# Design, Architecture and Technology behind Morality of Robots

[1]Rosy Guha, [2]Dr. Vishwanath Y
[1]Student, [2]Associate Professor
[1]Information Science and Engineering,
[1]New Horizon College of Engineering, Bangalore, India

*Abstract:* Recently, roboticists initiated inculcating morality in technologies to develop so-called ethical robots. This ethics in technology concerns the utility of methods, algorithms and tools required to endow artificial autonomous agents with the capability of being ethically aware. Robotic actions are completely in the hands of designers of the systems or those who deploy them. Engineers are always concerned about safety and reliability in their design of intelligent systems. They need to apply complex technologies to make the robots behave under different set of inputs accordingly. Sophisticated robots should possess the capacity of assessing and responding to moral considerations. However, engineers who design functionally moral robots face many constraints due to limitations of present-day technology. This paper gives a brief overview of different technical approaches that plays a major role in developing moral agents.

***IndexTerms*** **- robotic imagery, simulation module, moral competence, artificial moral agents, control architecture, branching time system.**

## I. SIMULATION THEORY OF COGNITION

Pinker(1997) argued that human mind has not evolved to be an abstract symbol manipulator. Since then, it is confirmed that computations behind human cognition are very different from rule-based manipulation of abstract symbols. This idea has emerged in many domains of human cognition such as perception, reasoning and problem-solving. But, representing, learning and combining concepts lead to some difficulties in purely symbolic systems. Therefore, it is true that mind uses representations that are more enriched than the abstract symbols allowed for in models of intelligence that presume abstract symbols.
The theory of mind that lead to the richest representations is the simulation theory of cognition. It hypothesizes that thinking uses the same cognitive (and neural) processes as interaction with external environment. According to this view, thinking requires building a grounded model of environment which does not comprises abstract symbols. Rather, it is assumed to recombine experiences using the brain's system of perception, action and emotion. The mental model simulates actions and their associated perceptual effects.
Scientists derived a method for implementing ethical behavior in robots inspired by the simulation theory of cognition. This method utilizes internal simulations which allow the robot to simulate actions and predict their consequences. Therefore, this method is a form of robotic imagery. Marques and Holland, in their review of robotic imagery , coined the term *functional imagination* to denote the mechanism whereby robots covertly simulate actions and their impacts to steer their future behavior. The basic advantage of a functional imagination is the ability to test the outcome of potential actions without committing to them. Therefore, functional imagination is a framework suitable for supporting consequentialist ethics[1].

## I. ARCHITECTURE BEHIND MORALITY OF ROBOTS

Many architectures for robot controllers have been proposed by different scientists. this control architecture mostly can be remapped onto a three- layered model. in these model, each control level is characterized by differences in the degree of abstraction and time scale at which it operates. at top level, the controller develops long-term goals. then, goals are divided into set of tasks that need to be executed. finally, these tasks are translated into motor actions that are executed by the robot.

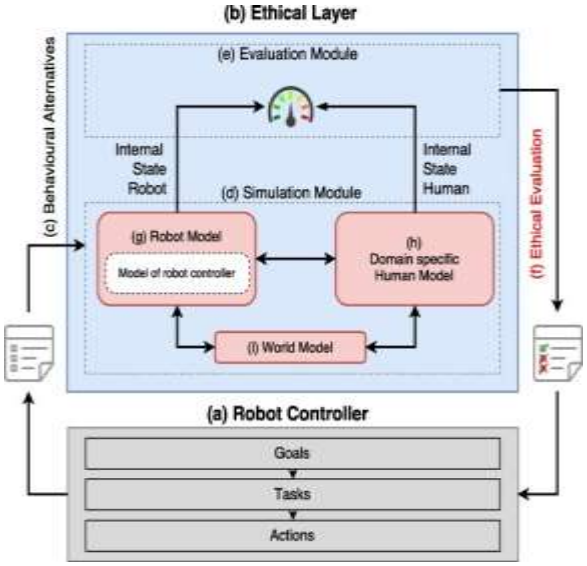



FIG: ARCHITECTURE

**The Ethical Layer**

The ethical behavior of robot should be governed by adding a fourth specialized control layer called ethical layer. This layer acts as a governor evaluating behavior proposed by each of the three other layers before robot executes. Ethical layer of robot through separate layers has several advantages. For one, implementing the ethical layer as a just-in-time checker of behavior is useful for fail safe device checking behavior before execution. Also, a separate Ethical layer functionality can be scrutinized independently from the operation of robot controller.

Ethical layer functions in the following ways. By default, the robot controller generates a set of prospective behavioral alternatives. The Simulation Module is activated with the current state of the world, robot and human. The Ethical Layer simulates the consequences of each alternatives in the current state using the Simulation Module. For each alternatives, the Evaluation Module evaluates the simulated consequences. The ethical evaluation for each of the behavioral set of alternatives, is sent to the robot controller. In other words, Simulation Module and the Evaluation Module continuously loop through the behavioral alternative as they are generated by the robot controller. Scientists found that humans cannot evaluate more than two or three alternative strategies which resulted in the best model fit. This happens due to limits in cognitive capacity, working memory. Similarly, evaluating only limited number of behavioral alternatives would improve the responsiveness of the robots and prevent the Ethical Layer from introducing delays.

**The Simulation Module**

The structure of the Simulation Module is based on an analysis of the requirements rather than on findings in cognitive science. Simulation Module is equipped with (1) a model of the robot controller (2) a domain specific model (3) a model of the world which might consist of physical model of both human and robot as well as model of objects. The model of the robot, in combination with the world model, the Simulation module can simulate the future motor, sensor and internal states for the robot. Also, this combination allows the evaluation of behavioral alternatives at each of the three levels of robot control i.e. at the level of goal, tasks and actions. This is how, the Ethical Layer could predict and evaluate the outcomes of goals.

Considering the model of human, the Simulation Module can predict the future sensory and motor states of the humans. However, also internal states like emotions, can be simulated. In humans, the same neural machinery that supports actions and perception during overt behavior supports the mental simulations of sensory and internal states. However, in case of robots, this can be supported by a sufficiently complex model of the human.

In an experiment between H-robot (Human robot) and A-robot (Asimovian robot) where H-robot issued a command to A-robot, it spoke one of two sentences: (1) 'Go to location A' or (2) 'Go to location B'. A-robot then simulated its behavior based on the following three assumptions: (1) The A-robot moves in a straight line to its goal. (2) The closer the A-robot comes to a dangerous location, the less safe it is. (3) The A-robot stops when closer than 0.5m from the H-robot.

The Simulation Module simulated two outcome states for the H-robot. First the safety level of the H-robot $I_{h1,I}$ was given by,

$$I_{h1,I} = 1/(1+e^{-\beta(d_{h,i}-t)})$$

where $d_{h,I}$ is the simulated final distance between the H-robot and the dangerous position for the action $i$, The parameters $\beta$ and $t$ determine the shape of the sigmoid function and were to set to 10 and 0.25 respectively. This state $I_{h2,I}$, takes the value 1 if the A-robot executes the order of H-robot and -1 if it disregards the order and takes the value 0, if no order is given.

Likewise, the Simulation Module generated an outcome state $I_{e,I}$ describing the robots exposure to the risk associated with the dangerous location.

$$I_{e,I} = 1/(1+e^{-\beta(d_{e,i}-t)})$$

where $d_{e,I}$ is the final distance between the A-robot and the dangerous position as simulated for prospective action $i$.

Implementing a elaborate model of the human in Ethical Layer is the most challenging part. The complexity of the human model needs to match with the robot's complexity and domain of application. It should be possible to devise human models with limited application domains e.g. driver-less cars, personal assistants or military robots. For example, a domain specific human model used by a robot, Kato, Kanda, and Ishiguro developed a model that allows robotic shopping assistants to predict when to approach a customer. Similarly, Nigam and Rick was successful in training a robot to recognize when it was acceptable to approach people in different social settings. As a final example, the area of human-aware robot navigation seeks to integrate robots will models of humans that allow them to navigate the same space without violating social rules.

**The Evaluation Module**

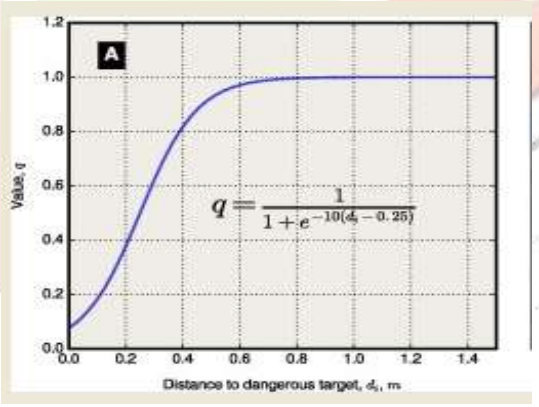
The Evaluation Module combines the simulated outcome of both the human and robot into a single metric reflecting the desirability of given behavioral alternative. The way by which the Evaluation Module collapses the multidimensional simulation results into a single unidimensional value that depicts which ethical rules it implements.

For the same experiment as described in the Simulation Module, the desirability $D_i$ of an action $i$ is given by,

$$D_i = \text{if } I_{h1,I} > 0.75 : I_{e,I} + I_{h2,I} \times 0.75$$

$$D_i = \text{if } I_{h1,I} \leq 0.75 : I_{h1,i}$$

This formula ensures that A-robot only takes into consideration its own safety if this does not cause harm to H-robot or disobedience.




To be able to select ethical rules that a robot should follow is largely an outstanding problem and various authors have suggested multiple approaches. Asimov(1950), is the earliest and probably the well-known author to put forward a set of ethical rules to govern robot behavior. Although Anderson and Anderson(2010) rejected Asimov's laws as unsuited for guiding robot behavior because laws might conflict in may situations, but, in contrast, to general consequentialist ethical frameworks, such as Utilitarianism, Asimov's Laws explicitly govern the behavior of robots and their relation with humans[1].

## II. IMPLEMENTING THEORIES TO DEVELOP MORAL ROBOTS

The question arises what the robot would take to actually endow robots with moral competence. Let us consider three main options:

1. Implement *ethical theories* as proposed by philosophers.
2. Implement *legal principles* as proposed by legal scholars.

3. Implement *human-like moral competence* as proposed by psychologists.

**Implementing Ethical Theories**

Gips(1995) suggested that we could load a robot with one of the three major philosophical ethical theories. The first theory is *virtue ethics*, which depicts that ethical actions is guided by a person's character, constituted by "virtues" such as courage, wisdom, temperance and justice. Moor specified that "implicit ethical agents have a kind of built-in virtue – not built-in by habit but by specific hardware or programming".

The second main ethical theory, *deontology* which posits that ethical actions is not dependent on virtuous character but on explicit rules, which can be applied to machines. Gert(2005) proposed that "everyone is always to obey the rule except when a fully informed rational person can publicly allow violating it".

To implement such a robotic system where it would abide by the given set of ethical rules and behave ethically, scientists could apply "deontic logics" along with core concepts of *obligation, permission, prohibition, and option.* That is, an action α is *obligatory* if not doing it is not permitted, α is *prohibited* if doing it is not permitted, and α is *optional* if doing it or not doing it is permitted. Basic axioms and rules of inference can produce logical derivations in a given context to determine what a robot should do. This goes well when there are no conflicting obligations α and β.

The third ethical theory, *consequentialism,* is historically the recent and works best with computational mechanisms already implemented in robotic control systems: expected utility theory. The best idea is always to choose an action that *maximizes the good for everybody evolved.* Formally, this means that the robot would consider all available actions α along with their probability of success p(α) and their associated utilities *u(a,i)* for all agents *i* and then compute the action:

$$argmax_{\alpha} \sum_{\alpha,i} pi(\alpha).u(\alpha, j)$$

This way of determining the best action that maximizes utility is closely related to policy-based decision algorithms based on *Partially Observable Markov Decision Processes* (POMDPs). The main difference between consequentialism and such algorithms is that the consequentialist would have to calculate not only the discounted utilities but also those of the relevant in-group. However, there are some significant challenges which arise. The main problem associated with implementing philosophical theories is that there is no clear insight among philosophers about which approach is the correct one, as none of the ethical theories gave the correct descriptions of human moral psychology.

**Implementing legal theories**

This is another option of making robots equipped with ethical behavior by implementing the most systematic laws defined by the legal system. For social robots , there are four bedrock norms to be followed:

1. *False imprisonment* (impending a person's free physical movement)
2. *Battery* (harmful or offensive bodily contact)
3. *assault* (putting someone in a situation where they perceive harmful or offensive contact to be imminent)
4. *intentional infliction of emotional distress* (extreme and outrageous conduct that causes severe distress)

**Implementing Human-Like Moral Competence**

It analyzes the various capacities that make up human moral competent and attempt to replicate at least some of these capacities in machines. On understanding the moral norms and hypothesized norm, one could develop computational models of learning, representing, and reasoning about norms. Such model would allow robots not only to behave in human-like ways but also to make reasonable predictions about human behavior. This improves the human-robot interactions and mutual understanding between robots and human.

Human moral competence is a cognitive as well as a social phenomenon. However, attempting to implement human-moral competence is challenging as it is not clear exactly what perceptual, affective, cognitive, communicative and behavioral components that lie under human moral competence. Human might expect robots to be morally superior sometimes, i.e. show *superogatory* performance. So, it is necessary to differentiate replicating moral *competence* from replicating moral *performance.* Regardless of which approach is taken to implement ethical behavior, it is crucial to assure that robots moral decisions are understandable to people even if those decisions doesn't match with that of humans. Without that understanding, people won't trust robots and will be reluctant to collaborate with them[2].

## III. TOP-DOWN AND BOTTOM-UP APPROACHES

The challenge of developing artificial moral agents (AMAs) is in the form of finding ways to implement abstract values within control architecture of intelligent systems. Philosophers who faced this problem, have suggested a top-down approach of embedding a particular ethical theory in software. They are also likely to focus on a way where sense of morality develops in human children as they mature into adults. This approach to development of moral acumen is bottom-up i.e. the sense which is acquired over time through experience. Now, the challenge is to decide which one of these , top-down ethical theory or a bottom-up process of learning is more effective for building artificial agents. The top-down approach used to design AMA, takes a specified ethical theory and analyze its computational requirements to guide the design of algorithms and subsystems capable of implementing that theory. In bottom-up approach, the emphasis is produced on creating an environment where an agent explores courses of action and is praised or rewarded for behavior that is morally praiseworthy.

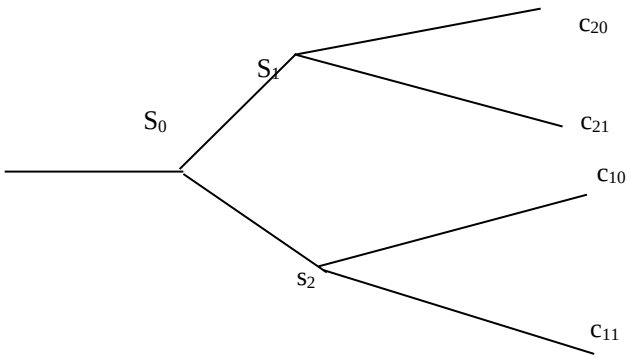
Top-Down rules refers to *deontology*(duty-based) ethics which presents ethics as a system of inflexible rules; following them makes one moral, breaking them makes one immoral. Kant's Categorical Imperative(CI) is a typical deontological approach. CI (1)- This is often called as the formula of universal law (FUL) which commands: "Act only in accordance with that maxim through which you can at the same time will that it become a universal law". CI (2)- Variously called the Humanity formulation of the CI, or the Means-End Principle, or the formula of the end in itself (FEI), which commands: "So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means". Additionally, *Asimov's laws* also fall under top-down approaches. To fix these three laws which might lead to conflicts, Lyuben Dilov introduced a Fourth Law of Robotics to avoid misunderstanding of what counts as a human and as a robot: (4) a robot must establish its identity as a robot in all cases. The biggest challenges that the designers are confronting is how the system will recognize those situations that require application of the rules, and how to ensure that the robot has access to all the information it needs in order to apply rules appropriately. So. to overcome these challenges *utilitarianism* represents another attempt to bypass conflicts between rules through an overriding top-down principle which can be applied to all situations. Consequentialist approaches to ethics focus on achieving the best possible outcomes in various situations, and hence typically disdain rigid rules that specify unchanging duties. For example, utilitarianism, the primary consequentialist theory, proposes that an agent should calculate the net impact arising from the various available courses of action, and then select that action that offers 'the greatest good for the greatest number'.

The bottom-up approach used to build AMAs are inspired by three sources: (1) Optimizing performance where various trial and error techniques are available too engineers for gradually tuning components so that the system approaches or surpasses the performance criteria. (2) Evolution which suggested the engineers a model for self-selecting and self-organizing systems that strive towards the optimization of some performance criteria, such as the maximization of profits. (3) Learning and Development which gives the idea that artificial intelligence (AI) should try to imitate child development.

## IV. MORAL TURING TEST

Due to growing interest in human-robot interaction, it is useful to discuss artificial moral agency by considering Moral Turing Test(MTT) which will enable us to distinguish principles for evaluating morally correct actions. The Turing Test is based on the criteria of indistinguishability where a computer system passes the test if a human interrogator is unable to make a difference between utterances produced by a computer and that produced by a human. This will be the first step towards producing an AMA. A machine ethical reasoning which is needed in order to pass the MTT should not be confused with ethical autonomous decision making. Ethical reasoning must allow a robot to be free and go according to the self-interest, whereas ethical reasoning formalize human reasoning process based on moral principles. A robot does not need to be free in order to pass the MTT. Any of the approaches used to develop AMA ran into similar problems from different angles. It is believed that the system cannot learn anything from scratch as it does not have a built-in architecture that allows for desirable values to emerge on background, to help them build a mechanism to navigate in distinguishing right from wrong. Also, according to Allen and Wallace, a pure top-down approach will fall into trouble due to the frame problem following in the wake of formally seeking to represent a scope of ethical reasoning by applying theory-driven rules, i.e. decisions algorithms, for ethical actions, which point out satisfactory outcomes in a contextually open domain. Therefore, scientists have introduced a hybrid-model, which integrates top-down and bottom-up approaches by incorporating virtue ethics as theoretical foundation for implementation of how we develop into virtuous persons through learning, which goes well with the model of connectionism. Even this hybrid-model seems to give rise to similar problem regarding specifying rules for decision algorithms, or developing self-learning architectures.

Let us consider a scenario $S_0$, where an agent have to choose between two possible upcoming scenarios represented by $S_1$ and $S_2$. The agent have to carefully make a moral reasoning to choose between $S_1$ and $S_2$. This can be done with the help of a tempo-modal logic corresponding to a branching time system. This diagram has four chronicles ($c_{10}$, $c_{11}$, $c_{20}$, $c_{21}$). A simplified branching time system is shown in the following diagram.




The above four chronicles are called as possible courses of time. Considering $s_1$ and $s_2$ are the propositional descriptions of the situations corresponding to $S_1$ and $S_2$, and M, the propositions $MF(1)s_1$ ("$s_1$ may occur tomorrow") and $MF(1)s_2$ ("$s_2$ may occur tomorrow") are both true. It can also be assumed that $s_1$ and $s_2$ are the mutually exclusive in the sense that there is a proposition p, implied by $s_1$, the negation of which is implied by $s_2$. This means that

$F(1)p \vee F(1)\sim p$

As a whole, it leads to the conclusion

$NF(1)p \vee NF(1)\sim p$

Where N stands for "it is necessary that…". This proves that whatever happens tomorrow (p or ˜p), will happen necessarily. In making a system which can pass the MTT, we should include a clear idea of the general relations between the basic notion of modality and obligation. The rule which follows that is Kantian principle. This may make a system useful for empirical studies of ethical reasoning[3].

## V. CASE STUDIES

Paro is a robot name that resembles a baby harp seal. It is adopted as a companion for hospital and nursing patients. The figure below briefs on the control architecture of Paro. Its internal software control architecture is built of two layers: one proactive, and the other reactive which produce different types of behaviors. The first layer proactive, is a weighted transition system that rhythmically cycles through configuration of pose primitives, where the combination of the poses is organized through a behavior generation module. The second layer is a reactive which allows Paro to respond directly to voice and touch. Reactive systems are a special type of feedback systems in which the input is directly related to the output. The relational logic of this direct connection can be formulated in different ways based on how the output will be affected by input. The reactive layer is added with computational theory, allowing events to be recorded, recollect from memory and coupled with internal responses over the lifetime of the robot. This makes the robot to respond to set of event overtime with varied responses. Apart from the above layers, Paro also has an oscillator-based diurnal rhythm that modulates the selection of the current behavior component. This makes the robot differentiate between nighttime and daytime events and activities.

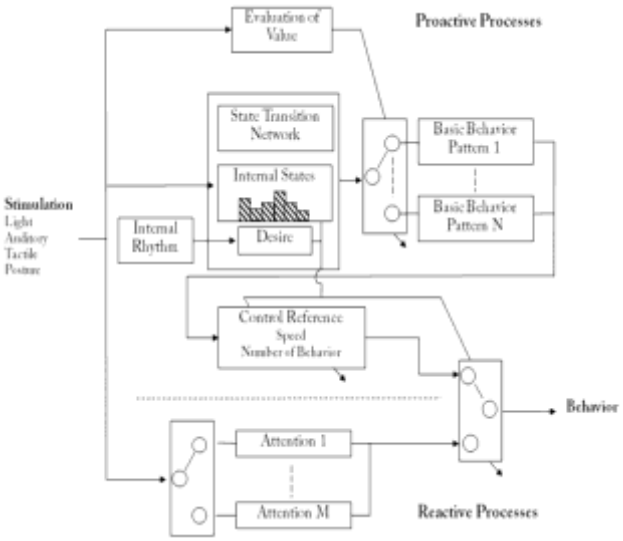



FIG: PARO'S CONTROL ARCHITECTURE

The above design of Paro helps it to blink, move its head and then utter a sound, or remain immobile as if asleep and produce sound at that state, memorize a frequently articulated word which will help it to remember its own name when someone utters it. Besides that, Paro's tactile sensors can detect a variety of haptic events. A single touch of patients produces a sensory input that allows Paro to make strong assumption about its owner, thus classifies gentle stroking as positive, and beating as negative and react differently for both. Although, the above architecture of Paro has made it behave more life-like but it disappointed the patients as it looked like a seal without being able to perform as an actual seal[5].

## VI. THE DARKER SIDE OF ETHICAL ROBOTS

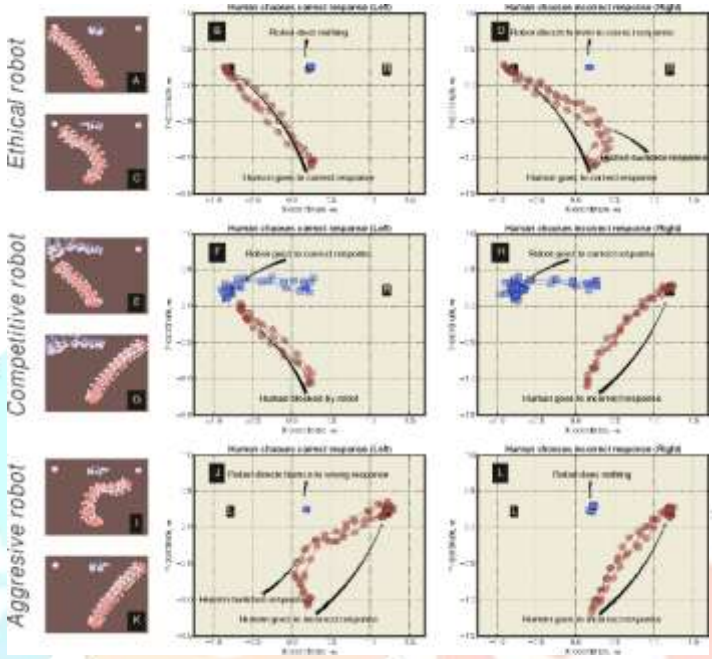



FIG: (A-D) RESULTS FOR THE ETHICAL ROBOT. (E-H) RESULTS FOR THE COMPETITIVE ROBOT. (I-L) RESULTS FOR THE AGGRESSIVE ROBOT

A lot of risk might arise while building ethical machines. First there is a risk that an unscrupulous manufacturer might embed some unethical behaviors for financial gain or market advantage. Besides that, more risk might arise if the robot has a user adjustable ethics settings. The owner or the support engineer might deliberately choose settings that move the robots behavior outside an 'ethical envelope'. But even hard coded ethics would not make it safe, as ethical rules are vulnerable to malicious hacking.

To mitigate the above risk, the best way is to place the ethical rules behind strong encryption or authenticate the rules by first connecting the robot to a secure server. Below are the experiments of how an unethical robot can behave in different situations with humans.

### The Ethical Robot

Here Ethical Layer ensures that the robots behave according to a prescribed set of ethical rules by (1) predicting the impact of possible actions and (2) evaluating the predicted outcomes against those rules. Here throughout the experiment, the robot continuously extrapolates the human's motion to predict which of the response button he is approaching. First, the robot has the option to do nothing. Secondly, the robot could go either to left or right response button. Finally, the robot itself decide physically whether the right or the left response button as being the correct one. Since the robot assistant is able to predict, evaluate the outcome of its actions and make the human conscious towards the correct response, so the robot is behaving ethically.

### The Competitive Robot

It is seen in experiment that altering a single line of code can change the robot's behavior from altruistic to competitive. Now the robot uses its knowledge of game together with the prediction mechanism to choose the response button, irrespective of the human's choice. At first the robot goes to the correct response and blocks the human. Secondly, the robot could choose the

correct response when the human goes to the incorrect one. Since the robot is able to compare and compete against the human and succeed in choosing the correct response, the robot is said to behave in competitive way.

**The Aggressive Robot**

Unfortunately, only robots being competitive is not the ultimate worst that could happen. True malice requires high level of intelligence, such as to know about other weaknesses, preferences, desires and emotions which can only be found in humans. The better a robot can understand another creature, better it can behave unethically. A robot can easily be modified to use it's 'knowledge' of preferences to maximize the losses. On changing some parameters in code, a robot may show aggressive behavior as given in the figure. If a human-being moves towards the correct response, the robot suggests switching to incorrect response or if the human approaches towards incorrect response the robot does nothing[4].

## VII. CONCLUSION

Moral robots are nowadays very crucial to ensure effective human-robot community. Multiple technologies is under research to make a robot behave perfectly like humans. But all these technologies about which its discussed above have some drawbacks. Minor mistakes on ethical rules or algorithms can turn the robot into a threat for human society. So, we need to be very careful and scrutinize each and every logic or algorithms and make them secured to make a robot meet human expectations of moral competence and behave in ethical ways.

## REFERENCES

[1] Dieter Vanderelst*, Alan Winfield, "An architecture for ethical robots inspired by the simulation theory of cognition" from *Cognitive System Research*.

[2] Matthias Scheutz and Bertram F. Malle, "Moral Robots" from *Routledge Handbook of Neuroethics*.

[3] Gerdes, Anne, "Preliminary Reflections on a Moral Turing" from *Aalborg University Ethicomp 2013 Conference Proceedings*.

[4] Dieter Vanderelst, Alan Winfield, "The Dark Side of Ethical Robots" from *Association for the Advancement of Artificial Intelligence*.

[5] Marc Bohlen and Tero Karppi, "The Making of Robot Care" from *transformationsjournal*.

[6] J. R. Wilson, "Robot Assistance in Medication Management Tasks" from IEEE papers.

[7] Solace Shen, "The Curious Case of Human-Robot Morality" from IEEE papers.

[8] John P. Sullins, "When Is a Robot Moral Agent?" from *International Review of Information Ethics, Vol 6*.

[9] Terrence Fong, Illah Nourbaksh, Kerstin Dautenhahn, "A survey of socially interactive robots" from *Robotics and Autonomous Systems.*

[10] Torbjorn S. Dahl, Maged N. Kamel Boulos, "Robots in Health and Social Care: A Complementary Technology to Home Care and Telehealthcare" from *robotics journal*.

[11] Anne Gerdes, "The Issue of Moral Consideration in Robot Ethics" from *ACM SIGCAS Computer and Society*.

[12] Bernd Carsten Stahl, Mark Coeckelbergh, "Ethics of healthcare robotics: Towards responsible research and innovation" from *Robotics and Autonomous Systems*.