

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

UNIT I 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 behaviour 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 superintelligence—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 right or wrong for a specific machine but also on how to approach important questions such as: How can 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 due to 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 or not 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 potential dangers 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 the dangers 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
- Deepfakes
- 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 explainability

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 to its 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 upskill 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 what it 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 and videos, AI voice changers as well as deepfakes 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 it can 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 to think 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 AI can 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 superintelligence, 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 PSYCHOLOGY**

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.

Problems of Anthropomorphisation

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 an evolutionary 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 of mind” 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 might not 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 Ordnance Disposal (EOD) teams within the military (Carpenter 2016). Her work offers a 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 (Fig. 7.1).

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 to such 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 thereby inhibit 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 tend to 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 from winning 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 robot led 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 the way 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 can carry 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.

Can Artificial Intelligence be given Legal Rights and Duties? Can legal personhood be given to AI? Can the locustadi be present for 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 the legal 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.

While we witness the extremely efficient technology of AI in sectors such as humanoid robots 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.

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 necessity for 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 can lead 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 of manufacturing.

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 events and 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 concerns is the amount of energy required to train and operate AI algorithms. Training an AI model can require 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, as farmers focus on maximizing yields rather than promoting ecosystem health.

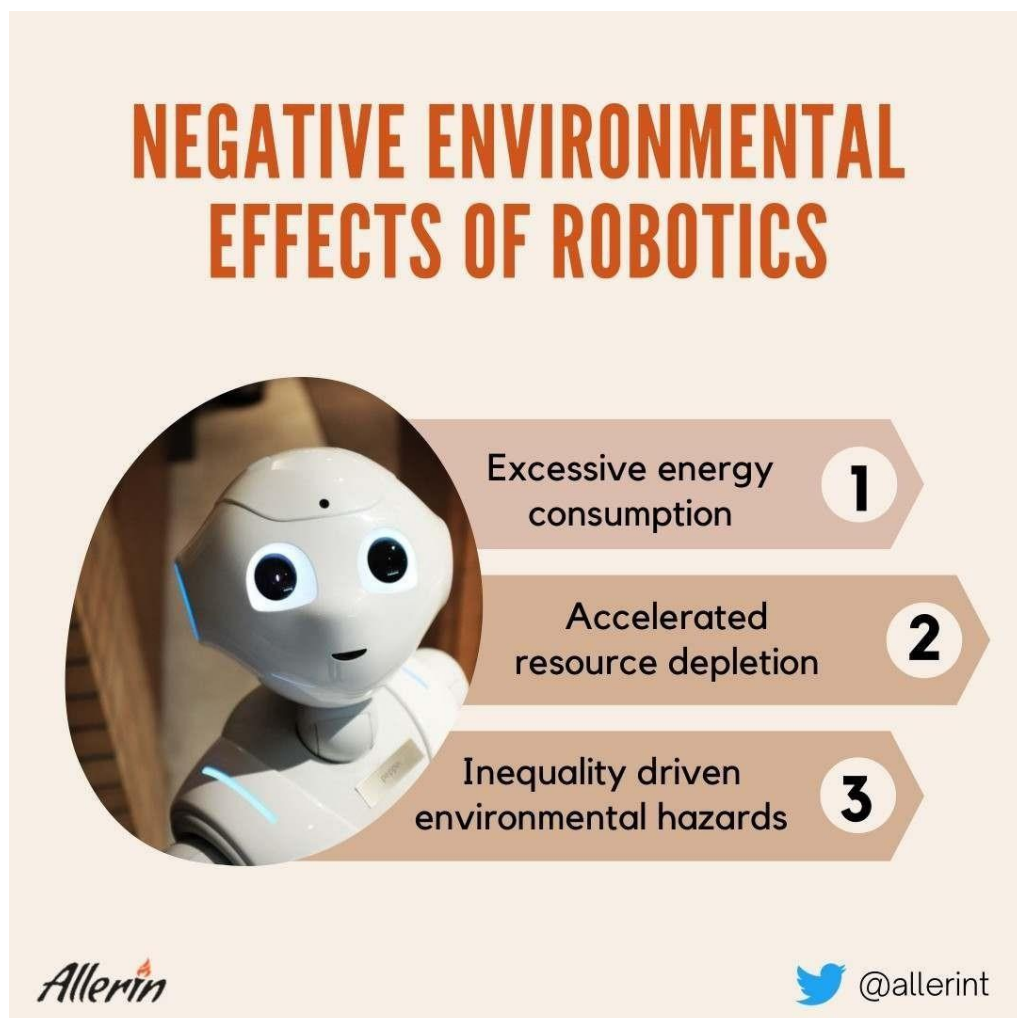
Finally, there are concerns about the ethical implications of using AI to manage the environment. AI algorithms are only as good as the data they are trained on, and biases in this data can lead to biased decision-making. For example, if an AI algorithm is trained on data that prioritizes economic growth 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 automation are well documented by now, we also need to focus on the ways in which technology can negatively impact our environment. The environmental impact of AI includes the environmental issues caused by robotics too.

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



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 some of 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 daily power 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 to planned obsolescence and depletion of natural resources.

Planned obsolescence involves the creation of products that become obsolete fast and need 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 a moral 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 have a 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 give users 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 a highway 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 good uncertainty 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 issue may 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, and the 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, but there have been bad books, too."

AI-enabled facial recognition has been used to profile certain ethnic groups and target political 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 most often 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 to identify 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 for certain 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 media posts and assess mental health using voice recognition.

Could AI 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't see this as a threat."

"On one hand, we have these novel machine-learning tools that display some autonomy from 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's really 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 the 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 AI 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 to create 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. "We have 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 that these 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 is not 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 people and 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."