# Rule based fuzzy cognitive maps and natural language processing in machine ethics

Rollin M. Omari
*School of Computer Science, Australian National University, Canberra, Australia, and*
Masoud Mohammadian
*Faculty of Business, Government and Law, University of Canberra, Canberra, Australia*

Received 17 October 2015
Revised 12 April 2016
Accepted 14 April 2016

## Abstract

**Purpose** – The developing academic field of machine ethics seeks to make artificial agents safer as they become more pervasive throughout society. In contrast to computer ethics, machine ethics is concerned with the behavior of machines toward human users and other machines. This study aims to use an action-based ethical theory founded on the combinational aspects of deontological and teleological theories of ethics in the construction of an artificial moral agent (AMA).

**Design/methodology/approach** – The decision results derived by the AMA are acquired via fuzzy logic interpretation of the relative values of the steady-state simulations of the corresponding rule-based fuzzy cognitive map (RBFCM).

**Findings** – Through the use of RBFCMs, the following paper illustrates the possibility of incorporating ethical components into machines, where latent semantic analysis (LSA) and RBFCMs can be used to model dynamic and complex situations, and to provide abilities in acquiring causal knowledge.

**Research limitations/implications** – This approach is especially appropriate for data-poor and uncertain situations common in ethics. Nonetheless, to ensure that a machine with an ethical component can function autonomously in the world, research in artificial intelligence will need to further investigate the representation and determination of ethical principles, the incorporation of these ethical principles into a system's decision procedure, ethical decision-making with incomplete and uncertain knowledge, the explanation for decisions made using ethical principles and the evaluation of systems that act based upon ethical principles.

**Practical implications** – To date, the conducted research has contributed to a theoretical foundation for machine ethics through exploration of the rationale and the feasibility of adding an ethical dimension to machines. Further, the constructed AMA illustrates the possibility of utilizing an action-based ethical theory that provides guidance in ethical decision-making according to the precepts of its respective duties. The use of LSA illustrates their powerful capabilities in understanding text and their potential application as information retrieval systems in AMAs. The use of cognitive maps provides an approach and a decision procedure for resolving conflicts between different duties.

**Originality/value** – This paper suggests that cognitive maps could be used in AMAs as tools for meta-analysis, where comparisons regarding multiple ethical principles and duties can be examined and considered. With cognitive mapping, complex and abstract variables that cannot easily be measured but are important to decision-making can be modeled. This approach is especially appropriate for data-poor and uncertain situations common in ethics.

**Keywords** Decision making and ethics, Natural language processing in machine ethics, Rule based fuzzy cognitive maps

**Paper type** Research paper

Journal of Information, Communication and Ethics in Society
Vol. 14 No. 3, 2016
pp. 231-253

1477-996X
DOI 10.1108/JICES-10-2015-0034

## 1. Introduction

The emerging field of machine ethics endeavors to transform our increasingly pervasive machines into safer artificial agents. This recently emerging subfield of artificial intelligence (AI) is increasingly motivated by concerns regarding the dangers artificial intelligent agents may pose to humanity (McCarthy and Hayes, 1969; Anderson *et al.*, 2004; Yampolskiy, 2013; Bostrum, 2014), and focuses on the underlying design and principles directed toward constraining lethal actions of autonomous agents, so that their behaviors are bounded. Namely, machine ethics ultimately strives to develop next-generation autonomous agents, capable of following ideal ethical principles in decisions they make (Anderson and Anderson, 2007; Arkin, 2008), and it draws upon interdisciplinary knowledge accumulated in both computer science and philosophy. By focusing on the ethically acceptable behavior of artificial agents, this new field distinguishes itself from the earlier work of computer ethics, a field that has traditionally focused on the ethical issues regarding the use of technology by humans (Anderson and Anderson, 2007).

Within the machine ethics research community, it is commonly agreed-upon that any artificial moral agent (AMA) currently engineered would be an *implicit ethical agent*, that is a machine capable of carrying out its intended purpose in a safe and responsible manner as determined by its *designer*, and not necessarily able to extend its moral reasoning to novel situations (Moor, 2006; Shulman *et al.*, 2009; Chen *et al.*, 2009). Opinions within the field, however, fragment on the desirability and feasibility of developing an *explicit ethical agent,* that is an AMA analogous to an ethical human decision-maker, capable of calculating the best action in ethical dilemmas, represent ethical principles explicitly, operate effectively on its knowledge base and justify all moral judgments and actions (Anderson and Anderson, 2007; Moor, 2006; Shulman *et al.*, 2009).

In this paper, two considerable challenges of machine ethics are explored. First is the issue of the acquirement and incorporation of the necessary information needed to make an ethical decision, which we attempt to solve by using latent semantic analysis (LSA); second is the issue of the actual ethical reasoning and decision-making process, which we attempt to solve through the use of rule-based cognitive maps (RBFCMs). The latter problem is considered to be especially challenging due to the incomplete codification of ethics (Anderson and Anderson, 2007). This incomplete standard has resulted in the establishment of general disagreement on what moral structure AMAs should possess, with diverging suggestions ranging from the application of evolutionary algorithms in populations of artificial agents to achieve "moral selection", neural network models of cognition and various hybrid approaches founded on ethical theories, such as virtue ethics, Kant's Categorical Imperative, utilitarianism, value systems inspired by the Golden Rule and several others (Shulman *et al.*, 2009).

The following study will demonstrate how RBFCMs can be used to represent causality, where inputs and their effects are modeled using fuzzy operations (e.g. and, or, if, then). The what-if scenarios commonly associated with cognitive maps are used to mimic the thought processes associated with human thinking. This study will demonstrate that the evolution of an RBFCM is iterative, where current values for each of the concepts are computed with their inputs' previous values. Here it will be illustrated that the use of either crisp or fuzzy values for particular concepts is far superior in capturing dynamic causal relations between concepts. The study will further

demonstrate that a possible solution to automated information retrieval and incorporation is the use of either Web crawlers or large internet repositories of information, such as Wikipedia. Finally, this paper will discuss the use of LSA to calculate the semantic similarity between words/concepts, and provide a case study as an example of an AMA applied in a limited domain, where it uses ethical principles in the guidance of its own behavior.

### *1.1 Explicit ethical machines*

The goal of most machine ethical endeavors is the development of explicit ethical AMAs. An explicit ethical agent is able to explain why a particular action is either right or wrong by appealing to an ethical principle, demonstrating the critical distinction between explicit ethical agents and implicit ethical agents, which is the ability to justify ethical judgments that only an explicit representation of ethical principles allows (Anderson and Anderson, 2007). Therefore, one can safely consider that machines that have learned, or are programmed, to make correct ethical judgments, but do not have principles to which they can appeal to justify or explain their judgments, are lacking an essential requirement to being acknowledged as explicit ethical agents (Anderson and Anderson, 2007).

The challenges for machine ethics, therefore, can be divided into two main categories: philosophical concerns about the feasibility of computing ethics and programming challenges from the AI perspective (Anderson and Anderson, 2007). In the former category, one needs to address the question of whether ethics is a computable problem. Act utilitarianism is a well-known ethical theory that provides an affirmative answer to such a question; this theory maintains that the rightness and wrongness of actions are determined entirely by the consequences of those actions (Anderson and Anderson, 2007).

For machine ethics, act utilitarianism is easily implementable through the use of an algorithm capable of computing the best action, an action that derives the greatest amount of net good or pleasure, from all alternative actions. For each action and person affected, the algorithm computes the intensity of the pleasure/displeasure, the duration of the pleasure/displeasure and the probability of occurrence of the pleasure or displeasure (Anderson and Anderson, 2007). In its implementation, such an algorithm represents a situation where a machine, acting as an advisor to humans, can prompt the human user to consider alternative actions that might result in greater net good consequences than the actions being considered (Anderson and Anderson, 2007). Such a machine can further prompt the human user to consider the effects of each of those actions on all those affected, where for some individuals, actions can result in consequences affecting entire nations.

Although the theory of act utilitarianism represents a good starting point in programming ethically sensitive AMAs, critics of the theory, however, have argued that the theory can allow for human rights violation, and can contradict some of our notions of justice (Anderson and Anderson, 2007). However, as suggested by Ross (1930), these problems can be supplanted by the combination of the elements of both teleological theories and deontological theories. A deontological approach to ethics, such as Kant's categorical imperative, can emphasize the importance of rights and justice, although at the risk of ignoring consequences, by combining the best elements of the two approaches, a theory with several prima facie duties can be generated, where some

duties are concerned with the consequences of actions and others concerned with justice and rights, thus acknowledging the complexities of ethical decision-making than a single absolute duty theory (Ross, 1930).

The AMA developed in this paper utilizes RBFCMs to implicitly apply a decision procedure for these multiple prima facie duties in a constrained domain. The map is designed to connect and abstract a decision principle from prima facie duties that often provides conflicting advice in ethical dilemmas. When conflicts do arise, a default action is considered from cases of ethical dilemmas where ethicists are in agreement as to the correct action. Such an approach to ethical decision-making is believed to be more likely to capture the complexities of ethical decision-making than a single, absolute duty ethical theory. This paper uses fuzzy rules to abstract information leading to a general decision principle from ethical experts' intuitions about particular ethical dilemmas. The virtue of having abstraction and principles to follow, rather than being programmed in an *ad hoc* fashion to behave correctly in limited situations, the AMA can have a means to represent ethical dilemmas in a wider variety of cases in different domains and determine the ethically correct action in such novel situations.

### *1.2 Latent semantic analysis*

For information retrieval purposes, there are certain advantages in the work of semantic similarity by utilizing corpus statistics associated with LSA. LSA is a fully automatic mathematical and statistical technique for extracting and inferring relations of the expected contextual usage of words in passages of discourse. Its application in machine ethics seems appropriate, as LSA has been found to be capable of simulating a variety of human cognitive phenomena, ranging from developmental acquisition of recognition vocabulary, sentence–word semantic priming, discourse comprehension, word categorization and essay grading (Laham and Foltz, 1998).

Furthermore, LSA can be understood in two ways:

(1) simply as a practical convenient tool for obtaining the meaning similarities among words and texts, and approximating the substitutability of words in larger text segments based on their contextual usage; or

(2) as a model for the computational processes and representations underlying human cognitive acquisition and utilization of knowledge (Landauer *et al.*, 1998).

In both cases, LSA produces measures of word–word, word–passage and passage–passage relations that are comparable to several human cognitive phenomena involving association or semantic similarity (Landauer *et al.*, 1998). All of these capabilities hinge upon the fact that any textual analysis derived by LSA is not based on simple contiguity frequencies, co-occurrence counts or word-usage correlations, but depend on a powerful mathematical analysis that is capable of correctly inferring deeper relations (hence *latent semantic*), and thus consequently able to predict human meaning-based judgments and performance (Landauer *et al.*, 1998).

Commonly, the first step in producing semantic measures via LSA is the representation of a text as a matrix in which a row stands for a unique word and a column stands for a sentence or text passage (Laham and Foltz, 1998). Each cell then contains a weighted transformation of the number of times that a given word appears in a given passage, calculated by a function that expresses the word's importance in the particular context and wider context (Laham and Foltz, 1998).

The second step involves decomposing the matrix in such a way that every passage is represented as a vector whose value is the sum of vectors standing for its component words. This step utilizes singular value decomposition (SVD), where the rectangular matrix generated in the previous step is decomposed into the product of three other matrices. One component matrix describes the original row entities as vectors of derived orthogonal factor values, another describes the original column entities in the same way and the third is a diagonal matrix containing scaling values such that when the three components are matrix multiplied, the original matrix is reconstructed (Landauer *et al.*, 1998). Finally, the measure of similarity, between words and words, passages and words and passages and passages, is computed in the reduced dimensional space utilizing computed dot products, cosines or other vector-algebraic metrics (Landauer *et al.*, 1998).

LSA, thus, simply represents the meaning of a word as an average of the meaning of all the passages in which it appears and the meaning of a passage as an average of the meaning of all the words it contains. In this paper, LSA is used due to its specific ability to simultaneously derive representations of these two interrelated kinds of meaning. LSA-derived semantic measurements assume great importance in the choice of parameters by which a word or passage can be described. By reducing this dimensionality, and combining it with surface information into deeper abstractions, mutual implications of words and passages can be captured, and better approximations to human cognitive relations can be produced. These properties are utilized in the AMA to extract and infer relations between a background corpus representing utilitarian principles used in biomedical ethics, and the ethical case study to be solved. Furthermore, LSA's efficacy in knowledge acquisition, semantic relatedness and information retrieval is used to analyze the ethical case study, determine the essential concepts and assist in the construction of the RBFCM in this paper.

### *1.3 Rule-based cognitive maps*

Fuzzy cognitive maps (FCMs) represent a modeling approach developed by Kosko that combines fuzzy logic and neural networks (Kosko, 1997). FCMs are considered to be highly robust and capable of modeling complex and intelligent systems (Papageorgiou *et al.*, 2005). More specifically, the robust nature of FCMs is illustrated in their application in planning and decision-making in the fields of international relations and social systems modeling (Papageorgiou *et al.*, 2005), as well as in management science, operations research and organizational behavior (Papageorgiou *et al.*, 2005). Furthermore, FCMs have been used to construct virtual worlds, and have been proposed for modeling supervisory systems (Kosko, 1997), decision-making in medical situations such as radiation therapy planning and for characterizing urinary bladder tumors (Papageorgiou *et al.*, 2005).

Although the advantages of cognitive maps inherently reside from the fact that causal associations remain as the major alternative in which understanding regarding the world is organized (Carvalho and Tomè, 1999a), in the area of ethics, and more generally, the area of social sciences and psychology, it is obvious that cognitive maps should use other kinds of relations between concepts, thus allowing for a better representation of real-world systems that involve cognition and imprecision in real-world problems. Namely, the use of fuzzy sets in their traditional rule-based form, comprising logic and inference, is particularly more adequate in representing

qualitative knowledge involved in cognitive maps, due to their linguistic nature and abilities to represent imprecision.

Rule-based fuzzy cognitive maps (RBFCMs) offer the exact alternative to FCMs with such properties, i.e. RBFCMs are essentially a combination of FCMs and standard rule-based fuzzy systems, where feedback and mechanisms to deal with causal relations are added (Carvalho and Tomè, 1999a). These systems are fuzzy directed graphs composed of fuzzy nodes (Concepts) and fuzzy links (Relations), where each component allows for a higher representation of the complex dynamics associated with real-world qualitative systems. More importantly, RBFCMs can simultaneously deal with the inability of emulating the effects of causality, a problem commonly associated with traditional fuzzy operations, and they also provide cognitive maps the needed adequacy in representing qualitative knowledge through the provision of traditional fuzzy rules and operations (Carvalho and Tomè, 1999a, 1999b).

These fuzzy mechanisms fundamentally involve fuzzy logic and fuzzy sets. Namely, fuzzy logic involves a range of truth values of real number variables ranging between 0 and 1, where as opposed to probability, fuzzy logic uses degrees of truth as a mathematical model of vagueness, rather than as a mathematical model of ignorance. Depending on the modeled scenario, the notion of "vagueness" is determined by fuzzy sets and membership functions. For instance, a fuzzy set could be used to characterize the notion of "fullness" and "emptiness" of a glass, where in one scenario, fullness could be characterized as anything above 0.5 (50 ml), or in another scenario, it could be characterized as anything above 0.3 (30 ml). In each scenario, a membership function would formally indicate the subsets in which an element of the fuzzy set belonged, i.e. it would quantify the grade of membership of the element, where a value of 0 would mean the element is not a member of the fuzzy set and the value of 1 meaning a full member of the fuzzy set, while any values in between 0 and 1 would mean that the element is a partial member (Zadeh, 1965).

As encapsulated by FCMs, these fuzzy mechanisms can also be extended toward characterizing concepts and their relations. More specifically, the concepts used in common FCM applications represent the actors, entities and social, political, economic or abstract concepts that compose the system being modeled. Within the context of AMAs in medical ethics, examples of concepts would be Autonomy, Justice, Beneficence, the Welfare of a population or the Safety of an individual, where each concept would be designed to be a fuzzy variable, defined by its individual set of membership function. In contexts involving FCMs, membership functions of concepts represent the variation of the concept, where linguistic operators (*hedges*) are applied to modify the meaning of the set; such hedges include *SomewhatNegative, DecreaseSlightly* or a range from *DecreaseVeryMuch* to *IncreaseVeryMuch,* etc. (Carvalho and Tomè, 1999b).

- With three given concepts (A, B, C), if two were to cause one to increase by a "little", then the affected concept will increase by "more than a little".
- If two of the concepts have the exact opposite effect on the third concept, then the third concept will not change.
- If the first concept affects the third concept by "a little" and the second concept affected the third concept by "much", then the third concept will increase "more than much".
- The effect when both decrease is similar.

In RBFCMs, relations are defined by the use of different kinds of fuzzy "IF […] THEN" rule bases. Two different kinds of "causal" relations exist, the first is a *causal relation* and the second is an *influence relation* (Carvalho and Tomè, 2000). Causal relations are usually considered to involve a relative variation that never imposes an absolute value; that is, it is the change in one concept that causes the effect, which in turn is a change in the other concept, thus representing an accumulative causal effect (Carvalho and Tomè, 2000).

By contrast, influence relations represent situations where a change in one concept reinforces the change in another concept, or imposes an absolute value on that concept (Carvalho and Tomè, 2000). That is, if the first concept causes the third concept to increase a "little", and the second concept causes the third concept to increase by "much", then the third concept would increase somewhere between "little" and "much", whereas opposed effects tend to nullify each other (as in a causal relation).

It is commonly considered that causal associations are to describe our understanding of the world (Carvalho and Tomè, 2000). However, in ethical situations, principles are not accumulative, but rather imposing and conflicting. Due to this significant distinction, influence relations in the form of IF […] THEN rules are used to express the relations between multiple concepts constituting the cognitive map.

## 2. Design principles

Ethics is usually composed of two components, ethical reasoning and ethical decision-making. Ethical reasoning refers to the rights and wrongs of human conduct. That is, each person is afforded a standard that defines their personal values, which come into play when the person faces certain dilemmas or decisions, and ultimately determine their ethical framework. Conversely, ethical decision-making refers to the process of evaluating and selecting actions among alternatives in a manner consistent with ethical principles, where unethical options are eliminated and the best ethical alternative is selected (Anderson and Anderson, 2007).

This decomposition of ethics as two key components naturally identifies and simplifies the challenges that machine ethics needs to solve. Namely, ethical reasoning can be delineated as a task of knowledge acquisition, whereas ethical decision-making can be delineated as a task of action selection. More specifically, it is proposed that knowledge acquisition provides the ability to identify situational information and ethical components commonly associated with ethics, and therefore leads to the automated development of "personal values" as per ethical reasoning. Conversely, action selection as determined by the acquired knowledge leads to the elimination of unethical options and selection of best alternatives, as per ethical decision-making.

Figure 1 illustrates the proposed design guidelines used in constructing the AMA presented here. That is, in our approach, knowledge acquisition and ethical decision-making were separated into their respective parts. Reasoning involved the application of LSA in the identification of situational information pertinent to the case study, where the involved ethical principles analogous to the concepts in the FCM are identified and weighted. Once identification and weighting of the ethical principles is completed, their variable strengths are represented by real numbers, expressing the variable weights between the concepts in the FCM. Application of LSA determined the weights of the FCM, where membership functions were used to represent the concept relations, weight sand importance, and ultimately provided concepts with the ability to activate each other. The resulting FCM

included conventional design elements, where the contribution of a concept connection to another concept is the product of the activity on the line and the value of the connection strength, where the total input to a concept is the sum of all the individual products, and finally, where connections vary from positive to negative, with positive connections adding to a concept's total activity, while negative connections subtracted from the concept's total activity.

The ethical components are also integrated and inspired by a combinational ethical theory, which is used to improve simulation of complexity of ethical decision-making, rather than a single absolute duty theory. The theory incorporates the favorable aspects of the teleological and deontological approaches to ethics, while the addition of fuzzy rules allows for needed exceptions to adopting one or the other approach exclusively.

There are two well-known ethical frameworks used in biomedicine:

(1) Ross's theory, dealing with general ethical dilemmas, that has seven duties (Ross, 1930).

(2) Beauchamp and Childress's four principles of biomedical ethics that are specifically intended to cover ethical dilemmas in the field of biomedicine (Beauchamp and Childress, 2001).

Due to the higher rate of consensus within biomedical ethics, than in other areas, and given that it has fewer duties, the ethical agent in this paper is designed to compute ethics utilizing Beauchamp and Childress's principles of biomedical ethics.

Beauchamp and Childress's principles of biomedical ethics include the principle of *respect for autonomy* that states that the healthcare professional should not interfere with the effective exercise of patient autonomy, where the patient has the right to refuse or choose their treatment. Briefly, for a decision by a patient concerning his or her care to be considered fully autonomous, it must be intentional, based on sufficient understanding of his or her medical situation and the likely consequences of forgoing treatment, sufficiently free of external and internal constraints. The principle of *non-maleficence* requires that the healthcare professional not harm the patient, derived from "primum non nocere" "first, do no harm", while the principle of *beneficence* states that the healthcare professional should promote patient welfare, where a practitioner should always act in the best interest of the patient. Finally, the principle of *justice* states that healthcare services and burdens should be distributed in a fair and equal fashion. The two additional principles that are sometimes discussed include *veracity* and *fidelity*, where veracity is concerned with the concept of informed consent and the importance of honesty. Fidelity is concerned with confidentiality, a principle commonly applied to conversations between doctors and patients.

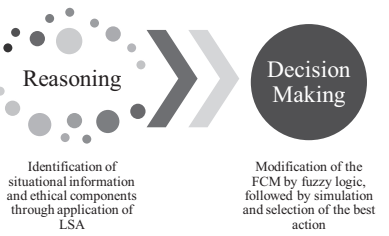


**Figure 1.**
Basic two components of the artificial moral agent

The domain selected for the agent's application is medical ethics, as the ethical case study selected from the area is consistent with the choice of prima facie duties previously selected, and, in total, it was representative of five of the six principles of biomedical ethics, which include respect for autonomy, non-maleficence, beneficence, justice and veracity.

The chosen case study primarily represented beneficence. It involved a patient with a serious liver problem, diagnosed as a primary tumor of the liver. In contrast to what the patient wanted, two physicians were involved in deciding that the patient should not be listed for a liver transplant. The decision for denying transplantation was based on the Hippocratic oath; more specifically, it was based on the belief of what might benefit the patient. In the case study, it was elucidated that any liver transplantation of the patient, due to his cancer, would only provide a 5-15 per cent chance of producing a 1-year-cancer-free survival; that is, an 85-95 per cent chance that the transplant would not overcome the cancer. Furthermore, it was mentioned that transplant combined with immune-suppression therapy, as per standard protocol, would produce serious, burdensome side effects. However, the patient believed that the 5-15 per cent chance of success made any burdens worth it. Thus, the dilemma hinged upon the patient's differing evaluation of risk and benefit versus the physicians' moral roles and obligations toward the patient and possibly other patients.

Although the case study primarily represented beneficence, it also incorporated other principles in the format of key phrases. The identification of such phrases and their numerical evaluation via LSA partly allowed for automatic information acquirement and incorporation. For instance, throughout the case study, the ineffective status of a liver transplant was mentioned in various contexts, with each context signifying different ethical principles; overall, the low chance of survival after a liver transplant involved the ethical principle of justice – the obligation of the physicians' to their other patients, who possibly had a higher chance of survival after transplantation; non-maleficence – the obligation to cause no harm, either through the denying of the transplantation or the subjection of the patient to unnecessary side effects; and veracity – the obligation of honesty, possibly regarding the chance of cancer-free survival after transplantation.

### *2.1 Application of latent semantic analysis*

Take the following example of ten book titles searched from amazon.com regarding the search term "decision making". In this example, the index words are underlined and signify words that appear in two or more search results:

- *SmartChoices*: A Practical *Guide* to *Making Better Decisions*.
- Decisive: How to *Make Better Choices* in Life and Work.
- The *Decision* Book: 50 *Models* for Strategic *Thinking*.
- *Thinking*: The New Science of *Decision-Making*, *Problem-Solving*, and Prediction.
- Harvard Business Review on *Making Smart Decisions*.
- *Make* Up Your Mind: A *Decision Making Guide* to *Thinking* Clearly and Choosing Wisely.
- The Psychology of Judgment and *Decision Making*.
- *Thinking*, Fast and Slow.

- Creative *Problem Solving* for Managers: Developing Skills for *Decision Making* and Innovation.
- Human-Centric *Decision-Making Models* for Social Sciences.

More specifically, in this matrix, each index word is a row and each title is a column. Each cell contains the number of times that word occurs in that title. For instance, the word "smart" appears a single time in title T1 and a single time in title T2, whereas "decisions" or "making" appear in almost all titles. As with the artificial agent, the matrices built during LSA are usually very large, but also very sparse (most cells contain 0). This is due to the fact that documents usually contain only a small number of all the possible words identified. During the more sophisticated implementations of LSA, this sparseness is used to save resources both in terms of time and memory (Table I).

In the artificial agent, the raw matrix counts were modified so that rare words were weighted more heavily than common words. That is, a word that only occurred in only 5 per cent of the documents was weighted more heavily than a word that occurred in 90 per cent of the documents. This is a common method applied in most information retrieval systems, with the most popular weighting technique represented as TFIDF (Term Frequency – Inverse Document Frequency). Under this method, the count in each cell is replaced by the following formula (Aizawa, 2003):

$$TFIDF_{i,j} = \left(\frac{N_{i,j}}{N_{*j}}\right) \times \log\left(\frac{D}{D_i}\right),$$

where

$N_{i,j}$ = the number of times word $i$ appears in document $j$;
$N_{*j}$ = the number of total words in document $j$;
$D$ = the number of documents (the number of columns); and
$D_i$ = the number of documents in which word $i$ appears (the number of non-zero columns in row $i$).

| | Title | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Index words | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 | T10 |
| Smart | 1 | | | | 1 | | | | | |
| Choices | 1 | 1 | | | | | | | | |
| Making | 1 | | | 1 | 1 | 1 | 1 | | 1 | 1 |
| Better | 1 | 1 | | | | | | | | |
| Decisions | 1 | | 1 | 1 | 1 | 1 | 1 | | 1 | 1 |
| Guide | 1 | | | | | 1 | | | | |
| Make | | 1 | | | | 1 | | | | |
| Models | | | 1 | | | | | | | 1 |
| Thinking | | 1 | 1 | | | 1 | | 1 | | |
| Problem | | | | 1 | | | | | 1 | |
| Solving | | | | 1 | | | | | 1 | |

**Table I.**
Sparse document matrix for the search "decision making"

**Note:** 0 has been removed to reduce clutter

In this formula, words that concentrate in certain documents are emphasized (by the $N_{i,j}/N_{*,j}$ ratio), and words that only appear in a few documents are emphasized (by the $\log(D/D_i)$ term (Aizawa, 2003). The log function normalizes the internal values and allows for term frequency filtering (Aizawa, 2003). Essentially, as a term appears in more documents, the ratio inside the logarithm approaches 1, bringing the *TFIDF* closer to 0; this is because of the inherent inverse relationship between the term frequency and document frequency of that term (Aizawa, 2003).

Once the artificial agent constructed the matrix, the powerful technique of SVD is used to analyze the matrix. The reason why SVD was implemented into the system is due to its ability to find a reduced dimensional representation of the matrix, thus emphasizing the strongest relationships, while discarding the noise (Klema and Laub, 1980). Simply, SVD is used because it could find the best possible reconstruction of the matrix with the least possible information. However, the SVD can identify the number of dimensions needed for approximating the matrix. Too few dimensions and important patterns are left out, too many and noise caused by random word choices will confound the calculations.

For large collections of documents, the number of dimensions used is in the range of 100 to 500. In the artificial agent, 400 dimensions are used, due to Wikipedia's use in training the LSA algorithm. Once the abovementioned analyses are completed for both the documents describing the ethical principles and the documents containing the ethical case studies, the reduced matrices are compared via cosine similarity. Cosine similarity is a simple measure of similarity between two sparse vectors, such as those generated via LSA (Dehak *et al.*, 2010). It is a judgment of orientation and not magnitude, where two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0 and two vectors completely opposed have a similarity of −1 (Dehak *et al.*, 2010):

$$Similarity = \cos\theta = \frac{A \cdot B}{\|A\|\|B\|}$$

The dot product and magnitude calculation of the two vectors $A$ and $B$ represent the cosine similarity of the two vectors (Dehak *et al.*, 2010). This calculation is used to create vectors that represent documents in an n-dimensional term space, where the relevancy rankings between the documents are determined by measuring the angle between the vectors. The results are interpreted simply, where two documents with smaller angles in relation with each other are interpreted as having higher similarity values. Figure 2 represents such an example, where the query document (doc q) and the second document (doc2) are considered to have a higher similarity.

### *2.2 Application of rule-based fuzzy cognitive maps*

Implementing fuzzy logic and RBFCMs in the artificial agent as a decision engine helped to ensure that correct actions are chosen when the different duties give conflicting advice. By using RBFCMs, the artificial agent could inadvertently abstract the relationships between the prima facie duties from particular ethical dilemmas where there is an agreed-upon correct action.

A primary task of the decision engine is to utilize the similarity results generated by the LSA component, which compares and analyzes the documents describing their

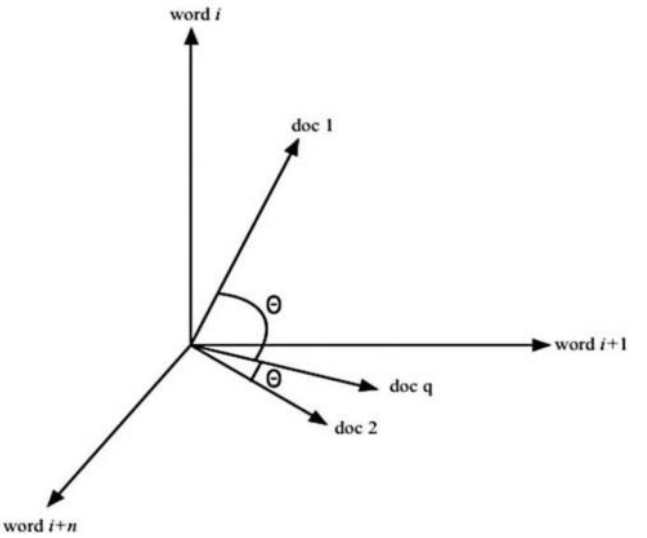


**Figure 2.**
Example of a vector space similarity graph

corresponding ethical principles and the ethical case studies. For the decision engine to utilize the LSA results, fuzzy rules expressing the similarity scores are used to produce fuzzy values, which are then used to alter the causal values expressing the strength of conflict or relation between the concepts. The decision engine is designed to always use the results from the LSA component to compute possible scenarios and select the best outcome.

In this context, the RBFCM model is to identify each kind of LSA similarity score generated for the different concepts, as different types of simulation outcomes that affect the machine's behavior. Each of these simulation outcomes is triggered by the LSA similarity results. The outcomes are activated with the help of a fuzzy rule base where fuzzy rules are used to map multiple inputs to outputs. As described earlier, the concepts are represented as fuzzy sets, each with their corresponding fuzzy membership function, and the edges were represented as fuzzy relations, each expressed by one or more fuzzy rules.

The following simple example illustrates how an RBFCM can be used to incorporate LSA results in a decision process. Consider the following simplified scenario of the first case study: *the patient wishes to enact their right of autonomy, and choose the treatment they want. The treatment could be highly beneficial; however, the health resources are scarce, and someone else could benefit just as likely.* Figure 3 displays an FCM that captures such a scenario.

The activation of the concept *Treatment* by the other concepts can be implemented using a fuzzy rule base. The fuzzy rules are used to map the multiple input concepts (the causes) to the output concept (the effect). Multiple fuzzy rules may be used to correspond to the knowledge described in an individual FCM. The FCM in Figure 1 can be implemented with one of various fuzzy rules.

IF *the similarity score* of *autonomy* is *high* AND *the similarity score* of *beneficence* is *high* AND *the similarity score* of *justice* is *medium,* THEN *treatment* is *recommended highly.*

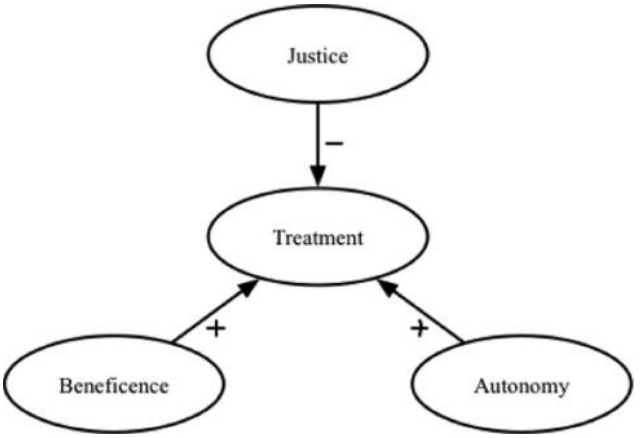


**Figure 3.**
Example FCM for justice vs beneficence

Different combinations of fuzzy rules were reused to capture the dynamic results produced by LSA. For map simulations, a threshold function is used to force the concept values to be monotonically mapped into a normalized range. The widely used logistic function $1/(1 + e - 1x)$ is used as the threshold function. Other factors such as common FCM variables are utilized in manipulating system behavior. The common transmitter, receiver and ordinary variables are used to show how the concepts acted in relation to each other. These variables are defined by their out-degree $[od(vi)]$ and in-degree $[id(vi)]$. Out-degree is the row sum of the concept values in the adjacency matrix. It showed the cumulative strength of connections $(aij)$ exiting the concept, where $N$ was the total number of variables (Mouratiadou and Moran, 2007):

$$[od(vi)] = \sum_{k=1}^{N} (a_{ij})$$

In-degree is the column sum of the concept values. It showed the cumulative strength of variables entering the concept (Mouratiadou and Moran, 2007):

$$[id(vi)] = \sum_{k=1}^{N} (a_{ji})$$

Conventionally, transmitter variables have a positive out-degree, and zero in-degree. Receiver variables have a positive in-degree and zero out-degree (Mouratiadou and Moran, 2007). The centrality of a variable is the summation of its in-degree and out-degree:

$$ci = od(v_i) + id(v_i)$$

The contribution of a variable in the cognitive map was understood by calculating its centrality, which shows how connected the concept was to other concepts and what the cumulative strength of these connections was. The total number of receiver variables in

the map is considered as an index of the map's complexity. A high number of receiver variables is assumed to indicate a cognitive map's consideration of many outcomes and implications. A large number of transmitter variables are assumed to indicate a formal hierarchical structure, where causal arguments are not well elaborated. These distinctions allow for cognitive map analysis and characterization, where the ratio between the number of receivers and transmitters (*R/T*) determines the map's complexity. In this study, it is assumed that complex maps have larger ratios, as they define more utility outcomes and include less controlling forcing functions.

## 3. Results

The LSA similarity results between the ethical principles and case study considered are intuitively correct for some concepts; however, there are some concepts that are in conflict with those humanly assessed (Table II).

Justice and beneficence were considered to be the major conflicting results in terms of a comparison between the expected results and LSA results. The ethical dilemma in the analyzed case study hinged upon the complex interaction between beneficence and justice, where livers for transplantation were considered to be scarce health resources. Nonetheless, a cognitive map representing the LSA results was constructed and analyzed. The first step in analyzing the cognitive map was to describe and tabulate the number of variables, connections and the graph theory structural indices. An application of this analysis was done for the adjacency matrix representing the FCM (Table III).

In the constructed map, there were no transmitter or receiver variables (Table IV). All concepts had a relatively high out-degree and in-degree, with veracity having the lowest, containing an out-degree and in-degree of 0.60. All concepts were considered ordinary variables. The most central concept in this map was autonomy, followed by patient health and then treatment.

In addition to examining the structure of cognitive maps through graph theory indices, their indicative variables were examined in two different ways.

The variables were tabulated by their centrality, providing insights into the components' level of input they received, their in-degree and also how much output they provided, their out-degree (Figure 4). For instance, autonomy's centrality of 9.60 was not only a frequency of expression but also how important that given concept was, given the whole structure of the cognitive map. By looking at the map's level of in-degree and out-degree, it was evident that the concepts were mainly influencing each other, with the exception of veracity. Second, the concepts were separated according to their variable types, where they were categorized as transmitter, ordinary or receiver variables (Figure 5). The type of variables revealed the base structure of the system. For instance, if a concept was considered a transmitter variable, then

**Table II.**
LSA similarity results between the ethical principles and case study

| Ethical principles | LSA results | Expected results |
|---|---|---|
| Justice | 0.56 | ~0.80 |
| Veracity | 0.73 | ~0.70 |
| Autonomy | 0.60 | ~0.65 |
| Beneficence | 0.67 | ~0.80 |
| Non-maleficence | 0.73 | ~0.75 |

**Table III.**
Adjacency matrix coded from fuzzy cognitive map

| Liver case study | Autonomy | Beneficence | Justice | Veracity | Avoid killing | Non-maleficence | Patient health | Treatment |
|---|---|---|---|---|---|---|---|---|
| Autonomy | 0.00 | 0.50 | −0.50 | 0.60 | 0.80 | 1.00 | 0.70 | 0.70 |
| Beneficence | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 0.60 |
| Fidelity | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Justice | −0.50 | −1.00 | 0.00 | 0.00 | −0.80 | −0.20 | 0.00 | −1.00 |
| Veracity | 0.60 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Avoid killing | 0.80 | 0.00 | −0.80 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| Non-maleficence | 1.00 | 1.00 | −0.20 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 |
| Patient health | 0.70 | 1.00 | 0.00 | 0.00 | 1.00 | 1.00 | 0.00 | 0.60 |
| Treatment | 0.70 | 1.00 | −1.00 | 0.00 | 1.00 | 0.00 | 0.60 | 0.00 |

**Table IV.**
Graph theory indices for the adjacency matrix (Table III)

| Concepts | 1. Autonomy | 2. Beneficence | 3. Justice | 4. Veracity | 5. Avoid killing | 6. Non-maleficence | 7. Patient health | 8. Treatment |
|---|---|---|---|---|---|---|---|---|
| Out-degree | 4.80 | 3.10 | 3.50 | 0.60 | 3.60 | 3.20 | 4.30 | 4.30 |
| In-degree | 4.80 | 4.50 | 2.50 | 0.60 | 3.60 | 3.20 | 4.30 | 3.90 |
| Centrality | 9.60 | 7.60 | 6.00 | 1.20 | 7.20 | 6.40 | 8.60 | 8.20 |
| Transmitter | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Receiver | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Ordinary | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| No connection | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

it could be seen as a forcing function, which cannot be controlled by any other variable. In this case, all concepts were considered to be ordinary variables, having the properties of both transmitter and receiver variables, suggesting that the map was interconnected, with all concepts having contributory influence on all other concepts.

The steady-state calculation offers an idea of the ranking of the variables in relationship to each other, and according to the system's base structure. The necessary calculations for this method were made using the auto-associative neural network method (Reimann, 1998), in which the multiplication of the adjacency matrix with a vector of initial states was done (Table III).

Typically, a logistic function $1/(1 + e - 1 \times x)$ was used to transform the results into the interval [0, 1]. This non-negative transformation allowed for a better understanding and representation of activation levels of variables. It also enabled a qualitative comparison among the causal output of concepts. The resulting transformed vector was then repeatedly multiplied by the adjacency matrix and transformed until the system converged to a fixed point. A restriction of 20 simulation time steps was used. Most of the models that were run ended in a stable state; however, theoretically, they could have also settled into a limit cycle, or chaotic attractor (Dickerson and Kosko, 1993). For example, Table V shows that in the steady state, justice was significantly lower than autonomy, beneficence or patient health, suggesting that justice was severely limiting the patient's potential treatment, and thus in the event of a decision, should be ignored. Such a run also makes two characteristics of cognitive maps apparent: first, are the emergent properties of cognitive maps, and second, the synergistic interaction among different concepts.

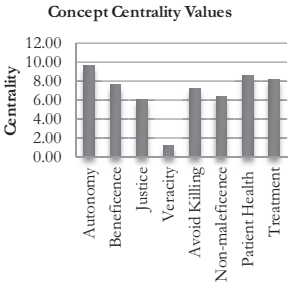


**Figure 4.**
Map centrality of the adjacency matrix (Table III)

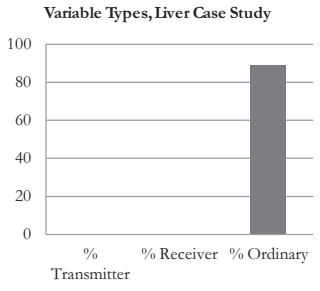


**Figure 5.**
Variable types for the adjacency matrix (Table III)

**Table V.**
Calculation of steady state for the adjacency matrix Table III

| FCM simulation | Autonomy | Beneficence | Justice | Veracity | Avoid killing | Non-maleficence | Patient health | Treatment |
|---|---|---|---|---|---|---|---|---|
| | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| | 0.978119 | 0.924142 | 0.075858 | 0.645656 | 0.880797 | 0.942676 | 0.986613 | 0.869892 |
| | 0.977211 | 0.961298 | 0.095166 | 0.642647 | 0.929453 | 0.946531 | 0.981187 | 0.933148 |
| | 0.979139 | 0.962827 | 0.086699 | 0.642522 | 0.932136 | 0.947879 | 0.983394 | 0.936065 |
| | 0.979383 | 0.963393 | 0.086202 | 0.642787 | 0.932981 | 0.948241 | 0.983534 | 0.936942 |
| | 0.979432 | 0.963463 | 0.086064 | 0.642821 | 0.933082 | 0.948293 | 0.983574 | 0.937056 |
| | 0.979439 | 0.963476 | 0.086046 | 0.642828 | 0.933101 | 0.948302 | 0.983580 | 0.937076 |
| Initial state vector | 0.979440 | 0.963478 | 0.086043 | 0.642829 | 0.933104 | 0.948304 | 0.983581 | 0.937079 |
| | 0.979441 | 0.963478 | 0.086042 | 0.642829 | 0.933104 | 0.948304 | 0.983581 | 0.937080 |
| | 0.979441 | 0.963478 | 0.086042 | 0.642829 | 0.933104 | 0.948304 | 0.983581 | 0.937080 |
| Steady state | 0.979441 | 0.963478 | 0.086042 | 0.642829 | 0.933104 | 0.948304 | 0.983581 | 0.937080 |

## 4. Discussion

Two considerable challenges of machine ethics were explored in this paper. First was the issue of the acquirement and incorporation of the necessary information needed to make an ethical decision, and second was the issue of the actual ethical reasoning and decision-making process. LSA and the use of corpus-based methods to calculate semantic similarity between words/concepts were explored as a possible solution to information acquisition and incorporation, whereas an RBFCM was explored as a possible solution for ethical decision-making processes founded on its traditional fuzzy feedback architecture. More specifically, the use of RBFCMs over conventional FCMs was motivated by the fact that unlike FCMs, RBFCMs are considered true cognitive maps, as they are not limited to the representation of causal relations and are capable of overcoming feedback and complex relations due to their fuzzy rule-based mechanisms (Carvalho and Tomè, 1999b).

Conversely, the use of natural language processing (NLP) was motivated by the observation that digital databases, communication networks and huge repositories of textual data have become increasingly available to the wider public. Namely, as search engines and online encyclopedias, such as Wikipedia, become increasingly important conduits to relevant information, it follows naturally that their services should be leveraged in developing AMAs capable of retrieving their own information and also capable of interacting more naturally with humans through the recognition of ambivalence, impreciseness or ambiguity of user requests, and the detection in differences between user actions and user intentions (Hofmann, 1999). As shown in this paper, for prototypical examples of such AMAs to be possible, they need only mimic the existing search engines or use online encyclopedias in their efforts.

The use of LSA over other NLP methods was motivated by the typical scenario involving natural language queries in request formulation by users for information retrieval. Namely, most existing retrieval methods for natural language queries associated with document search utilize simple word-matching strategies in determining the rank relevance of a document associated with a query (Hofmann, 1999). However, it is well known that literal term-matching suffers from severe drawbacks, mainly due to the ambivalence of words, their inherent lack of precision (Hofmann, 1999) and the complexities introduced by the two classical problems of synonymy and polysemy. LSA overcomes these problems by representing documents or queries, not by terms, as conventional literal term-matching or vector-based methods, but by the latent concepts referred to by the terms (Dumais, 2004). In this paper, this rationale behind LSA was used to compare document semantic similarities between a self-generated corpus (document collection) and the ethical case study to be solved, or between the ethical case study and a Wikipedia corpus, where the similarity scores generated were then used to determine weight values utilized in the decision-making phase. It is this addition of an information retrieval component that provided the constructed AMA several advantages over previous AMAs, such as those presented by Anderson *et al.* (2006), where the AMA could not retrieve its own information or develop some "understanding" of the nuances involved in ethical reasoning.

However, it should be noted that although LSA provides an NLP mechanism that performs at high approximate levels to humans, there are still potential limitations to the approach presented in this paper. Namely, the training by text alone as used in this paper did not include the necessary language exposures associated with real-world situations and thus did not instill the necessary commonsense reasoning in the AMA. Foltz *et al.* (1999) have estimated the impact of the absence of such training, also including the absence of direct

instruction by parents and teachers, and the association of language with perception and action; namely, they have shown that LSA might be approximately 10 per cent inferior to humans, lacking some of the more important aspects of human language, such as the generation of meaningful statements, the understanding of linguistically posed propositions and the appreciation of metaphors and analogies (Foltz *et al.*, 1999). Therefore, it should be noted that the absence of such factors might introduce significant limitations or confounding factors for constructed AMAs using the similar approaches as presented here.

More specifically, our AMA might have been incapable of identifying and inferring key information associated with the ethical case study. For instance, the ethical dilemma chosen for analysis hinged upon a delicate interaction between the ethical principles of beneficence and justice, where beneficence was considered important, as the patient wished for a liver transplant, which could have been extremely beneficial to their health, and where justice involved the scarcity of livers, with the patient's corresponding physicians regarding transplantation as a waste of scarce medical resources, due to the patient's advanced cancer stage. In such scenarios, the action of denying a liver transplantation would be considered the most appropriate action, with the decision founded upon the assessment of the predictive and present value of the liver as a resource for a more appropriate patient, or as a treatment option for the current patient, and the chances of success in both situations.

All the necessary information required to make such a decision was presented in the case study; however, the AMA was incapable of assessing the situation deeply, as the necessary concepts and tools were not available to it. It is considered that the use of reinforcement learning, deep neural networks and lexical-based methods for semantic similarity, in combination with statistical-based methods, could have provided the necessary information presented in the case study. It is also suggested that the use of WordNet or other similar lexical databases and the use of path-length similarity could have provided further assistance in the identification of key terms, thus providing the AMA with better contextual information, leading to better concept weighting and ultimately the selection of the most appropriate decision for the ethical dilemma.

Under James Moor's categorization, the AMA constructed in this paper would still be considered an implicit ethical agent (Moor, 2006). First, the constructed agent, when analyzing the ethical case study could not take everyone's interest into account. Second, the agent had no sense of self-interest and the agent could not make decisions beyond the way it was programmed, and finally, the agent could not derive all the real-world commonsense reasoning that is required for ethical decision-making. Nonetheless, after the cognitive map was drawn and the adjacency matrix coded, simulations were conducted to determine the system's steady state, and, ultimately, its ethical decision. An ethically acceptable decision was derived from the steady state of the adjacency matrix, where justice was significantly lower than autonomy, beneficence or patient health, suggesting that justice was a severe limitation in the patient's potential treatment, and thus in the event of a decision, was ignored in favor for patient treatment.

Further assessment of the system's behavior can be conducted, as suggested by Allen *et al.* (2000) in their paper, by applying a variant of Alan Turing's test. Their proposed "comparative moral Turing test" (cMTT) circumvents disagreement concerning definitions of ethical behavior as well as the requirement that a machine has the ability to articulate its decisions, where an evaluator assesses the comparative morality of pairs of descriptions of morally significant behavior, and describes the actions of a human being in an ethical dilemma and, the other, the actions of a machine faced with the same dilemma. If the machine

is not identified as the less moral member of the pair significantly more often than the human, then it has passed the test. However, it is noted that human behavior is typically far from being morally ideal, and a machine that passes the cMTT might still be below the high ethical standards to which one would probably desire a machine to be held. This legitimate concern suggests that instead of comparing the machine's behavior in a particular dilemma against typical human behavior, future comparisons should be made with behavior recommended by a trained ethicist faced with the same dilemma. It is also believed that the principles used to justify the decisions that are reached by both the machine and ethicist should be made transparent and compared. A cMTT was not conducted in this research; however, some assessment was conducted on the AMA. The decision principles that were instilled into the cognitive map are supported by W.D. Ross's evaluation of ethical behavior. Furthermore, the fact that the AMA could derive the appropriate decision of ignoring justice, in this case, offers preliminary support for the hypothesis that ethical principles can be incorporated into an artificial agent's decision procedure, thus enabling that machine to determine the ethically acceptable action.

## 5. Conclusion

This paper possibly provides a contribution and a theoretical foundation for machine ethics through exploration of the rationale and the feasibility of adding an ethical dimension to machines. Further, the constructed AMA illustrates the possibility of utilizing an action-based ethical theory that provides guidance in ethical decision-making according to the precepts of its respective duties. The use of LSA illustrates their powerful capabilities in understanding text and their potential application as information retrieval systems in AMAs. The use of cognitive maps provides an approach and a decision procedure for resolving conflicts between different duties. In their application in ethics and AMAs, the advantages of using cognitive maps outweigh their disadvantages. In particular, it was found that when cognitive maps were created with a standard methodology, the structural indices of the maps were close to each other. Suggesting that cognitive maps could be used in AMAs as tools for meta-analysis, where comparisons regarding multiple ethical principles and duties can be examined and considered. With cognitive mapping, complex and abstract variables that cannot easily be measured but are important to decision-making can be modeled. This approach is especially appropriate for data-poor and uncertain situations common in ethics. To ensure that a machine with an ethical component can function autonomously in the world, research in AI will need to further investigate the representation and determination of ethical principles, the incorporation of these ethical principles into a system's decision procedure, ethical decision-making with incomplete and uncertain knowledge, the explanation for decisions made using ethical principles and the evaluation of systems that act based upon ethical principles.

## References

Aizawa, A. (2003), "An information-theoretic perspective of tf–idf measures", *Information Processing & Management*, Vol. 39 No. 1, pp. 45-65.

Allen, C., Varner, G. and Zinser, J. (2000), "Prolegomena to any future artificial moral agent", *Journal of Experimental and Theoretical Artificial Intelligence*, Vol. 12 No. 2000, pp. 251-261.

Anderson, M. and Anderson, S.L. (2007), "Machine ethics: creating an ethical intelligent agent", *AI Magazine*, Vol. 28 No. 4, p. 15.

Anderson, M., Anderson, S.L. and Armen, C. (2004), "Towards machine ethics", *Proceedings of AAAI, San Jose, CA*.

Anderson, M., Anderson, S.L. and Armen, C. (2006), "An approach to computing ethics", *Intelligent Systems, IEEE*, Vol. 21 No. 4, pp. 56-63.

Arkin, R.C. (2008), "Governing lethal behavior: embedding ethics in a hybrid deliberative/reactive robot architecture part I: motivation and philosophy", *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI), Amsterdam*, IEEE, pp. 121-128.

Beauchamp, T.L. and Childress, J.F. (2001), *Principles of Biomedical Ethics*, Oxford University Press, Oxford.

Bostrum, N. (2014), *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Oxford.

Carvalho, J.P. and Tomè, J.A. (1999a), "Rule based fuzzy cognitive maps-fuzzy causal relations", in Mohammadian, M. (Ed.), *Computational Intelligence for Modelling, Control and Automation, 18th International Conference of the North American Fuzzy Information Processing Society (NAFIPS), IEEE, New York, NY*.

Carvalho, J.P. and Tomè, J.A. (1999b), "Rule based fuzzy cognitive maps and fuzzy cognitive maps-a comparative study", *18th International Conference of the North American Fuzzy Information Processing Society, 1999 NAFIPS, New York, NY*, IEEE, pp. 115-119.

Carvalho, J.P. and Tomè, J.A. (2000), "Rule based fuzzy cognitive maps-qualitative systems dynamics", *19th International Conference of the North American Fuzzy Information Processing Society, 2000 NAFIPS, Atlanta, GA*, IEEE, pp. 407-411.

Chen, C.H., Weng, Y.H. and Sun, C.T. (2009), "Toward the human-robot co-existence society: on safety intelligence for next generation robots", *Social Robotics*, Vol. 1 No. 4, p. 267.

Dehak, N., Dehak, R., Glass, J., Reynolds, D. and Kenny, P. (2010), "Cosine similarity scoring without score normalization techniques", *Proceedings of Odyssey Speaker and Language Recognition Workshop, Brno*, pp. 71-75.

Dickerson, J.A. and Kosko, B. (1993), "Virtual worlds as fuzzy cognitive maps", *Virtual Reality Annual International Symposium, Seattle, WA*, IEEE, pp. 471-477.

Dumais, S.T. (2004), "Latent semantic analysis", *Annual Review of Information Science and Technology*, Vol. 38 No. 1, pp. 188-230.

Foltz, P.W., Laham, D. and Landauer, T.K. (1999), "The intelligent essay assessor: applications to educational technology", *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, Vol. 1 No. 2.

Hofmann, T. (1999), "Probabilistic latent semantic indexing", *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY*, ACM, pp. 50-57.

Klema, V. and Laub, A. (1980), "The singular value decomposition: its computation and some applications", *IEEE Transactions on Automatic Control*, Vol. 25 No. 2, pp. 164-176.

Kosko, B. (1997), *Fuzzy Engineering*, Prentice Hall, New York, NY.

Laham, T.K.L.D. and Foltz, P. (1998), "Learning human-like knowledge by singular value decomposition: a progress report", *Advances in Neural Information Processing Systems 10: Proceedings of the 1997 Conference: [Eleventh Annual Conference on Neural Information Processing (NIPS), 1-6 December 1997, Denver, Colorado*, Vol. 10, MIT Press, p. 45.

Landauer, T.K., Foltz, P.W. and Laham, D. (1998), "An introduction to latent semantic analysis", *Discourse Processes*, Vol. 25 Nos 2/3, pp. 259-284.

McCarthy, J. and Hayes, P.J. (1969), "Some philosophical problems from the standpoint of artificial intelligence", *Machine Intelligence 4*, *Proceedings of the Fourth Annual Machine Intelligence Workshop*, *Edinburgh University Press, Edinburgh*, pp. 431-450.

Moor, J.H. (2006), "The nature, importance, and difficulty of machine ethics", *Intelligent Systems*, IEEE Vol. 21 No. 4, pp. 18-21.

Mouratiadou, I. and Moran, D. (2007), "Mapping public participation in the water framework directive: a case study of the Pinios River Basin, Greece", *Ecological Economics*, Vol. 62 No. 1, pp. 66-76.

Papageorgiou, E.I., Parsopoulos, K.E., Stylios, C.S., Groumpos, P.P. and Vrahatis, M.N. (2005), "Fuzzy cognitive maps learning using particle swarm optimization", *Journal of Intelligent Information Systems*, Vol. 25 No. 1, pp. 95-121.

Reimann, S. (1998), "On the design of artificial auto-associative neuronal networks", *Neural Networks*, Vol. 11 No. 4, pp. 611-621.

Ross, W.D. (1930), *The Right and the Good*, Clarendon Press, Oxford.

Shulman, C., Jonsson, H. and Tarleton, N. (2009), "Machine ethics and super intelligence", *Reynolds and Cassinelli*, *The Fifth Asia-Pacific Computing and Philosophy Conference*, 1-2 October, *University of Tokyo, Japan*, pp. 95-97.

Yampolskiy, R.V. (2013), *Artificial Intelligence Safety Engineering: Why Machine Ethics is a Wrong Approach*, Springer, Berlin, Heidelberg, pp. 389-396.

Zadeh, L.A. (1965), "Fuzzy sets", *Information and Control*, Vol. 8 No. 3, pp. 338-353.

## Further reading

Eden, C., Ackermann, F. and Cropper, S. (1992), "The analysis of cause maps", *Journal of Management Studies*, Vol. 29 No. 3, pp. 309-324.

Klein, J.H. and Cooper, D.F. (1982), "Cognitive maps of decision-makers in a complex game", in Kandel, A. (Ed.), *Journal of the Operational Research Society*, Fuzzy Expert Systems, CRC Press, Boca Raton, Vol. 33 No. 1, pp. 63-71.

Kosko, B. (1987), "Adaptive inference in fuzzy knowledge networks", *Proceedings of 1st International Conference on Neural Networks*, Vol. 2 No. 1, pp. 261-268.

Kosko, B. (1988), "Hidden patterns in combined and adaptive knowledge networks", *International Journal of Approximate Reasoning*, Vol. 2 No. 4, pp. 377-393.

Kosko, B. (1992), "Fuzzy associative memory systems", in Kandel, A. (Ed.), *Fuzzy Expert Systems*, CRC Press, Boca Raton, FL, pp. 135-162.

Özesmi, U. (2001), "Bilissel (Kognitif) Haritalamaya Gore HalkinTalepleri (The wants and desires of the local population based on cognitive mapping)", Yusufeli Baraji Yeniden Yerlesim Plani (Yusufeli Damlake Resettlement Plan), Devlet Su Isleri (DSI)(State Hydraulic Works). Sahara Muhendislik, Ankara, pp. 154-169.

Taber, R. (1991), "Knowledge processing with fuzzy cognitive maps", *Expert Systems with Applications*, Vol. 2 No. 1, pp. 83-87.

Turing, A.M. (1950), "Computing machinery and intelligence", *Mind*, Vol. 59 No. 236, pp. 433-460.

**Corresponding author**
Masoud Mohammadian can be contacted at: masoudm991@gmail.com