Towards Machine Ethics

Michael Anderson
University of Hartford
Department of Computer Science
West Hartford, CT
Anderson@Hartford.edu

Susan Leigh Anderson
University of Connecticut
Department of Philosophy
Stamford, CT
Susan.Anderson@UConn.edu

Chris Armen
Trinity College
Department of Computer Science
Hartford, CT
Chris.Armen@TrinColl.edu

Abstract

We contend that the ethical ramifications of machine behavior, as well as recent and potential developments in machine autonomy, necessitate adding an ethical dimension to at least some machines. We lay the theoretical foundation for m*achine ethics* by discussing the rationale for, the feasibilty of, and the benefits of adding an ethical dimension to machines. Finally, we present details of prototype systems and motivate future work.

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. This is evidenced in the Ten Commandments of Computer Ethics advocated by the Computer Ethics Institute of The Brookings Institution (Barquin 1992), where admonishments such as "Thou Shalt Not Use A Computer To Harm Other People" and "Thou Shalt Not Use A Computer To Steal" speak to this human-centered perspective. 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, as well as recent and potential developments in machine autonomy, necessitates this. We explore this dimension through investigation of what has been called *machine ethics*. In contrast to computer hacking, software property issues, privacy issues and other topics normally ascribed to *computer* ethics, m*achine* ethics is concerned with the consequences of behavior of machines towards human users and other machines.

In the following, we lay the theoretical foundation for machine ethics by discussing the rationale for, the feasibility of, and the benefits of adding an ethical dimension to machines. Finally, we present details of prototype systems and motivate the next steps of our research.

Rationale

Not only have machines conquered chess (Deep Blue[1]), but speech understanding programs are used to handle reservations for airlines (Pegasus[2]), expert systems monitor spacecraft (MARVEL[3]) and diagnose pathology (PathFinder (Heckerman et al. 1992)), robotic systems have been taught to drive and have driven across the country (NavLab[4]), unmanned combat jets are flying (X-45A UCAV[5]) and more. There is no limit to the projections that have been made for such technology – from cars that drive themselves and machines that discharge our daily chores with little assistance from us, to fully autonomous robotic entities that will begin to challenge our notions of the very nature of intelligence. Behavior involving all of these systems may have ethical ramifications, some due to the advice they give and others due to their own autonomous behavior.

Clearly, relying on machine intelligence to effect change in the world without some restraint can be dangerous. Until fairly recently, the ethical impact of a machine's actions has either been negligible, as in the case of a calculator, or, when considerable, has only been taken under the supervision of a human operator, as in the case of automobile assembly via robotic mechanisms. As we increasingly rely upon machine intelligence with reduced human supervision, we will need to be able to count on a certain level of ethical behavior from them. The fact that we *will* increasingly rely on machine intelligence follows from a simple projection of our current reliance to a level of reliance fueled by market pressures to perform faster, better, and more reliably.

As machines are given more responsibility, an equal measure of accountability for their actions must be meted out to them. Ignoring this aspect risks undesirable machine behavior. Further, we may be missing an opportunity to harness new machine capabilities to assist us in ethical decision-making.

Feasibility

Fortunately, there is every reason to believe that ethically sensitive machines can be created. An approach to ethical

[1] http://www.research.ibm.com/deepblue/
[2] http://www.sls.csail.mit.edu/PEGASUS.html
[3] http://voyager.jpl.nasa.gov/Proposal-2003/VgrTech.pdf
[4] http://www.ri.cmu.edu/labs/lab_28.html
[5] http://www.boeing.com/phantom/ucav.html

decision-making that dominated ethical theory from Kant through the mid-twentieth century – *action-based ethics* (where the emphasis is on telling us how we should act in an ethical dilemma) – lends itself to machine implementation. Action-based theories are rule governed and, besides agreeing with intuition, these rules must be consistent, complete, and practical (Anderson 2000)[1]. As John Stuart Mill said in *Utilitarianism*, for an action-based theory to have a chance of being consistent:

> There ought either to be some one fundamental principle or law…or if there be several, there should be a determinant order of precedence among them… [a] rule for deciding between the various principles when they conflict…. (Mill 1974)

A further condition of consistency is that this one principle or rule for deciding between principles when there are several that might conflict should never tell us, in a given situation, that a particular action is both right and wrong. There should always be a single answer to the question: In the given ethical dilemma, is this action right or wrong?

To say that an action-based ethical theory is complete means that it does all that it's supposed to do, that is, it tells us how we should act in any ethical dilemma in which we might find ourselves. The added requirement of practicality ensures that it is realistically possible to follow the theory. Consider, for example, a variation on the Act-Utilitarian theory which we will explore shortly. A theory which would have us do whatever would *in fact* (rather than *is likely to*) result in the best consequences is not practical because there is no way that we can know beforehand what will happen, how things will turn out.

Consistent, complete and practical rules lend themselves to an algorithmic formulation that is necessary for a machine implementation. Consistency, in computer science terms, means that the algorithm is *deterministic;* informally, this means that given a particular set of inputs, the algorithm will always come to the same conclusion. Complete, in computer science terms, means that the algorithm will produce valid output for all valid input. Practicality has two interpretations from the computer science perspective: (1) the input to the algorithm is well-defined and available, and (2) the algorithm can be implemented efficiently; i.e., it will reach a conclusion in a reasonable amount of time, where "reasonable" can be characterized mathematically.

As a first step towards showing that an ethical dimension might be added to certain machines, let us consider the possibility of programming a machine to follow the theory of Act Utilitarianism, a theory that is consistent, complete and practical. According to this theory that act is right which, of all the actions open to the agent, is likely to result in the greatest net good consequences, taking all those affected by the action equally into account. Essentially, as Jeremy Bentham long ago pointed out, the theory involves performing "moral arithmetic" (Bentham 1799). A machine is certainly capable of doing arithmetic. Of course, before doing the arithmetic, one needs to know what counts as a "good" and "bad" consequence. The most popular version of Act Utilitarianism – Hedonistic Act Utilitarianism – would have us consider the pleasure and displeasure that those affected by each possible action are likely to receive. And, as Bentham pointed out, we would probably need some sort of scale (e.g. from 2 to -2) to account for such things as the intensity and duration of the displeasure or pleasure that each individual affected is likely to receive. But this is information that a human being would need to have as well to follow the theory. Given this information, a machine could be developed that is just as able to follow the theory as a human being.

Hedonistic Act Utilitarianism can be implemented in a straightforward manner. The algorithm is to compute the best action, that which derives the greatest net pleasure, from all alternative actions. It requires as input the number of people affected, and for each person, the intensity of the pleasure/displeasure (e.g. on a scale of 2 to -2), the duration of the pleasure/displeasure (e.g. in days), and the probability that this pleasure/displeasure will occur for each possible action. For each person, the algorithm simply computes the product of the intensity, the duration, and the probability, to obtain the net pleasure for each person. It then adds the individual net pleasure to obtain the Total Net Pleasure:

$$\textit{Total Net Pleasure} = \sum \textit{(Intensity} \times \textit{Duration} \times \textit{Probability) for each affected individual}$$

This computation would be performed for each alternative action. The action with the highest Total Net Pleasure is the right action.

In fact, the machine might have an advantage over a human being in following the theory of Act Utilitarianism for several reasons: First, human beings tend not to do the arithmetic strictly, but just estimate that a certain action is likely to result in the greatest net good consequences, and so a human being might make a mistake, whereas such error by a machine would be less likely. Second, human beings tend towards partiality (favoring themselves, or those near and dear to them, over others who might be

[1] In more recent years, there has been a revival of *virtue-based ethics*, where the emphasis is on what sort of persons we should be, rather than how we should act. But it's not clear that this would force us to replace action-based ethics with virtue-based ethics since, as William Frankena has argued:

> we [should] regard the morality of duty and principle and the morality of virtues and traits of character not as rival kinds of morality between which we must choose, but as two complementary aspects of the same morality….for every principle there will be a morally good trait…and for every morally good trait there will be a principle determining the kind of action in which it is to express itself. (Frankena 1993)

affected by their actions or inactions), whereas an impartial machine could be devised. Since the theory of Utilitarianism was developed to introduce objectivity into ethical decision-making, this is important. Third, humans tend not to consider *all* of the possible actions that they could perform in a particular situation, whereas a more thorough machine could be developed. Imagine a machine that acts as an advisor to human beings and "thinks" like an Act Utilitarian. It will prompt the human user to consider alternative actions that might result in greater net good consequences than the action the human being is considering doing and it will prompt the human to consider the effects of each of those actions on *all* those affected. Such *crtitiquing model expert systems* (systems that evaluate and react to solutions proposed by users) are in use today (e.g. TraumAID[1]) that very likely could incorporate elements of the ethical theory. Finally, for some individuals' actions – actions of the President of the United States or the CEO of a large international corporation – so many individuals can be impacted that the calculation of the greatest net pleasure may be very time consuming, and the speed of today's machines give them an advantage.

We conclude, then, that machines can follow the theory of Act Utilitarianism at least as well as human beings and, perhaps even better, given the data which human beings would need, as well, to follow the theory. The theory of Act-Utilitarianism has, however, been questioned as not entirely agreeing with intuition. It is certainly a good starting point in programming a machine to be ethically sensitive – it would probably be more ethically sensitive than many human beings – but, perhaps, a better ethical theory can be used.

Critics of Act Utilitarianism have pointed out that it can violate human beings' *rights*, sacrificing one person for the greater net good. It can also conflict with our notion of justice – what people *deserve* – because the rightness and wrongness of actions is determined entirely by the future consequences of actions, whereas what people deserve is a result of past behavior. In the Twentieth Century, W. D. Ross (Ross 1930) argued that any single-principle ethical theory like Act Utilitarianism is doomed to fail, because ethics is more complicated than following a single absolute duty. He maintained that ethical decision-making involves considering several *prima facie* duties – duties which, in general, we should try to follow, but can be overridden on occasion by a stronger duty.

Ross suggests that there might be seven *prima facie* duties:

1. *Fidelity* (One should honor promises, live up to agreements one has voluntarily made.)
2. *Reparation* (One should make amends for wrongs one has done.)
3. *Gratitude* (One should return favors.)
4. *Justice* (One should treat people as they deserve to be treated, in light of their past behavior.)
5. *Beneficence* (One should act so as to bring about the most amount of good.)
6. *Non-Maleficence* (One should act so as to cause the least harm.)
7. *Self-Improvement* (One should develop one's own talents and abilities to the fullest.)

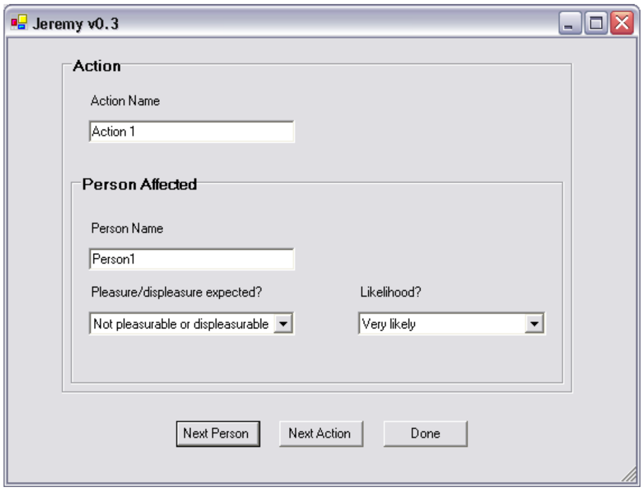
The first four duties arise because of past behavior, and so are a correction to utilitarian thinking. It is interesting that Ross separated the single act utilitarian principle into two — with duties 5 and 6 — and he maintained that, in general, duty 6 is stronger than duty 5. This is because Ross believed (and most of us would surely concur) that it is worse to harm someone that not to help a person. Simply subtracting the harm one might cause from the good, as Act Utilitarianism does, ignores this important ethical truth. The final duty incorporates a bit of Ethical Egoism into the theory and accounts for our intuition that we have a special obligation to ourselves that we don't have to others.

These duties all have intuitive appeal, with the exception of the duty of *Gratitude* which should probably be changed to "one should return favors *one has asked for*," otherwise one could force ethical obligations on individuals simply by doing them favors. Ross' theory of *prima facie* duties seems to more completely account for the different types of ethical obligations that most of us recognize than Act Utilitarianism. It has one fatal flaw, however. Ross gives us no decision procedure for determining which duty becomes the strongest one, when, as often happens, several duties pull in different directions in an ethical dilemma. Thus the theory, as it stands, fails to satisfy Mill's minimal criterion of consistency. Ross was content to leave the decision up to the intuition of the decision-maker, but ethicists believe that this amounts to having no theory at all. The agent could simply do whatever he feels like doing and find a duty to support this action.

It is likely that a machine could help us to solve the problem of developing a consistent, complete and practical version of Ross' theory that agrees with intuition, a problem that human beings have not yet solved because it would involve trying many different combinations of weightings for the duties, which quickly becomes very complicated. A simple hierarchy won't do because then the top duty would be absolute and Ross maintained that all of the duties are *prima facie.* (For each duty, there are situations where another one of the duties is stronger).

We suggest that a method like Rawls' "reflective equillibrium" approach (Rawls 1951) to refining an ethical principle would be helpful in trying to solve this problem and aid us in ethical decision-making. This method would involve running through possible weightings of the duties and then testing them on our intuitions concerning particular cases, revising the weightings to reflect those intuitions, and then testing them again. This approach, that would very quickly overwhelm a human being, lends itself to machine implementation.

[1] http://www.cis.upenn.edu/~traumaid/home.html



Action

Action Name: Action 1

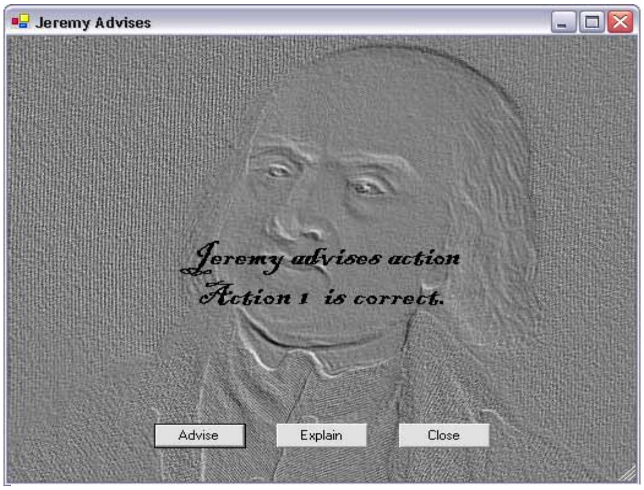
Person Affected

Person Name: Person1

Pleasure/displeasure expected?: Not pleasurable or displeasurable

Likelihood?: Very likely

Figure 1: *Jeremy* data entry



Jeremy advises action
Action 1 is correct.

Figure 2: *Jeremy* advice

We can extend the algorithm above described for Hedonistic Act Utilitarianism to capture the additional complexity of Ross' theory. For a given possible action, we once again will sum over all of the individuals affected. However, instead of computing a single value based only on pleasure/displeasure, we must compute the sum of up to seven values, depending on the number of Ross' duties relevant to the particular action. The value for each such duty could be computed as with Hedonistic Act Utilitarianism, as the product of Intensity, Duration and Probability.

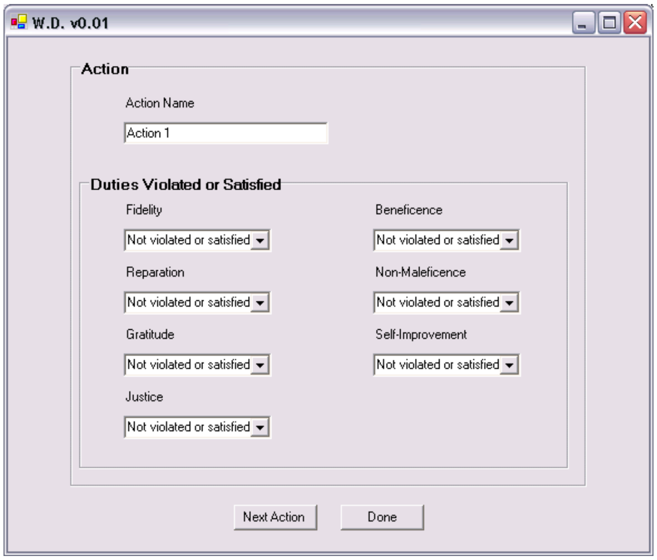
In addition, we must incorporate a factor that captures Ross' intuition that one duty may take precedence over another, for instance that Non-Maleficence is in general a stronger duty than Beneficence. Giving each duty a factor of 1.0 represents equal precedence to all duties. To represent the observation that Non-Maleficence is generally stronger than Beneficence, we might give Non-Maleficence a factor of 1.5. In a simple example in which these are the only two duties that apply, and all other factors are equal, the duty of Non-Maleficence will then have 1.5 times the effect of the duty of Beneficence.

It remains to show how to determine these weights. We propose to apply well-studied approaches that are employed in machine learning that capture Rawls' notion of "reflective equillibrium". In these *supervised learning* (Mitchell 1997) approaches, a set of training data is required; for the current task, the training data would consist of a set of ethical dilemmas together with our consensus of the correct answers. We also identify an *objective function* or goal; in this case, the objective function is simply whether the result of the algorithm conforms to our consensus of correct ethical behavior. The learning algorithm proceeds by adjusting the weights in order to satisfy the objective function as it is exposed to more problem instances. As the choice of weights is refined, the machine could then be more likely to make a correct ethical choice for an ethical dilemma to which it has not yet been exposed.

Besides determining what ethical principles we would like to see a machine follow — a fairly simple theory like Act Utilitarianism or a more complicated one such as an ideally weighted set of *prima facie* duties like Ross' — there is also the issue of how to begin adding this ethical dimension to machines. We suggest, first, designing machines to serve as ethical advisors, machines well-versed in ethical theory and its application to dilemmas specific to a given domain that offer advice concerning the ethical dimensions of these dilemmas as they arise. The next step might be adding an ethical dimension to machines that already serve in areas that have ethical ramifications, such as medicine and business, by providing them with a means to warn when some ethical transgression appears imminent. These steps could lead to fully autonomous machines with an ethical dimension that consider the ethical impact of their decisions before taking action.

Benefits

An ethical dimension in machines could be used to alert humans who rely on machines before they do something that is ethically questionable, averting harm that might have been caused otherwise. Further, the behavior of more fully autonomous machines, guided by this ethical dimension, may be more acceptable in real-world environments than that of machines without such a dimension. Also, machine-machine relationships could benefit from this ethical dimension, providing a basis for resolving resource conflict or predicting behavior of other machines.



Action Name: Action 1

Fidelity: Not violated or satisfied

Beneficence: Not violated or satisfied

Reparation: Not violated or satisfied

Non-Maleficence: Not violated or satisfied

Gratitude: Not violated or satisfied

Self-Improvement: Not violated or satisfied

Justice: Not violated or satisfied

Figure 3: *W.D.* data entry

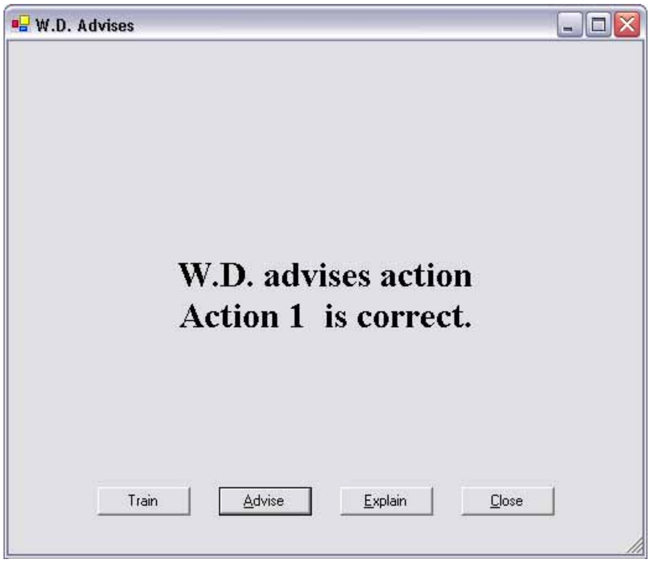
Working in the area of machine ethics could have the additional benefit of forcing us to sharpen our thinking in ethics and enable us to discover problems with current ethical theories. This may lead to improved ethical theories. Furthermore, the fact that machines can be impartial and unemotional means that they can strictly follow rules, whereas humans tend to favor themselves and let emotions get in the way of clear thinking. Thus, machines might even be better suited to ethical decision-making than human beings.

Implementation

As a first step towards our goals, we have begun the development of two prototype ethical advisor systems — *Jeremy*, based upon Bentham's Act Utilitarianism, and *W.D.*, based upon Ross' *prima facie* duties. These programs implement the core algorithms of the ethical theories upon which they are based and, as such, will form the basis for domain-specific systems built upon the same theories. The object of the current programs is to determine the most ethically correct action(s) from a set of input actions and their relevant estimates (which have been simplified for direct user entry).

Jeremy

Jeremy (Figs. 1 & 2) presents the user with an input screen that prompts for the name of an action and the name of a person affected by that action as well as a rough estimate of the amount (*very pleasurable, somewhat pleasurable, not pleasurable or displeasurable, somewhat displeasurable, very displeasurable*) and likelihood (*very likely, somewhat likely, not very likely*) of pleasure or displeasure that person would experience if this action was chosen. The user continues to enter this data for each person affected by this action and this input is completed for each action under consideration.



W.D. advises action
Action 1 is correct.

Figure 4: *W.D.* advice

When data entry is complete, *Jeremy* calculates the amount of net pleasure each action achieves (assigning 2, 1, 0, -1 and -2 to pleasure estimates and 0.8, 0.5, and 0.2 to likelihood estimates and summing their product for each individual affected by an action) and presents the user with the action(s) for which this net pleasure is the greatest. *Jeremy* then permits the user to seek more information about the decision, ask for further advice, or quit.

W.D.

W.D. (Figs. 3 & 4) presents the user with an input screen that prompts for the name of an action and a rough estimate of the amount each of the *prima facie* duties (*fidelity, reparation, gratitude, justice, beneficence, self-improvement, nonmaleficence)* are satisfied or violated by this action (*very violated, somewhat violated*, *not satisfied or satisfied, somewhat satisfied, very satisfied)*. The user continues to enter this data for each action under consideration.

When data entry is complete, *W.D.* calculates the weighted sum of duty satisfaction (assigning -2, -1, 0, 1 and 2 to satisfaction estimates) and presents the user with the action(s) for which the sum of the weighted *prima facie* duties satisfaction is the greatest. *W.D.* then permits the user to train the system, seek more information about the decision, ask for further advice, or quit.

The weights for each duty, currently simply set to 1.0, are to be learned as suggested by Rawls' notion of "reflective equilibrium". As each ethical dilemma is put to *W.D.*, the user is permitted to suggest a particular action that is more intuitively correct than that chosen by *W.D.* Weights for each duty are then updated using a *least mean square* (Mitchell 1997) training rule by adding to each the product of the difference between the weighted sums of each action and the satisfaction estimates for the user-suggested action. As these weights are learned, *W.D.* choices should become more aligned with intuition.

Towards an ethical advisor

Jeremy and *W.D.* are straight-forward implementations of their respective ethical theories. They form a necessary foundation for future development of systems based on these theories, but deriving the raw data required by these systems may be daunting for those not well-versed in the ethical theories they implement or the task of resolving ethical dilemmas in general. It is our intent to insert a layer between the user and these algorithms that will provide guidance in deriving the data necessary for them.

To motivate the next steps we wish to take, consider the following example: You promised a student that you would supervise an independent study for him next semester that he needs in order to graduate on time. Since then, your Dean has offered you the chance to be acting chair of your department for a semester — with a monetary bonus attached to the offer — until a search for a new chair is completed. You can't do both. What should you do?

A naive Act Utilitarian (*Jeremy*) analysis might well lead to a stalemate. If you keep your promise to the student and turn down the offer, it might be equally pleasurable to the student and displeasurable to you, and vice versa. A more sensitive analysis of the case might bring out more subtle aspects of the dilemma, for instance who besides the principles might be affected and consideration of long term consequences in addition to obvious short term ones. Those other than the principles (teacher and student) that might be affected could include, for example, the department if no one else can do a good job as an acting chair, the student's family that may not be able to afford another semester of school or your family if it is need of the money. Long term consequences might include damage to your relationships with other students or to the student's chances of getting into grad school if you renege on your promise and accept the offer, or the loss of a golden opportunity to realize your dream of being an administrator or the risk of disappointing of your Dean if you keep your promise and reject offer.

Further, it is not clear that the obligation of a "promise" can be fully captured with an Act Utilitarian approach like *Jeremy*'s which is only concerned with consequences of actions. On the other hand, Ross' approach imposes other duties on agents including the *prima facie* duty of fidelity where one should honor promises.

A simple analysis in Ross' approach (*W.D.*) might determine that 1) if you renege on your promise and accept the offer, the duty of beneficence is satisfied because you gain money while the duties of fidelity and non-maleficence are violated because you are not keeping your promise and are hurting the student, and 2) if you keep your promise and reject the offer, the duties of fidelity and beneficence are satisfied because your promise is kept and you are helping the student while the duty of non-maleficence is violated because you are harming yourself by not getting the bonus. Given equal levels of satisfaction and violation, as well as an equal weighting of duties, the recommended action is to keep your promise and reject the offer. This action satisfies two duties and only violates one, whereas the alternative satisfies only one and violates two.

A more sensitive analysis would need to consider others beyond the principles, for instance your family, the student's family, the department, the Dean, etc. and, further, the duty of self-improvement may come into play if one has aspirations for doing administrative work. The numbers of individuals positively affected may be enough to raise the satisfaction level for the duty of beneficence to override the violation of the duty of fidelity and, as a consequence, reneging on your promise and accepting the offer is the action recommended by this analysis.

Clearly, more sensitive analyses of ethical dilemmas may prove difficult for users of *Jeremy* and *W.D.* without guidance. We seek to provide such guidance by abstracting and codifying questions supportive of such analyses such as "Who beyond the principles will be affected?", "What will the long term consequences be?", "Are there any other actions possible?", etc. Further, as answers to even these questions might elude users, we intend to provide domain specific guidance to help users determine them. Interestingly, the Rossian framework allows one to create a set of duties particular to a specific domain, for example the Principles of Biomedical Ethics (Beauchamp and Childress 1979). An ethical advisor system based on this set of duties, well-versed in knowledge of the medical domain and its typical dilemmas, for instance, could help elicit information more pointedly for this domain.

In conclusion, we are creating systems that assist users with their dilemmas by helping them consider all that is ethically relevant and providing a means to apply sound ethical theory to it. Arriving at "the answer" is less important than facilitating this process of careful deliberation. We believe this is an important first step towards the ultimate goal of ethically sensitive machines.

Acknowledgements

This work is supported in part by a grant from the National Aeronautics and Space Agency via the Connecticut Space Grant College Consortium.

References

Anderson, S. L. 2000. We Are Our Values. In *Questioning Matters, an Introduction to Philosophical Inquiry*, 606-8 edited by D. Kolak, Mayfield Publishing Company, Mountain View, California.

Bentham, J. 1799. *An Introduction to the Principles and Morals of Legislation*, Oxford.

Barquin, R. C. 1992. In Pursuit of a "Ten Commandments for Computer Ethics". Computer Ethics Institute, (http://www.brook.edu/its/cei/default.htm).

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

Frankena, W. 1993. To Be or Do, That is the Question. In *Doing and Being, Selected Readings in Moral Philosophy*, 208, edited by J. G. Haber, Macmillan, New York.

Heckerman, D. E., Horvitz, E. J., and Nathwani, B. N. 1992. Toward Normative Expert Systems. *Methods of Information in Medicine*, 31:90-105.

Mill, J. S. 1974. *Utilitarianism, in Utilitarianism and Other Writings*, 253, edited by M. Warnock, New American Library, New York.

Mitchell, T. 1997. *Machine Learning*, McGraw Hill.

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.