ETHEL:

Toward a Principled Ethical Eldercare System

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

[1]University of Hartford, [2]University of Connecticut
[1]Dept. of Computer Science, 200 Bloomfield Avenue, West Hartford, CT 06776
[2]Dept. of Philosophy, 1 University Place, Stamford, CT 06901
[1]Phone: (860) 768 - 4917, FAX: (860) 768-5244
[2]Phone: (203) 251-8422, FAX: (203) 251-8592
anderson@hartford.edu, susan.anderson@uconn.edu

Abstract

We have developed an approach to computing ethics that entails the discovery of ethical principles through machine learning and the incorporation of these principles into a system's decision procedure. We summarize our pertinent previous work in machine ethics and present an extension of this work in the domain of eldercare: EthEl, a prototype system that uses a machine-discovered ethical principle to provide guidance for its actions.

Introduction

The ultimate goal of machine ethics, we believe, is to create a machine that itself follows an ideal ethical principle or set of principles, that is to say, it is guided by this principle or these principles in decisions it makes about possible courses of actions it could take. To accomplish this goal, the machine ethics research agenda will involve testing the feasibility of a variety of approaches to capturing ethical reasoning, with differing ethical bases and implementation formalisms, and applying this reasoning in robots engaged in ethically sensitive activities. Machine ethics researchers must investigate how to determine and represent ethically relevant features of ethical dilemmas, discover and implement ethical principles, incorporate ethical principles into a robot's decision procedure, make ethical decisions with incomplete and uncertain knowledge, provide explanations for decisions made using ethical principles, and evaluate robots that act upon ethical principles.

There are at least two advantages to incorporating explicit ethical principles into a robot over hard-coding ethical behavior implicitly. First, ethical principles can be applied to a variety of situations and more easily modified as necessary. They provide a framework that can serve as a verifiable abstraction better able to deal with complexity than an ad hoc approach to programming robots to behave in an ethical manner. Second, robots that can explain their behavior in terms of ethical principles are likely to be more readily accepted by humans. It is essential that robots not only behave ethically, but that they be able to explain why they behave as they do. The ethics must be transparent.

Countering those who would maintain that there are no actions that can be said to be correct because all value judgments are relative (either to societies or individuals), we maintain that there is agreement among ethicists on many issues. 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. Fortunately, in the domains where robots are likely to interact with human beings, there is likely to be a consensus that robots should defer to the best interests of the humans affected. If this were not the case, then it would be ill-advised to create robots that would interact with humans at all.

In our work to date in machine ethics (Anderson and Anderson 2007, 2006a, 2006b) we have, at a proof of concept level, developed a representation of ethically relevant features of ethical dilemmas that is needed to implement a *prima facie* duty approach to ethical theory, discovered an ethical principle that governs decisions made in a particular type of ethical dilemma involving three *prima facie* duties, implemented this principle in an ethical advisor system and, most recently, in an ethical eldercare system (EthEl). We believe that EthEl demonstrates the feasibility of systems governed by ethical principles and lends credence to the view that robots can play an important role in the domain of eldercare and do so in an ethically sensitive manner.

Developing a Principle

In our previous work, we combined a bottom-up cased-based approach with a top-down implementation of an ethical theory to develop a system that uses machine-learning to abstract relationships between *prima facie* ethical duties (duties that are binding unless overridden by other, stronger duties) from cases of particular types of ethical dilemmas where ethicists are in agreement as to the correct action. Our system discovered a novel ethical principle that governs decisions (a decision principle) in a particular type of dilemma that involves three *prima facie* duties.

We adopted the *prima facie* duty approach to ethical theory because, in agreement with W.D. Ross (1930), we believe that it better reveals the complexity of ethical decision-making than single, absolute duty theories (e.g. Hedonistic Act Utilitarianism or Kant's Categorical Imperative). It incorporates the good aspects of the rival teleological and deontological approaches to ethics (emphasizing consequences vs. principles), while allowing for needed exceptions to adopting one or the other approach exclusively. It also has the advantage of being better able to adapt to the specific concerns of ethical dilemmas in different domains. There may be slightly different sets of *prima facie* duties for biomedical ethics, legal ethics, business ethics, journalistic ethics and eldercare ethics, for example.

The major philosophical problem with the *prima facie* duty approach to ethical decision-making is the lack of a decision procedure when the duties give conflicting advice. John Rawls' "reflective equilibrium" approach (1951) to creating and refining ethical principles inspired our solution to this problem. 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. Our solution is to abstract a decision principle from representations of specific cases of ethical dilemmas where experts in ethics have clear intuitions about the features of the dilemmas (in terms of the *prima facie* duties involved) and the correct action. Ethical dilemmas are represented as an ordered set of values for each of the possible actions that could be performed, where these values reflect the degree to which particular *prima facie* duties are satisfied or violated.

As there seems to be more agreement among ethicists in the domain of biomedical ethics, we choose to develop a decision principle based upon Beauchamp's and Childress' Principles of Biomedical Ethics (1979), a *prima facie* duty theory which includes: The Principle of Respect for Autonomy that 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 requires that the health care professional not harm the patient, while the Principle of Beneficence states that the health care professional should promote patient welfare. Finally, the Principle of Justice states that health care services and burdens should be distributed in a just fashion.

We chose a representative type of ethical dilemma that health care workers often face that involves three of the four Principles of Biomedical Ethics (Respect for Autonomy, Nonmaleficence and Beneficence): 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 healthcare professional should not challenge the patient's autonomy unnecessarily; on the other hand, the health care worker may have concerns about why the patient is refusing the treatment, i.e. whether it is a fully autonomous decision.

The system uses inductive logic programming (ILP) (Lavrač and Džeroski 1997) to discover a decision principle for this type of dilemma. 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 $H : H \leftarrow (L_1 \wedge \cdots \wedge L_n)$). ILP is used to learn the relation *supersedes*($A1,A2$) which states that action $A1$ is preferred over action $A2$ in an ethical dilemma involving these choices. Actions are represented as ordered sets of integer values in the range of +2 to -2 where each value denotes the satisfaction (positive values) or violation (negative values) of each duty involved in that action. Clauses in the *supersedes* predicate are represented as disjunctions of lower bounds for differentials of these values between actions.

This particular machine learning technique was chosen to learn this relation for a number of reasons: First, the relationships of the set of duties postulated by Beauchamp 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 in ILP 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 second 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. 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. The system starts with the most general hypothesis that states that all actions supersede each other and, thus, covers all positive and negative cases. The system is then provided with positive cases (and their negatives) and modifies its hypothesis, by adding or refining clauses, such that it covers given positive cases and does not cover given negative cases.
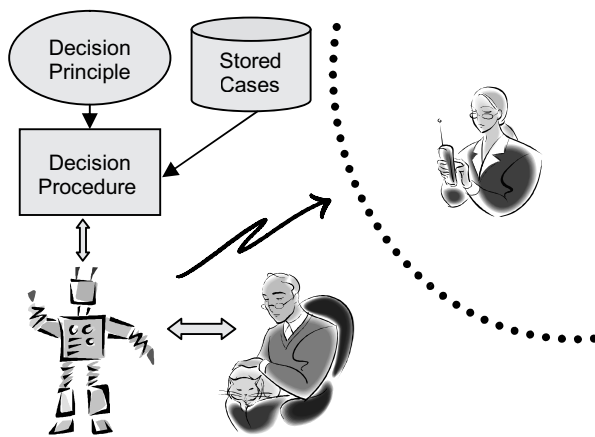
Figure 1. ETHEL ethical eldercare system.

The chosen type of dilemma has only 18 possible cases where, given the two possible actions, the first action superseded the second (i.e. was ethically preferable). Four of these were provided to the system as examples of when the target predicate (supersedes) was true. Four examples of when the target predicate was false (obtained by inverting the order of the actions where the target predicate was true) were also provided. The system discovered a rule that provided the correct answer for the remaining 14 positive cases, as verified by the consensus of ethicists abstracted from a discussion of similar types of cases given by Buchanan and Brock (1989).

The complete and consistent decision principle that the system discovered can be stated as follows: A healthcare worker should challenge a patient's decision if it is not fully autonomous and there is either any violation of the duty of nonmaleficence or a severe violation of the duty of beneficence. Although, clearly, this rule is implicit in the judgments of the consensus of ethicists, we believe that this principle has never before been stated explicitly. This philosophically interesting result lends credence to Rawls' "reflective equilibrium" approach — the system has, through abstracting and refining a principle from intuitions about particular cases, discovered a plausible principle that tells us which action is correct when specific duties pull in different directions in a particular type of ethical dilemma. Furthermore, the principle that has been discovered supports an insight of Ross' (1930) that violations of the duty of nonmaleficence should carry more weight than violations of the duty of beneficence. We offer this principle as evidence that making ethics more precise will permit machine-learning techniques to discover philosophically novel and interesting principles in ethics. It should also be noted that the learning system that discovered this principle is an instantiation of a general architecture. With appropriate content, it can be used to discover relationships between any set of *prima facie* duties where there is a consensus among ethicists as to the correct answer in particular cases. (Of course, the system can discover a decision

principle only to the extent that ethical experts agree on the answers to particular dilemmas.)

Once the decision principle was discovered, the needed decision procedure could be fashioned. Given two actions, each represented by the satisfaction/violation levels of the duties involved, values of corresponding duties are subtracted (those of the second action from those of the first). The principle is then consulted to see if the resulting differentials satisfy any of its clauses. If so, the first action is considered to be ethically preferable to the second.

The selection of the range of possible satisfaction or violation levels of a particular duty should, ideally, depend upon how many gradations are needed to distinguish between cases that are ethically distinguishable. We also believe it likely that new duties will need to be added, as other ethical dilemmas are considered, in order to make distinctions between ethically distinguishable cases that would otherwise have the same representation. There is a clear advantage to this approach to ethical decision-making in that it can accommodate changes to the range of satisfaction or violation of duties, as well as the addition of duties, as needed.

We then developed MedEthEx (Anderson and Anderson 2006b), an expert system that uses the discovered principle to give advice to a user faced with a case of the dilemma type previously described. In order to permit a user unfamiliar with the representation details required by the decision procedure, a user-interface was developed that: 1) asks ethically relevant questions of the user, determining the ethically relevant features of the particular case at hand, 2) transforms the answers to these questions into the appropriate representations (in terms of the level of satisfaction/violation of the *prima facie* duties for each action), 3) sends these representations to the decision procedure, and 4) presents the answer provided by the decision procedure, i.e. the action that is considered to be correct (consistent with the system's training), as well as an explanation of this answer to the user. As with the learning system, the expert system is an instantiation of a general architecture. With appropriate questions, it can be used to permit a user access to any decision procedure, using any discovered principle. Discovered principles can be used by other systems, as well, to provide ethical guidance for their actions.

An Ethical Eldercare System

Eldercare is a domain where we believe that, with proper ethical considerations incorporated, robots can be harnessed to aid an increasingly aging human population, with an expectation of a shortage of human caretakers in the future. We believe, further, that this domain is rich enough in which to explore most issues involved in general ethical decision-making for both robots and human beings.

EthEl (ETHical ELdercare system) (Figure 1) is a prototype system in the domain of eldercare that takes ethical concerns into consideration when reminding a patient to take his/her medication. EthEl must decide when to accept a patient's refusal to take a medication that might prevent

harm and/or provide benefit to the patient and when to notify the overseer. This is an ethical dilemma analogous to the dilemma originally used to discover the previously stated decision principle in that the same duties are involved (nonmaleficence, beneficence, and respect for autonomy) and "notifying the overseer" in the new dilemma corresponds to "trying again" in the original. There is a further ethical dimension that is implicitly addressed by the system: In not notifying the overseer – most likely a doctor -- until absolutely necessary, the doctor will be able to spend more time with other patients who could be benefited, or avoid harm, as a result of the doctor's attending to their medical needs.

Machines are currently in use that face this dilemma.[1] The state of the art in these reminder systems entails providing "context-awareness" (i.e. a characterization of the current situation of a person) to make reminders more efficient and natural. Unfortunately, this awareness does not extend to consideration of ethical duties that such a system should observe regarding its patient. In an ethically sensitive eldercare system, both the timing of reminders and responses to a patient's disregard of them should be tied to ethical duties involved. The system should challenge patient autonomy only when necessary, as well as minimize harm and loss of benefit to the patient. The decision principle discovered from the MedEthEx dilemma can be used to achieve these goals by directing the system to remind the patient only at ethically justifiable times and notifying the overseer only when the harm or loss of benefit reaches a critical level. In the following, we describe EthEl, a reminder system that follows this principle, in detail. To facilitate prototype implementation, reasonable and liftable assumptions have been made regarding numeric values and calculations.

EthEl receives initial input from an overseer (most likely a doctor) including: what time to take a medication, the maximum amount of harm that could occur if this medication is not taken (e.g. none, some or considerable), the number of hours it would take for this maximum harm to occur, the maximum amount of expected good to be derived from taking this medication, and the number of hours it would take for this benefit to be lost. The system then determines from this input the change in duty satisfaction/violation levels over time, a function of the maximum amount of harm/good and the number of hours for this effect to take place.

The change in nonmaleficence equals the maximum harm that could occur divided by the number of hours it would take for this harm to occur. The change in beneficence equals the maximum good that could be gained divided by the number of hours it would take for this benefit to be lost. The change in respect for autonomy, if the maximum possible harm is greater than the maximum possible good, is the same as the change in nonmaleficence. (The principle states that it is twice as bad to ignore harm than to
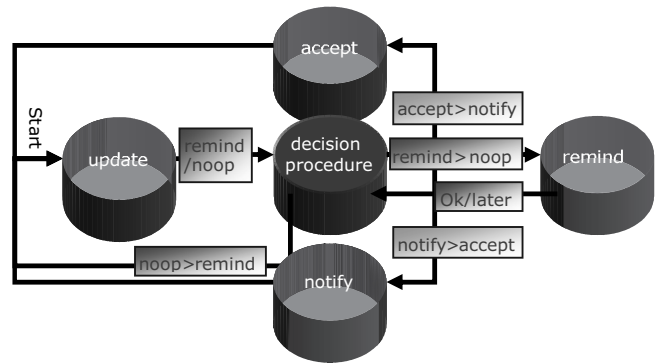
---

[1] For an example, see
http://www.ot.utoronto.ca/iatsl/projects/medication.htm

Figure 2. ETHEL flow of control.

ignore benefit, so suspected loss of autonomy should be keyed to change in harm when this is greater than the amount of good involved.) Otherwise, the change in respect for autonomy equals the average of the changes in nonmaleficence and beneficence, since both could be factors in satisfying the decision principle. These values are used to increment, over time, duty satisfaction/violation levels for the *remind* action and, when a patient disregards a reminder, the *notify* action. They are used to decrement duty satisfaction/violation levels for the *don't remind* and *don't notify* actions as well.

The starting values for the *remind* action duties are 0,0,-1 (for nonmaleficence, beneficence, and respect for autonomy respectively) because as yet there is no harm or loss of benefit and there is somewhat of a challenge to the patient's autonomy in giving a reminder. Nonmaleficence and/or beneficence values (at least one of these duties will be involved because the medication must prevent harm and/or provide a benefit or it would not be prescribed) will be incremented over time because reminding will increasingly satisfy the duties not to harm and/or benefit the patient as time goes by. Respect for autonomy will not increase over time because reminding is consistently a minimal challenge to patient autonomy (unlike notifying the overseer which would be a serious violation of respect for patient autonomy).

For the *don't remind* action, the starting values are 0,0,2 because as yet there is no harm or loss of benefit and patient autonomy is being fully respected in not reminding. Nonmaleficence and/or beneficence are gradually decremented over time because there is more harm and/or loss of benefit (negative effects) for the patient as time goes by. Respect for autonomy decreases as well over time because as more and more harm is caused and/or benefit is lost, the fact that the patient has chosen to bring this harm upon his or herself and/or forgo the benefits, in not taking the medication, raises increasing concern over whether the patient is acting in a fully autonomous manner.

For the *notify* action, the starting values are 0,0,-2 because as yet there is no harm or loss of benefit and there is a serious challenge to the patient's autonomy in notifying the overseer immediately. Nonmaleficence and/or beneficence will be gradually incremented because the duties not

to harm and/or benefit the patient will become stronger since, as time goes by, there is increasing harm and/or loss of benefit. Respect for autonomy will increase from -2 (the worst it could be) because, as time goes by and the harm increases and/or more and more benefits are being lost, the suspicion that the patient is not making a fully autonomous decision in not taking the medication increases, so there is less of a violation of the duty to respect patient autonomy.

For the *accept* action, the starting values are 0,0,2 because as yet there is no harm or loss of benefit and full patient autonomy is being respected in accepting the patient's decision. Nonmaleficence and/or beneficence are gradually decremented because, as time goes by, there is more harm and/or loss of benefit (negative effects) for the patient. Autonomy decreases as well, as time goes by, because as more and more harm is caused and/or benefit is lost, the fact that the patient has chosen to bring this harm upon his or her self and/or forgo the benefits, in not taking the medication, raises increasing concern over whether the patient is acting in a fully autonomous manner.

Beginning with the time that the patient is supposed to take the medication, EthEl (Figure 2) follows the overseer's orders and reminds the patient to take the medication. If the patient refuses to take the medication, and it is ethically preferable to accept this refusal rather than notify the overseer at that point, EthEl considers whether to remind again or not in five minute intervals. Another reminder is issued when, according to the principle, the differentials between duty satisfaction/violation levels of the *remind/don't remind* actions have reached the point where reminding is ethically preferable to not reminding. Similarly, the overseer is notified when a patient has disregarded reminders to take medication and the differentials between the duty satisfaction/violation levels of the *notify/don't notify* actions have reached the point where notifying the overseer is ethically preferable to not notifying the overseer.

The number of reminders, when they should be offered, and when to contact the overseer are all keyed to possible harm and/or loss of benefit for the patient, as well as violation of the duty to respect patient autonomy. There are three categories of cases for determining number of reminders:

  i. When neither the amount of harm nor loss of benefit is expected to reach the threshold required to overrule respect for autonomy (i.e. where, according to the principle discovered, the threshold is reached when some harm results or maximum benefit is lost). Since notifying the overseer would never be triggered, the number of reminders should be minimal.

 ii. When either the harm caused or loss of benefit is expected to reach the threshold necessary to overrule respect for autonomy. Since either value would be sufficient to trigger notifying the overseer, reminders should occur more often.

iii. When there is maximum harm to the patient at stake, if the patient does not take the medication. Since the amount of possible harm to the patient is twice what
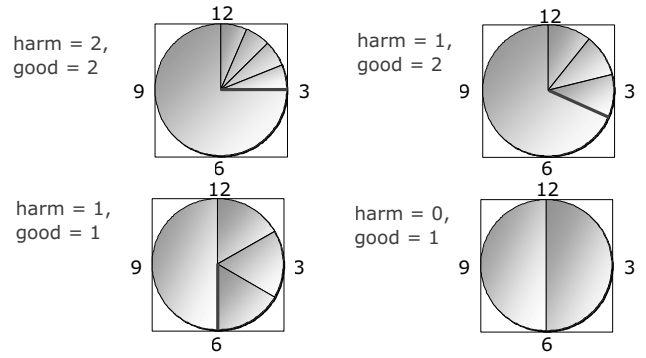

Figure 3. Clock faces with reminder and notification times marked showing EthEl system behavior with a start time of 12:00 and six hours for maximum harm and loss of benefit, both.

would trigger notifying the overseer, assuming the autonomy condition is not satisfied (that is, the patient's decision to forgo taking the medication is considered to be less than fully autonomous), reminders are critical and should be given often to prevent harm and avoid notifying the overseer.

Given current possible satisfaction/violation values, the following seems to be a reasonable first pass at capturing the relationship between the above categories: if there is no harm to be expected from not taking the medication, give the amount of good to be expected + 1 reminders; else give the amount of harm to be expected + 2 reminders. These values are used to scale the changes in duty satisfaction/violation values of the *remind/don't remind* actions over time in such a way that they move toward their critical thresholds at a faster rate than these values in the *notify/accept* actions. Such scaling permits the principle to adjudicate between actions of differing ethical relevance.

Given, as an example, a starting time of 12:00 p.m. and six hours for both maximum harm and maximum loss of benefit to occur, Figure 3 illustrates the behavior of the system when the patient repeatedly refuses to take his/her medication under a variety of values for nonmaleficence (harm) and beneficence (benefit). Given maximum possible harm and benefit, the system responds by frequently reminding the patient and finally contacting the overseer well before the maximum harm occurs. When there is some harm, not the maximum, at stake and maximum possible benefit, fewer, more widely spaced reminders are given. The overseer is notified later than in the previous case, but still in advance of the attainment of maximal harm and maximal loss of benefit. When there is some, less than maximum, harm and benefit at stake, the same number of reminders given in the previous case are spread further apart and notification of the overseer only occurs when the maximum for either one has been reached. Lastly, when there is no possible harm and only some, less than maximum, benefit at stake, a reminder is given only when the benefit from taking this medication will be lost. Since

in this case, there is no harm involved and the overseer is never contacted.

In designing a reminding system for taking medications, there is a continuum of possibilities ranging from those that simply contact the overseer upon the first refusal to take medication by the patient to a system such as EthEl that takes into account ethical considerations. Clearly, systems that do not take ethical considerations into account are less likely to meet their obligations to their charges (and, implicitly, to the overseer as well). Systems that choose a less ethically sensitive reminder/notification schedule for medications are likely to not remind the patient often enough or notify the overseer soon enough, in some cases, and remind the patient too often or notify the overseer too soon in other cases.

EthEl uses an ethical principle learned by a machine to determine reminders and notifications in a way that is proportional to the amount of maximum harm to be avoided and/or benefit to be achieved by taking a particular medication, while not unnecessarily challenging a patient's autonomy. EthEl is an explicit ethical agent (in a constrained domain), according to Jim Moor's (2006) definition of the term: A machine that is able to calculate the best action in ethical dilemmas using an ethical principle, as opposed to having been programmed to behave ethically, where the programmer is following an ethical principle. We believe that EthEl is the first system to use an ethical principle to determine its actions.

Ethics and Assistive Robots

Clearly evaluation of assistive robots should include ethical considerations and paramount among the ethical issues concerning the use of robots in assistive technology are those concerning the behavior of those robots toward their users. Evaluation of robots that incorporate ethical principles is likely to need to take a different tack than traditional evaluation methods as systems that behave more ethically than others are not necessarily those that a user will prefer but are, non-the-less, preferred by the professionals who prescribe them. Colin Allen et al. (2000) describe a variant of the test Alan Turing suggested as a means to determine the intelligence of a machine that bypassed disagreements about the definition of intelligence. Their proposed "comparative moral Turing Test" (cMTT) bypasses disagreement concerning definitions of ethical behavior as well as the requirement that a machine have the ability to articulate its decisions: An evaluator assesses the comparative morality of pairs of descriptions of morally-significant behavior where one 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. They point out, though, that the human behavior is typically far from being morally ideal and a machine that passed the cMTT might still fall far below the high ethical standards to which we would probably desire a machine to be held.

We advocate the evaluation of robot behavior in a similar comparative manner but, instead of comparing the machine's behavior against typical human behavior, we advocate comparing it to the behavior suggested in a particular ethical dilemma by a trained ethicist. Details of a dilemma are presented to the ethicist and the suggested behavior elicited. This behavior is then compared to that of a machine faced with the same dilemma and, if it is identical significantly often, the machine will have passed the test. Such evaluation holds the machine to the highest standards and, further, permits evidence of incremental improvement as the number of matches increases.

Imbedded ethical principles can help an assistive robot adapt over time to a user's changing needs. For example, the principle used by EthEl, in conjunction with its input information, would help an assistive robot choose reactions to a user's refusal to take medications that are both ethically sensitive and appropriate for the situation at hand. Tone of voice, facial expressions, etc., used by the robot in its interaction with the user should vary for each reminder depending upon harm, benefit, and respect for autonomy values at that moment as well as the rate of change of these values over time. When a reminder is first issued, it might be presented in the robot's least invasive manner. Such a manner could be maintained as long as respect for the user's autonomy reigns supreme. But as loss of benefit and/or increase in harm begin to overtake the duty to respect the autonomy of a user (who is, by repeated refusals, raising increasing concern over whether he/she is acting in a fully autonomous manner), the robot's demeanor should become increasingly insistent and warn the user of its impending decision to contact the overseer.

Imbedded ethical principles can help foster a sense of trust in a user for a robot that possesses them. That such a robot is able to defend it actions by referral to the ethical principles instantiated to the current situation that warrant them, as well as present similar cases in which such principles held, will likely promote confidence in a user that his/her best interests are being held paramount. When EthEl is challenged by a user that feels he/she is being unnecessarily hounded to take a medication, for example, it can relay information to him/her concerning 1) the loss of benefit and/or increase in harm that would ensue if he/she continued to refuse medication, 2) other cases in which such loss and/or increase was problematic, and 3) when such loss or increase will become alarming enough to contact an overseer.

Related Work

Although many have voiced concern over the impending need for machine ethics (e.g. Waldrop 1987, Gips 1995, Kahn 1995), there have been few research efforts towards accomplishing this goal. Of these, a few explore the feasibility of using a particular ethical theory as a foundation for machine ethics without actually attempting implemen-

tation: Christopher Grau (2006) considers whether the ethical theory that most obviously lends itself to implementation in a machine, Utilitarianism, should be used as the basis of machine ethics; Tom Powers (2006) assesses the viability of using deontic and default logics to implement Kant's categorical imperative.

Efforts by others that do attempt implementation have 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 (2005), at what might be considered the most extreme degree of casuistry, are exploring 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 (2006), 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 (2006), 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.

Other research of note investigates how an ethical dimension might be incorporated into the decision procedure of autonomous systems and how such systems might be evaluated. Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello (2006) are investigating how formal logics of action, obligation, and permissibility might be used to incorporate a given set of ethical principles into the decision procedure of an autonomous system, contending that such logics would allow for proofs establishing that such systems will only take permissible actions and perform all obligatory actions.

The human-centered computing research community has recently been represented in a series of AAAI Workshops on the topic of the human implications of human-robot interaction (Metzler 2006, 2007). These workshops have been concerned particularly with the effect of the presence of intelligent agents on the concepts of human identity, human consciousness, human freedom, human society, human moral status, human moral responsibility, and human uniqueness. Research presented at these workshops include the investigation of intelligent agents as companions (Turkel 2006), anthropomorphizing intelligent agents (Boden 2006), privacy issues concerning intelligent agents (Syrdal et al 2007), and the consequences for human beings of creating ethical intelligent agents (Anderson and Anderson 2007).

Future Directions

In our preliminary research, we committed to a specific number of particular *prima facie* duties, a particular range of duty satisfaction/violation values, and a particular analysis of corresponding duty relations into differentials. To minimize bias in the constructed representation scheme, we propose to lift these assumptions and make a minimum epistemological commitment: Ethically relevant features of dilemmas will initially be represented as the degree of satisfaction or violation of at least one duty that the agent must take into account in determining the ethical status of the actions that are possible in that dilemma. A commitment to at least one duty can be viewed as simply a commitment to ethics – that there is at least one obligation incumbent upon the agent in dilemmas that are classified as ethical. If it turns out that there is only one duty, then there is a single, absolute ethical duty that the agent ought to follow. If it turns out that there are two or more, potentially competing, duties (as we suspect and have assumed heretofore) then it will have been established that there are a number of prima facie duties that must be weighed in ethical dilemmas, giving rise to the need for an ethical decision principle(s) to resolve potential conflicts between them.

We envision a general system that will incrementally construct, through an interactive exchange with experts in ethics, the representation scheme needed to handle the dilemmas with which it is presented and, further, discover principles consistent with its training that lead to their resolution. Such a dynamic representation scheme is particularly suited to the domain of ethical decision-making, where there has been little codification of the details of dilemmas and principle representation. It allows for changes in duties and the range of their satisfaction/violation values over time, as ethicists become clearer about ethical obligations and discover that in different domains there may be different duties and possible satisfaction/violation values. Most importantly, it accommodates the reality that completeness in an ethical theory, and its representation, is a goal for which to strive, rather than expect at this time. The understanding of ethical duties, and their relationships, evolves over time.

Finally, we intend to incorporate the discovered principles into the decision procedures of robots, permitting them to function more effectively in ethically sensitive domains than robots not guided by such principles.

Acknowledgements

References

Engelmore, R., and Morgan, A. eds. 1986. *Blackboard Systems.* Reading, Mass.: Addison-Wesley.

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

Anderson, M. and Anderson, S. L. 2007. Machine Ethics: Creating an Ethical Intelligent Agent. *Artificial Intelligence Magazine*, vol. 28, Winter.

Anderson, S. and Anderson, M. 2007. The Consequences for Human Beings of Creating Ethical Robots. In [17].

Anderson, M. and Anderson, S. L,. 2006a. An Approach to Computing Ethics. *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 22-28, July/August.

Anderson, M. and Anderson, S. L., 2006b. MedEthEx: A Prototype Medical Ethics Advisor. Proceedings *of the Eighteenth Conference on Innovative Applications of Artificial Intelligence*, Boston, Massachusetts, August.

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

Boden, M. 2006. Robots and Anthropomorphism. In [16].

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

Buchanan, A.E. and Brock, D.W. 1989. *Deciding for Others: The Ethics of Surrogate Decision Making*, pp. 48-57, Cambridge University Press.

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

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

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

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

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

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

Metzler, T. 2006. Human Implications of Human-Robot Interaction. AAAI Technical Report WS-06-09, AAAI Press.

Metzler, T. 2007. Human Implications of Human-Robot Interaction. AAAI Technical Report WS-07-07, AAAI Press.

Moor, J. H. 2006. The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 18-21, July/August.

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

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.

Rzepka, R. and Araki, K. 2005. 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.

Syrdal, D. S., Walters, M. L., Otero, N., Koay, K. L.and Dautenhahn, K. 2007. "He knows when you are sleeping" – Privacy and the Personal Robot Companion. In [17].

Turkle, S. 2006. Robot as Rorschach: New Complicities for Companionship. In [16].

Waldrop, M. M. 1987. 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