AI & Soc (2020) 35:103–111
DOI 10.1007/s00146-017-0768-6

ORIGINAL ARTICLE

The problem of machine ethics in artificial intelligence

Rajakishore Nath[1] · **Vineet Sahu**[2]

Received: 5 June 2017 / Accepted: 12 October 2017 / Published online: 19 October 2017

Abstract The advent of the intelligent robot has occupied a significant position in society over the past decades and has given rise to new issues in society. As we know, the primary aim of artificial intelligence or robotic research is not only to develop advanced programs to solve our problems but also to reproduce mental qualities in machines. The critical claim of artificial intelligence (AI) advocates is that there is no distinction between mind and machines and thus they argue that there are possibilities for machine ethics, just as human ethics. Unlike computer ethics, which has traditionally focused on ethical issues surrounding human use of machines, AI or machine ethics is concerned with the behaviour of machines towards human users and perhaps other machines as well, and the ethicality of these interactions. The ultimate goal of machine ethics, according to the AI scientists, is to create a machine that itself follows an ideal ethical principle or a set of principles; that is to say, it is guided by this principle or these principles in decisions it makes about possible courses of action it could take[a]. Thus, machine ethics task of ensuring ethical behaviour of an artificial agent. Although, there are many philosophical issues related to artificial intelligence, but our attempt in this paper is to discuss, first, whether ethics is the sort of thing that can be computed. Second, if we are ascribing mind to machines, it gives rise to ethical issues regarding machines. And if we are not drawing the difference between mind and machines, we are not only redefining specifically human mind but also the society as a whole. Having a mind is, among other things, having the capacity to make voluntary decisions and actions. The notion of mind is central to our ethical thinking, and this is because the human mind is self-conscious, and this is a property that machines lack, as yet.

Keywords Artificial intelligence · Artificial moral agent · Moral agency · Mind · Subjectivity

1 Introduction

The main aim of AI is to understand, recreate and possibly surpass human intelligence in artificial entities. But unlike philosophy, which is concerned with the use of intelligence alone, AI strives to build intelligent entities as well as understand them. There have been various definitions of AI. Haugeland claims that AI is, "the exciting new effort to make computers think …. machines with minds, in the full and literal sense".[1] On the other hand, according to Winston, it is "the study of the computations that make it possible to perceive, reason, and act".[2] Let us look at these two definitions from different perspectives. Here, Haugeland and Winston point out that artificial intelligence is concerned with thought process and reasoning. That is to say, computers do think. People with widely varying backgrounds and professional

[a]Anderson, S. & Anderson, M., "The Consequences for Human Beings of Creating Ethical Robots" in Proceedings of AAAI Workshop Human Implications of Human-Robot Interaction, Vancouver, BC, Canada, July, 2007.

✉ Rajakishore Nath
rajakishorenath@iitb.ac.in

Vineet Sahu
vineet@iitk.ac.in

[1] Department of Humanities and Social Sciences, Indian Institute of Technology Bombay, Mumbai, India

[2] Department of Humanities and Social Sciences, IIT Kanpur, Kanpur, India

[1] Haugeland (1989, p. 2).

[2] Winston (1984, p. 2).

knowledge are contributing new ideas and introducing new tools into this discipline. Cognitive science minded psychologists have developed new models of the mind based on the fundamental concepts of artificial intelligence, symbol systems and information processing. Linguists are also interested in these fundamental notions while developing different models in computational linguistics. And philosophers, in considering the progress, problems and potential of this work towards non-human moral intelligence, have sometimes proposed solutions to the age-old problems of the nature of mind and its moral knowledge.

However, we know that artificial intelligence is a part of computer science in which intelligent systems are designed that exhibit the characteristics we associate with intelligence in human behaviour, understanding language, learning, reasoning, problem-solving, ethical behaviour, and so on. It is believed that insights into the nature of the mind can be gained by studying the operation of such systems. Artificial intelligence researchers have invented scores of programming techniques that support some sort of intelligent behaviour. Artificial intelligence research may have impact on science, technology, and society in general in the following way:

1. It can solve some difficult problems in chemistry, biology, geology, engineering and medicine.
2. It can objectify the social, moral, and legal behaviour.
3. It can manipulate robotic devices to perform some useful, repetitive, sensory-motor tasks.

Besides, artificial intelligence researchers investigate different kinds of computation and different ways of describing computation in an effort not just to create intelligent artefacts, but also to understand what intelligence is. According to Tanimoto,[3] their basic tenet is to create computers which think. Thus, AI expands the field of intelligent activity of human beings in various ways.

The hypothesis of artificial intelligence and its corollaries are empirical whose truth, or falsity is to be determined by experiment and empirical tests. The method of testing the results of artificial intelligence is of the following types:

1. In the narrow sense, artificial intelligence is part of computer science, aimed at exploring the range of tasks over which computers can be programmed to behave intelligently. Thus, it is the study of the ways computers can be made to perform cognitive tasks, which generally human beings undertake.
2. In the wider sense, artificial intelligence is aimed at programs that simulate the actual processes that human beings undergo in their intelligent behaviour. And these simulation programs are intended as theories describing and explaining human performance. But they are tested by comparing the computer output with the human behaviour to determine whether both the result and also the actual behaviour of computers and persons are closely similar.[4]

Therefore, the wide sense of the method of AI claims that AI can behave similarly to persons/human. If this is so, then AI can give rise the moral behaviour the way humans behave morally in the societal interaction. This is possible because AI is also an example of a physical symbol system, a system that is capable of inputting, outputting, storing, etc., following different courses of operation. These systems are capable of producing intelligence depending on the level of mechanical sophistication they are. Thus, these capabilities in AI behave intelligently like human beings so that there is something called AI ethics. But before discussing the problem of ethics in AI, it is imperative to discuss the place of mind in AI because it is mind, which gives rise to conscious moral behaviour in the world.

2 Mind in artificial intelligence

This section would explore the states of mind in artificial intelligence. As we know the main aim of artificial intelligence is not only to develop advanced technology to solve difficult problems for the society but also reproduce mentality in machines. That is to say that AI aims at producing machines with mind. Therefore, artificial intelligence is the discipline that aims to understand the nature of human intelligence through the construction of computer programs that imitate intelligent behaviour. It also emphasizes the functions of the human brain and the analogical functioning of the digital computer. If we say that machines have minds, then we have to ascribe certain 'belief', 'knowledge', 'free will', 'intention', 'observations', etc., to a machine. In that case, the machines will perform intelligent tasks and thus will behave like human beings. According to one extreme view, the human brain is just a digital computer, and the mind is a computer program. This view, as John Searle calls it is strong artificial intelligence.[5]

According to strong artificial intelligence, "the appropriately programmed computer with the right inputs and outputs literally has a mind in exactly the same sense that you and I do".[6] This tells that not only the devices would just referred to indeed be intelligent and have minds, etc., but mental qualities of a sort can be attributed to teleological functioning of any computational device, even the

[3] Tanimoto (1987, pp. 6–7).

[4] Simon (1987, p. xi).

[5] Searle (1996, p. 41).

[6] Searle (1987, p. 210).

very simplest mechanical ones such as a thermostat. Here, the idea is that mental activity is simply the carrying out of some well-defined operations, frequently referred as an algorithm. We may ask here as to what an algorithm actually is. It would be adequate to define an algorithm simply as a calculation procedure of some kind. But in the case of a thermostat, the algorithm is extremely simple: the device registers whether the temperature is greater or lesser than the setting, and then, it arranges for the circuit to be disconnected in the former case and to remain connected in the latter. For understanding any significant kind of mental activity of a human brain, a very complex set of algorithms has to be designed to capture the complexity of the human mental activities. The digital computers are approximations to the complex human brain.

The strong artificial intelligence view is that there are differences between the essential functioning of a human being (including all its conscious manifestations) and that of a computer and it lies only in the much greater complication in the case of brain. All mental activities such as thinking, feeling, intelligence, etc., are to be regarded, according to this view, merely as aspects of this complicated functioning of the brain; that is to say that they are the features of the algorithm being carried out by the brain. The brain functions like a Turing machine[7]that carries out all sets of complicated computations. And these brain style computations naturally designed like a computing machine to think, calculate, and carry out algorithmic activities. Again, these algorithmic activities give rise to all mental phenomena like thinking, feeling, willing, decision making, etc.

Moreover, the supporters of strong AI argue that we have every reason to believe that eventually, computers will truly have minds. As Winston says that, "intelligent robots must sense, move and reason".[8] Accordingly, intelligent behaviour is interpreted as giving rise to abstract automation. That is to say that an artificial, non-biological system could thus be a sort of thing that could give rise to conscious experience. They claim that humans are indeed the same kind as machines in general, and in particular, our mental behaviour is finally the result of the mechanical activities of the brain. The basic idea of the computer model of the mind is that mind is the software and the brain is the hardware of a computational system. The slogan is: "the mind is to the program, the brain is to the hardware".[9] For strong AI, there is no distinction between brain processes and mental processes. Because the process is a happening in the brain is a computational process, the mind is the alternative name of the brain, which is ultimately the same as a machine.

However, the theory of computation deals wholly with abstract objects such as Turing machine, Pascal program, finite state automation, and so on. These abstract objects are formal structures, which are implemented, in formal systems. But the notion of implementation is the relation between abstract computational objects and physical systems. Computations are often implemented on non-biological substrates—such as synthetic silicon-based computers. Whereas the computational systems are abstract objects with a formal structure determined by their states and state transition relations, the physical systems are concrete objects with a causal structure determined by their internal states and the causal relations between the states. It may be pointed out that a physical system implements a computation when the causal structure of the system mirrors the formal structure of the computation. The system implements the computation if there is a way of mapping states of the system onto states of the computations so that the physical states that are causally related map onto the formal states that are correspondingly formally related.[10]

The fact is that there is rich causal dynamics inside computers, as there is in the brain. There is real causation going on between various units of brain activity precisely mirroring patterns of causation between the neurons. For each neuron, there is a specific causal link with other neurons. It is the causal patterns among the neurons in the brain that are responsible for any conscious experiences that may arise. The brain, as Marvin Minsky says, "happens to be a meat machine".[11]He points out that the brain is an electrical and chemical mechanism, whose organization is enormously complex, whose evaluation is barely understood and which produces complex behaviour in response to even more complex environment. These evaluations and behaviour may be moral/social/legal. Then the questions are: What would the world be like if we had intelligent machines? What would the existence of such machines say about the nature of human beings and their relation to the world around them? What would be machine rights or duties? Can machines take moral as well as legal decision? Can there be an artificial moral agent? Let us discuss and find out the philosophical issues that are embedded in the very idea of artificial moral agent.

3 Artificial moral agent

There are many ethical issues related to artificial intelligence because as we have seen according to AI scientists, there are no distinctions between the mind and the machine. Not only this identification between the mind and

[7] Turing (1950, pp. 433–460).

[8] Winston (1984, p. 380).

[9] Searle (1990, p. 21).

[10] Chalmers (1996, p. 321).

[11] This view quoted by McCorduck (1979, p. 70).

the machines brings moral issues but also ascribing mental qualities to machines. Could an artificial intelligence learn the difference between right and wrong? Is there anything called artificial moral agent? Is there anything called machine ethics? And similar questions have been raised in the last section. But for a section of AI scientists, machine ethics is possible because the ultimate goal of artificial moral agent[12] is to create a machine that itself follows an ideal ethical principle; that is to say, it is guided by these principles in decisions it makes about possible courses of action it could take. In AI, there are two kinds of ethical agents, i.e., implicit ethical agent and explicit ethical agent. According to James Moor, a machine that is an implicit ethical agent is one that has been programmed to behave ethically, or at least avoid unethical behaviour, without an explicit representation of ethical behaviour. It is constrained in its behaviour by its designer's ethical principles. A machine that is an explicit ethical agent, on the other hand, is able to calculate the best action in ethical dilemmas using ethical principles.[13] Using Moor's terminology, most of those working on machine ethics would say that the ultimate goal is to create a machine that is an explicit ethical agent or artificial moral agent.

It is important to see how a machine would gather information to make a decision and incorporate it into general behaviour. Ethics can be seen as both easy and hard. It appears easy because we all take ethical[14] decisions on a daily basis. But this does not mean that we are all experts in ethics. It is hard because it is a field that requires much study and experience. It is easy because value decisions are part and parcel of human existence—we all make moral choices. But it is difficult to make explicit the criteria of ethical decision making, thereby to codify it or teach it to a machine. Unlike many other knowledge bodies, ethics in intricately connected to our ability to have first-person perspectives. AI researchers must engage with moral philosophers just as moral philosophers must engage with AI researchers. And there is a fundamental distinction in their approaches. Therefore, machine ethics is an interdisciplinary field because AI scientists believe that it is possible to do research in machine ethics. As we know, ethics, by its very nature, is the most practical branch of philosophy. It is concerned with how agents ought to behave when faced with ethical dilemmas. Research in machine ethics has the potential to discover problems with current theories, perhaps even leading to the development of better theories.[15] For example, the way moral philosophers are spending time discussing actual cases that occur in the field of biomedical ethics, in the same way, AI researchers working with moral philosophers might find it helpful to begin with this domain, discovering a general approach to computing ethics that not only works in this domain, but could be applied to other domain as well.

Now the question is: how may we feel confident that an artificial moral agent would make the right decision in situations that were not anticipated? According to AI scientists, an explicit ethical agent is able to explain why a particular action is right or wrong by appealing to an ethical principle. A machine that has learnt, or been programmed, to make correct ethical judgements, but does it not have principles to which it can appeal to justify or explain its judgements, there seems to something lacking, something essential missing, for being accepted as an ethical agent. An ethical agent does not blindly implement rules or principles, the agent accommodates the detailed context and is able to offer an explanation for choosing a particular course of action.

The non-mathematical value ethical theories are based on habit or customs or free choices and seem to fit better with the training regimens of artificial neural mechanism. The artificial neural network mechanism talks about moral decision-making in different choices in different environments. Now the philosophical question is: Is morality derived from computing artificial neural network? Does this artificial neural network follow different procedure? The choice of particular food pattern in a restaurant is a decision no different in kind from deciding whether or not to steal property that my friends so in a slapdash fashion left unthinking.

However, here we could make the distinction between the sense of preferences and the sense of moral decisions as we have pointed out in the above discussion and thereby inviting consideration a question for artificial moral agent. There is a conflict between inclination and duty, between what one wants to do and what one ought to do, rather than as a conflict between mere preferences or wants.

Immanuel Kant made a similar point when he distinguished between an agent that acts from a sense of duty, rather than merely in accordance with duty. Although, there are strong distinctions between Kantian moral agent and artificial moral agent, but there is possible of creation of Kantian artificial moral agent, which will be capable of performing moral action. But Kant said, "Act only on those maxims whereby you can at the same time will that they should become universal laws".[16] This maxim sanctions the development of the artificial moral agent, or ethical robot is applied to the categorical imperative. As we know that Kant's moral philosophy is based on deontological

[12] Allen et al. (2000).

[13] Moor (2006).

[14] Ethical in the sense of 'having an ethical property' not a morally commendable/right decision.

[15] Anderson and Anderson (2007, p. 16).

[16] Kant (1993, p. 30).

framework, that is, the idea that doing what is right is not different from one's duty. But we have to take into account that the duty should be exercised through rationality.

Furthermore, an important aspect of Kant's ethics is that ethics is transcendental and which has ontological status, because for him God is a moral concept. Again, he is arguing that morality and God cannot be separated. This does not presuppose a belief in God. It merely contends that it gives an adequate description of the requisite conditions we need to make sound judgments. But a human judge may commit some kind of partial judgments out of emotion or sympathy, even if she or he is well acquainted with the Godly character of moral actions. Then a case might be made for saying that an unemotional AI could be considered a better moral judge than a human.[17] As we have seen in the above discussion that such machine could store and retrieve more factual data, not to be disturbed by human passion and interests, it could perform from detached choices, and thus AI follows on a deontological framework.

Again, according to him "Act in such a way that you treat humanity, whether in your person or the person of any other, never merely as a means to an end, but always at the same time as an end".[18] The persons are capable of moral evaluation, must be considered as ends in themselves rather than as means to other ends. That is to say that moral action is free from any kind of determinate factors. Thus, Kantian categorical imperatives are done out of moral duty but committed to freedom by moral agents who are rational and free. Otherwise, these maxims are not applicable. If this is so, it is impossible to have Kantian artificial moral agent. Then it is impossible to have an artificial moral agent from Kantian perspectives. This is because of the fact that artificial moral agent violates the categorical imperatives, freewill, and rational agent, which are laid down by Kantian ethics.

The fundamental difference underlies all of Kant's ethical metaphysics and is used to mark crucial moments in his moral philosophy. It provides the bedrock for the distinction between heteronomy and autonomy, that is, between being determined to act by something outside of the self and being self-determined, and it serves to demarcate the difference between the various kinds of imperatives that allows Kant to single out the categorical imperative as the moral one.

However, The 'moral agent' in Kant's picture is the person suspended between inclination and duty, where inclination is derived from desire which, in turn, is always fixed to something outside the self and where duty is determined by pure practical reason as action in conformity with the 'universality of a law as such'.[19] This suspension, to be clear, does not make the person moral, but rather *able to be* moral, and success as a moral agent is left to be decided by whether she follows inclination or duty in a situation where the two diverge.[20] Again, the categorical imperative by itself does not specify enough to agree to the AI scientists to implement Kantian morality in an AI machines, because the acts from reason alone is not a real option for artificial moral agent at all.

If we believe that machines could play a role in improving the lives of human beings, we must feel confident that these machines will act in a way that is ethically acceptable. The ethical component of machines that affects human lives must be transparent, and principles that seem reasonable to human beings provide that transparency.

There are many philosophical objections to machine ethics. Moral philosophers need to ask whether ethics is the sort of thing that can be computed or whether machines are type of entities that can behave ethically. To be a moral agent, who must be capable of acting with intentionality, this requires consciousness and freewill. Only a being that has feelings would be capable of understanding the feelings of others. Since, many doubt that machines will ever be conscious, have freewill, or emotions, this would seem to rule them out as being moral agents. In reply to above statements, according to Anderson and Anderson, "This type of objection, however, shows that the critic has not recognised an important distinction between performing the morally correct action in a given situation, including being able to justify it by appealing to an acceptable ethical principle, and being held morally responsible for the action. Yes, intentionality and free will in some sense are necessary to hold a being morally responsible for its actions, and it would be difficult to establish that a machine possesses these qualities, but neither attribute is necessary to do the morally correct action in an ethical dilemma and justify it. All that is required is that the machine act in a way that conforms with what would be considered to be the morally correct action in that situation and be able to justify its action by citing an acceptable ethical principle that it is following".[21]

Moreover, the connection between emotionality and being able to perform the morally correct action in an ethical dilemma is more complicated. Even if a robot can learn from its own prior actions, it is not necessarily moral. The complex quality of judiciousness is still needed for several reasons. The judicious quality allows the agent to recognize when it has encountered another agent or an appropriate object of moral reasoning. It allows the

[17] LaChat (1986).

[18] Kant (1993, p. 36).

[19] Ibid., p. 30.

[20] Beavers (2009).

[21] Anderson and Anderson (2007, p. 19).

artificial moral agent to understand the potential needs and desires of another, as well as what might cause harm to the other. This requires at least a rudimentary theory of mind, that is, a recognition that another entity exists with its own thoughts, beliefs, values, cultures, and needs. This theory of mind need not take an extremely complex form, but for an agent to behave morally, it cannot simply act as though it is the only entity that matters. The moral agent must be able to develop a moral valuation of other entities, whether human, animal or artificial. It may have actuators and sensors that give it the capacity to measure physical inputs from body language, stress signs, and tone of voice, to indicate whether another entity is in need of assistance and behave morally in accordance with the needs it measures. Judicious, and not merely rationality, is critical for developing and evaluating moral choices; just as emotion is inherent to human rationality, it is necessary for machine morality.[22]

Certainly one has to be sensitive to the suffering of others to act morally. It is not clear how a machine could be trained to take into account the suffering of others in calculating how it should behave in an ethical dilemma, without having emotion itself. Furthermore, it is important to recognise that having emotions can actually interfere with a being's ability to determine, and perform the right action in an ethical dilemma. Humans are prone to getting 'carried away' by their emotions to the point where they are incapable of following moral principles. Therefore, emotionality can be viewed as a weakness of human beings that often prevents them from doing the 'right thing'.[23] One way to view the puzzle of machine ethics is to consider how we might program intelligent system that would themselves refrain from evil and perhaps promote good. Here, there may be a kind of altruistic behaviour in the case of artificial moral agent. As we know that altruism is a kind of selflessness for the betterment of others at the cost of oneself, and in the case of artificial moral agent action, AI sacrifices for human society without any returns for itself. Now a fundamental question is: Is there any free choice in artificial moral agent? As we have seen in the above discussion that the notion of 'freedom of choice' plays a vital role being a moral agent. In the case of artificial moral agent, it is 'hard' to imagine the idea of freedom of choice, which functions in a particular deterministic way. Indeed, if there are no choices, then actions, however 'good', say that of an artificial moral agent, can hardly be seen as moral choices, simply because they are not chosen, but merely implemented.

There are two clear approaches to developing machine ethics—first to develop an ethics-framework for machines to work because they work so closely with humans. The other is to develop an ethics-framework to accommodate the possibility of moral artificial agents, who stand on equal moral footing with human persons, exhibiting self consciousness and thereby deserving moral rights. Then, the research in machine ethics agenda will involve testing the feasibility of a variety of approaches to capturing ethical reasoning, with differing ethical bases and implementation formalisms, and applying this reasoning in systems engaged in ethical sensitive activities. This research will investigate how to determine and represent ethical principles into system's decision procedure, make ethical decisions with incomplete and uncertain knowledge, provide explanations for decisions made using ethical principles, and evaluate systems that act based upon ethical principles. The question that arises is that, if machines were to become moral intelligent agent, what moral obligations we would have toward them. Would we treat them as slaves or as equal? Should they have rights? Should computers be allowed to make battlefield decisions in war? Could it be autonomous? Is it ethical to ascribe mentality to machine? And also there are many other moral issues like rationality, freedom, and value related to artificial intelligence. Any scientific investigation always strives towards the difference between value and scientific fact.

A discussion of the ethics in self-driving cars could bring these issues to the fore. Self-driving cars or autonomous vehicle technology has advanced significantly, and it is only political and ethical considerations that stand between it and the real world. A significant reason for these political and ethical considerations is that humans are free to err, but self-driving cars are not. Any collision involving a self-driving car is conceived less like an accident and more as a programming/learning error. As Gunkel points out 'What needs to be decided, therefore, is at what point, if any, might it be possible to hold a machine responsible and accountable for an action?"[24] The ethical issues regarding self-driving cars are twofold—first, can there be a foolproof programming for self-driving cars, the errors in which can cost human lives; and second, if self-driving cars (machine)learn, can they be granted responsibility or accountability for their decisions? The second question anticipates the question of the artificial moral agent, and the first question sticks to the conventional instrumental understanding of technology wherein the flaws and successes of a machine are directly credited to the human designer.

There are many philosophical challenges to artificial intelligence because human minds have greater standing in virtue of their higher rational capacities, particularly the perhaps unique capacity for subjective experience. The

[22] DeBaets (2014).

[23] *Ibid.*

[24] Gunkel (2012, p. 18).

subjective experience has a more varied range of emotional responses in virtue of their rationality. Human beings are the product of natural conceptions, and the morality has close relationship with consciousness, rationality, and our ability to have first-person perspectives. Therefore, the subjective notion of moral agency plays very vital role in the case of moral act.

4 The subjective notion of moral 'agency'

Each subjective being has a uniqueness of experiencing things, and it is important to understand the very nature of their subjective experiences. Thus, consciousness seems to involve something that is essentially subjective. In case of a conscious mind, there is a subjective point of view, which is accessible only to the conscious being itself. Consciousness is a phenomenon, which cannot be measured, observed or experienced in public because it is a personal matter. Even if it is personal, the notion of 'privacy' is not applicable to it because as we know that the notion of 'privacy' as we know from Wittgenstein's private language argument does not apply to the *personal subjective experience* (PSE) in the sense that the PSE are inter-subjectively intelligible and that they are available for interpersonal communication.[25] It can be known only from a first-person perspective, but not from the third person perspective, i.e., objective or AI perspective. As Thomas Nagel shows that subjectivity is a fundamental feature of consciousness. It is because of the fact that consciousness is what makes the mind–body problem intractable, as 'subjectivity' is its most conceptually troublesome feature. Self is the subjectivity, which encompasses our feelings, thinking, and perception. The qualitative character of experience is what it is like for it's subject to have the experience. In his article, '*What it is like to be a bat?*' Nagel presents the notion of subjectivity, which proves the irreducibly subjective character of experience. He writes, "Conscious experience is a widespread phenomenon. It occurs at many levels of animal life, though we cannot be sure of its presence in the simpler organisms, and it is very difficult to say in general what provides evidence of it… no matter how the form may vary, the fact that an organism has conscious experience *at all* means, basically, that there is something it is like to *be* that organism…But fundamentally an organism has conscious mental states if and only if there is something it is like to *be* that organism—something it is like *for* the organism".[26]

We can know the physical facts about a bat, but we cannot know what it is to be like a bat. The Nagelian thesis is that we cannot fully comprehend the bat's experience; because we cannot adopt its point of view of the world. The subjective experiences of the bat are beyond our comprehension. The objective facts regarding the organism do not and cannot explain the subjective character of the bat's experiences. AI knowledge body cannot answer the question- 'what is it like to be an artificial moral agent?' Thus, Nagel sees the subjectivity of consciousness as a challenge to AI and at the same time to AI ethics. It is to AI because AI theories cannot explain one's phenomenal consciousness. Thus, subjectivity is just not captured in third person vocabulary. In Nagel's word, subjectivity is, "…the subjective character of experience. It is not captured by any of the familiar, recently devised reductive analyses of the mental, for all of them are logically compatible with its absence. It is not analyzable in terms of any explanatory system of functional states, or intentional states since they could be ascribed to robots or automata that behaved like people though they experienced nothing".[27]

However, conscious experience is the representation of subjectivity. Facts about conscious experience, therefore, do not exist independently of a particular subject's point of view. Objective phenomena have a reality independent of appearances, but subjective phenomena are phenomenological appearances. Nagel claims that AI stands little chance of providing an adequate third-person account of consciousness, as there is no objective nature to phenomenal experience. Phenomenal experience cannot be observed from multiple points of view. Hence, from the subjective point of view, we know what it is to be like us, but we do not know what it is to be like a bat. We do not know what it is like to have sonar experiences. Sonar experiences imply a subjective perspective, and we must occupy that particular point of view to know that sonar experiences. For example, we must be in the bat's position to know the bat's sonar experiences. Nagel writes, "…we may ascribe general *types* of experience on the basis of the animal's structure and behaviour. Thus, we describe bat sonar as a form of three-dimensional forward perception; we believe that bats feel some versions of pain, fear, hunger, lust and that they have other, more familiar types of perception besides sonar. But, we believe that these experiences also have in each case a specific subjective character, which it is beyond our ability to conceive. And if there is conscious life elsewhere in the universe, it is likely that some of it will not be describable even in the most general experiential terms available to us."[28] In contrast to subjective experience, the experience of knowing the square root of 144 as 12 or that table salt is a compound of sodium and chlorine does not require any kind of

[25] Wittgenstein (1976, Part I. Sec 243–244).

[26] Nagel (1998, p. 519).

[27] Ibid.

[28] Ibid., p. 521.

experience. There is simply no phenomenological or qualitative feel to these knowledge claims. This is not to deny that it may require some experience. It could be that anyone who has this knowledge must also have experience. And what makes mathematical and scientific knowledge objective is not the particular kind of experience accompanying that knowledge. However, to know what it is like to see red entails having a particular kind of experience, which is the experience of seeing red. In Nagel's word "In the case of experience, on the other hand, the connection with a particular point of view seems much closer. It is difficult to understand what could be meant by the *objective* character of an experience, apart from the particular point of view from which its subject apprehends it".[29]

This subjective character of experience cannot be captured by any functional or causal analysis. Therefore, we do not know how AI can explain consciousness and be ethical. AI rules out the subjective point of view, and therefore, fails to explain human's ethical experiences. That is to say that subjective conscious experiences itself cannot be explained on the basis of what we observe about the machine effects. While rejecting AI model of subjective artificial moral agent, we would like to point out that conscious states are simply not, *qua* conscious states, potential objects of perception; they depend upon the brain, but they cannot be observed by directing the senses onto the brain. You cannot see a brain state *as* a conscious state.[30]

In case of subjectivity, experiences are representations. For example, my visual experience of my blue shirt is a mental representation of the shirt as being blue. When I introspect on my visual experience, I form a second-order representation of the first-order representation of the shirt. Other people have syntactically similar second-order representations. But each individual can introspect only with her/his own experiences. For Lycan, this is the ultimate explanation of subjectivity. He analyses Nagel's view and replies that, "…seeing someone's brain in a state of sensing-blazing-red is nothing at all like sensing blazing red oneself."[31] Similarly, in case of the bat's sonar sensation S: We do not have the sonar sensation S; we cannot ourselves feel S. We do not know what it is like to have S (we do not have cognitive access to S) in the way the bat does.[32] These facts are obviously true and accepted even by the materialists. When we observe the bat, at that time, we observe only some physical or functional state, but thereby we do not have that conscious state ourselves; we do not have the same perspective with respect to it.

The individual consciousness can be understood or reported only from the first-person point of view and not from the third-person objective point of view. An objective representation can be described in an objective way. This representation or concept is a function from the world to the individuals. AI takes it as an objective fact and tries to describe it as functions of mind. An experience is held to be a conscious experience, which is likely for the subject of the experience to have it. Thus, we have to accept the qualitative feel of experience. This qualitative feel, unique to every distinguishable experience, is supposed to be what it is like for the subject of the experience to have the experience.

Subjectivity is the most important feature of a moral agent and the judgments are taken as 'subjective' when their truth or falsity is not a matter of fact or 'objective' criteria but depends on certain attitudes and feelings of the maker of the judgment. For Searle, the term 'subjective' is an ontological category. The statement 'Someone is feeling pain in his/her leg' is completely objective, because it is true by the existence of a fact and is not dependent on the attitude or opinion of the observer. But the actual pain itself has a subjective mode of existence, which implies that consciousness is subjective. The term 'pain' is subjective as it is not equally accessible to any observer. Therefore, every conscious state is always someone's conscious state.[33] Someone has a special relation to her/his own conscious states, which is not related with other people's conscious states. He says, "Subjectivity has the further consequence that all of my conscious forms of intentionality that give me information about the world independent of myself are always from a special point of view. The world itself has no point of view, but my access to the world through my conscious states is always perspectival, always from my point of view".[34]

A theory of consciousness needs to explain how a set of neurobiological processes can cause a system to be in a subjective state of sentience or awareness. We accept the view that subjectivity is a ground floor, irreducible phenomenon of natural science. So being objective cannot explain how this is possible. Searle says that 'consciousness' stands these subjective states of sentience or awareness that we possess when we are conscious, that is, during the period we are not in coma or are not unconscious. Thus, consciousness is a subjective qualitative phenomenon. It is not a mechanical state or a certain kind of set of a computer program as many philosophers believe. There are two most common mistakes about consciousness such as that it can be analyzed computationally. The AI shows that conscious mental states are mechanical or computational states. This

[29] Ibid., p. 523.

[30] McGinn (1997, p. 533).

[31] Lycan (1987, p. 76).

[32] Ibid.

[33] Searle (1994, p. 95).

[34] Ibid.

is the main reason that machines remain lifeless and inert devices, even if they are manipulated intelligently by human designers. The robot is simply a machine, which is essentially distinct from the human in its behavioural aspects. It gives us the view that for a system to be conscious, it is both necessary and sufficient that it has the right computer program or set of programs with the right inputs and outputs. There is no logical connection between the inner, subjective, qualitative mental states and the external, publicly observable output. Our mental states cannot be fully represented in a machine or in a computer. We claim that the domain of ethics is built on our ability to have the first person perspective.[35] Because we can imagine what the other may be feeling, we have the input to be more judicious in our actions and policies. It is because of this subjective feeling, that morality is subjective. A most common tool in ethics is to imagine the situation from someone else's perspective and then decide. This ability is built upon two capacities—first, to have a first-person perspective and, second, to imagine the other's first-person perspective. These two capacities lay the foundation for our ethical domain. And it is here that AI does not clearly harbour these two capacities so as to have the moral ability.

5 Conclusion

The subjective feeling is necessary to be an ethical being. It is because of the fact that to be ethical, depends on the subject, but not in the case of artificial moral being, which lack subjective moral feeling. It is 'I', who is an 'agent' that feels such moral emotion. The 'I' is the central problem of consciousness. AI tries to explain how conscious experience arises from the mechanical processes of an artificial agent. Even if they can prove conscious states to be caused by the mechanical states of the machine, they cannot show how and why the conscious states belong to the 'subject'. Even if the 'subject' that has consciousness is not identical with the brain states either. The 'subject' is distinct from the body.[36] If the mental world is irreducible and we have a reasonable assurance that mind at any cost stands beyond the horizon of the physical world, we can make a safe bet that mind has a reality of its own and that AI theory of all sorts fail to understand the inner dynamics of the mind. If it fails to explain the subjective character, it fails to become an artificial moral agent. Thus, the very idea of an artificial moral agent or machine ethics fails to be a moral agent and it is because of the fact that the question of 'why be moral?' is not applicable to an artificial moral agent; rather it is applicable to subjective beings only.

[35] See, Nath (2016).

[36] Nath (2009).

References

Allen C, Varner G, Zinser J (2000) Prolegomena to any future artificial moral agent. J Exp Theor Artif Intell 12(3):251–261

Anderson M, Anderson SL (2007) Machine ethics: creating an ethical intelligent agent. AI Mag 28(4):15–26

Beavers AF (2009) Between angels and animals: the question of robot ethics, or is Kantian moral agency desirable? In Association for practical and professional ethics. Eighteenth annual meeting, Cincinnati, Ohio, March 5–8. http://faculty.evansville.edu/tb2/PDFs/Robot%20Ethics%20-%20APPE.pdf. Retrieved on 1 Oct 2017

Chalmers DJ (1996) The conscious mind. Oxford and New York: Oxford

DeBaets AM (2014) Can a robot pursue the good? Exploring artificial moral agency. J Evolut Technol 24(3):76–86

Gunkel DJ (2012) The machine question: critical perspectives on AI, robots, and ethics. MIT Press, Cambridge

Haugeland J (1989) Artificial intelligence: the very idea. The MIT Press, Cambridge

Kant I (1993) [1785] Grounding for the metaphysics of morals. Translated by Ellington, James W (3rd Edn), Hackett

LaChat MR (1986) Artificial intelligence and ethics: an exercise in the moral imagination. AI Magz 7(2):70–79

Lycan WG (1987) Consciousness. The MIT Press, Massachusetts

McCorduck P (1979) Machines who thinks. W.H. Freeman and Company, San Francisco

McGinn C (1997) Can we solve the mind–body problem? In: Ned B, Owen F, Güven G (eds), The nature of consciousness, The MIT Press, Cambridge

Moor JH (2006) The nature, importance and difficulty of machine ethics. IEEE Intell Syst 21(4):18–21

Nagel T (1998) What it is like to be a bat. In: Ned B, Owen F, Güven G (eds), The nature of consciousness, The MIT Press, Cambridge, MA

Nath R (2009) Philosophy of artificial intelligence. A critique of the mechanistic theory of mind. The Universal Publishers, Boca Raton

Nath R (2016) Can naturalism explain consciousness: a critique. Artif Intell Soc J Knowl Cult Commun. doi:10.1007/s00146-016-0671-6

Searle JR (1987) Minds and brains without programs. In: Blakemore C, Greenfield S (eds), Mindwaves: thoughts on intelligence, identity, and consciousness, Bail Blackwell, Oxford

Searle JR (1990) Is the brain a digital computer? Proc Address Am Philos Assoc 64(3):21–37

Searle JR (1994) The rediscovery of the mind. Harvard University Press, Cambridge

Searle JR (1996) Minds, brains and science. Harvard University Press, Cambridge

Simon HA (1987) Guest foreword. In: Stuart CS (ed), Encyclopedia of artificial intelligence, vol 1, Wiley, New York

Tanimoto SL (1987) The elements of artificial intelligence. Computer Science Press Inc, Maryland

Turing AM (1950) Computing machinery and intelligence. Mind 49:433–460

Winston PH (1984) Artificial intelligence. Addison-Wesley Publishing Company, London

Wittgenstein L (1976) Philosophical investigations. Anscombe GEM (translated), Basil Blackwell, Oxford