value-based principles for the 'responsible stewardship of trustworthy AI', summarised below: • AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being. • AI systems should respect the rule of law, human rights, democratic values and diversity, and there should include appropriate safeguards to ensure a fair society. • There should be transparency around AI to ensure that people understand outcomes and can challenge them. • AI systems must function in a robust, secure and safe way throughout their life cycles and risks should be continually assessed. • Organisations and individuals developing, deploying or operating AI systems should be held accountable. These principles have been agreed by the governments of the 36 OECD Member States as well as Argentina, Brazil, Colombia, Costa Rica, Peru and Romania (OECD, 2019a). The G20 human-centred AI Principles were released in June 2019 and are drawn from the OECD Principles (G20, 2019). United Nations The UN has several initiatives relating to AI, including: • AI for Good Global Summit- Summits held since 2017 have focused on strategies to ensure the safe and inclusive development of AI. The events are organised by the International Telecommunication Union, which aims to 'provide a neutral platform for government, industry and academia to build a common understanding of the capabilities of emerging AI technologies and consequent needs for technical standardisation and policy guidance.'

UNICRI Centre for AI and Robotics

The UN Interregional Crime and Justice Research Institute (UNICRI) launched a programme on AI and Robotics in 2015 and will be opening a centre dedicated to these topics in The Hague (UNICRI, 2019). • UNESCO Report on Robotics Ethics - The UNESCO World Commission on the Ethics of Scientific Knowledge and Technology (COMEST) has authored a report on 'Robotics Ethics', which deals with the ethical challenges of robots in society and provides ethical principles and values, and a technology-based ethical framework (COMEST, 2017). World Economic Forum The World Economic Forum (WEF) formed a Global AI Council in May 2019, co-chaired by speech recognition developer Kai-Fu Lee, previously of Apple, Microsoft and Google, and current President

of Microsoft Bradford Smith. One of six 'Fourth Industrial Revolution' councils, the Global AI Council will develop policy guidance and address governance gaps, in order to develop a common understanding among countries of best practice in AI policy (World Economic Forum, 2019a). In October 2019, they released a framework for developing a national AI strategy to guide governments that are yet to develop or are currently developing a national strategy for AI.

The WEF describe it as a way to create a 'minimum viable' AI strategy and includes four main stages:

1) Assess long-term strategic priorities

2) Set national goals and targets

3) Create plans for essential strategic elements

4) Develop the implementation plan.

The strong European representation in this analysis is reflective of the value of the unifying EU framework, as well as Europe's economic power. The analysis also praises the policy strategies of individual European nations, which, importantly, have been developed in a culture of collaboration. Examples of this collaborative approach include the EU Declaration of Cooperation on AI, in which Member States agreed to cooperate on boosting Europe's capacity in AI, and individual partnerships between Member States, such as that of Finland, Estonia and Sweden, working together to trial new applications of AI.

Singapore ranked highest of all nations while Japan, the second country in the world to release a national strategy on AI, ranked 10th. China's position as 21st in the global rankings is expected to improve next year as its investments in AI begin to pay off. Progress in Asia overall has been unbalanced, with two countries in the region also ranking in the bottom ten worldwide, reflecting the income inequality in the region. Despite the comparatively slow development of their national strategy, the USA ranked 4th, with Canada not far behind. Both nations are supported by their strong economies, highly skilled workforces, private sector innovation and abundance of data, to a level at which regions missing from the top 10 – Africa, South America and Australasia – are unable to compete. This framework provides a highly useful metric by which to assess the ability of governments to capitalise on AI's potential in the coming years.