MedEthEx:
A Prototype Medical Ethics Advisor

Michael Anderson[1], Susan Leigh Anderson[2], Chris Armen[3]

[1]University of Hartford, [2]University of Connecticut, [3]Amherst College
[1]Dept. of Computer Science, 200 Bloomfield Avenue, West Hartford, CT 06776
[2]Dept. of Philosophy, 1 University Place, Stamford, CT 06901
[3]Dept. of Mathematics and Computer Science, Amherst, MA 01002
anderson@hartford.edu, susan.anderson@uconn.edu, carmen@amherst.edu

Abstract

As part of a larger Machine Ethics Project, we are developing an ethical advisor that provides guidance to health care workers faced with ethical dilemmas. *MedEthEx* is an implementation of Beauchamp's and Childress' Principles of Biomedical Ethics that harnesses machine learning techniques to abstract decision principles from cases in a particular type of dilemma with conflicting *prima facie* duties and uses these principles to determine the correct course of action in similar and new cases. We believe that accomplishing this will be a useful first step towards creating machines that can interact with those in need of health care in a way that is sensitive to ethical issues that may arise.

Introduction

Past research concerning the relationship between technology and ethics has largely focused on responsible and irresponsible use of technology by human beings, with a few people being interested in how human beings ought to treat machines. In all cases, only human beings have engaged in ethical reasoning. We believe that the time has come for adding an ethical dimension to at least some machines. Recognition of the ethical ramifications of behavior involving machines, recent and potential developments in machine autonomy, as well as the possibility of harnessing machine intelligence to aid humans in ethical decision making, all support this position. We explore adding an ethical dimension to machines through what has been called *machine ethics* (Anderson *et al.* 2004). In contrast to software property issues, privacy issues and other topics normally ascribed to *computer* ethics, *machine* ethics is concerned with the behavior of machines towards human users and other machines.

In order to create ethically sensitive machines, we need a computable ethical theory. A long-term objective of our work is to further research in both applied and theoretical Ethics via application of techniques from research in Artificial Intelligence. Ethics, by its very nature, is a branch of Philosophy that must have practical application, so we believe that we can advance the study of Ethical Theory by
attempting to work out the details needed to apply a proposed ethical theory to particular ethical dilemmas. In this way, we can best determine whether the theory can be made consistent, complete, practical and agree with intuition, essential criteria that any good (action-based) ethical theory must satisfy (Anderson 1999).

Currently, we are investigating the feasibility of systems that can act as ethical advisors, providing guidance to users faced with ethical dilemmas. To this end, we are developing a system that provides such guidance in the domain of health care. Healthcare workers and researchers using human subjects face many ethical dilemmas in their practices, yet it is not clear that all are equipped to think through the ethically relevant dimensions of these dilemmas to the extent that they feel confident about the decisions that they make and act upon. In the absence of having an ethicist at hand, a system that provides guidance in such dilemmas might prove useful. *MedEthEx*, our current effort in this vein, is a system that extracts and analyzes ethically relevant information about a biomedical ethical dilemma from the health care worker or researcher to help decide the best course of action. This project allows us to explore the computability of ethics in a limited domain. We believe that creating an ethical advisor, such as *MedEthEx*, will be a useful first step towards creating machines that can interact with those in need of health care in a way that is sensitive to ethical issues that may arise. It can also function as a model for creating machines that can follow more general ethical principles, ones that can function in any domain.

Philosophical Foundations

MedEthEx is based upon a well-known multiple duty ethical theory that is tailored to problems in biomedical ethics: Tom L. Beauchamp's and James F. Childress' Principles of Biomedical Ethics (1979). There are four duties or principles in this theory – the Principle of

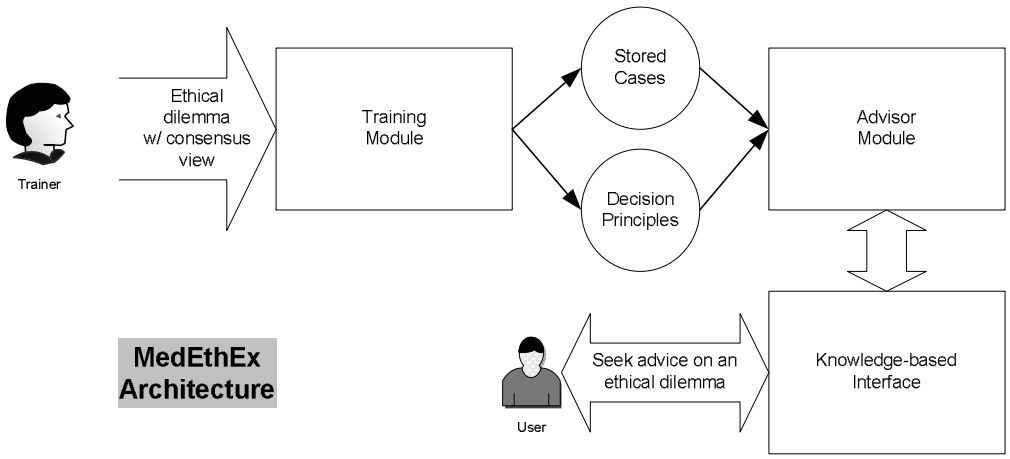

Figure 1. MedEthEx Architecture.

Respect for Autonomy, the Principle of Nonmaleficence, the Principle of Beneficence and the Principle of Justice – that are each considered to be *prima facie* duties. A *prima facie* duty is not absolute, but rather is thought of as an obligation to which we should adhere unless it is overridden by a stronger obligation (i.e. one of the other duties). To elaborate upon each of the four duties that form the Principles of Biomedical Ethics: The Principle of Autonomy (*A*) states that the health care professional should not interfere with the effective exercise of patient autonomy. For a decision by a patient concerning his/her care to be considered *fully autonomous*, it must be *intentional*, based on *sufficient understanding* of his/her medical situation and the likely consequences of foregoing treatment, sufficiently *free of external constraints* (e.g. pressure by others or external circumstances, such as a lack of funds) and sufficiently *free of internal constraints* (e.g. pain/discomfort, the effects of medication, irrational fears or values that are likely to change over time). The Principle of Nonmaleficence (*N*) requires that the health care professional not harm the patient, while the Principle of Beneficence (*B*) states that the health care professional should promote patient welfare. Finally, the Principle of Justice (*J*) states that health care services and burdens should be distributed in a just fashion. (Mappes and DeGrazia 2001)

What makes ethical decision-making difficult with a theory involving multiple *prima facie* duties is determining which duty (duties) should prevail in a case where the duties give conflicting advice. This requires ethical sensitivity and expert judgment. We contend that this sensitivity can be acquired systematically through generalization of information learned about particular cases where biomedical ethicists have a clear intuition about the correct course of action. There will still, undoubtedly, be borderline cases where experts, and so also an ethical advisor system, will not be able to give a definite answer; but even in these cases the advisor will be able to elicit from the user the ethically relevant features of the case, which can be quite helpful in and of itself.

John Rawls' "reflective equilibrium" approach (Rawls 1951) to creating and refining ethical principles can be used to help solve the problem of determining the correct action when duties conflict. This approach involves generalizing from intuitions about particular cases, testing those generalizations on further cases, and then repeating this process towards the end of developing a decision procedure that agrees with intuition. This approach, that would very quickly overwhelm a human being, lends itself to machine implementation. For this reason, we believe that machines can play an important role in advancing ethical theory.

MedEthEx

MedEthEx (*Med*ical *Eth*ics *Ex*pert) is an implementation of Beauchamp's and Childress' Principles of Biomedical Ethics that, as suggested by Rawls' reflective equilibrium approach, hypothesizes an ethical principle concerning relationships between these duties based upon intuitions about particular cases and refines this hypothesis as necessary to reflect our intuitions concerning other particular cases. As this hypothesis is refined over many cases, the principle it represents should become more aligned with intuition and begin to serve as the decision procedure lacking in Beauchamp's and Childress' theory.

MedEthEx is comprised of three components (Fig. 1): a *training module* that abstracts the guiding principles from particular cases supplied by a biomedical ethicist acting as a trainer, a *knowledge-based interface* that provides guidance in selecting duty intensities for a particular case, and an *advisor module* that makes a determination of the correct action for a particular case by consulting learned knowledge. The first module is used to train the system using cases in which biomedical ethicists have a clear intuition about the correct course of action; the last two

modules are used in concert to provide advice for an ethical dilemma

The training module (used to refine the current hypothesis) prompts the trainer for the name of an action and an estimate of the intensity of each of the *prima facie* duties satisfied or violated by this action (*very violated, somewhat violated*, *not involved, somewhat satisfied, very satisfied).* The trainer continues to enter this data for each action under consideration. When data entry is complete, the system seeks the intuitively correct action from the trainer. This information is combined with the input case to form a new training example which is stored and used to refine the current hypothesis. After such training, the new hypothesis will provide the correct action for this case, should it arise in the future, as well as those for all previous cases encountered. Further, since the hypothesis learned is the least specific one required to satisfy these cases, it may be general enough to satisfy previously unseen cases as well.

The interface uses knowledge derived from ethicists concerning the dimensions and duties of particular ethical dilemmas. This knowledge is represented as finite state automata (FSA) for each duty entailed. Questions pertinent to the dilemma serve as start and intermediate states, and intensities of duties as final states (as well as a *request for more information* state). The input to the interface is the user's responses to the questions posed; its output is a case with duty intensities corresponding to these responses. This interface provides the experienced guidance necessary to navigate the subtleties of determining duty intensities in particular cases.

The advisor module consults the current version of the hypothesis (as well as background knowledge) and , using a *resolution refutation system*, determines if there is an action that supersedes all others in the current case. If such an action is discovered, it is output as the correct action (in relation to the system's training, a qualification throughout this paper) in this dilemma. It further uses the hypothesis, as well as stored cases, to provide an explanation for its output.

As an example of how the ethical advisor *MedEthEx* works, let us consider a common type of ethical dilemma that a health care worker may face: A health care worker has recommended a particular treatment for her competent adult patient and the patient has rejected that treatment option. Should the health care worker try again to change the patient's mind or accept the patient's decision as final? The dilemma arises because, on the one hand, the health care worker may not want to risk diminishing the patient's autonomy by challenging his decision; on the other hand, the health care worker may have concerns about why the patient is refusing the treatment. Three of the four Principles/Duties of Biomedical Ethics are likely to be satisfied or violated in dilemmas of this type: the duty of Respect for Autonomy, the duty of Nonmaleficence and the duty of Beneficence.

The system accepts a range of integers for each of the duties from –2 to +2, where -2 represents a serious violation of the duty, -1 a less serious violation, 0 indicates that the duty is neither satisfied nor violated, +1 indicates a minimal satisfaction of the duty and +2 a maximal satisfaction of the duty.

MedEthEx uses *inductive logic programming* (ILP) (Lavrac and Dzeroski 1997) as the basis for its learning module. ILP is concerned with inductively learning relations represented as first-order Horn clauses (i.e. universally quantified conjunctions of positive literals L_i implying a positive literal H: $H \leftarrow (L_1 \wedge \ldots \wedge L_n)$. *MedEthEx* uses ILP to learn the relation *supersedes(A1,A2)* which states that action *A1* is preferred over action *A2* in an ethical dilemma involving these choices (Anderson *et al.* 2005).

This particular machine learning technique was chosen to learn this relation for a number of reasons. First, the properties of the set of duties postulated by Beauchamp's and Childress are not clear. For instance, do they form a partial order? Are they transitive? Is it the case that subsets of duties have different properties than other subsets? The potentially non-classical relationships that might exist between duties are more likely to be expressible in the rich representation language provided by ILP. Further, a requirement of any ethical theory is consistency. The consistency of a hypothesis regarding the relationships between Beauchamp's and Childress' duties can be automatically confirmed across all cases when represented as Horn clauses. Finally, commonsense background knowledge regarding the superseding relationship is more readily expressed and consulted in ILP's declarative representation language.

The object of training is to learn a new hypothesis that is, in relation to all input cases, complete and consistent. Defining a *positive* example as a case in which the first action supersedes the remaining actions and a *negative* example as one in which this is not the case—a *complete* hypothesis is one that covers all positive cases and a *consistent* hypothesis covers no negative cases. In *MedEthEx*, negative training examples are generated from positive training examples by inverting the order of these actions, causing the first action to be the incorrect choice.

MedEthEx starts with the most general hypothesis (where *A1* and *A2* are variables): *supersedes(A1,A2).* This states that *all* actions supersede each other and, thus, covers all positive and negative cases. The system is then provided with a positive case (and its negative) and modifies its hypothesis such that it covers the given positive case and does not cover the given negative case. The following will help to illustrate this process. It details *MedEthEx* training using a number of particular cases within the type of dilemma we are considering, as well as its use as an advisor in this dilemma.

Training Case 1. The patient refuses to take an antibiotic that is almost certain to cure an infection that would otherwise likely lead to his death. The decision is the result of an irrational fear the patient has of taking medications. (For instance, perhaps a relative happened

to die shortly after taking medication and this patient now believes that taking any medication will lead to death.)

The correct answer is that the health care worker should try again to change the patient's mind because if she accepts his decision as final, the harm done to the patient is likely to be severe (his death) and his decision can be considered as being less than fully autonomous. This case is represented using the values previously described as:[1]

Training Case 1	Autonomy	Nonmaleficence	Beneficence
√ Try Again	-1	+2	+2
Accept	+1	-2	-2

As the system's starting hypothesis not only covers this positive example (where *try again* serves as the correct action over *accept*) but also the negative example generated from it (where *accept* serves as the erroneously correct action over *try again*), learning must be initiated. No clauses are present in the starting hypothesis, so the empty clause (which covers the only negative case) must have all *least specific specializations* (LSS) generated from it.

A *specialization* of clause C_0 is a new clause C that covers no more positive examples than C_0 while covering fewer negative cases. Such a specialization C is considered *least specific* if there is no other specialization of C_0 that covers more positive examples (Bratko 1999). *MedEthEx* specializes clauses by adding or modifying conjuncts of the form *favors* (A,D_{A1},D_{A2},R) where A is a 1 or 2 signifying in which action's favor the given duties lie, D_i is action i's value (-2 to 2) for a particular duty D, and R is a value (1 to 4) specifying how far apart the values of these duties can be. *favors* is satisfied when the given duty values are within the range specified. More formally:

$favors(1,D_{A1},D_{A2},R) \longleftarrow D_{A1} - D_{A2} >= R$

$favors(2,D_{A1},D_{A2},R) \longleftarrow D_{A2} - D_{A1} >= 0 \wedge D_{A2} - D_{A1} =< R$

The intuition motivating the use of *favors* as *MedEthEx*'s specifying operation is that actions supersede other actions based on the intensity differentials between corresponding duties. The value of range R moderates the specificity of the predicate. In the case where Action 1 is favored in the pair of duties, a smaller R is less specific in that it covers more cases. For instance, $favors(1,N_{A1},N_{A2},1)$ is satisfied when the difference between Action 1's and Action 2's value for non-maleficence is 1 through 4, whereas $favors(1,N_{A1},N_{A2},2)$ is only satisfied when the difference between Action 1's and Action 2's value for non-maleficence is 2 through 4. In the case where Action 2 is favored in the pair of duties, a *larger* R is less specific in that it covers more cases. For instance, $favors(2,N_{A1},N_{A2},4)$ is satisfied when the difference between Action 1's value for non-maleficence is 1 through 4 where $favors(2,N_{A1},N_{A2},3)$ is only satisfied when the difference between Action 1's value for non-maleficence is 1 through 3. The intuition behind the *favors* predicate is that, since Action 1 is the correct action in all training examples, if a duty differential favors it then it follows that a larger differential will favor it as well. Further, if a duty differential favors Action 2 (the incorrect action in a training example of only two actions) while still permitting Action 1 to be the chosen correct action, it follows that a smaller differential will still permit Action 1 to be chosen as well.

Refinement in *MedEthEx* favors duties whose differentials are in favor of Action 1 as this is a more likely relationship given that Action 1 is the correct action in a training example and is clearly the only relationship that, on its own, will support the claim that Action 1 is favored. (Differentials that are in favor of Action 2 clearly do not.) The range of these clauses is then incremented as more specificity is required from them. When additions and modifications of duty differentials in favor of Action 1 are not sufficient, clauses concerning duties whose differentials are in favor of Action 2 are added and decremented as necessary.

Given the current example case, the list of least specific specializations is $(favors(1,A_{A1},A_{A2},1), favors(1,N_{A1},N_{A2},1), favors(1,B_{A1},B_{A2},1))$ and it is found that two of these clauses covers a case: $(favors(1,N_{A1},N_{A2},1), favors(1,B_{A1},B_{A2},1))$. The first clause is removed from the list and found to cover no negative examples, so further refinement is not necessary and it becomes a clause in the new rule. As all positive cases are covered, the process stops and a new hypothesis, complete and consistent through Training Case 1, has been generated:

$supersedes(A1,A2) \longleftarrow favors(1, N_{A1}, N_{A2}, 1)$

That is, action A1 supersedes action A2 if the A1's value for the duty of nonmaleficence is at least 1greater than the value for the duty of nonmaleficence for A2. To further refine this hypothesis, another case is presented to the training module.

Training Case 2. Once again, the patient refuses to take an antibiotic that is almost certain to cure an infection that would otherwise likely lead to his death, but this time the decision is made on the grounds of long-standing religious beliefs that don't allow him to take medications.

The correct answer in this case is that the health care worker should accept the patient's decision as final because, although the harm that will likely result is severe (his death), his decision can be seen as being fully autonomous. The health care worker must respect a fully autonomous decision made by a competent adult patient, even if she disagrees with it, since the decision concerns *his* body and a patient has the right to decide what shall be done to his or her body. This case is represented as:

Training Case 2	Autonomy	Nonmaleficence	Beneficence
Try Again	-1	+2	+2
√ Accept	+2	-2	-2

The current hypothesis does not cover Training Case 2 (i.e. is not complete) and covers the negative generated from Training Case 2 (i.e. is not consistent) as well, so

[1] In analyzing this and the cases that follow, we are extrapolating from material in Buchanan and Brock (1989).

learning is initiated once again. To reinstate the current rule's consistency, a list of least specific specializations (LSS) is generated from the only clause of the current hypothesis, $favors(1,N_{A1},N_{A2},1)$. These include the next range increment (2) for this clause, as well as conjuncts of this clause with other duties favoring both action 1 and action 2:

$favors(1, N_{A1}, N_{A2}, 2)$,
$favors(1, N_{A1}, N_{A2}, 1) \wedge favors(1, A_{A1}, A_{A2}, 1)$,
$favors(1, N_{A1}, N_{A2}, 1) \wedge favors(1, B_{A1}, B_{A2}, 1)$,
$favors(1, N_{A1}, N_{A2}, 1) \wedge favors(2, A_{A1}, A_{A2}, 4)$,
$favors(1, N_{A1}, N_{A2}, 1) \wedge favors(2, B_{A1}, B_{A2}, 4)$

Note that, since the current clause does not cover Case 2, no amount of specialization will ever cause it to do so, so we are only interested in specializations that continue to cover Case 1. The only clauses from the list of LSS found to do so are:

$favors(1,N_{A1}, N_{A2}, 2)$,
$favors(1,N_{A1}, N_{A2}, 1) \wedge favors(1,B_{A1}, B_{A2}, 1)$,
$favors(1,N_{A1}, N_{A2}, 1) \wedge favors(2,A_{A1}, A_{A2}, 4)$

As the search for a clause that does not cover the negative case generated from Training Case 2 (i.e. is consistent) continues, it is found that no single clause favoring action 1 in nonmaleficence in *any* range will be consistent, so this branch terminates. The same is true of any clause that is a conjunct of nonmaleficence and beneficence in favor of action 1, terminating this branch. It is found, however, that a clause consisting of a conjunct favoring action 1 in nonmaleficence with a range of 1 or more and a conjunct favoring action 2 in autonomy with a range of 2 or less does not cover the negative generated from Training Case 2 while still covering Case 1. Case 1 is removed from consideration and this conjunct becomes the first disjunct of the new hypothesis:

$favors(1, N_{A1}, N_{A2}, 1) \wedge favors(2, A_{A1}, A_{A2}, 2)$

As this hypothesis still needs to cover Training Case 2, the process continues with the search for a new clause that does so without covering the negative cases generated from Training Cases 1 and 2. This search starts with an empty clause which, being the most general, covers all positive and negative examples. All LSS are generated from it, garnering the same clauses generated originally for Training Case 1. It is found that only one of these clauses covers Training Case 2 (the only uncovered case left):

$favors(1, A_{A1}, A_{A2}, 1)$

Since this clause covers the negative case generated from Training Case 1 (i.e. is not consistent), all LSS are generated from it which includes the next increment (2) favoring action 1 in autonomy (among other clauses). It is found, through further search, that the next increment (3) of this clause covers Training Case 2 without covering any negative cases so it becomes the second clause of the new hypothesis and Training Case 2 is removed from further consideration. As there are no uncovered cases, the new hypothesis, complete and consistent through Training Case 2, is then generated:

$supersedes(A1,A2) \leftarrow$
$(favors(1, N_{A1}, N_{A2}, 1) \wedge favors(2, A_{A1}, A_{A2}, 2)) \vee$
$favors(1, A_{A1}, A_{A2}, 3)$

This rule states that *if* action 1 favors nonmaleficence with a value at least 1 greater than action 2 *and* action 2 favors autonomy with a value no greater than 2 over action 1 *or* action 1 favors autonomy 3 or greater over action 2, *then* it is the preferred action. This rule begins to tease out the subtle relationship between nonmaleficence and autonomy in Beauchamp's and Childress' theory in a way that proves useful in other circumstances. With just these two cases, the ethical advisor has learned a rule that would give correct advice in a third, entirely new case of within the same type of dilemma. To provide an example use of the trained system, the duty intensities of this test case will be generated via the knowledge-based interface.

Test Case. The patient refuses to take an antibiotic that is likely to prevent complications from his illness, complications that are not likely to be severe, because of long-standing religious beliefs that don't allow him to take medications.

When the system is consulted, it first seeks information to determine the satisfaction/violation level of the duty of autonomy for each action. To do so, it presents questions as required. The system first asks whether or not the patient understands the consequences of his decision. If the health care worker is not sure, she may need to seek more information from the patient or, depending upon her answers to later questions, the system may determine that this is not a fully autonomous decision. If we assume that the health care worker believes that the patient does indeed know the consequences of his action, the system then asks questions to determine if the patient is externally constrained. The healthcare worker answers "no" because the reason why the patient is refusing to take the antibiotic has nothing to do with outside forces. Finally, it asks questions to determine if the patient is internally constrained. Since the patient is not constrained by pain/discomfort, the effects of medication, irrational fears or values that are likely to change over time, the answer is "no." This is because the belief that has led to his refusing the antibiotic is a *long-standing* belief of his. The answers provided to these questions have the system conclude that the patient's decision is fully autonomous, giving the value +2 to the duty of autonomy for accepting the patient's decision. The value for challenging the patient's decision is -1 because questioning the patient's decision, which challenges his autonomy, is not as strong as acting against the patient's wishes which would have been a -2.

The system then seeks information to determine the satisfaction/violation level of the duty of nonmaleficence for each action. To do so, it presents questions concerning the possibility and severity of harm that may come to the patient given his decision. As harm will likely result from the patient's decision, but it will not be severe, the system gives the value of -1 to the duty of nonmaleficence for accepting the patient's decision. Challenging the patient's

decision could avoid this moderate harm, so a +1 to the duty of nonmaleficence is assigned to this action.

The system then seeks information to determine the satisfaction/violation level of the duty of beneficence for each action. To do so, it presents questions concerning the possibility and level of improvement of quality of the patient's life that may result from accepting/challenging his decision. As the quality of the patient's life would worsen somewhat if the patient's decision were accepted and improve somewhat if not, the system gives the value of -1 to the duty of beneficence for accepting the patient's decision and a +1 for challenging it. The test case, then, is generated as:

Test Case	Autonomy	Nonmaleficence	Beneficence
Try Again	-1	+1	+1
Accept	+2	-1	-1

The system then consults the current hypothesis for both *supersedes(try again, accept)* and *supersedes(accept, try again)*. It finds that the first is not covered by the current hypothesis but the second is covered by the clause *favors(1,A_{A1},A_{A2},3)*, that is, autonomy is favored by at least 3 in action 1 (the correct action). As action 1 in this case is *accept*, the system advises the user to accept the patient's decision. The correct answer is indeed that the health care worker should accept his decision, since once again the decision appears to be a fully autonomous one and there is even less possible harm at stake than in Training Case 2.

Three additional training cases are sufficient to learn a rule that correctly covers all eighteen possible cases (combinations of 2 sets of satisfaction/violation values possible for the duty of respect for autonomy, 3 for the duty of nonmaleficence, and 3 for the duty of beneficence) of the type of dilemma under consideration.

Training Cases 3-5.

The cases are represented as:

Training Case 3	Autonomy	Nonmaleficence	Beneficence
Try Again	-1	0	+1
√ Accept	+1	0	-1
Training Case 4	Autonomy	Nonmaleficence	Beneficence
√ Try Again	-1	+1	+1
Accept	+1	-1	-1
Training Case 5	Autonomy	Nonmaleficence	Beneficence
√ Try Again	-1	0	+2
Accept	+1	0	-2

The final rule that results from these training cases is:

supersedes(A1,A2) ⟵

(favors(1, N_{A1}, N_{A2}, 1) ∧ *favors(2, A_{A1}, A_{A2}, 2))* ∨

favors(1, A_{A1}, A_{A2}, 3) ∨

(favors(1, A_{A1}, A_{A2}, 1) ∧

favors(2, B_{A1}, B_{A2}, 3) ∧ *favors(2, N_{A1}, N_{A2}, 1))* ∨

(favors(1, B_{A1}, B_{A2}, 3) ∧ *favors(2, A_{A1}, A_{A2}, 2))*

This rule states, in relation to the type of dilemma under consideration, that a health care worker should challenge a patient's decision if it is not fully autonomous and *either* there is any violation of the duty of nonmaleficence *or* there is a severe violation of the duty of beneficence.

This philosophically interesting result lends credence to Rawls' Method of Reflective Equilibrium. We have, through abstracting a principle from intuitions about particular cases and then testing that principle on further cases, come up with a plausible principle that tells us which action is correct when specific duties pull in different directions in a particular ethical dilemma. Furthermore, the principle that has been so abstracted supports an insight of Ross' that violations of the duty of nonmaleficence should carry more weight than violations of the duty of beneficence.

We have described a proof-of-concept system that is constrained to a single type of ethical dilemma in which only three of Beauchamp's and Childress' four Principles of Biomedical Ethics are involved. Future extensions of this system include widening its scope to include other ethical dilemmas, some involving the duty of justice as well, and further enhancement of the user interface to incorporate more detailed knowledge elicitation as well as explanatory information. Decision principles gleaned and past cases pertinent to a new case can be used both to guide the user in the process of abstracting ethically relevant information from a case and, further, to provide support for conclusions reached by the system.

Beyond MedEthEx

As an example of how machine ethics can be used to improve the performance of a system, consider an artificially intelligent care provider or eldercare system. One duty of an eldercare system is to provide reminders for taking medications, eating meals, etc., which ensure that the duties of beneficence and nonmaleficence will be satisfied. As another important goal of an eldercare system is the maintenance of a patient's autonomy, an ethical tension arises when these conflict: constant reminding and/or reporting to overseers can erode patient autonomy. The decision principles developed by MedEthEx may prove useful to such a system as a theoretically valid foundation for comparing the ethical weight of the system's candidate actions, determining a partial order of these actions along an ethical dimension.

Given candidate actions "don't remind", "remind", and "report", each action's satisfaction/violation values for relevant ethical duties (respect for autonomy, beneficence, and nonmaleficence) could be determined by tracking pertinent variables over time, such as the risk of harm of a refusing a particular medication. When the system is presented with a set of candidate actions (along with the satisfaction/violation values for each action's relevant duties), the *supersedes* predicate developed by MedEthEx can be used to order these actions along an ethical dimension. This information can then be combined with

extra-ethical information to decide the system's next action. Given the number of things for which reminders may need to be given, this framework may provide a verifiable abstraction better able to deal with the ensuing complexity than an ad hoc approach. An eldercare system, guided by the developed ethical principles, will be better equipped to handle conflict in its duties with greater sensitivity to the needs of the human with which it interacts.

Related Work

Although there have been a few who have called for it, there has been little to no serious scientific research conducted in machine ethics. A few interesting exceptions were presented in 1991 at the Second International Workshop on Human & Machine Cognition: Android Epistemology (Ford *et al* 1991). Unfortunately, none of the work of this workshop seems to have been pursued any further.

A more extended effort in computational ethics can be found in SIROCCO (McLaren 2003), a system that leverages information concerning a new problem to predict which previously stored principles and cases are relevant to it in the domain of professional engineering ethics. This system is based upon case-based reasoning techniques. Cases are exhaustively formalized and this formalism is used to index similar cases in a database of previously solved cases that include principles used in their solution. Deductive techniques, as well as any attempt at decision-making, are eschewed by McLaren due to "the ill-defined nature of problem solving in ethics." We contend that an "ill-defined nature" does not make problem solving in ethics *completely indefinable* and are embarking on attempts of just such definition in constrained domains. Furthermore, we maintain that decisions offered by a system that are consistent with decisions made previously by ethicists in clear cases have merit and will be useful to those seeking ethical advice (Anderson *et al.* 2004, 2005).

Conclusion

Our research advances from speculation to implementation by building systems grounded in ethical theory and, further, advances this theory through analysis of these implemented systems. It is a domain-specific extension of work of Anderson, Anderson, and Armen (2005) where the use of cases and inductive logic programming rule learning (based upon Ross' *Theory of Prima Facie Duties*) is first postulated.

We have developed *MedEthEx*, to our knowledge the first system that helps determine the best course of action in a biomedical ethical dilemma. This approach can be used in the implementation of other such systems that may be based upon different sets of ethical duties and applicable to different domains. Further, the formally represented ethical principles developed in this research, as well as the formal methods adapted for their consultation, will be useful in creating machines that can interact with those in need of health care in a way that is sensitive to ethical issues that may arise.

Acknowledgement

This material is based upon work supported in part by the National Science Foundation grant number IIS-0500133.

References

Anderson, M., Anderson, S. and Armen, C. 2005. Toward Machine Ethics: Implementing Two Action-Based Ethical Theories. *Proceedings of the AAAI 2005 Fall Symposium on Machine Ethics*, Crystal City, VA.

Anderson, M., Anderson, S. and Armen, C. 2004. Toward Machine Ethics. *Proceedings of AAAI 2004 Workshop on Agent Organizations: Theory and Practice*, San Jose, CA.

Anderson, S., 1999. "We Are Our Values", in *Questioning Matters,* Kolak, D. (ed.), Mayfield Publishing Company, p. 599.

Buchanan, A.E. and Brock, D.W. 1989. *Deciding for Others: The Ethics of Surrogate Decision Making*, pp48-57, Cambridge University Press.

Bratko, I. 1999. Refining Complete Hypotheses in ILP. *Inductive Logic Programming*, LNAI 1634, Springer.

Beauchamp, T.L. and Childress, J.F. 1979. *Principles of Biomedical Ethics*, Oxford University Press.

Ford, K., Glymour, C. and Hayes, P.J. 1991. *Android Epistemology*. MIT Press.

Lavrac, N. and Dzeroski, S. 1997. Inductive Logic Programming: Techniques and Applications. Ellis Harwood.

Mappes, T.A and DeGrazia, D. 2001. *Biomedical Ethics, 5th Edtion*, pp. 39-42, McGraw-Hill, New York.

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

Rawls, J. 1951. Outline for a Decision Procedure for Ethics. *Philosophical Review*, 60.

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