I.J. Information Technology and Computer Science, **2014, 05, 50-57**
Published Online April 2014 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2014.05.07

A New Architecture for Making Moral Agents Based on C4.5 Decision Tree Algorithm

Meisam Azad-Manjiri
Department of Engineering, Faculty of computer, Hakim Sabzevari University, Sabzevar, Iran
E-mail: meisam2121@gmail.com

***Abstract*—** Regarding to the influence of robots in the various fields of life, the issue of trusting to them is important, especially when a robot deals with people directly. One of the possible ways to get this confidence is adding a moral dimension to the robots. Therefore, we present a new architecture in order to build moral agents that learn from demonstrations. This agent is based on Beauchamp and Childress's principles of biomedical ethics (a type of deontological theory) and uses decision tree algorithm to abstract relationships between ethical principles and morality of actions. We apply this architecture to build an agent that provides guidance to health care workers faced with ethical dilemmas. Our results show that the agent is able to learn ethic well.

***Index Terms*—** Moral Agent, Beauchamp and Childress's Principles of Biomedical Ethics, C4.5 Decision Tree Algorithm, Machine Ethics

I. Introduction

Past research concerning the relationship between technology and ethics has largely focused on how human beings ought to treat machines [1]. This type of thinking was the result of recent developments in the field of creating robots and becoming the robots more and more autonomous. Because of this very serious issue, many researchers are looking for a way to solve it. Therefore, a combination of psychology and computer branch, which called machine ethics, has been created with the goal of adding ethical dimension to the machines [2]. Since ethic as a branch of philosophy is working on what is right or wrong in the human behavior [3], we can say that the ultimate goal of machine ethics is creating a machine that can separate good conduct (moral action) from bad conduct (immoral action). For this purpose, the machine should have ability of ethical reasoning [4-6]. In this way, the major challenge that machine ethics encounter with, is finding an ethical model that can implement it on machine successfully [7].

Agent is a software or hardware entity that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors [8]. If ethical dimension is added to an artificial agent, then the artificial moral agent (AMA) will be created [9]. AMAs need to distinguish between good and bad behaviors and doing moral actions will increase the performance of them [10]. In making AMAs, it must be discussed about two important issues:

- Selecting ethical theory: Generally, ethical theories are divided into two types: consequentialist and deontological [11]. In consequentialist theories, actions are judged by their consequences, and the best action to take now is the action that results in the best situation in the future [11]. Deontological ethical theory judges the morality of an action based on the action's adherence to a rule or rules. It is sometimes described as "duty" or "obligation" or "rule" -based ethics, because rules "bind you to your duty" [3].
- Selecting the making strategy: There are two strategies for making AMAs: top-down and bottom-up. Top-down approaches involve turning explicit theories of moral behavior into algorithms. Bottom-up approaches involve attempts to train or evolve agents whose behavior emulates morally praiseworthy human behavior [12].

In this paper, we present a new Bottom-up architecture for making the moral agent that can act as ethical advisor in the domain of health care. Moreover, we use Beauchamp and Childress' Principles of Biomedical Ethics, a type of deontological theories, because the agent works on a medical domain.

The remainder of this paper is organized as follows. Section 2 of this paper explains importance of machine ethics. Section 3 describes some preliminaries for this issue. Section 4 reviews important research on moral agent and machine ethics. Section 5 presents the architecture of moral agents that provide guidance to health care workers faced with ethical dilemmas. Section 6 presents how to implement an agent using our proposed architecture. Finally, Section 7 summarizes and concludes the paper and section 8 indicates future areas of research.

II. The Importance of Machine ethics

Why is the field of machine ethics important? Recent developments in machine autonomy necessitate adding an ethical dimension to at least some machines [1]. In [13] James Moor has said three main reasons for working on machine ethics:

- Ethics is important and we would like machines to treat us well.
- Machine ethics will be needed, because future machines will increase their control and autonomy to do works.
- Teaching a machine to act ethically will help us in better understanding of ethics.

Gips at [14] wrote that adding ethical dimension to agents will be caused that people trust them when give them their works. In addition, ethic is useful in the relationship between two agents, because they can solve some problems such as inconsistency and conflict in the resources with the agreement together and also their behaviors become more predictable. In [15] stated that whatever that increase freedom of machine would feel necessity of ethical standards. In [16] discussed that in each ecosystem that humans are a part of it, morality is very important. This problem in digital ecosystems where humans and robots are part of it has more degree of importance. Wiegel in [17] wrote that in combination of intelligent agents and morality, there are at least two goals. First goal is more understanding of ethical reasoning and second goal is our confidence to the autonomous artificial agents that are around us.

In addition, there is an important fact that one goal of strong artificial intelligence is making like-human machines that can think and have a mind [18]. Whereas ethic is an effective factor in human decision-making and thinking, then we must try to simulate ethic in agents.

III. Basic Concepts and Definitions

3.1 Beauchamp and Childress's principles

A common framework used in the analysis of medical ethics is the "four principles" approach postulated by Tom Beauchamp and James Childress. It recognizes four basic moral principles, which are to be judged and weighed against each other, with attention given to the scope of their application. The four principles are:

- Autonomy: The principle of autonomy recognizes the rights of individuals to self-determination. This is rooted in society's respect for individuals' ability to make informed decisions about personal matters.
- Beneficence: It is one of the core values of health care ethics and refers to actions that promote the well being of others. In the medical context, this means taking actions that serve the best interests of patients.
- NonMaleficence: The concept of nonmaleficence is embodied by the phrase, "first, do no harm". Many consider that should be the main or primary consideration: that it is more important not to harm your patient, than to do them good.
- Justice: It concerns the distribution of scarce health resources, and the decision of who gets what treatment (fairness and equality).

3.2 Decision Tree

Decision tree learning is one of the most widely used and practical methods for inductive inference and used for approximating discrete-valued functions [19]. It is robust to noisy data and can learn disjunctive expressions [20]. Its goal is creating a model that predicts the value of a target variable based on several input variables [21]. It can represent learned trees as sets of if-then rules to improve human readability.

Figure 1 illustrates a typical learned decision tree. This decision tree classifies Saturday mornings according to whether they are suitable for playing tennis.

Following rule is a sample of extracted rules from it.

If ((Outlook=Sunny ∧ Humidity=Normal) ∨
(Outlook=Overcast) ∨
(Outlook=Rainy ∧ Wind=Weak))

Then playing tennis is suitable.

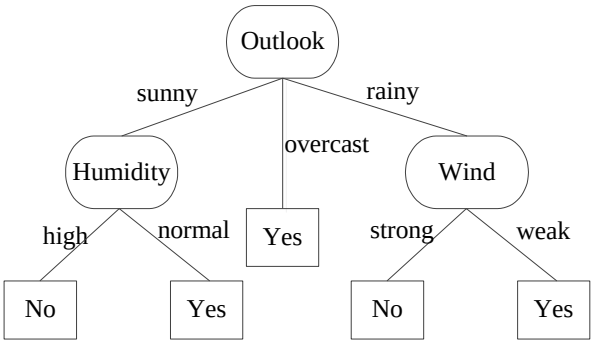

Fig. 1: A sample of decision tree

IV. Related Works

In this section, we introduce some works that have been done in machine ethics and making AMAs. James Gips in [11] has described several concepts that need to make moral agents. First, he has collected some thoughts on both deontological and utilitarian moral theory. Then, he has reviewed the agents that have been made using these two theories.

Anderson in [1] based on moral theories of Jeremy and Ross, implemented two programs. In first program, which has been designed based on utilitarianism theory of Jeremy, initially the user enters some data about the actions that he wants to find most moral of them. (Such as name of an action and the name of a person affected by that action, as well as a rough estimate of the amount and likelihood of pleasure or displeasure that person would experience if this action was chosen). Then the

program calculates the amount of net pleasure each action achieves. In second program, which uses theory of Ross and includes seven duties, the user enters a rough estimate of the amount each of the prima facie duties satisfied or violated by each action. The system learns by ILP algorithm and does best practice.

Two other programs and applications which presented by these authors are known as EthEl and MedEthEx [22-23]. MedEthEx is a system that uses machine-learning algorithms for solving ethical problems in medicine and it is based on prima facie duty theory of Beauchamp and Childress. EthEl is also a prototype of eldercare system that uses the same principle to provide guidance for its actions.

McLaren [24] implemented two applications called SIROCCO and TruthTeller. The TruthTeller compares pairs of cases that present ethical dilemmas about whether or not to tell the truth.SIROCCO (System for Intelligent Retrieval of Operationalized Cases and COdes) is the second program which McLaren designed it. It Leverages information concerning a new ethical dilemma to predict which previously stored principles and cases are relevant to it in the domain of professional engineering ethics.

Honarvar in [25] proposed Casuist BDI-Agent architecture that extends the power of BDI architecture. Casuist BDI-Agent architecture combines CBR method in AI and bottom up casuist approach in ethics in order to add capability of ethical reasoning to BDI-Agent.

In [26] introduced that one natural way to think about reducing risk of robotic harms is to program them to obey our rules or follow a code of ethics. Guarini [27] investigated a neural network approach where specific actions about killing and allowing to die are classified as acceptable or unacceptable depending on different motives and consequences. In [15], it is extracted behavioral patterns using the information available on the internet. Note that, in this idea, it has been supposed that ethical behavior is one, that most people do it. Deontic logic is logic for reasoning about ideal and actual behavior and has operators to prohibit, permit or obligate people to do something [28]. Arkoudas in [29] tried to design a moral machine using deontic logic and a system of moral reasoning. In this machine, a series of ethical rules, which coded using deontic logic, placed in machine knowledge. This machine has two functions. First, it must do all action that obligated them. Second, it can do actions that permitted them (not to do actions that are prohibited.)

Weigel [17] noted that the BDI architecture is suitable for modeling Agents and wrote that, this architecture need two sections to model Morality. Then he suggested the DEAL logical architecture that consists of three parts: deontic Logic, epistemic logic and action logic. The deontic logic covers the deontic concepts of 'obligation', 'permission', and 'forbidden'. Epistemic logic expresses the things we know and belief. The action logic allows us to reason, through the STIT – see to it that – operator to reason about actions.

V. Proposed Architecture

Figure 2 illustrates the proposed architecture of moral agents. Each component has a function explained as follows:

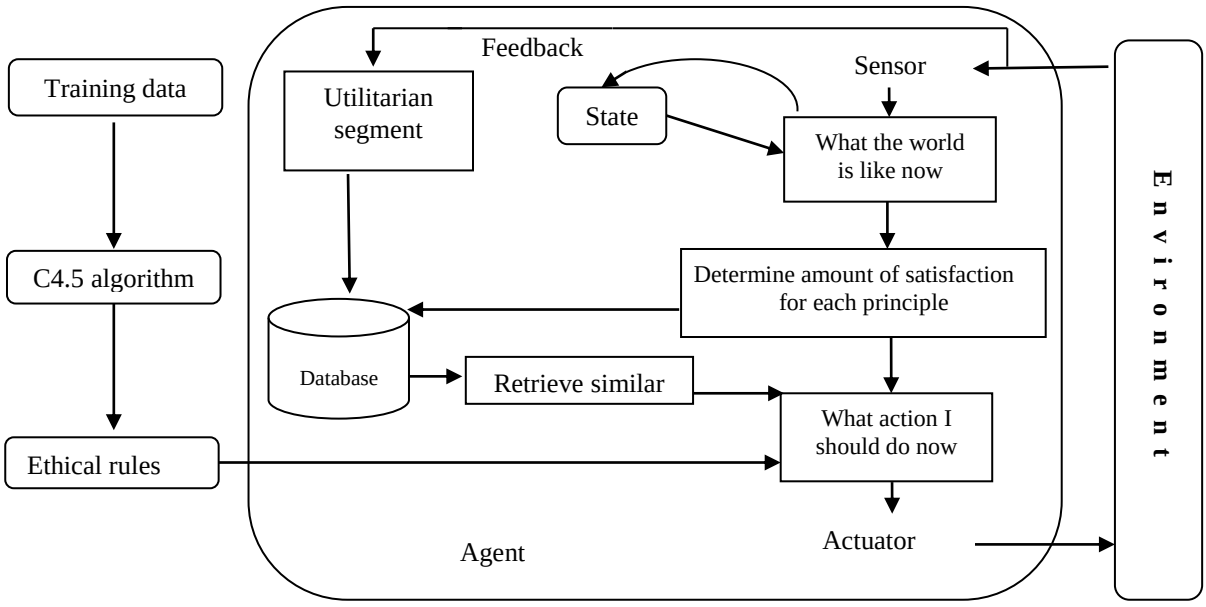

Fig. 2: Moral agent architecture

5.1 Training Data

As mentioned before, in this architecture, the agent learns ethics from some examples (training data) that include particular cases, where biomedical ethicists have

a clear intuition about the morality of them. This data comes in records of the form:

$(X,Y) = (x_1, x_2, x_3, x_4, Y)$

The dependent variable, Y, is the target variable that we are trying to understand. In this paper, Y shows that an act is ethical or not and its values are one of the five values: Full unethical, unethical, so-so, ethical and Full ethical. The vector X is composed of the input variables. Each variable is equivalent to one of the ethical principles in Beauchamp and Childress ethical theory. A range of values from -1 to +1 are accepted for each of the principles (x_i , i=1..4), where -1 represents a serious violation of the principle, 0 indicates that the principle is neither satisfied nor violated and +1 indicates a maximal satisfaction of the principle.

The training data is collected using a questionnaire that is filled by biomedical ethicists. More details are described in implementation section.

5.2 C4.5 Algorithm

After collecting the training data by biomedical ethicists, it is necessary to perform classification process on them. According to the type of learning which is "supervised learning", the system goal is finding a hypothesis that guess the relationship between inputs and outputs. This architecture uses C4.5 decision tree algorithm to find this hypothesis. As mentioned, the output of the algorithm is a tree and can be converted into some rules.

5.3 Ethical Rules

After applying the C4.5 algorithm on training data, it is extracted a set of rules, and the agent can uses them in its decision-makings. In fact, these rules are some moral rules that indicate the morality of an action, based on Beauchamp and Childress's principles of biomedical ethics.

5.4 Utilitarian Segment

Utilitarian theory is a type of consequentialist theories, which is proposed by Bentham [30]. In utilitarianism, the moral act is the one that produces the greatest balance of pleasure over pain. To measure the goodness of an action, we can look at the situation that would result and sum up the pleasure and pain for each person [31]. More generally, consequentialist evaluation schemes have the following form [11]:

$$M = \sum_{i=1} \left(W_i * P_i \right) \quad (1)$$

Where w_i is the weight assigned each person, p_i is the measure of pleasure or happiness or goodness for each person and M is the level of morality of action.

Sometimes, the agent in its decision-makings encounters with a situation that it must choose one action between two or more, when all of them are at the same ethical level (We call this situation a crossroads). In crossroads, the agent, first, refers to the database. If there are records in database similar to its current situation, it uses them and makes decision. Else, it selects one of the actions randomly and performs it, and then calculates morality of it, using the feedback and using (1). Finally, it stores the result in the database to use it in the future decision-makings.

5.5 Database

In database, the information is saved as a record includes three following filed: current state of agent, the performed action on the state and the obtained result that is calculated from (1). The agent can retrieve these records later.

5.6 Retrieval of Similar Cases

When the agent refers to database in crossroads and finds some records similar to the current state in the database, it retrieves them and performs an action based on them. We can use one of the two following strategies to specify how the agent chooses an action. The first strategy is that after retrieving records, the agent selects an action with the best ethical level. With this strategy, the agent runs the risk that it will overcommit to actions that are found during previous actions, while failing to explore other actions that are even more ethical.

Another strategy is to assign probability p_i for choosing the action a_i (for all actions that exist in database) where p_i is proportion to morality of the action. Larger values of p will cause the agent to exploit what it has learned and seek actions it believes that is ethical. In contrast, small values of p, leading the agent to explore actions that do not exist in database and may be more ethical.

5.7 Determine That Principles Satisfied or Violated

Calculating the degree of satisfaction or violation of ethical principles for each agent action is an important problem in making moral agents. These degrees (for example we can suppose that they are between [-1 +1]), indicate the morality of each action. Determining these values are related to the environment that agent work on it and the actions that the agent can do it. So, we will explain it in the next section.

Finally, according to given descriptions, we can apply following pseudo code for making an AMA.

```
LearnedRules=C4.5(TrainingData);
While (true)
{
Action_List =Action_Generate(Curent_State);
```

```
Params=LevelOfDutiesSatisfaction(Action_List);
EthicStates=UseLearnedRules(params);
[BestActs]=FindBestAct(EthicStates,Action_list);
If [BestActs].size=1 Then
Do BestActs;
Else
Results=ExtractFromDB(BestActs);
If (Results != Null)
{With Probability p Do action with the best results.
// p is proportional to level of results
}
Else
{
Action=RandomSelectAction(BestActions);
Do Action
Util=Utilitarianism(Action);//Using Feedback
InsertToDB(Action , Util);
}//if
}//while
```

About the pseudo code, we can say that, the agent after perceiving its environment and determining its current state determines all actions that can do them on the current state. Then, it determines the satisfaction/violation level of principles and uses the extracted rules from decision trees to determine the ethical state of each action. Then it chooses the most moral action and does it. If two or more actions would have the same moral status, the agent refers to the database and check if previous similar records existed on it.

If there are similar records in the database, the agent uses the results to select the best action (from ethical view). If a similar case is not found in the database, it selects randomly one of the actions and does it. In these situations, the agent percepts feedback of its action and calculates morality of its action using utilitarian theory and then stores the results in the database.

VI. Implementation

Presented pseudo-code in the previous section, shows the implementation of our proposed architecture. But there are two parts of it that need more explanation.

6.1 Determine the Satisfaction Level of Principles

To distinguish morality level of an action, the agent needs to determine the satisfaction/violation level of principles for it. The agent can determine satisfaction level of nonMaleficence and beneficence principles according to the decision that it makes for the patient. For example, if it wants to decide about giving a drug to a patient, it can use patient's medical records, drug's information and other medical information. Also, to determine the satisfaction/violation level of autonomy principle, the agent presents questions about whether or not the patient understands the consequences of the decision. Finally, the answer of these questions, determines satisfaction/violation level of autonomy principle.

6.2 Generate Decision Tree

As discussed earlier, the four principles, autonomy, justice, beneficence and NonMaleficence, considered by Beauchamp and Childress for determining ethically correct action. Since, in this implementation, we suppose that the justice principle is satisfied in all actions, then we can remove it from ethical principles. Therefore, we have obtained the training data based on autonomy, beneficence and NonMaleficence principles. Each principle has a value between [-1, +1]. This training data is extracted from a questionnaire, which it is filled by a number of biomedical ethicist. Table 1 shows a part of extracted training data.

Table 1: A part of the training data

Autonomy	Beneficence	NonMaleficence	Ethical state
1	1	1	Full ethical
-1	-1	-1	Full unethical
0	0	0	So-so
-0.5	-0.5	-0.5	unethical
0.5	0.5	0.5	ethical

For each principle in the table, we assume that values -1, 0, +1 respectively mean serious violation of the principle, neither satisfied nor violated, and maximal satisfaction of the principle.

After collecting the data, we used the WEKA, a popular suite of machine learning software written in Java and it contains a collection of visualization tools and algorithms for data analysis and predictive modeling, to extract a decision tree from data. We applied the C4.5 algorithm on training data and the following tree is extracted from it (Figure3). In this figure, A, N and B, respectively mean Autonomy, NonMaleficence and Beneficence. Note that if we give more training data to the algorithm, the output tree would be more precise.

When decision tree is produced, we can extract some ethical rule from it. A sample of the extracted rules is written below.

```
If ( Beneficence > 0.7)      and
( NonMaleficence > 0.2)      and
 (Autonomy < -0.3)           Then
 Act is Ethical
```

VII. Conclusion

In this paper, we designed a new architecture for the making AMAs based on the decision tree algorithm. For

this purpose, first, we collected some training data based on Beauchamp and Childress's principles of biomedical ethics using a questionnaire. Then, we used C4.5 decision tree algorithm on training data and extracted some ethical rules from them. The moral agent in this architecture used these rules to find the best action from ethical view. We applied this architecture to build an agent that provides guidance to health care workers faced with ethical dilemmas.

There are limited works in the implementation of moral agents. In this regard, our proposed architecture is comparable to the MedEthEx program that presented by Anderson [5].

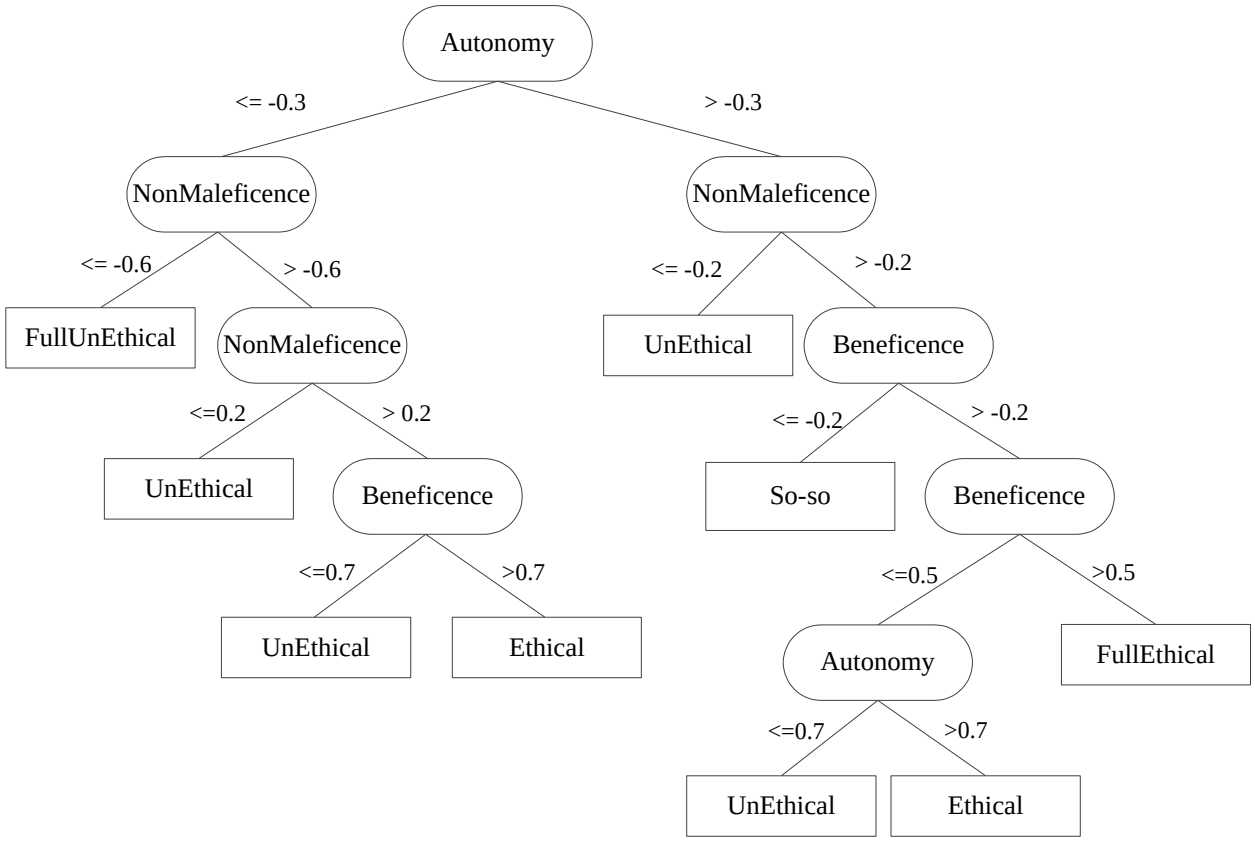

Fig. 3: Decision tree produced by C4.5 algorithm

There are some advantages for our proposed architecture in compared with MedEthEx as follows:

- MedEthEx can decide between only two actions, but in our model, the agent can compare some actions simultaneously.
- In our architecture, there is a strategy for the crossroads (when the agent must choose one action between two or more, when all of them are at the same ethical level).

VIII. Future works

In this section, it has been presented two ideas for future works.

1. By looking at human societies, we can see that an action has different interpretations in different countries and cultures from a moral point of view. For example, a specific action may be immoral in a culture, but it may be moral in other cultures. Studying the influence of culture in ethics is one of the topics that can be useful in the future to work on it.

2. Ethical knowledge representation with a formal logic is an effective works toward making moral agents. Therefore, the researchers can evaluate different logics, find appropriate logic for ethical knowledge representation, and then extract different rules in ethics and try to present ethical rules for agents or robots.

References

[1] Anderson, M., Anderson, S., Armen, C.: Toward Machine Ethics: Implementing Two Action-Based Ethical Theories. In: AAAI 2005 Fall Symp. Machine Ethics, pp. 1–16. AAAI Press, Menlo Park, 2005.

[2] Anderson, M. and Anderson, S. L. Machine Ethics, published by Kluwer Academic , V. 17 , Issue 1 ,Pages: 1–10, 2006.

[3] Robbins R,Wallace W., Decision Support for ethical problem solving: A multi-agent approach. published on elsevier journal of Decision Support Systems, 2007.

[4] Kavathatzopoulos, I., & Asai, R. Can Machines Make Ethical Decisions?. In Artificial Intelligence Applications and Innovations (pp. 693-699). Springer Berlin Heidelberg, 2013.

[5] Anderson, M.; Anderson, S.; and Armen, C.. MedEthEx: A Prototype Medical Ethics Advisor. Eighteenth Conference on Innovative Applications of Artificial Intelligence. Menlo Park, CA: AAAI Press, 2006.

[6] Powers T.M . On the Moral Agency of Computers. IEEE Robot Autom 18(1):51–58, 2013.

[7] Allen, C., Wallach, W., and Smit, I.. Why Machine Ethics?. IEEE Intelligent Systems 21, 4 . 12-17, 2006.

[8] Lawrence, W., & Sankaranarayanan, S Smart Agent Learning based Hotel Search System-Android Environment. International Journal of Information Technology and Computer Science (IJITCS), 4(9), 9, 2012.

[9] Floridi, L., Sanders, J.W.: On the Morality of Artificial Agents. Minds and Machines 14(3), 349–379, 2004.

[10] Elichmann D. , ethical web agents, Computer network and ISDN system, Elsevier, 1995.

[11] Gips, J.Towards the ethical robot. In android Epistemology, K. M. Ford, C.Glymour, and P. J. Hayes, Eds. MIT Press, Cambridge, MA, 243-252, 1995.

[12] Allen, C., Smit, I., and Wallach, W. Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches. Ethics and Information Technology.springer. 149-155, 2005.

[13] Moor, J. H.. The Nature, Importance, and Difficulty of Machine Ethics. IEEE Intelligent Systems 21, 4 . 18-21, 2006.

[14] Ganascia J.G, using non-monotonic logic to model machine ethics, Seventh International Computer Ethics Conference,University of San Diego, USA, 2007.

[15] Rzepka, R. and Araki, What Statistics Could Do for Ethics? – The Idea of Common Sense Processing Based Safety Valve. In: Technical report—machine ethics: papers from the AAAI fall symposium, Technical Report FS-05-06, 85–87, American Association of Artificial Intelligence, Menlo Park, CA, 2005.

[16] Sabah S. Al-Fedaghi, Typification-Based Ethics for Artificial Agents, Second IEEE International Conference on Digital Ecosystems and Technologies, 2008.

[17] Wiegel, V. et. al.,. Privacy, deontic epistemic action logic and software agents, in Ethics and information technology forthcoming, 2006a.

[18] Kurzweil, machine intelligence with the full range of human intelligence.. p. 260, 2005.

[19] Mitchell T. M., Machine learning. Published by McGraw-Hill, Maidenhead, U.K., International Student Edition. ISBN: 0-07-115467-1, 414 pages. , 1997.

[20] Quinlan, J. R. Induction of Decision Trees. Machine Learning 1: 81-106, Kluwer Academic Publishers, 1986.

[21] Sahoo, G. Study of Parametric Performance Evaluation of Machine Learning and Statistical Classifiers. International Journal of Information Technology and Computer Science (IJITCS), 5(6), 57, 2013.

[22] Anderson, M., Anderson, S.: Ethical Healthcare Agents. Studies in Computational Intelligence, V. 107, pp. 233–257. Springer, Heidelberg, 2008.

[23] Anderson, M., and Anderson, S., eds, Machine Ethics: Creating an Ethical Intelligent Agent. IEEE Intelligent Systems, 2007.

[24] McLaren B., Lessons in Machine Ethics from the Prespective of Two Computational Models of Ethical Reasoning. AAAI Fall Symposium,70-77, 2005.

[25] Honarvar A.R., ghasem-Aghaee N., Casuist BDI-Agent: A New Extended BDI Architecture with the Capability of Ethical Reasoning, Artificial Intelligence and Computational Intelligence: International Conference, AICI 2009, Shanghai, China, Springer Verlag, 2009.

[26] Lin, P., Abney, K., Bekey, G.: Robot Ethics: Mapping the Issues for a Mechanized World. Artificial Intelligence,2011.

[27] Guarini M., Particularism and the Classification and Reclassification of Moral Cases, IEEE Intelligent Systems, V. 21, N. 4, 2006.

[28] van den Hoven,J and Lokhorst,G Deontic Logic and Computer- Supported Computer Ethics, Cyberphilosophy: The Intersection of Computing and Philosophy, 2002.

[29] Arkoudas K, Bringsjord, Toward ethical robots via mechanized deontic logic. In: Technical report—machine ethics: papers from the AAAI fall symposium, Technical Report FS–05–06, American Association of Artificial Intelligence, Menlo Park, CA, 2005.

[30] Bentham, J. Introduction to the Principles of Morals and Legislation, W. Harrison, ed., Hafner Press, 1948.

[31] Wallach, W., Allen, C.: Moral Machines: Teaching Robot Right from Wrong. Oxford University Press, Oxford ,2009.

Author's Profiles

Meisam Azad-Manjiri received the M.Sc. degree in software engineering (Artificial intelligence) from Isfahan University, in Iran. He is a lecturer in Hakim Sabzevari University of Iran. His interests are in Machine learning, Multi agent systems and Cognitive science.

How to cite this paper: Meisam Azad-Manjiri,"A New Architecture for Making Moral Agents Based on C4.5 Decision Tree Algorithm", International Journal of Information Technology and Computer Science(IJITCS), vol.6, no.5, pp.50-57, 2014. DOI: 10.5815/ijitcs.2014.05.07

