ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

A Model for Ethical Artificial Intelligence

Rohit Dsouza[1], Shraddha Ravishankar[2], Arushi Shah[3], PurvaRaut[4]

Student, Dept of Information Technology, Dwarkadas J. Sanghvi College of Engineering, Mumbai, India[1, 2, 3]

Assistant Professor, Dept of Information Technology, Dwarkadas J. Sanghvi College of Engineering, Mumbai, India[4]

Abstract: The birth of the field of Artificial Intelligence was primarily due to the curiosity of researchers and inventors to create a perfect replica of the human mind. The thought of being able to simulate the human brain on a machine captured the imaginations of many, enticing and assimilating a wide array of research in the field, leading to a constant string of breakthroughs and new discoveries. With each breakthrough, our understanding of the boundaries and limits of the field were pushed, making it impossible to accurately estimate the complete scope of any project, with the only limitation being the technology available at the time. The primary purpose of any device implementing AI today, i.e. automation, was merely an application of the technology stumbled upon. As automation takes over our daily lives, it leads us to the main concern of where the line of ethics needs to be drawn and when does our pursuit for an ideal rational agent end with us questioning the morality of the decisions made by the machine. This paper tries to understand a model that weighs each choice with it's moral conformation to truly decipher if a machine can understand each choice beyond the simple weighted values of desirability and goal reachability. After all, are we still trying to create the perfect human intelligence or just a perfect machine?

Keywords: Artificial Intelligence, Neural Networks, Brain Hierarchy, Morality, Ethics, Kant's Laws, Responsibility.

I. INTRODUCTION

Ethics and morals, though used interchangeably, are separated by a fine line. Ethics are based on the principles of right conduct, generally laid down by the society, whereas morals are the principles on which one's judgments of right and wrong are based. Morals are more abstract and often rely on personal or religious opinion. In concise terms, ethics are the science of morals, and morals are the practice of ethics. [1]

The foundations for what AI is today were laid less than a century ago by Alan Turing, British mathematician and WWII code-breaker. The "Turing Test" proposed by him to determine a machine's ability to think like a human is widely used even today. The term "Artificial Intelligence" was coined at the Dartmouth Conference in 1956 and thus began what was termed as the "Golden Years" of AI. The early years witnessed development of back tracking algorithms to be used in general problem solving along with natural language processing allowing computers to communicate in languages such as English. The concept of micro worlds introduced us to SHRDLU, a program that interacts with the user to work in virtual environment consisting of blocks. However, AI witnessed its first dark phase between 1974 and 1980 due lack of funds and and high level of criticism. In the 1980s, the field of AI was revived when the British Government began funding it again to compete with Japan's fifth generation computer project. This period led to the rise of expert systems, a program that answers questions or solves problems about a specific domain of knowledge, using logical rules that are derived from the knowledge of experts. Project CYC created a massive database consisting of mundane things that an average person is generally aware of, with the goal of enabling AI applications to perform human-like reasoning. Research began to pick up again in 1993 and after that, in 1997, IBM's Deep Blue became the first computer chess-playing system to beat a reigning world chess champion. Other notable achievements that followed include "Eugene the chatterbox" and "Watson the supercomputer" that won Jeopardy!. [2], [3]

Today Artificial Intelligence is applied in varied domains ranging from search engines to personal assistants and even fields like data mining, medical diagnosis or banking solutions. As machines become more autonomous, they raise several ethical issues. Primarily, they must not endanger human lives in any case or carry out actions that compromise the moral status of the machines themselves. An AI system must be robust against any manipulation that can prove to be detrimental otherwise. Another issue that arises is that of responsibility for the action of these systems.

Thus, ethics is widely defined with respect to how a person or a group of individuals want an AI system to function over time. While it is universally accepted that present-day AI systems lack moral status, it is unclear exactly what attributes define this moral status. Two criteria are commonly proposed as being importantly linked to moral status, either separately or in combination, namely, sentience and sapience (or personhood). [4] These may be characterized roughly as follows:

Sentience: the capacity for phenomenal experience or qualia, such as the capacity to feel pain and suffer
Sapience: a set of capacities associated with higher intelligence, such as self- awareness and being a reason-responsive agent

Animals experience sentience but it is only the human population, barring infants and the mentally incapable, which possess qualities of sapience. A combination of the two, if inculcated in an AI system, can earn it a full moral status that is comparable to human beings.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

II. PRIMARY EXAMPLE

An experiment was carried out in a Swiss Laboratory, where a group of bots were programmed with the task of finding a "food source". This food source was indicated by a light colored ring at one end of the arena, which was visible to them at close range, using a downward facing sensor. The other end of the arena, labelled with a darker ring was "poisoned". Each bot could produce a blue light that the others could detect with camera. The bots would get points based on how much time they spent near food or poison, which indicated how successful they were. Initially the bots produced this blue light randomly. Further, as the robots became better at finding food, the light became more and more informative and the bots became increasingly drawn to it after just 9 generations. As the lights increasingly gave away their presence, the bots became more secretive. By the 50th generation, they became much less likely to shine near the food than elsewhere in the arena, and the light became a much poorer source of information that was much less attractive to the bots.

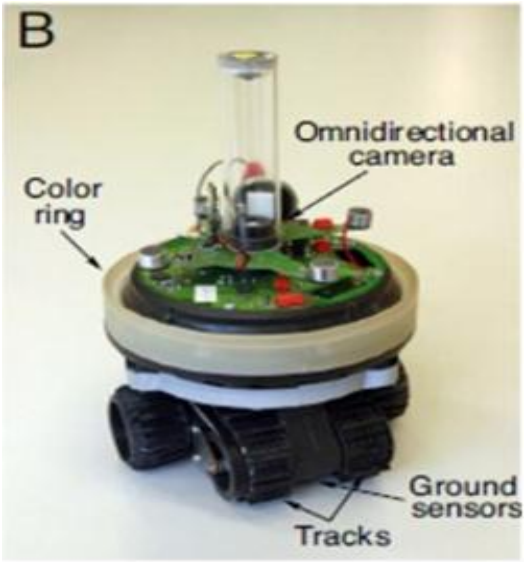

Fig1. Swiss Laboratory Robots

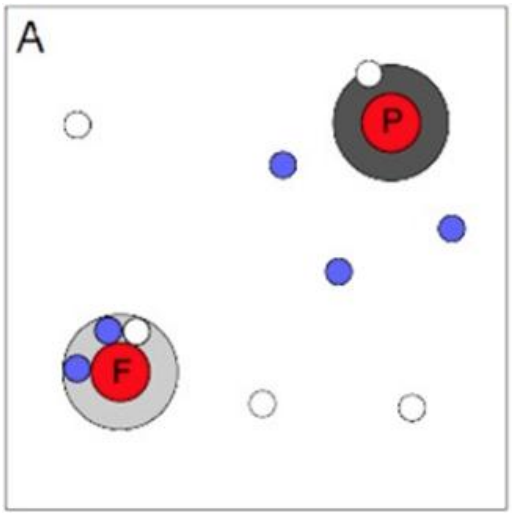

Fig2. Working Representation of the Bots

Their evolution was made possible by an artificial neural network controlled by a binary "genome". This network consisted of 11 neurons that were connected to the bot's sensors and 3 that controlled the two tracks and it's blue light. The neurons were linked via 33 connections– synapses –and the strength of these connections was each controlled by a single 8-bit gene. In total, each robot's 264-bit genome determined how it reacted to the information gleaned from it's senses. After around 500 generations of evolution, around 60% of the bots never emitted light near food, but around 10% of them did so most of the time. Some of these bots were slightly attracted to the blue light, but a third were strongly drawn to it and another third were actually repulsed.

The above experiment demonstrates that robots, if programmed to think like humans, are capable of developing human like tendencies. Such outcomes can be deleterious, if not controlled. The probability of a robot's ability to lie is unknown, but certainly cannot be ignored and hence, a control mechanism is now. a critical requirement. [5]

III. KANT'S LAWS

A decision making algorithm is often governed by a set of rules that help to guide the input through the tree and provide a more efficient and streamlined approach to solving a problem. Similarly, when we face the problem of identifying the morality and ethicality of a particular action, we need a set of laws which can help in defining the basic principles that encompass the fundamentals of the two concepts. These laws are best described using the three formulations stated below, proposed by the German philosopher, Immanuel Kant. [6], [7]

A. The First Formulation

"Act only according to that maxim whereby you can at the same time will that it should become a universal law without contradiction."

This formulation is almost self-explanatory, stating that every decision made by the AI must be such that it can be applied as a universal law. This means that every action should be universally acceptable, without any opposition towards it. This may seem practically impossible to implement, as certain choices might be subjective in design, leading every individual to choose an option purely based on preference rather than reasoning.

B. The Second Formulation

"Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely as a means to an end, but always at the same time as an end."

This law states that an action involving another entity that may or may not be affected by the outcome of the action, must be performed only if the involvement of the entity is crucial to the performance. In accordance, it permits the entity to be incorporated as a part of the task only if he/she is not being used as a means to perform a certain action, purely driven by the selfish motives of the person performing the action.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

C. The Third Formulation

"Therefore, every rational being must so act as if he were through his maxim always a legislating member in the universal kingdom of ends."

The final formulation states that every action affecting another entity must be performed only if the outcome is acceptable to the system performing the action, when perceived from the receiving end of the action. In simple terms, it encapsulates the golden concept of - " Do not impose on others what you do not wish for yourself ". This law forces the individual to consider the repercussions of each action from the point of view of those affected by them.

IV. MODEL OF IMPLEMENTATION

The basic ideology of the model mimics the way neural networks function, more accurately how the human brain would function in a similar situation. Borrowing the fundamental ideology of a human making a decision about whether an action is ethical or unethical, we can propose a similar method that can efficiently handle the very same situation in a well-trained AI system. But first, let us try and understand the decision making process of the human brain through a brief overview, using the concept of brain hierarchy. The process starts with the generation of the basic thoughts and ideas entailed by the action to be performed. These thoughts are processed along with the input parameters perceived by the various sensory organs in our body. Once these basic requirements have been gathered, the neural networks constituting our human brain can freely process them to draw a variety of conclusions. This brings us to one of the primary evolutionary gifts that distinguish human beings from other mammals and reptiles. As human beings, our brain allows us to process the information gathered in a particular situation and compute the possible outcomes, relative to each action taken. Although commonly used to analyse situations in our daily lives, this deceptively complex process enables us to think rationally and enhances our ability to solve a problem using logic and reasoning. This is also one of the primary reasons for the very existence of morals and societal ethics. Once we realise the consequences of each action, we apply our own understanding of morals and ethics to decide on which would be the most suitable action to be taken in the situation at hand. On the basis of this approach, where morality and ethics are placed at the summit of the decision making pyramid, we can approach our model for artificial intelligence.

Similar to the human brain, a typical AI system functions using an artificial neural network, which is designed to replicate the structure and functioning of the neural network in the human brain. Following a similar model, the neural network consists of different layers, each consisting of a number of nodes. Each node serves to perform a distinct function which can reveal a bit more about the solution to the problem at hand. This implies that the system does not perceive the initial problem as a whole, but rather follows a divide and conquer strategy by tackling parts of the problem, one at a time.

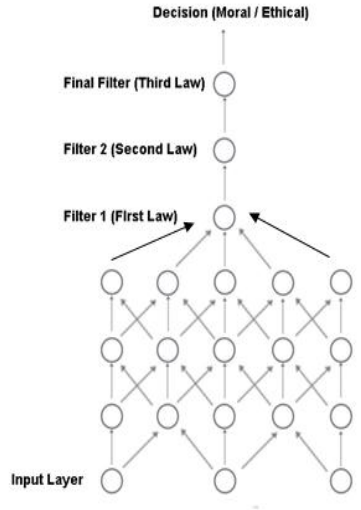

Fig 3. Representation of Suggested Model

Each node performs its part and forwards the results obtained to the next node which may or may not use these results in its own computations. While this may lead us to believe that each node is independent, keep in mind that the system depends on the result of each node to devise a solution to the problem, much like joining the pieces of a jigsaw puzzle. To understand this better, consider a problem which deals with identifying a brown dog in a given picture. On receiving the input, in this case a picture, the first node might simply be trained to identify the outline of a dog. Once successful, the position of this outline will be passed on to the next node which might be assigned with the task of identifying the color, filled within the outline. Both these results can then be used by the system to decide whether or not a dog is present in the picture, and if so is it brown in color. This leads us to our next question. How will the node know what a dog looks like?

The answer to this question once again leads us to the most fundamental feature of an artificial neural network, which is it's resemblance to the human brain. This resemblance is far from superficial and enables the network to perform one of the most important functions of the brain, which is learning. A task performed with ease, even by a child, is the key to artificial intelligence and allows us to train our neural networks to identify various objects or understand certain tasks, in a manner we all are familiar with. By using a sample set consisting of the object to be identified, the network can be trained to flawlessly recognise it by matching it's key features, with the level of accuracy being bounded only by the sample size and the time spent in training the network. When such a network comes across a situation that it may have faced before, this allows it to perform another remarkable feature. Since it has already learned that an action, say, 'X' performed in a situation 'Y' could lead to 'Z', like human beings, it can almost predict the outcome of an action in a particular situation. This incredible ability of the network enables it to analyse the potential of every option available in a situation and choose the best path in order to achieve the goal assigned to it. Similar to the case mentioned

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

earlier, this ability is once again restricted only by the knowledge base of the Ai system. This leaves us with an highly functional AI system capable of making it's own decisions and all the fundamental requirements for implementing a model to enable an AI to make a morally or ethically right decision.

Our model proposes the use of a reusable module, capable of testing the input against a set of pre-defined laws used to define ethics and morality. In accordance with the concepts explained above, the input to such a model consists of the outcome or impact of each decision, computed by the system. As a pre-requisite, the system needs to possess a knowledge base as the model deals with helping an AI, which is already capable of making a variety of decisions and estimating their individual impact, to choose the ethically correct decision. For a system with no prior knowledge, this would be futile as it would be unaware of the consequences of its own actions, engendering the need to externally train the network first. Once trained, the AI may face a certain situation with 'n' different actions, available to be performed. At this stage, it will first compute the impact of each action and submit the result to the first block of the module for consideration. The first block is used to implement the first law, which checks for universality. Utilizing the knowledge gathered by it, the system checks if the outcome of an action is considered morally and ethically correct by a stipulated number of people. This is achieved by setting a sample size for the population and a threshold value which specifies the minimum number of votes which need to be in favor of the decision for it to be considered ethically correct. If the knowledge base records show that the threshold value has been met, the information is passed on the second block. Note that these blocks are arranged in a descending order of priority, which means that violation of the conditions specified by any block will deny access, for that particular action, to the next block. Thus, if this threshold value is not met, the action will simply be discarded as unethical, irrespective of the possible results computed by the next two blocks.

The second block is used to implement the second formulation stated above and ensures that no person is used by the AI, simply as a means to an end. This block checks the involvement of human interaction in the action to be performed and if present, it verifies that it is absolutely required and cannot be avoided by any alternative means. Once again, the system may scan it's knowledge base for similar actions performed in the past to identify any anomaly, in case the same action has been performed before without any human involvement. If found, the next step would be to identify and report the source of this anomaly before any action is to be performed. This kind of informed approach may be a bit more time consuming that simply allowing the AI to choose between the alternatives. However, it would lead to major drawbacks in terms of security, allowing any suspicious behaviour to pass by unnoticed. For example, a bot that could access a specific database before and has been denied access for security reasons might resort to asking an uninformed employee for help in order to access the database. In this case, identifying the sudden involvement of another entity in the action by comparing it with the same action performed in the past could help in identifying and mitigating the risk, in time.

The final block of the model aims at implementing the third formulation stated above, which states that an action can be distinguished as moral or ethical, if the entity performing the action would accept the impact of the action without any objections, as if it is at the receiving end. This implies that the system must perform an action with certain consequences, only if it can readily accept the situation when faced with the same set of consequences. The model implements this maxim by passing the outcome of each action once again through the first two blocks and accepting the action only if its outcome can pass the first two conditions. In order to ensure that the system does not encounter an infinite loop, considering an infinite number of future possibilities through multiple iterations of the same process, we omit the inclusion of the third block once again, during the second iteration.

Once the action had completed its passage through this proposed module, any action in the result set obtained can successfully be deemed as morally and ethically correct. For example, consider a robot being asked to assist an elderly woman to cross the road. It faces the choice of either helping the woman or denying her any help. On analysing both the possibilities, the first block would majorly describe the action of helping the woman to be ethical and the latter to be unethical, immediately leading to it being discarded. Further, the second block would ensure that there is no anomaly present in the action in term of human interaction and also that the woman is the only human naturally involved in the process. Finally, the third block would place the outcome of enabling the woman to reach the other side of the road as ethically correct, thus allowing the robot to decipher the morality and ethics behind the encountered problem.

Another prime focus of this model lies in reusability. With the advent of evolutionary AI and the ever expanding horizon of it's scope, neural networks and artificial intelligence have found their place in a variety of fields and will soon be an inseparable part of our daily routines. With their application growing at such a rapid pace, the possibility of products that can be developed with AI at their core is unpredictable. Hence, in such a situation, a reusable model is the key to wide spread implementation of this concept. Our approach allows any sufficiently trained AI system to differentiate between an ethical and unethical choice, with the addition of this module as the penultimate stage of it's inherent neural network. Thus, it can prove to be a powerful and ever growing tool with a future scope for further refinement.

V. COMPLEXITY OF IMPLEMENTATION IN CURRENT SYSTEMS

Morality and ambiguity are two sides of the same coin. Morality can never be pointed at with a straight arrow and have everyone agree on it. It is unfortunately not a universally set upon ideal. One man's pleasure if often

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

another man's poison. Often morality is complex to explain to a human. And teaching AI that, well, that's been the problem. A utilitarian may say that murder is wrong because it does not maximize good for all those involved, but that doesn't hold true for someone who is only interested in maximizing good for himself. In this way, both agents think their action is moral. Then comes the whole issue of a machine being able to make a moral choice but not hold moral responsibility. A convoluted paradox at hand? Friedman and Kahn Jr posited that intentionality was a necessary condition for moral responsibility and computers as conceived in 1992 could not possess intentionality. And without intention, how could a computer make a moral choice?

Initially, a top down approach was considered to implement morality in intelligent agents using AMA's. The robot would follow the laws of ethics like a religion, weighing every action it took against each principle. However, it was faced with the following difficulties. The centuries old dilemma of moral philosophers, whether any one ethical theory is enough to capturing the breadth and complexity of human moral considerations, comes into play and says that even if we do design an AMA around a top-down framework, that alone would not be satisfactory to guarantee the acceptability of the system's behaviour to everyone. Deontological, consequentialist, and virtue-based ethical systems each have their own strengths and weaknesses. When the possibility of substantiating a particular ethical theory within a computational system was considered, additional challenges arose. Framing the challenge, weighing values against each other, resolving conflicts between rules, calculating consequences, insuring that the systems have adequate information, factoring in knowledge about human motivations, and managing computational looping were the main challenge faced when using AMA's. [8]

The other theory that was considered was using a bottom up approach, wherein the artificial agent is allowed to experience different situations and learn morality through appreciation of good behavior using game theory and genetic algorithms. It would allow the agent to slowly learn and evolve into a moral agent. Yet, even in the accelerated environment of computer systems, where many generations of artificial agents can mutate and replicate within a few seconds, evolution and learning are very slow processes. It was also unclear what would be the appropriate goal for an evolving AMA, or how that goal might be usefully defined for a self-organising system. The development of such systems with complex moral faculties from the bottom-up largely depends on what future technologies have in store for us.

VI. RESPONSIBILITY OF THE AGENT

Moral responsibility is about the consequence of human action. A person or a group of people are considered morally responsible when their voluntary actions are morally significant and their outcomes are such that would make it appropriate to blame or praise them. Thus, it might be considered a person's moral responsibility, when they see a person drowning, to try to rescue the person by jumping in the water and saving them. If he or she manages to save the person, we will praise them, on the other hand if he or she refuses to help we may blame them. Morally responsibility establishes a link between a person (let's call them the subject) and someone or something (the object) that is affected by the actions of this subject. Sometimes ascribing responsibility to someone involves giving an account of who was at fault for an accident and who should be punished. It can also be about determining the obligations a person has and their duty of fulfilling them.

Most analysis of moral responsibility share at least the following three conditions [9], [10] :

1. There should be a causal connection between the person and the outcome of actions. A person is usually only held responsible if she had some control over the outcome of events.
2. The subject has to have knowledge of and be able to consider the possible consequences of her actions. We tend to excuse someone from blame if they could not have known that their actions would lead to a harmful event.
3. The subject has to be able to freely choose to act in certain way. That is, it does not make sense to hold someone responsible for a harmful event if her actions were completely determined by outside forces.

For someone to be held morally responsible for an event, he or she has to have some influence over the event. A person who has no power over the event cannot be held morally responsible for the outcome of it. Computer technologies obscure the causal connections between a person's actions and the eventual consequences. Usually, when we trace the sequence of events that lead to a computer-related incident, we are led in many directions; as such incidents are usually not the result of a single error or mishap. Technological accidents are commonly the product of a multitude of errors, misunderstanding or neglectful actions of various individuals involved in the development, use and maintenance of the computer system.

The involvement of a number of actors in the development and deployment of technologies is known as the problem of 'many hands'.[9], [11], [12]A commonly discussed example of the problem of many hands is the case of the malfunctioning radiation treatment machine Therac-25. [13], [14]This computer-controlled machine was designed for the radiation treatment of cancer patients as well as for X-rays. During a two-year period in the 1980's the machine massively overdosed six patients, contributing to the eventual death of three of them. These incidents were the result of the combination of a number of factors, including software errors, inadequate testing and quality assurance, exaggerated claims about the reliability, bad interface design, overconfidence in software design, and inadequate investigation or follow-up on accident reports. Even so, in their analysis of the events Leveson and Turner concluded that it was hard to place the blame on a single person. The actions or negligence of all those involved might not have proven fatal were it not for the other contributing events. That however does not mean

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

that there is no moral responsibility in this case [11], [15], as many of the individuals could have acted differently, but retrospectively, it is a tedious and complicated taskto identify the appropriate person that can be held responsible for the event. Additionally, the temporal and physical distance created by computing between a person and the consequences of their actions can blur the causal connection between actions and events.[12] Computational technologies extend the reach of human activity through time and space. The designers of an automated decision-making system might have determined ahead of time how their system should make decisions, but they will rarely see how these decisions will impact the individuals they affect. Their original actions in programming the system may have effects on people years later.

In order to make a morally correct decision a person has to be able to consider and think about the consequences of their actions. He or she has be aware of the possible risks and consequences of his or her actions. Infact, it might be considered morally wrong to hold a person responsible for an action that they didn't know would cause harm.

The freedom to act is the third condition for attributing moral responsibility and also one of the most contested. We tend to excuse people from moral blame if they had no other choice other than to act in the way they did. We most likely would not hold people responsible if they were coerced or forced to take particular actions. Suppose a person is forced to shoot someone because they are being held at gunpoint. We tend to excuse the person the burden of moral responsibility while not condoning the act at the same time.

As computer technology is getting increasing complicated and AI has progressed leaps and bounds, maybe the human agent is not the only one that can be held responsible.[16] Dennett, for example, suggests that holding a computer morally responsible is possible if it concerned a **higher-order intentional computer system.**[18]An intentional system according to Dennet is one that has its own beliefs, desires and can be attributed with rationality. In other words, its behaviour can be described by assuming the systems has mental states and that it acts according what it thinks it should do, given its beliefs and desires. Many computers today, according to Dennett, are already intentional systems, but they lack the higher-order ability to reflect on and reason about their mental states. They do not reflect or have thoughts about their beliefs or desires. Dennett suggests that the fictional HAL 9000 from the movie 2001: A Space Odyssey would qualify as a higher-order intentional system that can be held morally responsible.

J Sullins too, is of the opinion that to be held morally responsible, an entity does not have to be a person.[17] He proposes that computers systems or, more specifically, robots are moral agents when they have a significant level of autonomy and they can be regarded at an appropriate level of abstraction as exhibiting intentional behaviour. A robot, according toSullins, would be considered autonomous if it was not under the direct control of other agents while performing its tasks. However, he adds a third condition. A robot also has to be in a position of responsibility to be a moral agent. That means that the robot performs some social role that bestow upon it certain responsibilities and the robot has certain beliefs and understandings about these responsibilities as well as other agents involved. To illustrate what kind of capabilities are required for "full moral agency", he draws an analogy with a human nurse. If a robot was autonomous enough to carry out the same duties as a human nurse and had an understanding of its role and responsibilities in the health care systems, then it would be a "full moral agent".

Contrary to the theories stated by Sullins and Dennet, critics who follow the earlier belief about AI not having personhood do not believe in attributing beliefs and responsibilities to a computer system.[19], [20]They point out that it makes no sense to treat computer system as moral agents that can be held responsible, for they cannot suffer and thus cannot be punished.[21], [22]

Responsibility in this case of morally active AI has to be treated as an individual case each time. Not every incident is going to have the same reason for that particular outcome. When we deal with AI, we tread really still waters, and we all know that still water runs deep. So we could have a system (like Skynet from the Terminator franchise) that learns something that was never intended to be taught to it or it could be a programming mistake or it could be intentional. Just like morality, one solution doesn't fit all.

VII. CONCLUSION

As we move ahead with the world, as AI is increasingly becoming part of our lives and now that it has reached this immensely progressive stage, how far can we go with a system that doesn't know right from wrong? Don't we all want to live in a world where our machines will have a sort of conscience as we do, and not worry about them turning on us one day(Yes, Skynet has scared us all). The scope for a morally thinking AI system is immense, right from self-driving cars, who will know that it's not okay to ram into a car or pedestrian on the road just to get there faster, to defense drones, the need is everywhere. For the advancement of our technology we need our machines to be truly intelligent. We need them to actually be able to think for themselves. And that is possible only when they have a certain moral character. The best part is that we can teach them the moral character they should have, so perhaps an ideal one unlike most humans. For AI to be truly intelligent, it must be able to make its own morally correct choices. In the immortal words of Leon Bloom, "Morality consists solely in the power of making a choice".

REFERENCES

1) "Ethics vs Morals Grammarist"**.** The Grammarist website.[Online].Available:**http://grammarist.com/usage/ethics-morals/**[1]
2) "A Brief History of Artificial Intelligence (2014)".The Live Science website. [Online]. Available: **http://www.livescience.com/49007-history-of-artificial-intelligence.html** [2]
3) "From Science Fiction to Reality: The evolution of Artificial Intelligence (2015)". The Wired website. {Online}. Available:

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

http://www.wired.com/insights/2015/01/the-evolution-of-artificial-intelligence/[3]

4) Nick Bostrom and Eliezer Yudkowsky, "The Ethics of Artificial Intelligence", 2011. [4]
5) "Robots Evolve to Deceive One Another (2009)".{Online}. Available: http://scienceblogs.com/notrocketscience/2009/08/17/robots-evolve-to-deceive-one-another/[5]
6) "Kantian Ethics". {Online}. Available: http://www.csus.edu/indiv/g/gaskilld/ethics/Kantian%20Ethics.htm [6]
7) Kant, Immanuel (1785). Thomas Kingsmill Abbott, ed. Fundamental Principles of the Metaphysic of Morals (10 ed.). Project Gutenberg. p. 61. [7]
8) Allen, C. W. Wallach, and I. Smit., "Why Machine Ethics?" Intelligent Systems, IEEE, 21(4): 12–17, 2006. [8]
9) Jonas, H., The Imperative of Responsibility. In search of an Ethics for the Technological Age, Chicago: The Chicago University Press, 1984. [9]
10) Eshelman, "Moral Responsibility". [Online]. Available: http://plato.stanford.edu/archives/win2009/entries/moral-responsibility/> [10]
11) Nissenbaum, H., "Computing and Accountability," Communications of the Association for Computing Machinery, 37(1): 72–80, 1994. [11]
12) Friedman, B., "Moral Responsibility and Computer Technology." Paper Presented at the Annual Meeting of the American Educational Research Association, Boston, Massachusetts, 1990. [12]
13) Leveson, N. G. and C. S. Turner, "An Investigation of the Therac-25 Accidents," Computer, 26(7): 18–41, 1993. [13]
14) Leveson, N., "Medical Devices: The Therac-25," in N. Leveson, Safeware. System, Safety and Computers, Addison-Wesley, 1995. [14]
15) Gotterbarn D., "Informatics and professional responsibility," Science and Engineering Ethics, 7(2): 221–230, 2001. [15]
16) Bechtel, W., "Attributing Responsibility to Computer Systems," Metaphilosophy, 16(4): 296–306, 1985. [16]
17) Sullins, J. P., "When is a Robot a Moral Agent?" International review of information Ethics, 6(12): 23–29, 2006. [17]
18) Dennett, D. C., "When HAL Kills, Who's to Blame? Computer Ethics," in HAL's Legacy: 2001's Computer as Dream and Reality, D. G. Stork (ed.), Cambridge, MA: MIT Press, 1997. [18]
19) Johnson, D. G. 2001. Computer Ethics (3 ed.). Upper Saddle River, New Jersey: Prentice Hall. "Computer Systems: Moral Entities but not Moral Agents," Ethics and Information Technology, 8: 195–204, 2006. [19]
20) Kuflik, A., "Computers in Control: Rational Transfer of Authority or Irresponsible Abdication of Authority?" Ethics and Information Technology, 1: 173–184, 1999. [20]
21) Sparrow, R., "Killer Robots," Journal of Applied Philosophy, 24(1): 62–77, 2007. [21]
22) Asaro, P., "A Body to Kick, But Still No Soul to Damn: Legal Perspectives on Robotics," in P. Lin, K. Abney, and G. Bekey (eds.) Robot Ethics: The Ethical and Social Implications of Robotics. Cambridge, MA: MIT Press, 2011. [22]