Vol. 18: 49–60, 2018
https://doi.org/10.3354/esep00186

ETHICS IN SCIENCE AND ENVIRONMENTAL POLITICS
Ethics Sci Environ Polit

Published September 18

Can humanoid robots be moral?

Sanjit Chakraborty*

Department of Philosophy, Jadavpur University, Kolkata, 700032, India

ABSTRACT: The concept of morality underpins the moral responsibility that not only depends on the outward practices (or 'output', in the case of humanoid robots) of the agents but on the internal attitudes ('input') that rational and responsible intentioned beings generate. The primary question that has initiated extensive debate, i.e. 'Can humanoid robots be moral?', stems from the normative outlook where morality includes human conscience and socio-linguistic background. This paper advances the thesis that the conceptions of morality and creativity interplay with linguistic human beings instead of non-linguistic humanoid robots, as humanoid robots are indeed docile automata that cannot be responsible for their actions. To eradicate human ethics in order to make way for humanoid robot ethics highlights the moral actions and adequacy that hinges the myth of creative agency and self-dependency, which a humanoid robot can scarcely express.

KEY WORDS: Humanoid robots · Morality · Responsibility · Docility · Artificial intelligence · Conscience · Consciousness

INTRODUCTION

This paper begins with prolegomena on the claim of the interrelated nature of moral agency and ethical conscience. Who can be a 'moral agent'? The simple answer would be a conscious 'human being' who has moral conscience (the sense of right and wrong) and who intends to act morally following rules and conduct. Morality supervenes on rational agency. Rationality and consciousness are the necessary properties of moral agency that should be guided by responsibilities. Moral responsibility is not only an outward practice but a feeling generated by the intentional states of mind. In line with the query, the most prominent question that has engendered a tremendous debate in the history of moral ethics is whether moral values are subjective or objective. This is a vexing issue involving vast areas of study. Morality identifies in ascribed conduct and particular codes that are mutually followed by members of society. It is interesting to consider that values are much closer to the general features of the objects that an agent can favour or disfavour from different contexts and situations. Moral values seem subject-centric, as they depend on the rationality and conscience of the particular agent. Moral values are subjective, as they differ according to the different intentions, motivations and choices of individuals. Subjective value is reliant on our method of valuing or judging of it. Subjective morality is closer to a 'knowing how' process (in an epistemic sense) where moral values depend on the agent's skills and uses. Thus, a moral value becomes context sensitive, or relative, since a subject can accept a moral judgement as being right at a particular time and in a given context, but according to the change of time and context, s/he can view the same moral judgement as being wrong (Chakraborty 2017).

If we make a reasonable observation of the idea of objectivity, then it is noticeable that although our knowledge of the external world is prejudiced by subjective experiences, it still remains objective. This is because knowledge of the external world is open to 'inter-subjective' verification and agreements. Moral values are based on universalisability and provide inalienable conditions that firmly stand for the objectivity of moral values. The principle of universalisability attains a strong position here. Moreover,

*Corresponding author: cogitosanjit@gmail.com

Publisher: Inter-Research · www.int-res.com

moral values are objective because moral values are valued and agreed upon by all people across a number of cultures, and universalisability has some universal applicability (Mackie 1977). The parameter of morality depends on the universalistic sense that is called the prescriptive mould of morality. David Hume, for example, held that the prescriptive term 'ought' cannot be deduced from a descriptive premise like 'is' that sounds more factual (Stove 1978). Hume's inducibility thesis of an ethical statement based on the factual proclamation may be a fantasy. We can deduce 'ought' conclusions from factual sentences like 'Any moral act is permissible', for instance.

One can argue that some natural principles may determine our moral values. A wide range of scientific proclamations maintains the distinction between fact and value (Putnam 2002, Chakraborty 2018). The objective ground-based science aims to restrict any subject-centric value in its theorems, but the 'inescapability value-laden' hypothesis of Graham (1981) emphasises an intertwined relationship between science and ethics. The conception of autonomy of science mingles in the realm of ethics because of its applicability in our moral and social life. Scientific values prop up social orders in a universalistic sense. A deformity arises when we claim that one can only derive moral values from the objective set of natural principles, as without the conscious being or rational agents, the conception of morality does not exist. Only rational human beings can think about morality and behave accordingly.

Traditionally taken as an aspect of human endeavour, moral values actually impart on normative stances (Zimmerman 2015). It seems true that morality appears through the objective affairs of the observer, but quite naturally, morality is subject-centric as it depends on rational human beings, i.e. morality supervenes on agents rather than on objects. Rawls (1971) considered that moral values look for the preference of the agent, but not in the sense of subjective preference, and stressed an internal reflective method that rationally justifies values as a general principle. Alternatively, we may stress the constant interaction between subjectivity and intersubjectivity. One can spell out moral values as 'intersubjective'. Although it is no doubt an intentional state of the believers, morality in fact centres around what a subject believes (regarding moral values). The subject's belief inclines not only towards subjectivity, but also towards objectivity and others' choices, where subjective intentional states are referentially interrelated to the objective world (Chakraborty 2017). In short, we impose values on the objects, but to take them universally, we inflict a kind of 'collective illusion' upon them. This socio-biological thought is nourished by a Darwinian facet of objectivity of 'collective illusion' (Ruse 1986) by way of which we confirm moral values intersubjectively so that subjective preferences play a pertinent role in objective values.

HUMANOID ROBOTS AND THE IDEA OF MORALITY

In 1956 (during a 2 month workshop at Dartmouth College), computer scientist John McCarthy perhaps first introduced the term 'artificial intelligence' (AI) to articulate the study of intelligence in a computer program that is proficient in thinking non-numerically (McCarthy & Minsky 2006). Later, McCarthy's high-level computer language (LISP) introduced in the 'MIT AI Lab Memo No 1' (Arkin 1998) became a prominent AI programming language. Humanoid robots have been the monumental achievement in AI since the 1980s. AI primarily challenges human intelligence by introducing humanoid robots anchored in the model of the human brain and cognition in order to show how AI could go beyond human intelligence. This research takes rationality as a capacity to weigh a current measure against an ideal performance of a cognitive machine (Russell & Norvig 2016). For these authors, AI depends on the 4 categories of thinking humanly, thinking rationally, acting humanly and acting rationally. A system becomes rational if it goes towards right things. AI researchers try to defend these mentioned criteria intended for the conception of the rationality of machines (Russell 1997). The cognitive procedures of machines set up computer models of the brain like the human brain, as they aim to promote artificial intelligence that can think humanly (following cognitive ways) and accordingly act rationally. The Turing test invents the representational knowledge along with reasoning perusing a rational agent (which may be the computer) to attain the best outcome or good decisions of a 'child machine' (Turing 1950, p. 456). In this regard, Winston (1992) stated, 'The study of computations that make it possible to perceive, reason, and act.'

Now the critical question is: 'What is a humanoid robot?' Humanoid robots are indeed not human beings but have an exterior resemblance to humans. AI induces humanoid robots that have extraterrestrial life forms somewhat comparable to human beings maintained by synthetic biology.[1] Machine-embodied cognition procedures control the seem-

ingly human behaviours of humanoid robots. It is significant to clarify here that humanoid robots can imitate the human brain in terms of 'thinking', but the concept of 'mindfulness' is beyond the ability of humanoid robots, as the AI algorithms measure its software and sensors that cannot grasp human consciousness. The spectrum of the humanoid robot's thoughts is not in any way competing with the consciousness that can be computable. 'Thinking machine' is a misnomer in the sense of being an intelligent-rational individual. Russell & Norvig (2016, p. 37) defined a rational agent as follows, 'For each possible percept sequence, a rational agent should select an action that is expected to maximise its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has.'

Therefore, this definition also dents the idea of learning and autonomy of AI because an autonomous rational agent can comprehend knowledge through experience and argumentation that may surpass any imposed or *a priori* knowledge. The quest for AI returns to philosophy, as it looks for a module of the human brain to find out how the mind works in the realm of the physical brain, or what the decision-making procedure is that calculates the utility of an agent's decision and some other related issues. My aim here is not to stress the idea of the language acquisition of humans and the machine language-learning procedures in detail, since the main interest of the paper is to emphasise the morality of AI or, more specifically, the question: Can humanoid robots be moral?

In philosophy, we like to see morality in the light of human 'intention'. 'Intention' is often understood as an act guided by one's choice or preference influenced by desires, suitable both in humans and in robotic systems as the module of machine advance. However, the main question again is: Can humanoid robots be moral? This is because the concept of humanoid robots initiates the problem of AI that becomes a conundrum for philosophers. The popular stance of robots or humanoid robots disinclines any intention or rationality in robots or humanoid robots, so there is no question about the morality of humanoid robots. Recent robotic science has claimed that humanoid robots not only have choices, but they have certain beliefs or degrees of beliefs comprised through the numerical forms along with propositions (Scassellati 2002). Bayesian diagnostic systems strive to articulate robots' beliefs in this way. Even behaviouristic psychology accomplishes this procedure by introducing the technique of 'reinforcement' (Wilson & Keil 1999). Behaviourist accounts insinuate the reinforcement of the robot representing behaviours by replacing several goal-oriented processes that the robots or humanoid robots intend to maintain. For the behaviourist, the notion of the mind is as an information-processing machine. Niv (2009, p. 330) wrote:

> In recent years computational accounts of these two classes of conditioned behavior have drawn heavily from the framework of reinforcement learning (RL) in which models all share in common the use of a scalar reinforcement signal to direct learning. Importantly, RL provides a normative framework within which to analyze and interpret animal conditioning. That is, RL models 1) generate predictions regarding the molar and molecular forms of optimal behavior, 2) suggest a means by which optimal prediction and action selection can be achieved, and 3) expose explicitly the computations that must be realized in the service of these.

Here, the elemental question is: What is considered 'thinking'? In short, thinking is a biological process that only conscious human beings and animals can perform. Thinking is doubtlessly allied to intelligence. Similarly, 'intelligence' seems nothing but some rationale-based decisions or choices that stand for reality and has some argumentative grounds behind it. Moreover, intelligence is a kind of toolbox by which an agent can accomplish his/her aim. An intelligent human being would not want to disobey the set of rules or customs that a society has extensively practiced. A creature that has intelligence cannot perform anything without having a goal and cannot comprehend something without having passions and emotions. However, if we consider that humanoid robots are moral, then the most pressing question would be: Can humanoid robots think about morality? We often incur the same question about some of our fellow humans who appear careless about moral principles but still enjoy life in their own way within society. The argument is that humanoid robots cannot be moral for the following 2 reasons:

(1) Humanoid robots cannot construct a distinction between truth and justification regarding the concepts of morality. Only human cognitive abilities can ensure these concepts as they have to pass through a lengthy history (socio-cultural based) and could learn the knowledge through constant practices (in their case of communication) that is highly lacking in humanoid robots.

[1]The biology of the machine is obviously maintained by 'synthetic biology', a quest for artificial designing of human biological components or living organisms etc. See http://hudsonrobotics.com/products/applications/synthetic-biology/ (accessed 16 March 2018)

(2) Davidson (2001) considered that the discrimination and the physical phenomenon of different objects that relate to causal inputs rely on the creatures' perceptual capacities where not only perception is required, but 'prerogative sentiment' is also necessary. Davidson hinted at a difference between classification and discrimination. Classification is the process where concept possession takes an imperative role, and discrimination is the process that can be applied to non-linguistic creatures. Davidson considered classification to be an anthropocentric process, where to identify a concept is to classify its objective properties, events etc. The concept of discrimination is a sort of 'disposition' that is beyond any normative force. Thus, the discriminating process cannot recognise any informed mistake, as the creature does not have any knowledge relating to correct and incorrect behaviours. Davidson (1984, p.170) stated that 'Someone cannot have a belief unless he understands the possibility of being mistaken, and this requires grasping the contrast between truth and error – true belief and false belief.' Humanoid robots cannot grasp these processes and especially the concepts of absurdity, rationality, frustration, sorrow and all other human sentiments. Here, the problem is that without these sentiments and intentional attitudes, it would be difficult for a humanoid robot to concern itself with morality. Davidson (1984) was correct when he claimed that no one can have a belief about an object or identify a belief as a belief without a holistic web of beliefs. This thesis hints that without having language and the causal history of reference, a creature cannot insist on or construe a descriptive thought.

AI researchers (especially Tanimoto 1987) think that language is no longer a barrier for humanoid robots. Computer translation and imagination might take a prominent position in the history of our modern era. Therefore, creating a barrier between biological and technological domains becomes less prominent in our day-to-day life. However, it is doubtful that a humanoid robot can recognise conceptual systems and will be able to develop new concepts accordingly. Humanoid robots may have proto-concepts, i.e. the concepts that they develop are quite similar to human concepts. But in no way can they grasp human concepts, as they do not pass through the cultures and linguistic practices that educate a human being on how to use words (creativity of language use) and concepts in linguistic communication. Chomsky (1962), in reply to Harris (1951), stated that a child (language learner) can hear the utterances of the people correctly as 'grammatical' and 'ungrammatical' forms. A child develops his/her grammar from the collective data that are satisfied with some innate constraints. Putnam (2013, p. 758) argued against this view: 'The trouble with this view is that the factual premise is false. People don't object to all and only ungrammatical sentences. If they object at all, it is to deviant sentences — but they do not, when they correct each other, clearly say (in a way that a child can understand) whether the deviance was syntactic, semantic, discourse-theoretic or whatever.'

Putnam (2013) challenged the Chomskyan view that a child can extrapolate from the 2 lists of 'grammatical' and 'ungrammatical' sentences. Putnam (2013) showed that grammar is a property of language, not something intrinsic that exclusively locates in the speaker's brain. He did so by defending in favour of language as a system of strings with an inductive definition of predicates, that is easily parallel with semantics, deductive logic and also inductive logic, whereas grammar is a property of language. Putnam seems right that a child does not like to learn a branch of syntactic rules that looks very crazy. The ways of learning 'semantic rules' do not indicate any concern about the uninterrupted strings of gestures. The learning process of 'semantic rules' for a child can be possible in a 'structure-dependent' notion. A child can 'internalise' a structure-dependent rule and even be able to build up an 'inner representation' of abstract structural notions like sentences, verbs, nouns and so on in the case of language acquisition. In *Renewing Philosophy*, Putnam (1992, p. 15) claimed:

> The view that language learning is not really learning, but rather the maturation of an innate ability in a particular environment (somewhat like the acquisition of a bird call by a species of bird that has to hear the call from an adult bird of the species to acquire it, but which also has an innate propensity to acquire that sort of call) leads in its extreme form, to pessimism about the likelihood that human use of natural language can be successfully stimulated on a computer- which is why Chomsky is pessimistic about projects for natural language computer processing, although he shares the computer model of the brain, or at least the 'language organ' with AI researchers.

Putnam has foisted a misleading conception that he called 'innateness propensity' about his critic Chomsky's notable idea of 'innate language', while Chomsky argued that the computational model only works for I-language (generative grammar) but not for language use, as it has some productive senses. Linguists like Chomsky consider that we can program I-language and 'universal grammar' in the internal processor of humanoid robots, as language is not

fully innate. However, the faculty of language acquisition is as innate as our visual faculty. The process of language acquisition and the structure of grammar are doubtlessly an innate-based biological adaptation that no humanoid robots or machine can ever achieve. There are good reasons for being sceptical about computer modelling of language use, primarily (but not solely) because of 'the creative aspect of language use' (N. Chomsky pers. comm.).

MORALITY DOES NOT LOOK BACK TO MECHANICAL MINDS, BUT TO CONSCIOUS HUMAN BEINGS

Robotic scientists like Nick Bostrom and Ronald Arkin, who struggle with superintelligence, are concerned about its accuracy and philanthropic attitudes towards moral values, so that the superintelligent humanoid robots can certainly evaluate the best outcomes in the case of any moral decision and benefit human and non-human creatures. The robotic scientist generally promotes the man–machine interface in a more friendly way. The cultural dimension of man and the atmosphere of machines are not similar, but they both are engaged in the general welfare of the community and the world. Like human activities, humanoid robots' acts have some social implications. However, the idea of superintelligence, an advanced version of AI, assures that we cannot anthropomorphise superintelligence in a productive sense. It is a technologically advanced module for robots, supercomputers. The artificial minds of the superintelligent humanoid robots may be able to copy the human brain and more than that. The humanoid robots' autonomous agency that leads them to perform some cognitive actions is entirely different (I suspect) from human motivations. The reason is that humanoid robots do not have human-like psyches and cognitive states. There may be some similar cognitive states found between human beings and superintelligent humanoid robots, but subjectivity (if we can call it the 'inner conscious life of humanoid robots') is very different from the human consciousness, as Bostrom (2016) pointed out. Humanoid robots can have the power of imitation and can follow the commands that are installed in their software systems without being guided by any self-understanding. Humanoid robots do not have any emotion, conscience, rationality or self-knowledge (privileged access) that can convey any effect to their moral judgments.

Language acquisition and mathematical intelligence rest on the process of maturation that one can attain externally through reference borrowing and linguistic practices. In our practical experience, we see that 2 + 2 = 4, but there are some exceptions that we (human beings only) may find, for example in the case of liquids, where one drop plus one more drop of water remains one as the drops unite with each other. Robots are unable to pursue this kind of genuinely mathematic-based intelligence. Machines only use some adaptable commands that may quickly change over time. Human beings can acquaint with this type of exception, yet they learn the process of proper counting (not fuzzy counting) just by eliminating the mentioned exception from their thoughts. Accepting the process will direct us towards the 'descriptive theory of morality'. This sounds problematic since it does not fit with often used 'thick ethical terms' in language, such as cruelty, emotion, love etc. The important question is whether these terms sound like fact-based concepts or normative concepts. I borrowed the concept of 'thick ethical terms' from my mentor Hilary Putnam's excellent work *The Collapse of the Fact/Value Dichotomy* (2002). Putnam (2002, p. 35–36) wrote,

> The sort of entanglement I have in mind becomes obvious when we study words like 'cruel'. The word 'cruel' obviously-or at least it is obvious to most people, even if it is denied by some famous defenders of the fact/value dichotomy-has normative and, indeed, ethical uses. If one asks me what sort of person my child's teacher is, and I say, 'He is very cruel,' I have both criticised him as a teacher and criticised him as a man. I do not have to add, 'He is not a good teacher,' or, 'He is not a good man.' I might, of course, say 'When he isn't displaying his cruelty he is a very good teacher'. But, I cannot simply without distinguishing the *respects* in which or, the *occasions* on which he is a good teacher; and, the respects in which or, the occasions on which he is very cruel and then say 'He is a very cruel person and a very good teacher.' Similarly, I cannot simply say, 'He is a very cruel person and a good man,' and be understood. 'Cruel' can also be used purely descriptively, as when a historian writes that a certain monarch was exceptionally cruel, or that the cruelties of the regime provoked some rebellions. 'Cruel' simply ignores the supposed fact/value dichotomy and cheerfully allows itself to be sometimes used for a normative purpose and sometimes as a descriptive term. (Indeed, the same is true of the term 'crime.') In the literature, such concepts are often referred to as 'thick ethical concepts'. That the thick ethical concepts are counterexamples to the idea that there exists an absolute fact/value dichotomy has long been pointed out, and the defenders of the dichotomy have offered three central responses.

One may disagree with Putnam's case of 'cruel person and a good teacher', as the term 'good' can be entirely descriptive in the sentence that 'he is a cruel person and a good teacher'. If an agent says 'cruel person and a good teacher', then clearly the agent

confers not much significance on the teacher's eminence, but that is possible. Most of these words have purely descriptive uses. Some evaluative words have an adjective like 'outrageous' or 'admirable' that may have the same values. This is an unacceptable opinion. Facts and beliefs are associated with the propositions that we classify as true or false. Values seem quite unclear, but one could assert that they are standards for conveying good/bad verdicts to the states of affairs signified by the propositions. In the case of decision analysis or decision theory, these are represented as utility functions, which assign numbers (utilities) to outcomes, analogously assigning probabilities to beliefs. This is the way to carve up the world; we impose this fact/value distinction on it, just as we inflict other distinctions, such as choice options vs. states of nature (things I control vs. the things I do not control). Now the interesting query is that in the case of humanoid robots, the conception of choice takes different outlooks. Here, the society does not instruct the processes that may control or are under control of their choices. Actually, the robotic system is controlled by programmed systems that have been already inculcated or injected into the robots' software instructions. Rationality-based choices cannot be suitable for robots, as they do not undergo the framework of social life and customs that human beings always exchange with each other. Another point is that machines or humanoid robots cannot expand their knowledge. Expanding knowledge relates to the rationality and the subjectivity that enhance a sort of creativity in it. No machine can design the system independently.

Before entering into further debates, let me explain the query that concerns 'what makes intelligence and consciousness possible?' Intelligence seems nothing but some rational centric decisions that correspond to reality and have some argumentative grounds behind them. Obviously, intelligence has another criterion that follows the rules and social customs idiosyncratically. An intelligent being would not want to violate the set of rules and customs that society has extensively promoted. A creature that has intelligence cannot serve anything without any goals, passions and emotions. Pinker (1997, p. 60) stated that:

> Intelligence, then, is the ability to attain goals in the face of obstacles using decisions based on rational (truth-obeying) rules. The computer scientists Allen Newell and Herbert Simon fleshed this idea out further by noting that intelligence consists of specifying a goal, assessing the current situation to see how it differs from the goal, and applying a set of operations that reduce the difference.

If we eliminate belief-desire from human behaviour like a behaviourist who believes in stimulus–response theory, then the notion of thinking in humanoid robots and even in human beings can be under question. Consider a woman who runs out of her flat. Scientific analyses of this incident can ensure us that she may have heard the fire alarm, saw smoke and received a call from the security guard who perhaps told her that the flat is on fire. Not all physical incidents that are mapped by the physicist can capture what is going on in the woman's mind, as the threat of fire may personally frighten her by believing that 'she is in danger' now. If she believes that a naughty child has set off the fire alarm or the smoke spewed out of the kitchen when food was being prepared or that the call that she received was a prank, then she would not have left the flat and obviously would be reluctant to believe that 'she is in danger'. Perhaps this is why all beliefs are somehow mingled with the stimulus that can stimulate a person to set a group of beliefs oriented around the probable situations. It also shows that there are doubtlessly physical incidents, but these physical incidents do not underrate the importance of intuitive psychology or belief-desire psychology of an agent. That is why intelligence is attuned to the belief-desire psychology of human beings that we cannot undermine. Humans have the ability to respond to danger, maladroit situations, all other things related to our physical world and especially the mental states like praise, emotion, love, hate, beauty etc. that humanoid robots cannot, as they lack the explanatory tools like belief-desire along with rationality or common sense.

All of the mentioned tools that humans use are derived from their social communications where intuitive psychology plays a pertinent role. Let us consider a case wherein I have received an invitation from the Indian Institute of Advanced Study, Shimla, for a talk at the Institute on the Philosophy of Hilary Putnam in the coming month's conference. The authority of the institute has confirmed in the same invitation that they will provide my air travel allowance along with boarding and lodging facilities for the 3 day conference. I persuasively inform the institute that I happily accept their invitation and will present my talk on Putnam's 'The Collapse of Fact/Value Dichotomy'. Accordingly, on the mentioned date, the organisers of the conference will arrange everything for me. Now the question is that this communication procedure between an invitee and the organiser does not depend on any personal understanding of knowing each other or meeting in the past. No scientific calculus can determine the preci-

sion of the conversation and can predict its consequence. In fact, this is a sort of process allied to the intuitive psychology that has some common sensual understanding. If an institute invites a person for a talk and the person accepts the invitation, then the invitee will be present at the Institute on the proposed date of the talk. There is a big clause that both the institute authority and the invitee informally know and admit, and that is the concept of 'ceteris paribus' ('all other things being equal') clause that I call 'if and only if terms and conditions'. This clause is informally assigned to the authority and the employer demands to be liable to obey his/her promises if and only if all the surrounding conditions remain favourable for him/her. These conditions do not involve any unwanted accidents, medical crises of both the speakers and organisers or the possibility of any natural disasters, road strikes etc. that might avert a person from fulfilling his/her promises. The process that depends on the 'ceteris paribus' clause is strictly an intuitive psychology-based science that Fodor (1987) introduced in his work *Psychosemantics*. No humanoid robot can grasp this process of intuitive psychology that involves the 'ceteris paribus' clause.

Another form of intelligence that the process of communication expresses is 'correspondence' that focusses on the meaningful states of the world, i.e. the informative communication process has a relation to the meaningful states of the world and corresponds suitably to the physical events and its configurations. A computer processor can maintain these processes well. Alan Turing was the first thinker who introduced these processes of 'correspondence' through an input–output system related to a machine table. Putnam (1979, p. 300) mentioned the importance and non-importance of the 'Turing machine' and his theory of machine 'functionalism' in this way:

> I think that machines have both a positive and negative importance. The positive importance of machines was that it was in connection with machines, computing machines in particular that the notion of the functional organisation first appeared. Machines forced us to distinguish between an abstract structure and its concrete realisation. The distinction does not come to the world for the first time with the machines. But in the case of computing machines, we could not avoid rubbing our noses against the fact that what we had to count as, to all intents and purposes the same structure could be realised in a bewildering variety of different ways; that the important properties were not physical-chemical. That the machines made us catch on to the idea of the functional organisation is extremely important. The negative importance of machines, however, is that they tempt us to oversimplification. The notion of functional organisation became clear to us through systems with a very restricted, particular functional organisation. So the temptation is present to assume that we must have that restricted and specific kind of functional organisation.

Mathematical equations or rational–functional machines can expressively correspond to the rules of mathematics and logic by carrying rational thoughts (based on grammatical sentences by following the rules and norms of the communication procedures) comparable with the human brain. These commercial procedures rely on artificial intelligence that has a model of the human brain. It distinctively challenges the human mind and human intelligence. John Searle's prominent argument against the artificial intelligence theory was given the name 'Chinese room argument' (Searle 1980) and very clearly and argumentatively elucidates that understanding of the sense of an extension of human intelligence can in no way be compared to manipulation of the symbols that humanoid robots perform. The program systems (internal software of the AI, machine or humanoid robots) fail to grasp any intentionality that is the best weapon of human intelligence in their process of communication and thoughts.

The most attractive part of the debate emerges when we are concerned about the concept of human language. The biological process of communication and the productivity of the thoughts secure the best challenge against AI or humanoid robots. Language is the medium that helps communication among intelligent social beings where imagination, experience and social interface remain engaged with significant roles. Computer models and humanoid robots can only generate propositional calculus and internal language (generative grammar[2] as Chomsky claimed and symbols that mathematicians or logicians uphold), but humanoid robots cannot pursue the creative use of the language of common sense. The propositional calculus-oriented computer systems become deviant as they preserve common

[2]Chomsky (1980) defined 'generative grammar' by stating that language is a set of sentences that is generated by grammar. The number of words in our vocabulary is limited, but because of the generative grammar we can construct unlimited grammatical sentences. This is a sort of logical structure of language that we may call grammar, which preserves certain numbers of recursive rules. According to Chomsky (1980, p. 220), 'The grammar of the language determines [the] properties of each of the sentences of the language...Language is the set of sentences that are described by the grammar...When we speak of the linguist's grammar as a 'generative grammar', we mean only that it is sufficiently explicit to determine how sentences of the language are in fact characterized by the grammar.'

sense or socio-linguistic background that previous computer systems didn't have. Their inductive module systems sort out the propositional calculus by corresponding to the intrinsic grammar or logical symbols that mainly manipulate their interpreted system processes. Moreover, the genetic-based language faculty strongly produces 'creativity of language,' a process that I believe partly depends on the propensity of innateness and partly on the agent–world relation which is lacking in any other organisms (like robots, computers etc.) except human beings.

Dennett (2012) argued that when we claim that a person knows, we spell out that a person knows something or specify precisely a few things that he knows. However, this specification depends on some indefinite numbers of assumptions. In the ordinary sense, the term 'know' refers to 'know as true. It may well be possible that an agent can claim to know a proposition as *P,* but this *P* (proposition) somehow turns out as false. Here, the twist is that the agent had a belief of *P,* but he did not know that *P* properly. The procedure of knowing depends on the psychological states like truth, justification, beliefs, etc. that we may call the second-order belief of an agent. Dennett cautioned that there are some cases where we find incompatibility between the 2 different notions of knowledge – truth conditions and the knowledge of belief. Dennett (2012, p. 202–203) claimed that:

> When called upon to produce one's knowledge one can do no better than to produce what one believes to be true, and whether or not what one believes to be true is true does not affect it being one of those things one will produce as knowledge when asked, or will otherwise act on as if one knew them… A thing (a fact or proposition or whatever) could not occupy a special psychological position (e.g. have a special functional potential in the direction of behaviour) in virtue of its truth, so knowing something cannot be purely a matter of being in a particular psychological state.

Here, the process of knowing depends on 2 different tasks. First, we need to determine what a person knows or exhibits as knowledge; then our task would be which of these can be true. Secondly, we have in our mind that a person can be regarded as a store of information and misinformation.

Now the query is: 'Can we specify the content of an agent's store with any precision?' Following Dennett, we can argue that the storage of information is not only the constitutive part of knowing of an agent because the libraries and dictionaries have lots of stored information, but they are unaware of these facts. The notable claim of Dennett raised in favour of 'knowing' is that 'knowing requires understanding' (Dennett 2012, p. 204) and understanding of a word does not rely on the understanding of sentences. Moreover, it may even be possible that one can understand a sentence without understanding the person's utterances or saying. The ability to produce paraphrases is a procedure to understand a sentence, because a computer program can certainly produce paraphrases of English to translate into Bengali sentences. However, the process does not show that computers can understand the sentences. In the case of a computer, although it has some verbal connection (input–output systems), it lacks acquaintance with the object to which the word referred. The process of the conceptual scheme and the perceptual apparatus does not work with a computer. Dennett (2012), in the process of understanding, gives importance to the concept of behavioural capacities of an agent. If an agent X claims that 'Y (another person) is here', it shows that the person must be able to assert and know the other consequences like 'Y is a friend of X' and the term 'here' means in town or not in another place, etc. It sounds true that there may be some cases where the first-order belief (word-object related belief) is wrong, yet the second-order belief (like 'I believe that whales are fishes') turns out to be true. This is because it does not depend on the referential relation to the object of the world, as the agent is the first-person authority of his/her beliefs. Moreover, the conceptual apparatus of the agent has privileged access to the particular knowledge. It may well be possible that here the content of the particular beliefs may turn out to be false (whales are indeed mammals), but the believer's beliefs remain unchanged. A third person may justifiably critique the content of the agent's particular belief, but the third person has no direct knowledge (without inference) of the agent's particular belief like 'whales are fishes' whether the agent is believing or deceiving himself by believing that 'whales are fishes'. No humanoid robots can ever acquire this sort of non-inferential and first-person based sound knowledge of an agent.

There may be the conceptual device of the brain that is linked to our thinking parts (intrinsic or internal part of the brain), but the language device of the brain makes the 'thinking about thinking' possible, as language represents what is going on in the mind (skin in) outside of the physical boundary of the subject. The concept of shareability challenges the idea of intrinsic concepts that preserve the *a priori* hypothesis. The mutual understanding, interpretability and interchangeability of the concepts felicitate language as a social art, and humanoid robots have no privilege to interact with the social art of language, as they have no causal history or lifeworld.

The Nagelean model seems interesting because of its biological-cum-theoretical approaches to ethics. Nagel (2013, p. 146) seemed inventive when he claimed:

> Ethics, though more primitive, is similar. It is the result of a human capacity to subject innate or conditioned pre-reflective motivational and behavioural patterns to criticism and revision and to create new forms of conduct. The capacity to do this presumably has some biological foundation. But the history of the exercise of this capacity and its continual reapplication in criticism and revision of its products is not part of biology. Biology may tell us about perceptual and motivational starting points, but in its present state it has little bearing on the thinking process by which these starting points are transcended.

Certain biological approaches are based on the behavioural patterns that are adapted to human emotions and motivation. The theoretical approaches to ethics instead rely on the methods of rational criticism and justification. It sounds true that the biological pattern-based ethics would endure the development processes with constant reassessment and rectification of the previous experiences, as it intends to trace a deeper attitude of understanding. An agent who would like to engage in this cannot be an automaton that only produces decisions or results, but the agent has the capabilities of rational choice and critical thought directed by common sense-oriented behavioural psychology. Critical thoughts and rationality intermingle with the biological resources like emotion, motivation, absurdity etc. This intermingled process leads to a sharp demarcation between the conception of the morality of robots and human beings. Humans' pre-reflective intuition-based beliefs are not only conjecturing numbers of mathematics but apply reason and formulate different methods that express the creative aspects of the human brain from a biological-cum-theoretical base. This process of rationality-based reasoning is exclusively determined partly by the individual brain and partly by society or the background history of the individual. No humanoid can grasp this method, as it has no motivation, which is ruled by rationality and common sense.

WHY NOT HUMAN ETHICS?

Indeed, critical thought and rationality interact with the conscious–biological resources. This intermingled process constructs a sharp demarcating line between the conception of the morality of humanoid robots and human beings. Humans' pre-reflective intuition-based beliefs not only think about the numbers of mathematics, but may also consider common sense by formulating different methods that cohere with the creative aspects of the human brain from a biological-cum-theoretical level. No humanoid robots can grasp this method, as they have no consciousness-related rational intention, which is controlled by rationality and ingenious practical intelligence. AI tries to show that the intelligence-cum-mentality is somehow related to computational machines; hence, we may say computers have minds. It looks like an amazing achievement for a thinking machine. However, there is a keen difference between the conceptions of 'in principle' and 'in practice'. Human beings have some creative and flexible behaviour that depends on the functional capacities of 'knowing how' processes, which a computer cannot grasp. The 'intelligence' that a machine possesses is nothing more than injected intelligence, and the 'knowledge' that a machine expresses is a kind of 'knowing that' knowledge (in a descriptive sense), like 2+2=4. Lady Ada Lovelace interestingly argued, 'The Analytical Engine has no pretensions to originate anything. It can do *whatever we know how to order it* to perform' (Turing 1950, p. 450).

In fact, moral values cannot be fully attained through social interaction; basically, conscience or morality is something deeply ingrained in our self-realisation and understanding, an inner part of the mind that is called responsibility. Responsibility is an anthropocentric construct that strictly relates to the conception of self-ascription and other-ascription. If we place morality in the realm of objectivity and figure out its universal stand setting on lives and society, then the human-made conception of values that has some fact-centric experiences must interact with the subject's choices. This process averts the expedition of artificial intelligence to threshold into the field of intended responsible morality and its quest for subjectivity and objectivity.

My philosophical view of 'person' is enriched by Singer's (1979) startling remarks on 'person' that refers to a being with 'personal identity', whereas most animals, foetuses and even infants are not 'persons' per se, although they do have relevant experiences. Non-persons are morally germane because, in Bentham's (2000) phrase, they can endure pain and suffering. Bentham (2000, p. 37) wrote, 'The pleasures of malevolence are the pleasures resulting from the view of any pain supposed to be suffered by the beings who may become the objects of malevolence: to wit, 1. Human beings. 2. Other animals.' However, they are no more than 'experience machines', so that, if we reinstate one with another then we have done

no impairment. The question of ruling out children and animals not only shows that they do not have propositional attitudes, but they cannot be regarded as a 'person' who has moral conscience. The same thing would be applied to humanoid robots. I differ from Singer (1979) and Bentham (2000) as I mean to put identity as a sort of theory of the self, arising from a person's theory of their history and projecting into the future, creating goals and values explicit to that person. An example of this sort of constructing values is the decision to become a leader, which cannot be made merely by considering one's existing preferences because one knows that one's preferences will alter as a result of one's decision. The same goes for the choice of a career or job. The conception of morality confers to the linguistic human being instead of the non-linguistic humanoid robots. It looks motivating that the moral conscience is partly *a priori* based attitudes that are linked to us biologically and we partly have attained it through social interactions and practices. Morality cannot be fully attained from society and others; morality is something deeply ingrained in our self-realisation and understanding, an inner matter of mind. However, it looks promising that 'conscience is part of what docility is about' as Jonathan Baron (pers. comm.) argued. Docility is a propensity to be influenced by others. Docility is not only associated with the linguistic personality, but it is also allied to non-linguistic personalities, like deaf children. Personality is in no way associated with humanoid robots, as they do not have any moral sense, and they are indeed docile automata who have no potency to be influenced by other machines. The programs that are installed in their software only can manipulate the robots devoid of any realising sagacity.

Morality emphasises on moral responsibility that is not an outward practice (or 'output', in the case of humanoid robots) of an agent, but a sort of thought (internal part) that only a rational and a responsible intentioned human being can perform. Responsibility is something that is strictly linked to the conception of self-ascription and other-ascription. If we place morality in the realm of objectivity and figure out its universal stand setting on lives and society, then the human-made conception of values that has some fact-centric experience must interact with the subjects' choices. This process thwarts the expedition of artificial intelligence to threshold into the field of intended responsible morality and its quest for subjectivity and objectivity. Here, the major question — Can humanoid robots be moral? — is deciphered from the normative outlook where morality underpins a human-centric outlook instinctively. Humanoid robots are able to differentiate between the 'right-making' and 'wrong-making' properties of moral values through their sensor, but the idea of liberty, justice, equality, open society, post-modernism etc. etymologically are not attainable and configured by the sensor or monitor related to decision-making procedures like a humanoid robot. These conceptions originate from human feelings, rationality and reasoning. Moreover, the principle of universal applicability of intersubjective stance on moral values could not be followed by robotic intelligence, as a robot cannot attain the concept of 'otherness' and the relation between the subjective, objective and inter-subjectivity.

CONCLUSION

Let us assume that some humanoid robots have a moral conscience and consequently can discriminate between right and wrong. The humanoid robots do not think that they have moral conscience, as they probably lack the experiences of seeing, feeling and reasoning that causes them to act.[3] Let us be hopeful like Arkin (1998) that some robots have a moral conscience and consequently can discriminate between right and wrong. We can programme humanoid robots to display empathy and an accurate sense of right and wrong. Humanoid robots (in particular sense) do not think or self-realise as a first-person authority that they have moral conscience as they probably do not have the experiences of realisation, self-knowledge, and humanistic feeling that would cause them to act intentionally and rationally. Eliminating morality from humans elevates the concept of morality that could be maintained by robotic AI, and that is in no way an easy task.

Here, the key question is: Can we reconstruct moral ethics and values in favour of humanoid robots so that the gap between the theoretical and practical differences can be marginalised? We know that the morality of the agents (human beings) and their motivation towards a moral sense frequently become irre-

[3]Rene Descartes is credited (by Haldane & Ross 1934, p. 116) with stating that 'For while reason is a universal instrument which can serve for all contingences, these (mechanical) organs have need of some special adaptation for every particular action. From this it follows that it is morally impossible that there should be sufficient diversity in any machine to allow it to act in all the events of life in the same way as our reason causes us to act'

sponsible, inconsistent, confusing and subject-centric choices etc. However, the process would be highly challenging, as morality is not a mere internalised structure of personal experience or subjective choices. Morality promotes the culture and customs of our past generations and becomes somewhat static because of their content and evaluation that are highly accepted and followed by the majority of people from generation to generation. One cannot step beyond the life-world framework because of the constraint of language and thought. Frank Wilczek, Nobel laureate in physics, wrote, 'Artificial intelligence is not the product of an alien invasion. It is an artefact of particular human culture and reflects the values of that culture' (Wilczek 2015, p. 122).

This is an appealing point that is indeed acceptable. One can say that without depending on the model of human values and moral conscience, it is implausible to prompt any ethical conducts or moral conscience for humanoid robots. The interesting point seems to me is that if we inject rational moral agency and autonomous creativity into humanoid robots, then we cannot treat them as a means. We should treat it as an end. Here, the problem is that humanoid robots need to be responsible for their actions as they are bound by the moral codes and conducts of our society. We know that self-interest can be at variance with other interests. Their presence would elevate the question of whether they should be responsible for their self-interest-intended acts. Artificial ethics should accept responsibilities, and humanoid robots should be the first-person authority of their actions. Humanoid robots can hardly think of themselves as autonomous agents! Morality is not merely the following of codes and conducts, but it requires some motives (intentions) to produce the conducts or rules independently that no humanoid robots can perform.

However, one thing that computer scientists can do is modify human ethics by discarding vagueness and puzzling moral norms. Humanoid robots are unable to think for themselves or for human beings from a responsible ground, as they do not have any aspiration or self-awareness and responsibility that can evolve their intelligence. Humanoid robots need to be morally and culturally systematic, empathetic and embryonic like moral human beings, but they should not follow the internal inadequacy of human ethics. If computer scientists construct superintelligent moral humanoid robots, then I am afraid of the extent of their 'creative capacities'. The conception of 'creativity' resumes when there is a strong list of alternative options like good and bad, moral and immoral, construction and destruction, harm, help etc. We may call someone a moral agent when the person constantly performs moral actions, but s/he has the alternative possibilities to also perform immoral actions. If we program software for every good possibility into superintelligent moral humanoid robots by claiming that a humanoid robot will perform only good moral actions for the benefit of society and the environment, then it would be impractical to say that humanoid robots have any creative sense or are moral agents.[4] Humanoid robots may take over the task more elegantly and error-free than humans. Nevertheless, we should not map out this sort of dehumanising quest as a key impulsion to innovation; rather, human innovation depends on the improvement of the ability to think and do more reasonable tasks guided by moral values and responsibility. Saudi Arabia can offer citizenship to Sophia (www.independent.co.uk/life-style/gadgets-and-tech/news/saudi-arabia-robot-sophia-citizenship-android-riyadh-citizen-passport-future-a8021601.html), one of the latest humanoid robots, but should not contravene human laws and ethical paradigms for the sake of humanoid robots. Indeed, the 'technological providence' of humanoid robots has no liable sense or the first-person authority over their actions which we can blame or punish for not following state laws.

Acknowledgements. My sincere thanks go to my mentors Noam Chomsky and the late Hilary Putnam for their valuable guidance over the years. I am indebted to Professor Jonathan Baron for introducing me to the concept of 'docility'. Thanks to the anonymous reviewers and the editors for their helpful comments that developed my argumentation.

LITERATURE CITED

Arkin RC (1998) Behavior-based robotics. MIT Press, Cambridge, MA

Bentham J (2000) An introduction to the principles of morals and legislation. Batoche Books, Kitchener

Bostrom N (2016) Superintelligence. Oxford University Press, Oxford

Chakraborty S (2017) Understanding moral values: subjective or objective. In: Majhi RM, Patra BP, Sahoo BC (eds) Morality, objectivity and defeasibility. Concept Publishing Company, New Delhi, p 93–103

Chakraborty S (2018) The fact/value dichotomy: revisiting Putnam and Habermas. Philosophia (in press), doi:10.1007/s11406-018-9977-6

[4]'That Beauty, Good, and Knowledge, are three sisters / That doat upon each other, friends to man, /Living together under the same roof...' I quoted the beautiful verse from Tennyson's poem 1832–1833

Chomsky N (1962) Explanatory models in linguistics. In: Nagel E, Suppes P, Tarski A (eds) Logic methodology and philosophy of science. Stanford University Press, Stanford, CA, p 528–550
Chomsky N (1980) Rules and representations. Basil Blackwell, Oxford
Davidson D (1984) Inquiries into truth and interpretation. Oxford University Press, New York, NY
Davidson D (2001) Subjective, intersubjective, objective. Clarendon Press, Oxford
Dennett D (2012) Content and consciousness. Routledge, London
Fodor J (1987) Psychosemantics: the problem of meaning in the philosophy of mind. The MIT Press, Cambridge, MA
Graham LR (1981) Between science and values. Columbia University Press, New York, NY
Haldane ES, Ross GRT (trans) (1934) The philosophical works of Descartes, Vol 1. Cambridge University Press, Cambridge
Harris Z (1951) Structural linguistics. University of Chicago Press, Chicago, IL
Mackie JL (1977) Ethics: inventing right and wrong. Penguin Books, London
McCarthy J, Minsky ML (2006) A proposal for the Dartmouth summer research project on artificial intelligence. AI Mag 27:12–14
Nagel T (2013) Moral questions. Cambridge University Press, New Delhi
Niv Y (2009) Reinforcement learning in the brain. J Math Psychol 53:139–154
Pinker S (1997) How the mind works. Penguin Books, New York, NY
Putnam H (1979) Mind, language and reality. Philosophical Papers Vol 2. Cambridge University Press, Cambridge
Putnam H (1992) Renewing philosophy. Harvard University Press, Cambridge, MA
Putnam H (2002) The collapse of the fact/value dichotomy. Harvard University Press, Cambridge, MA
Putnam H (2013) What is innate and why: comments on the debate. In: Beakley B, Ludlow P (eds) The philosophy of mind: classical problems/contemporary issues. New Phil Learning Private Limited, Delhi, p 757–771
Rawls J (1971) A theory of justice. Harvard University Press, Cambridge, MA
Ruse M (1986) Taking Darwin seriously. Basil Blackwell Publishing, Oxford
Russell S (1997) Rationality and intelligence'. Artif Intell 94: 57–77
Russell S, Norvig P (2016) Artificial intelligence: a modern approach, 3rd edn. Pearson, Delhi
Scassellati B (2002) Theory of mind for a humanoid robot. Autonomous Robots 12:13–24
Searle JR (1980) Minds, brains, and programs. Behav Brain Sci 3:417–457
Singer P (1979) Practical ethics. Cambridge University Press, Cambridge
Stove D (1978) On Hume's is-ought thesis. Hume Stud 4: 64–72
Tanimoto S (1987) The elements of artificial intelligence. Computer Science Press, Rockville, MD
Turing AM (1950) Computing machinery and intelligence. Mind 59:433–460
Wilczek F (2015) Three observations on artificial intelligence. In: Brockman J (ed) What do you think about machines that think? Harper Perennial, New York, NY, p 121–123
Wilson RA, Keil FC (1999) The MIT encyclopedia of the cognitive sciences. The MIT Press, Cambridge, MA
Winston PH (1992) Artificial intelligence. Addison-Wesley, Boston, MA
Zimmerman MJ (2015) Value and normativity. In: Hirose I, Olson J (eds) The Oxford handbook of value theory. Oxford University Press, Oxford, p 13–29

Editorial responsibility: Darryl Macer, Scottsdale, Arizona, USA

Submitted: October 24, 2017; Accepted: June 16, 2018
Proofs received from author(s): August 30, 2018