Toward ethical intelligent autonomous healthcare agents: a case-supported principle-based behavior paradigm

Michael Anderson[1] and Susan Leigh Anderson[2]

1. Dept. of Computer Science, University of Hartford, W. Hartford, CT, USA
anderson@hartford.edu
2. Dept. of Philosophy, University of Connecticut, Stamford, CT, USA
susan.anderson@uconn.edu

Abstract. A paradigm of case-supported principle-based behavior is proposed to help ensure ethical behavior of intelligent autonomous machines. The requirements, methods, implementation, and tests of this paradigm are detailed.

1 INTRODUCTION

Systems capable of producing change in the environment require particular attention to the ethical ramifications of their behavior. *Autonomous* systems not only produce change in the environment but can also monitor this environment to determine the effects of their actions and decide which action to take next. *Self-modifying* autonomous systems add to this the ability to modify their repertoire of environment changing actions. Ethical issues concerning the behavior of such complex and dynamic systems are likely to elude simple, static solutions and exceed the grasp of their designers. We propose that behavior of intelligent autonomous systems (IAMs) should be guided by explicit ethical principles abstracted from a consensus of ethicists. We believe that in many domains where IAMs interact with human beings (arguably the most ethically important domains), such a consensus concerning how they should treat us is likely to emerge and, if a consensus cannot be reached within a domain, it would be unwise to permit such systems to function within it.

Correct ethical behavior not only involves *not* doing certain things, but also *doing* certain things to bring about ideal states of affairs. We contend that a paradigm of *case-supported principle-based behavior* (CPB) will help ensure the ethical behavior of IAMs, serving as a basis for action selection and justification, as well as management of unanticipated behavior.

We assert that ethical decision-making is, to a degree, computable [2]. Some claim that no actions can be said to be ethically correct because all value judgments are relative either to societies or individuals. We maintain however, along with most ethicists, that there is agreement on the ethically relevant features in many *particular* cases of ethical dilemmas and on the right course of action in those cases. Just as stories of disasters often overshadow positive stories in the news, so difficult ethical issues are often the subject of discussion rather than those that have been resolved, making it seem as if there is no consensus in ethics. Although, admittedly, a consensus of ethicists may not exist for a number of domains and actions, such a consensus is likely to emerge in many areas in which intelligent autonomous systems are likely to be deployed and for the actions they are likely to undertake.

We contend that some of the most basic system choices have an ethical dimension. For instance, simply choosing a fully awake state over a sleep state consumes more energy and shortens the lifespan of the system. Given this, to ensure ethical behavior, a system's possible ethically significant actions should be weighed against each other to determine which is the most ethically preferable at any given moment. It is likely that ethical action preference of a large set of actions will need to be defined intensionally in the form of rules as it will be difficult or impossible to define extensionally as an exhaustive list of instances. Since it is only dependent upon a likely smaller set of ethically relevant features that actions entail, action preference can be more succinctly stated in terms of satisfaction or violation of duties to either minimize or maximize (as appropriate) each such feature. We refer to intensionally defined action preference as a *principle* [5].

A principle can be used to define a binary relation over a set of actions that partitions it into subsets ordered by ethical preference with actions within the same partition having equal preference. This relation can be used to order a list of possible actions and find the most ethically preferable action(s) of that list. This is the basis of CPB*:* a system decides its next action by using its principle to determine the most ethically preferable one(s). As principles are explicitly represented in CPB, they have the further benefit of helping justify a system's actions as they can provide pointed, logical explanations as to why one action was chosen over another. Further, as these principles are discovered from cases, these cases can be used to verify system behavior and provide a trace to its origin.

CPB requirements include a formal foundation in ethical theory, a representation scheme, a defined set of ethically significant actions, and a number of particular cases of ethical dilemmas with an agreed upon resolution. A method of discovery, as well as methods to determine representation details and transcribe cases into this representation, is helpful for facilitating the abstraction of principles from cases. Implementation of the paradigm requires means to determine dynamically the value of ethically relevant features of actions as well as to partition a set of ethically significant actions by ethical preference and to select the most ethically preferable. Finally, means to validate discovered principles and support and verify selected actions are needed. These aspects of CPB are detailed in the following.

2 REQUIREMENTS

An ethical theory, or at least an approach to ethical decision-making, is needed to provide a formal foundation for the system. Single absolute duty theories that have been proposed that are either teleological or deontological, such as Utilitarianism or Kant's Categorical Imperative, have been shown to have exceptions, not fully capturing the complexities of ethical decision-making. The *prima facie duty* approach to ethics [15] is ideal for combining multiple ethical obligations, both teleological and deontological, and can be adapted to many different domains. A *prima facie* duty is a duty that is binding unless it is overridden or trumped by another duty or duties. There are a number of such duties that must be weighed in ethical dilemmas, often giving rise to conflicts, necessitating the need for an ethical principle to resolve the conflicts. Although defenders of this approach have not given such decision principles, they have maintained that in particular cases it is intuitively obvious which duty/duties should prevail. We have devised a procedure for inferring an ethical decision principle from information about cases of ethical dilemmas of a particular type in a specific domain where there is a consensus among ethicists concerning the correct action.

Relevant data types must be established and representation schema for these defined. Ethical action preference is ultimately dependent upon the *ethically relevant features* that actions involve such as harm, benefit, respect for autonomy, etc. Such features are represented as a descriptor that specifies the degree of its presence or absence in a given action. For instance, it might be the case that one degree of harm is present in the action of not notifying an overseer that an eldercare robot's charge is refusing to take his/her medication.

For each ethically relevant feature, there is a *duty* incumbent of an agent to either minimize that feature (as would be the case for harm) or maximize it (as would be the case for, say, respect for autonomy).

An *action* is represented as a tuple of the degrees to which it satisfies (positive values) or violates (negative values) each of duty. For instance, given the previous example, it might also be the case that not notifying an overseer exhibits the presence of one degree of respect for autonomy and, combined with its one degree of presence of harm, the tuple representing this action would be (-1, 1) where the first value denotes the action's violation of the duty to minimize harm and the second value denotes the action's satisfaction of the duty to maximize respect for autonomy.

Given this representation for an action, a *case* involving two actions can be represented as a tuple of the differentials of their corresponding duties. In a *positive case* (i.e. where the first action is ethically preferable to the second), the duty satisfaction/violation values of the less ethically preferable action are subtracted from the corresponding values in the more ethically preferable action, producing a tuple of values representing how much more or less the ethically preferable action satisfies or violates each duty. For example, consider a case involving the previously represented action and another action in which an overseer is notified when the robot's charge refuses to take his/her medication. This new action would be represented as (1, -1) (i.e. satisfying the duty to minimize harm by one degree, and violating the duty to respect autonomy by the same amount) and, given that it is more important to prevent harm in this case and the ethically preferable action is this new one, the case would be represented as ((1- -1) (-1 - 1)) or (2, -2). That is, the ethically preferable action satisfies the duty to minimize harm by two more degrees than the less ethically preferable action and violates the duty to maximize respect for autonomy by the same amount.

A representation for a *principle* of ethical action preference can be defined as a predicate p in terms of lower bounds for duty differentials of cases:

$$\begin{array}{l} p(a_1, a_2) \leftarrow \\ \Delta d_1 \geq v_{1,1} \;\wedge \cdots \wedge\; \Delta d_m \geq v_{1,m} \\ \vee \\ \vdots \\ \vee \\ \Delta d_n \geq v_{n,1} \;\wedge \cdots \wedge\; \Delta d_m \geq v_{n,m} \end{array}$$

where Δd_i denotes the differential of a corresponding duty of actions a_1 and a_2 and $v_{i,j}$ denotes the lower bound of that differential such that $p(a_1, a_2)$ returns true if action $a_1$is ethically preferable to action a_2.

Ethically significant actions must be identified. These are the activities of a system that are likely to have a non-trivial ethical impact on the system's user and/or environment. It is from this set of actions that the most ethically preferable action will be chosen at any given moment.

Lastly, to facilitate the development of the principle, cases of a domain specific dilemma type with determinations regarding their ethically preferred action must be supplied.

3 METHODS

Given the complexity of the task at hand, computational methods are brought to bear wherever they prove helpful. To minimize bias, CPB is committed only to a knowledge representation scheme based on the concepts of ethically relevant features with corresponding degrees of presence/absence from which duties to minimize/maximize these features with corresponding degrees of satisfaction/violation of those duties are inferred. The particulars of the representation are dynamic—particular features, degrees, and duties are determined from example cases permitting different sets in different domains to be discovered.

As the representation is instantiated, cases are constructed in CPB from the values provided for the actions that comprise it. From features and the degrees to which these are present or absent in one of the actions in question, duties are inferred to either maximize or minimize these features and the degree to which the cases satisfy or violate each of these duties is computed.

As it is likely that in many particular cases of ethical dilemmas ethicists agree on the ethically relevant features and the right course of action, generalization of such cases can be used to help discover principles needed for ethical guidance of the behavior of autonomous systems [2][3]. A principle abstracted from cases that is no more specific than needed to make determinations complete and consistent with its training can be useful in making provisional determinations about untested cases. CPB uses *inductive concept learning* [11] to infer a principle of ethical action preference from cases that is complete and consistent in relation to these cases. The principles discovered are *most general specializations,* covering more cases than those used in their specialization and, therefore, can be used to make and justify provisional determinations about untested cases.

The suite of methods described above has been implemented in GenEth [5] and has been used to develop ethical principles in a number of different domains.
(See http://uhaweb.hartford.edu/anderson/Site/GenEth.html).

For example, the system, in conjunction with an ethicist, instantiated a knowledge representation scheme in the domain of medication reminding to include: the ethically relevant features of harm, interaction, benefit, and respect for autonomy and the corresponding duties (and the specific degrees to which these duties can be satisfied or violated) to minimize harm (-1 to +1), maximize benefit (-2 to +2), and maximize respect for autonomy (-1 to +1). The discovered principle is complete and consistent with respect to its training cases and is general enough to cover cases not in this set:

$$\begin{gathered} p(\text{notify}, \text{do not notify}) \rightarrow \\ \Delta\text{min harm} \geq 1 \\ \vee \\ \Delta\text{max benefit} \geq 3 \\ \vee \\ \Delta\text{min harm} \geq -1 \wedge \Delta\text{max benefit} \geq -3 \wedge \Delta\text{max autonomy} \geq -1 \end{gathered}$$

4 IMPLEMENTATION

The discovered principle is used to choose which ethically significant action the system should undertake next. All ethically significant actions need to be represented in terms of their current ethically relevant feature values. As time passes and circumstances change these values are likely to change. They can be computed from original input data, sensed from the environment, elicited from a user, etc. At any given moment, the set of these values comprise the current *ethical state* of the system.

At each point where the system needs to decide which ethically significant action to undertake, the current ethical state is determined and actions are partitioned into the partial order defined by the principle. Those actions that comprise the most ethically preferable partition represent the set of high-level goals that are best in the current ethical state. Being equally ethically preferable, any of these goals can be chosen by the system. This goal is then realized using a series of actions not in themselves considered ethically significant.

This implementation was instantiated at the prototype level in a Nao robot [4], the first example, we believe, of a robot that uses an ethical principle to determine which actions it will take.

5 TESTING

A *case-supported principle based behavior* paradigm provides a means of justification for, as well as a means of ascribing responsibility to, a system's actions. To validate principles we advocate an *Ethical Turing Test,* a variant of the test Alan Turing [17] suggested as a means to determine whether the term "intelligence" can be applied to a machine that bypassed disagreements about the definition of intelligence. This variant tests whether the term "ethical" can be applied to a machine by comparing the ethically-preferable action specified by an ethicist in an ethical dilemma with that of a machine faced with the same dilemma. If a significant number of answers given by the machine match the answers given by the ethicist, then it has passed the test. Such evaluation holds the machine-generated principle to the highest standards and, further, permits evidence of incremental improvement as the number of matches increases (see [1] for the inspiration of this test). We have developed and administered an Ethical Turing Test based upon the principles discovered using GenEth.

As an action is chosen for execution by a system, clauses of the principle that were instrumental in its selection can be determined and used to formulate an explanation of why that particular action was chosen over the others. As clauses of principles can be traced to the cases from which they were abstracted, these cases and their origin can provide support for a selected action.

6 SCENARIO

To make the CPB paradigm more concrete, the following scenario is provided. It attempts to envision an eldercare robot of the near future whose ethically significant behavior is guided by an ethical principle. Although the robot's set of possible actions is circumscribed in this scenario, it serves to demonstrate the complexity of choosing the ethically correct action at any given moment. The case-supported principle-based behavior paradigm is an abstraction to help manage this complexity.

EthEl (Ethical Eldercare Robot) is a principle-based autonomous robot who assists the staff with caring for the residents of an assisted living facility. She has a set of possible ethically significant actions that she performs, each of which is *represented as a profile of satisfaction/violation degrees of a set of prima facie duties. These degrees may vary over time as circumstances change. ETHEL uses an ethical principle to select the currently ethically preferable action from among her possible actions including charging her batteries, interacting with the residents, alerting nurses, giving resident reminders, and delivering messages and items. Currently ETHEL stands in a corner of a room in the assisted living facility charging her batteries. She has sorted her set of ethically significant actions according to her ethical principle and charging her batteries has been deemed the most ethically preferable action among them as her prima facie duty to maintain herself has currently taken precedence over her other duties. As time passes, the satisfaction/violation levels of the duties of her actions (her ethical state) vary according to the initial input and the current situation. Her batteries now sufficiently charged, she sorts her possible actions and determines that she should interact with the patients as her duty of beneficence ("do good") currently overrides her duty to maintain herself.*

She begins to make her way around the room, visiting residents in turn, asking if she can be helpful in some way—get a drink, take a message to another resident, etc. As she progresses and is given a task to perform, she assigns a profile to that task that specifies the current satisfaction/violation levels of each duty involved in it. She then resorts her actions to find the most ethically preferable one. One resident, in distress, asks her to alert a nurse. Given the task, she assigns a profile to it. Ignoring the distress of a resident involves a violation of the duty of nonmaleficence ("prevent harm"). Sorting her set of actions by her ethical principle, ETHEL finds that her duty of nonmaleficence currently overrides her duty of beneficence, preempting her resident visitations, and she seeks a nurse and informs her that a resident is in need of her services. When this task is complete and removed from her collection of tasks to perform, she resorts her actions and determines that her duty of beneficence is her overriding concern and she continues where she left off in her rounds.

As ETHEL continues making her rounds, duty satisfaction/violation levels vary over time until, due to the need to remind a resident to take a medication that is designed to make the patient more comfortable, and sorting her set of possible actions, the duty of beneficence can be better served by issuing this reminder. She seeks out the resident requiring the reminder. When she finds the resident, ETHEL tells him that it is time to take his medication. The resident is currently occupied in a conversation, however, and he tells ETHEL that he will take his medication later. Given this response, ETHEL sorts her actions to determine whether to accept the postponement or not. As her duty to respect the patient's autonomy currently overrides a low level duty of beneficence, she accepts the postponement, adjusting this reminder task's profile and continues her rounds.

As she is visiting the residents, someone asks ETHEL to retrieve a book on a table that he can't reach. Given this new task, she assigns it a profile and resorts her actions to see what her next action should be. In this case, as no other task will satisfy her duty of beneficence better, she retrieves the book for the resident. Book retrieved, she resorts her actions and returns to making her rounds. As time passes, it is determined through action sorting that ETHEL's duty of beneficence, once again, will be more highly satisfied by issuing a second reminder to take a required medication to the resident who postponed doing so previously. A doctor has indicated that if the patient doesn't take the medication at this time he soon will be in much pain. She seeks him out and issues the second reminder. The resident, still preoccupied, ignores ETHEL. ETHEL sorts her actions and determines that there would be a violation of her duty of nonmaleficence if she accepted another postponement from this resident. After explaining this to the resident and still not receiving an indication that the reminder has been accepted, ETHEL determines that an action that allows her to satisfy her duty of nonmaleficence now overrides any other duty that she has. ETHEL seeks out a nurse and informs her that the resident has not agreed to take his medication. Batteries running low, ETHEL's duty to herself is increasingly being violated to the point where ETHEL's the most ethically preferable action is to return to her charging corner to await the next call to duty.

What we believe is significant about this vision of how an ethical robot assistant would behave is that an ethical principle is used to select the best action in a each situation, rather than in just determining whether a particular action is acceptable or not. This allows for the possibility that ethical considerations may lead a robot to aid a human being or prevent the human being from being harmed, not just forbid it from performing certain actions. Correct ethical behavior does not only involve not doing certain things, but also attempting to bring about ideal states of affairs.

7 RELATED RESEARCH

Although many have voiced concern over the impending need for machine ethics for decades [18] [7] [10], there has been little research effort made towards accomplishing this goal. Some of this effort has been expended attempting to establish the feasibility of using a particular ethical theory as a foundation for machine ethics without actually attempting implementation: Christopher Grau [8] considers whether the ethical theory that best lends itself to implementation in a machine, Utilitarianism, should be used as the basis of machine ethics; Tom Powers [14] assesses the viability of using deontic and default logics to implement Kant's categorical imperative.

Efforts by others that do attempt implementation have largely been based, to greater or lesser degree, upon casuistry—the branch of applied ethics that, eschewing principle-based approaches to ethics, attempts to determine correct responses to new ethical dilemmas by drawing conclusions based on parallels with previous cases in which there is agreement concerning the correct response. Rafal Rzepka and Kenji Araki [16], at what might be considered the most extreme degree of casuistry, have explored how statistics learned from examples of ethical intuition drawn from the full spectrum of the World Wide Web might be useful in furthering machine ethics in the domain of safety assurance for household robots. Marcello Guarini [9], at a less extreme degree of casuistry, is investigating a neural network approach where particular actions concerning killing and allowing to die are classified as acceptable or unacceptable depending upon different motives and consequences. Bruce McLaren [12], in the spirit of a more pure form of casuistry, uses a case-based reasoning approach to develop a system that 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 without making judgments.

There have also been efforts to bring logical reasoning systems to bear in service of making ethical judgments, for instance deontic logic [6] and prospective logic [13]. These efforts provide further evidence of the computability of ethics but, in their generality, they do not adhere to any particular ethical theory and fall short in actually providing the principles needed to guide the behavior of autonomous systems.

Our approach is unique in that we are proposing a comprehensive, extensible, domain-independent paradigm grounded in well-established ethical theory that will help ensure the ethical behavior of current and future autonomous systems.

8 CONCLUSION

It can be argued that *machine ethics* ought to be the driving force in determining the extent to which autonomous systems should be permitted to interact with human beings. Autonomous systems that behave in a less than ethically acceptable manner towards human beings will not, and should not, be tolerated. Thus, it becomes paramount that we demonstrate that these systems will not violate the rights of human beings and will perform only those actions that follow acceptable ethical principles. Principles offer the further benefits of serving as a basis for justification of actions taken by a system as well as for an overarching control mechanism to manage unanticipated behavior of such systems. Developing principles for this use is a complex process and new tools and methodologies will be needed to help contend with this complexity. We offer the case-supported principle-based behavior paradigm as an abstraction to help mitigate this complexity.

ACKNOWLEDGEMENTS

This material is based in part upon work supported by the National Science Foundation under Grant Numbers IIS- 0500133 and IIS-1151305.

REFERENCES

[1] Allen, C., Varner, G. and Zinser, J. Prolegomena to Any Future Artificial Moral Agent**.** Journal of Experimental and Theoretical Artificial Intelligence 12, pp. 251-61, 2000.

[2] Anderson, M., Anderson, S. & Armen, C. MedEthEx: A Prototype Medical Ethics Advisor. Proceedings of the Eighteenth Conference on Innovative Applications of Artificial Intelligence, Boston, Massachusetts, August 2006.

[3] Anderson, M. and Anderson, S. L., Machine Ethics: Creating an Ethical Intelligent Agent, Artificial Intelligence Magazine, 28:4, Winter 2007.

[4] Anderson, M. and Anderson, S. L., "Robot be Good", Scientific American Magazine, October 2010.

[5] Anderson, M. and Anderson, S. L., GenEth: A General Ethical Dilemma Analyzer, 11th International Symposium on Formalizations of Commonsense Reasoning, Ayia Napa, Greece, May 2013.

[6] Bringsjord, S., Arkoudas, K. and Bello, P. Towards a General Logicist Methodology for Engineering Ethically Correct Robots. IEEE Intelligent Systems ,vol. 21, no. 4, pp. 38-44, July/August 2006.

[7] Gips, J. Towards the Ethical Robot. Android Epistemology, Cambridge MA: MIT Press, pp. 243–252, 1995.

[8] Grau, C. There Is No "I" in "Robot": Robots and Utilitarianism. IEEE Intelligent Systems , vol. 21, no. 4, pp. 52-55, July/ August 2006.

[9] Guarini, M. Particularism and the Classification and Reclassification of Moral Cases. IEEE Intelligent Systems , vol. 21, no. 4, pp.22-28, July/ August 2006.

[10] Khan, A. F. U. The Ethics of Autonomous Learning Systems**.** Android Epistemology, Cambridge MA: MIT Press, pp. 253–265, 1995.

[11] Lavrač, N. and Džeroski, S. Inductive Logic Programming: Techniques and Applications. Ellis Harwood, 1997.

[12] McLaren, B. M. Extensionally Defining Principles and Cases in Ethics: an AI Model, Artificial Intelligence Journal, Volume 150, November, pp. 145- 181, 2003.

[13] Pereira, L.M. and Saptawijaya, A. Modeling Morality with Prospective Logic, Progress in Artificial Intelligence: Lecture Notes in Computer Science, vol. 4874, p.p. 99-111, 2007.

[14] Powers, T. M. Prospects for a Kantian Machine. IEEE Intelligent Systems ,vol. 21, no. 4, pp. 46-51, July/August 2006.

[15] Ross, W.D., *The Right and the Good*, Oxford University Press, Oxford, 1930.

[16] Rzepka, R. and Araki, K. What Could Statistics Do for Ethics? The Idea of Common Sense Processing Based Safety Valve. Proceedings of the AAAI Fall Symposium on Machine Ethics, pp. 85- 87, AAAI Press, 2005.

[17] Turing, A.M. Computing machinery and intelligence. Mind, 59, 433-460, 1950.

[18] Waldrop, M. M. A Question of Responsibility. Chap. 11 in Man Made Minds: The Promise of Artificial Intelligence. NY: Walker and Company, 1987. (Reprinted in R. Dejoie et al., eds. Ethical Issues in Information Systems. Boston, MA: Boyd and Fraser, 1991, pp. 260-277.)