The Intersection of Morality and Machine Intelligence: A Literature Review

1st Muhammad Osama Nusrat
Dept of Computing
Fast Nuces
Islamabad, Pakistan
i212169@nu.edu.pk

2nd Saad Ahmed Jamal
Dept of Geoinformatics
Paris Lodron Universität Salzburg
Salzburg, Austria
saad.jamal@stud.plus.ac.at

***Abstract*—The deployment of Artificial Moral Agents (AMAs) in various environments has become a prominent topic of discussion. However, the introduction of these agents into public domain comes with inherent risks that must be considered. Autonomous robots and vehicles, operating in real-world scenarios, face the challenge of making critical decisions. There is a concern that these machines might make choices conflicting with human moral values, raising questions about accountability and responsibility. For instance, if an autonomous vehicle is involved in a fatal accident, the question arises as to who should be held accountable: the manufacturer or the owner. This review paper addresses these pressing issues surrounding the complexities associated with imbuing machines with human-like decision-making capabilities. While humans possess the ability to make decisions in critical situations, the training of machines to emulate human cognition remains a challenging and multifaceted endeavour.**

I. Introduction

Machine ethics is a complex and evolving issue that has been both depicted in fictional movies and explored in theoretical frameworks. As far back as our childhood, we encountered robots in movies like 'Terminator' and 'Transformers', introducing the concept of machines with the potential for both good and bad actions. However, the idea of autonomous moral agents was not yet a foreseeable reality. In 1942, Issac Asimov proposed what he called "Asimov's laws of robotics," consisting of three fundamental principles. These laws dictated that machines or robots should not harm humans, must obey their owners' orders, except when they conflict with the first law, and prioritize their own safety while adhering to the first two laws. Although these laws served as captivating storytelling elements, they were largely regarded as fictional constructs without practical application. The introduction of drone missiles in modern warfare serves as evidence that Asimov's laws face challenges in real-world scenarios. These machines, designed for military purposes, unavoidably harm human beings in the course of war, highlighting the limitations of the first law. Additionally, the second law is not immune to exploitation, as humans can instruct robots to carry out malevolent actions or engage in illegal activities, rendering it impractical for real-world implementation. In essence, the complexities and nuances of real-life ethical dilemmas render Asimov's laws inadequate for guiding the behavior of autonomous machines. The evolving landscape of machine ethics requires a more sophisticated and comprehensive approach to ensure responsible and ethical AI deployment.

The study of machines possessing independent thinking abilities emerged in the 1950s when Alan Turing introduced the Turing test. In this test, a human, a machine, and an interrogator engaged in communication through a teletypewriter. The interrogator was unaware of the machine's presence and was tasked with distinguishing between the human and the machine, based on the premise of identifying their genders. The robot's objective was to convincingly appear human-like to the interrogator, rendering itself indistinguishable as a machine. Success in fooling the interrogator would demonstrate the machine's ability to think and be considered intelligent. However, the Turing test faced criticism as it equated machine intelligence with human-like imitation. The flaw lay in assuming that humans always think intelligently, disregarding the fact that humans can make errors and exhibit non-intelligent behavior. Consequently, a machine imitating a non-intelligent human would still pass the test, leading to a misjudgment of true machine intelligence. The goal of creating Artificial Moral Agents (AMAs) that behave like humans remains distant and uncertain. Despite progress in machine ethics, numerous ambiguities persist, demanding resolution. The integration of intelligent AMAs can yield both positive and negative effects on humanity. On the positive side, AMAs can enhance productivity and save time for individuals. Additionally, they can play critical roles in military applications, contributing to strategic advantages. Conversely, their introduction may have adverse consequences, such as job displacement and increased unemployment, leading to potential socioeconomic challenges like poverty and an upsurge in crime rates.

In 2006, James Moor proposed a taxonomy of four types of ethical agents: ethical impact agents, implicit ethical agents, explicit ethical agents, and entirely ethical agents. Ethical impact agents are those that can have a positive or negative impact on human ethics, regardless of whether they are programmed to do so. For example, a digital watch that helps a student get up early is an ethical impact agent because it can have a positive impact on the student's ethics by encouraging them to be more responsible. Implicit ethical agents are those that are programmed to avoid unethical behavior, but they do not have the ability to make ethical decisions on their own.

For example, an ATM is an implicit ethical agent because it is programmed to only dispense money to people who have entered the correct PIN. However, an ATM cannot make ethical decisions on its own, such as whether or not to dispense money to someone who is trying to withdraw more money than they have in their account. Explicit ethical agents are those that are programmed to make ethical decisions on their own. For example, a self-driving car is an explicit ethical agent because it is programmed to make decisions about how to navigate the road based on a set of ethical principles. However, explicit ethical agents can be hacked or manipulated to make unethical decisions.Entirely ethical agents are those that have the same moral qualities as humans, such as free will, intentions, and consciousness. Entirely ethical agents are still hypothetical, but they are the ultimate goal of artificial intelligence research. Moor's taxonomy provides a useful framework for understanding the different ways in which artificial agents can interact with human ethics. It is important to consider the ethical implications of artificial agents when they are being designed and used.

Bringing autonomy to machines will be very beneficial to humanity. It will save us time and energy. Driving for a long time can be stressful because we have to focus all the time. This drains our energy and makes us less productive at work. Self-driving cars will help us by taking over the driving task and freeing us up to do other things.

Machines can also help us diagnose diseases. If we have a machine that can tell us what disease we have by analyzing our symptoms, we can save money by not having to go to the doctor and we can save time by getting the diagnosis faster. AI can also be used to detect cancer and other tumors.

Corti is a moral machine that listens to emergency calls and detects heart attack risk. It does this by correlating the breathing rate and speech patterns of the caller. This is just one example of how machines can be used to help people in emergencies.

Robots can also be used to increase productivity in the workplace. They can work all the time and they can be more productive than humans. This can save businesses money on labor costs and it can also free up humans to do more creative and strategic work. These are just some of the reasons why machine ethics is important. We need to make sure that machines are programmed with ethical principles so that they can be used for good and not for harm.

Scientists have come up with three ways to make machines ethical: deontological ethics, utilitarian ethics, and virtue-based ethics. Utilitarian ethics says that we should do things that will cause the greatest good for the most people. This means that we should avoid doing things that will cause harm to society.

When building an AMA (artificial moral agent), we should keep in mind that it should follow a utilitarian approach. However, this approach has some drawbacks. For example, if we have to choose between making a large group of people happy or making a small group of people happy, utilitarian ethics would say that we should make the large group happy. This is because utilitarian ethics favors the greater good.

However, this approach is not always appreciated. For example, if a train had to choose between killing five people who are breaking the rules or killing one innocent person, utilitarian ethics would say that the train should kill the one innocent person to save the five people. This is not acceptable to many people because it means that the train would be killing an innocent person.

Virtue ethics is a type of ethics that focuses on the character of the machine. In this approach, the machine learns what is good or bad by experiencing different situations. This is a bottom-up approach, as the machine learns from its own experiences. One concern with virtue ethics is that the machine might learn values that are not considered good by humans. For example, the machine might learn that it is good to lie if it helps it achieve its goals. This is why it is important to ensure that the machine is also taught human values.

Deontological ethics is a type of ethics that focuses on the actions of the machine. In this approach, the machine is taught a set of rules that it must follow. These rules are not based on the consequences of the actions, but on the actions themselves. A good example of deontological ethics is Asimov's three laws of robotics. These laws state that robots must not harm humans, must obey orders given by humans, and must protect their own existence as long as it does not conflict with the first two laws. While Asimov's laws seem very friendly, they can fail in some situations. For example, if a robot is ordered to kill a human by a human, the robot would have to choose between following the first or second law. This is a difficult choice, and there is no clear answer as to what the robot should do. Both virtue ethics and deontological ethics have their own strengths and weaknesses. It is important to consider both approaches when developing ethical machines.

There are three ways to build AMAs (artificial moral agents): top-down, bottom-up, and hybrid. The hybrid approach is not used yet.

The top-down approach gives AMAs a set of rules to follow. These AMAs can't make their own decisions. They just follow the rules. AMAs that use the bottom-up approach are left in an environment to learn on their own. They learn from their experiences what actions are good or bad, just like a child learns from their mistakes. The hybrid approach combines the top-down and bottom-up approaches. This approach can be used when we can't build an AMA using just one approach.

The advantage of the top-down approach is that it's rule-based. This means that the AMAs have to follow the rules. Some examples of rule-based top-down approaches are Asimov's laws of robotics and Utilitarianism. The disadvantage of the top-down approach is that the rules can sometimes conflict. For example, the first two of Asimov's laws conflict with each other. The advantage of the bottom-up approach is that the AMAs learn from their experiences and evolve over time. This is just like how humans learn from their mistakes and become more mature over time. The disadvantage of the bottom-up approach is that it takes a long time for the AMAs to learn and evolve. They need ample amount of time to become mature.

Alan Turing said in his 1950 patent explained that if a computer is trained morally like a child, it will eventually be able to compete with humans in every aspect of life. The hybrid approach has not yet been used because it's difficult to combine the ethical theories used in the top-down and bottom-up approaches.

In this survey, Machine ethics had been critically reviewed in four sections. The first is the introductory section. In the second section, the need of machine ethics is discussed. The third section presents the review of the techniques and methods used to imbue ethics by different philosophers and scientists. In the next section, experiments done to imbue ethics and the results are discussed. Then we will discuss the best strategy in my view. Finally, we will conclude by discussing the negative side of Machine ethics and AI and what threats it can cause to humanity with prospects of what can be done to improve machine ethics further.

II. Literature Review

In [1] , Stuart Russell did a small experiment to check whether an artificial moral agent can reach its goal without hurting anyone in the surroundings or not. It passed and failed depending on the situation. The problem was that people had the authority to turn off the switch, so in that case, it did not reach its goal. If the was power given to the robot, there was an element of fear that it might hurt anyone or break the rules to reach its goal destination.

In [2 **??** and [3], Wendell Wallach and Smit discussed different approaches used to imbue ethics in moral agents. These are top-down, bottom-up and hybrid approaches. They also discussed the pros and cons of these approaches. Top-down approaches are rule-based. There is no element of fear in it that these moral agents can do something which is not ethical to humans, but these agents are sort of dumb. They cannot think on their own, which a negative side is. On the other hand, in bottom-up approaches, the moral agent will learn from its experiences as we learn from our experiences, so this is an intelligent agent. The cons of the bottom-up method are that if there are multiple goals, it will not work correctly; in simple words, it will get confused about which plan should he complete first.

In [4], James Moor discussed four ethical agents: moral impact agents, implicit and explicit agents, and fully honest agents. He further elaborated that implicit agents are most common and used in daily life, such as washing machines, dishwashers, etc. Explicit agents include self-driving cars. Full ethical agents have human-like ethics. They are hypothetical and yet to be explored.

In [5], Russel discussed the negative side of AI and moral agents. According to him, if a machine becomes more super intelligent than a human, then the off switch problem will fail, and it will know the potential dangers which can stop it from reaching its goal so that it will destroy the switch and it will hurt anyone who will prevent it from reaching its destination.

Winfield [6] discusses the potential benefits and risks of using AI for medical diagnosis. He argues that AI can be used to improve the accuracy, speed, and cost-effectiveness of diagnosis, but that there are also risks of AI wrongly diagnosing patients, discriminating against patients, and violating patient privacy. He argues that autonomous ethical systems are needed to ensure that AI systems are used in a safe and ethical manner.

In [7], Suzanne and Anderson reviewed different ethical theories potentially used to imbue ethics and concluded which one is best. Kant's theory, Ross theory and Utilitarianism were discussed. Kant's theory was considered the best because it was not based on emotions. Rather than it mainly focuses on whether performing a certain action will yield a certain result or not. Researchers opposed Utilitarianism because it was unfair in some situations, such as the trolley dilemma where five people who are not following the rules are saved because of their number, and one innocent person is killed to protect the greater good.

In [8], Slavkovik has implemented Asimov's 1st law of robotics with the help of a small experiment. The authors did coding in Java, and an A* search algorithm was used for finding the shortest path. Two agents were used, called robot agents and the human agent. A 15x15 grid was used in which there is a danger zone called lava tile. The robot is responsible for saving the human agent from moving into the lava tile, and the human agent has to reach its goal. So when there was one human agent, the robot agent successfully saved the human agent, but when the number of human agents increased, it failed to protect both human agents.

Dennis [9] discusses the development of a database of ethical dilemmas that can be used by researchers to test the ethical behavior of AI systems. The database includes ethical dilemmas from a variety of different domains, and it can be used by researchers to test the ability of AI systems to identify ethical dilemmas, to generate ethical arguments, and to make ethical decisions.

Lindner [10] discusses the development of HERA, a hybrid ethical reasoning agent that can be used to implement ethical theories in robots. HERA is a promising approach to machine ethics, as it allows machines to learn ethical behavior from human users. HERA was evaluated by presenting it with multiple scenarios and asking human users to answer what they would do in that particular dilemma. The results of the evaluation showed that HERA was able to learn ethical behavior from human users.

Charisi [11] provides a comprehensive overview of the field of machine ethics. They discuss the different approaches to machine ethics, the challenges that need to be addressed, and the opportunities that exist in this field. They argue that machine ethics is a rapidly growing field, and they predict that it will become increasingly important in the years to come. They also argue that it is important to address the challenges that exist in this field, and to seize the opportunities that exist.

Slavkovik's [12] provides a comprehensive overview of the field of machine ethics. She discusses the history of machine ethics, the different approaches and methods that have been used, and the ethical dilemma tools that have been developed.

She argues that machine ethics is a rapidly growing field, and she predicts that it will become increasingly important in the years to come. She also argues that it is important to develop ethical machine-learning algorithms that can be used to train machines to make ethical decisions.

In [13] Thomas M. Powers discusses the characteristics of a Kantian machine. He argues that a Kantian machine could be programmed to act according to the principles of Kantian ethics, such as the categorical imperative, respect for persons, and autonomy. Powers believes that a Kantian machine would have a number of advantages over other types of machines, such as being less likely to harm humans and making fair and impartial decisions. However, he also acknowledges that there are challenges involved in creating a Kantian machine, such as programming the machine to understand the nuances of Kantian ethics.

Anderson [14] and Anderson introduced GenEth, a user-friendly tool that codifies every ethical dilemma. It is based on a number of ethical frameworks and can be used to analyze ethical dilemmas and identify the best course of action. GenEth is still under development, but it has the potential to be a valuable tool for helping people to make ethical decisions.

Anderson [15] has discussed Prima Facie duties. A prima facie duty is a duty that is obligatory and necessary, e.g. we must tell the truth, follow traffic rules obey elders but sometimes we have to lie, such as information in our office is private, and we cannot share it with others so in the case of a moral machine there are some cases where the machine cannot share its private information, so ross theory has a drawback in this case.

In [16], Anderson introduced us to Ethel ethical eldercare system. The Ethel ethical eldercare system uses a Nao robot to remind elderly patients about their medication time. However, if the patient refuses to take their medicine, the Nao robot may need to handle the situation in a way that is both ethical and effective. One way to address this issue is for the Nao robot to be programmed with a number of different responses that it could use depending on the patient's mood.

Anderson [19] discussed the need for friendly AI, which are intelligent explicit agents that are aligned with human values, transparent, and robust. She argues that these agents are needed as AI technology becomes more sophisticated and powerful. The development of friendly AI is essential for ensuring that AI technology is used for good and not for harm.

James Gips [22] discussed ethical theories that could be used to guide the development of AI, such as virtue-based consequentialism and deontological theory. He argues that these theories can help to ensure that AI agents are programmed with the virtues and moral rules that are consistent with human values. However, he also acknowledges that no ethical theory is perfect and that it is important to use a combination of different theories to address the different ethical issues that arise in the development and use of AI.

III. IMPLEMENTATION IN MACHINE ETHICS

A. Turing Test

In the 1950s, when Alan Turing proposed a Turing test. In this test, he used three people, one human being, one machine and an interrogator, and they communicated with each other using a teletypewriter. The interrogator was not told there was a machine also in the game, and he was told that he aimed to distinguish between which of the two is a male or female. Similarly, the robot's task was to make itself unrecognizable that it is not a machine from the human player. If the person interrogating could not distinguish which one is a male or female, it proves that the robot can think and be considered intelligent. It has some limitations because, according to Allen Turing, this test would confirm that a machine can imitate itself as a human and be considered wise. Still, the criticism it faced was that a human does not always need to think intelligently. Human makes mistakes. According to the Turing test, a machine must imitate a human but in the case when the device mimicking a human who is not intelligent will fail the test. Similarly, a machine can solve complex mathematical problems which are difficult for humans to solve and as we know, in a Turing test, an interrogator asks two persons some questions to distinguish between a machine and a human, so if the machine solves the complex problem, then interrogator will conclude it is a machine which has solved the problem, so it fails the Turing test. Further, some scientists believe that we are wasting time to pass the Turing test as there are easier ways to test a problem.

B. Asimov's First Law

According to Asimov's first law, a moral agent cannot hurt people in the environment. Secondly, it must obey its owner but keep in mind the first law. Lastly, it must save itself by keeping the above two laws. The first law is practically implanted by checking an agent in a two-dimensional grid, applying the first law's ethical behavior in it, and checking the results. They conducted an experiment in which a 15x15 grid was used. Two moral agents were used, one human and a robot. Both agents can move in the grid. There are some dangerous tiles in the grid, called lava, and they are red. The human agent and robot both have to reach the goal state by following Asimov's law of robotics. To reach the goal state by shortest path, an A* search algorithm was used. According to Issac Asimov's law, a robot must not harm humans, and it will protect humans from getting harm, so in this scenario, its duty will be to protect the human agent from going into the danger zone. Two scenarios were implemented in the first scenario, which we discussed above. The robot agent saves human agents from going into the danger zone. However, in the second scenario, when there are two human agents and one robot agent, the complexity part here the human agent who is more near to the danger zone will be first saved by the robot instead of the second human agent. If both human agents are equally near falling in the lava, then the machine will save the last human agent. The drawback in this implementation was

that when the human agent is going to the goal but in his path comes the danger tile which is not directly in its path but near him, the robot will think that the human agent is in danger and it must be safe it because the robot has to follow Asimov's law so it will prevent the human agent from reaching its goal state. Adding more another drawback is that if a human agent suddenly changes its intention and follows a dangerous path, the robot agent will not have time to react to save it from going in the lava tiles, so it is another con in this implementation.

C. HERA

A software library for moral decision making in robots has been introduced, called HERA. Many ethical theories such as utilitarianism, the principle of double effect and the Pareto principle are used by HERA. For implementing the proposed Hera approach, an IMMANUEL robot was used. The main goal is to imbue all ethical theories in IMMANUEL to decide which ethical theory is best for a particular situation. When we provide different scenarios or dilemmas to IMMANUEL, it thinks and makes the best possible decision. Several scenarios were presented to the robot, and it gave answers according to the theories imbued in it.

D. Ethel

Ethel stands for (ethical elder care robot). This ethical robot named Nao is responsible for assisting older people at home to inform and ask them about the medicine.

IV. STRENGTHS AND WEAKNESSES OF TOP DOWN, BOTTOM UP AND HYBRID APPROACHES

The top-down approach is self-programmed. This approach is often used in self-driving vehicles, as it ensures that the vehicle will follow the rules of the road, even if it means harming a pedestrian. However, the top-down approach does not allow for the machine to learn and adapt to new situations. This could be a problem in cases where the ethical rules are not clear or unambiguous.For example, if a self-driving vehicle is programmed to follow the rule "Do not harm humans," it may not be able to decide what to do if it is faced with a situation where it must harm one human to save another. In this case, the top-down approach would not be able to provide a clear answer, as the ethical rules are not clear-cut.

The bottom-up approach to machine ethics is a more flexible approach, as it allows the machine to learn ethical behavior from experience. This approach is often used in robots that interact with humans, as it allows the robot to learn what is considered acceptable behavior from the humans it interacts with. However, the bottom-up approach can be more difficult to implement, as it requires the machine to be able to understand and interpret human behavior.

One of the challenges of the bottom-up approach is that it can be difficult to ensure that the machine learns the correct ethical principles. For example, if a robot is allowed to interact with humans and learn from their behavior, it may learn that it is acceptable to lie or cheat in certain situations. This could lead to the robot making unethical decisions in the future.

Another challenge of the bottom-up approach is that it can be difficult to predict how the machine will behave in new situations. This is because the machine is learning from experience, and it is not always clear how the machine will apply what it has learned to new situations. This could lead to the machine making unethical decisions in unexpected situations.

Despite these challenges, the bottom-up approach to machine ethics has the potential to be a powerful tool for teaching machines ethical behavior. By allowing the machine to learn from experience, the bottom-up approach can help the machine to develop a more nuanced understanding of ethics. This could lead to the machine making more ethical decisions in the future.

The hybrid approach to machine ethics is a combination of the top-down and bottom-up approaches. This approach allows the machine to learn ethical behavior from experience, but it also allows the machine to be programmed with a set of ethical rules. This approach is still under development, but it has the potential to offer the best of both worlds.

The hybrid approach has the potential to address some of the challenges of the top-down and bottom-up approaches. For example, the hybrid approach could help to ensure that the machine learns the correct ethical principles, while also allowing the machine to be flexible and adaptable. However, the hybrid approach is also challenging to implement. It requires the machine to be able to understand and interpret both ethical rules and human behavior. This is a complex task, and it is not yet clear how to implement it effectively. In Table 1 we have compared different research in machine ethics that have been conducted till now.

TABLE I
SUMMARY OF LITERATURE REVIEW

Study	Technique	Purpose of finding	Strength	Weakness
Russel Et al[1]	Markov Decision	Can an AMA reach its goal state without causing harm	AMA will not hurt anybody in the surrounding to complete its goal	If anyone turns the switch off it will not reach its goal.
Wallach Et al[2]	Approaches used for ethical decision making in machines	What methods we can use to imbue ethics in machines	Top-down approach is rule-based.	Bottom-up approach as the agent will learn from experience so it will take a lot of time to learn.
Smit Et al[3]	Methods discussed for solving ethics issues	Finding best technique to solve machine morality.	We can solve problem in chunks and then combine.	Bottom up works good only when there is a single goal.
Smit Et al[3]	Methods discussed for solving ethics issues	Finding best technique to solve machine morality.	We can solve problem in chunks and then combine.	Bottom up works good only when there is a single goal.
Moor Et al[4]	Discussion of moral agents and its types	Which agents are the best among the 4.	Implicit agents are safe.	Explicit agents can sometime make decision against human ethic.
Russell Et al[5]	Oracle AI	How can AI be beneficial and unsafe	Alignment with human values, Transparency, Robustness	A super-intelligent machine will destroy its off switch.
Winfield Et al[6]	Artificial Intelligence	Why there is a need for ethical autonomous systems.	Expenses reduced	In medical diagnosis an algorithm can make a mistake in detecting disease
Louis Et al[7]	Survey	Compare approaches used by different researchers	Explicit agents will be able to use different ethical theories in different situations	Ethical theories can conflict with each other which will result in poor decision making of machine.
Slavkovik Et al[8]	Java, A* search algorithm	Implementation of Asimov 1st law	Robot agent will save the human from falling into the lava tiles	Robot agent fails to protect human agent when he changes its intention suddenly.
Dennis Et al[9]	SQL,Python	Create a repository with different ethical dilemmas implemented	A database with implementation of moral dilemmas will help researchers to access data easily and improve it	We need to promote these databases as it is an open source platform so more input is given by them to improve solutions.
Linder Et al[10]	HERA,Java	IMMANUEL (robot) will answer the questions what he will do ethically when presented different scenarios	Each ethical theory is imbued in IMMANUEL so it learns on its own which theory to apply in which scenario	People do not have trust with machines. They are yet not comfortable with machines whether they are friendly or not.
Dennis Et al[11]	Survey	Give awareness of the problems in machine ethics	Symbolic AI techniques are the safest to use.	Soft AI techniques need more time to learn.
Powers Et al[13]	Kantian Moral Approach	It provides us the characteristics of a Kantian machine	It is based on good intentions	Not generalized
Anderson Et al[14]	GenEth	Ethical theories in the shape of a code	User friendly	Awareness required to use this platform
Susan Et al[15]	Inductive Logic Programming	Steps in creating an explicit moral agent	Explicit ethical agents can lead to more transparent and understandable interactions between humans and robots.	Both philosophers and researchers should be at one page.
Anderson Et al[16]	ETHEL	How robots can be used to take care of old age people	Takes care of patient by notifying patient to take medicine	Patients will not be able to interact with humans
Briggs Et al[19]	DIARC ADE	Situations when agent will reject our order due to danger it feels.	Moral agent knows it has to save itsef from getting damage so it will not obey his owner order if it is unethical for it.	Sometimes the action seems unethical but consequence is ethical, in that situation it has some cons in it
Susan Et al[20]	Machine Learning	How can we built explicit ethical agents who are friendly	A robot learns from experience what order should be first obeyed at priority.	Teaching robots to apply ethical principles in unpredictable and novel situations is a highly complex task.

V. CONCLUSION AND FUTURE WORK

Autonomous systems have been present in the world for 50 years in the form of automatic washing machines, vending machines, etc., but they are not dangerous because they are isolated. The debate and the main concern is autonomous self-driving cars because they are not isolated. There will be many cars besides this, so there must be accountability in case of any mishap. That is why scientists feel reluctant to bring these autonomous agents into daily life, where they will have to interact with people. There are many ethical theories such as Kant's, Ross's, Utilitarianism, Deontological, principle of double effect, etc., but there is not a single one among them that is generalized. Every theory is good in different scenarios, so there is another trouble with a single ethical theory in the agents.

Moreover, it is not a simple task to replace human beings with machines. People will take much time to accept this. Furthermore, there is a lack of trust in humans towards machines. Much time will be taken to develop trust among humans. We as a human learn things from our surroundings as we grow, we become more mature, but how can we expect the same thing with machines except that we also leave machines to learn, but it is very risky learning must be done in an isolated environment away from humans. We have to teach each and everything to machines. For example, we cannot say to a machine that brings me a cup of coffee from a market. However, we will have to thoroughly guide that brings it without hurting anyone following the rules; sometimes, it becomes very tedious. If we ask the same thing by humans, it will be understood that you have to follow the rules and not disturb anyone to complete a goal. Who will be held accountable if a mishap occurs, the manufacturer or the programmer? These parameters are yet to decide. AI can be used intentionally for doing evil works such as robbery, killing innocent and we cannot control it. How are we so sure that the machine will think in the same way we do if it does not. Maybe the things which are not ethical in our eyes are ethical for machines. We need to find a generalized ethical theory for AMA'S or for exceptional cases and situations; a different theory should be made, such as killing is okay when you are saving your life, so for situations like these when the action looks wrong. However, the result is good. We must train robots in such situations. A final question arises are we expecting too much from AMAs. In an airplane mishap, nearly all people die same for bus accidents and train accidents, so why expect 100 percent from these nonhuman things. We also need to give a small room for mistakes.

REFERENCES

[1] Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell. "The off-switch game." In Workshops at the Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[2] Allen, Colin, Iva Smit, and Wendell Wallach. "Artificial morality: Top-down, bottom-up, and hybrid approaches." Ethics and information technology 7, no. 3 (2005): 149-155.

[3] Wallach, Wendell, Colin Allen, and Iva Smit. "Machine morality: bottom-up and top-down approaches for modelling human moral faculties." In Machine Ethics and Robot Ethics, pp. 249-266. Routledge, 2020.

[4] Moor, James H. "The nature, importance, and difficulty of machine ethics." IEEE intelligent systems 21, no. 4 (2006): 18-21.

[5] Russell, Stuart. "Provably beneficial artificial intelligence." Exponential Life, The Next Step (2017).

[6] Winfield, Alan F., Katina Michael, Jeremy Pitt, and Vanessa Evers. "Machine ethics: The design and governance of ethical AI and autonomous systems [scanning the issue]." Proceedings of the IEEE 107, no. 3 (2019): 509-517.

[7] Tolmeijer, Suzanne, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. "Implementations in machine ethics: a survey." ACM Computing Surveys (CSUR) 53, no. 6 (2020): 1-38.

[8] Alvarez, Mateo, Øyvind Berge, Audun Berget, Eirin Bjørknes, Dag VK Johnsen, Fredrik Madsen, and Marija Slavkovik. "Implementing Asimov's first law of robotics." In Norsk IKT-konferanse for forskning og utdanning. 2017

[9] Bjørgen, Edvard P., Simen Madsen, Therese S. Bjørknes, Fredrik V. Heimsæter, Robin Håvik, Morten Linderud, Per-Niklas Longberg, Louise A. Dennis, and Marija Slavkovik. "Cake, death, and trolleys: dilemmas as benchmarks of ethical decision-making." In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 23-29. 2018.

[10] Lindner, Felix, Martin Mose Bentzen, and Bernhard Nebel. "The HERA approach to morally competent robots." In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 6991-6997. IEEE, 2017.

[11] Charisi, Vicky, Louise Dennis, Michael Fisher, Robert Lieck, Andreas Matthias, Marija Slavkovik, Janina Sombetzki, Alan FT Winfield, and Roman Yampolskiy. "Towards moral autonomous systems." arXiv preprint arXiv:1703.04741 (2017).

[12] Dennis, Louise A., and Marija Slavkovik. "Machines that know right and cannot do wrong: the theory and practice of machine ethics." IEEE Intelligent Informatics Bulletin 19, no. 1 (2018): 8-11.

[13] Powers, Thomas M. "Prospects for a Kantian machine." IEEE Intelligent Systems 21, no. 4 (2006): 46-51.

[14] Anderson, Michael, and Susan Leigh Anderson. "GenEth: A general ethical dilemma analyzer." Paladyn, Journal of Behavioral Robotics 9, no. 1 (2018): 337-357.

[15] Anderson, Michael, and Susan Leigh Anderson. "Machine ethics: Creating an ethical intelligent agent." AI magazine 28, no. 4 (2007): 15-15.

[16] Anderson, Michael, and Susan Leigh Anderson. "EthEl: Toward a principled ethical eldercare robot." (2008).

[17] Anderson, Michael, Susan Leigh Anderson, and Vincent Berenz. "A value-driven eldercare robot: Virtual and physical instantiations of a case-supported principle-based behavior paradigm." Proceedings of the IEEE 107, no. 3 (2018): 526-540.

[18] Bogosian, Kyle. "Implementation of moral uncertainty in intelligent machines." Minds and Machines 27, no. 4 (2017): 591-608.

[19] Briggs, Gordon Michael, and Matthias Scheutz. ""Sorry, I can't do that": Developing Mechanisms to Appropriately Reject Directives in Human-Robot Interactions." In 2015 AAAI fall symposium series. 2015.

[20] Anderson, Michael, and Susan Leigh Anderson. "Robot be good." Scientific American 303, no. 4 (2010): 72-77.

[21] Anderson, Michael, Susan Leigh Anderson, and Chris Armen. "Towards machine ethics." In AAAI-04 workshop on agent organizations: theory and practice, San Jose, CA. 2004.

[22] Gips, James. "Toward the ethical robot." (1994).

[23] Allen, Colin, Wendell Wallach, and Iva Smit. "Why machine ethics?." IEEE Intelligent Systems 21, no. 4 (2006): 12-17.

[24] Tonkens, Ryan. "A challenge for machine ethics." Minds and Machines 19, no. 3 (2009): 421.

[25] Brundage, Miles. "Limitations and risks of machine ethics." Journal of Experimental & Theoretical Artificial Intelligence 26, no. 3 (2014): 355-372.

[26] Cave, Stephen, Rune Nyrup, Karina Vold, and Adrian Weller. "Motivations and risks of machine ethics." Proceedings of the IEEE 107, no. 3 (2018): 562-574.

[27] Goodall, Noah J. "Machine ethics and automated vehicles." In Road vehicle automation, pp. 93-102. Springer, Cham, 2014.

[28] Anderson, Michael, and Susan Leigh Anderson. "The status of machine ethics: a report from the AAAI Symposium." Minds and Machines 17, no. 1 (2007): 1-10.
[29] Muehlhauser, Luke, and Louie Helm. "The singularity and machine ethics." In Singularity Hypotheses, pp. 101-126. Springer, Berlin, Heidelberg, 2012.
[30] Deng, Boer. "Machine ethics: The robot's dilemma." Nature News 523, no. 7558 (2015)
[31] Nath, Rajakishore, and Vineet Sahu. "The problem of machine ethics in artificial intelligence." AI & SOCIETY 35, no. 1 (2020): 103-111.
[32] Jiang, Liwei, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. "Delphi: Towards machine ethics and norms." arXiv preprint arXiv:2110.07574 (2021).
[33] Tavani, Herman T. "Levels of trust in the context of machine ethics." Philosophy & Technology 28, no. 1 (2015): 75-90.
[34] Köse, Utku. "Are we safe enough in the future of artificial intelligence? A discussion on machine ethics and artificial intelligence safety." BRAIN. Broad Research in Artificial Intelligence and Neuroscience 9, no. 2 (2018): 184-197.
[35] Powers, T. "Deontological machine ethics." In 2005 AAAI Fall Symposium on Machine Ethics, pp. 79-86. 2005.