

A *Prima Facie* Duty Approach to Machine Ethics and Its Application to Elder Care

Susan Leigh Anderson Michael Anderson

University of Connecticut
Department of Philosophy
Storrs, CT
susan.anderson@uconn.edu

University of Hartford
Department of Computer Science
West Hartford, CT
anderson@hartford.edu

Abstract

Having discovered a decision principle for a well-known *prima facie* duty theory in biomedical ethics to resolve particular cases of a common type of ethical dilemma, we developed three applications: a medical ethics advisor system, a medication reminder system and an instantiation of this system in a Nao robot. We are now developing a general, automated method for generating from scratch the ethics needed for a machine to function in a particular domain, without making the assumptions used in our prototype systems.

Introduction

In our early work on attempting to develop ethics for a machine, we first established that it is possible to create a program that can compute the ethically correct action when faced with a moral dilemma using a well-known ethical theory. (Anderson et al. 2004) The theory we chose, Hedonistic Act Utilitarianism, was ideally suited to the task since its founder, Jeremy Bentham (1781), described it as a theory that involves performing “moral arithmetic”. Unfortunately, few contemporary ethicists are satisfied with this teleological ethical theory that bases the rightness and wrongness of actions entirely on the likely future consequences of those actions. It does not take into account justice considerations, such rights and what people deserve in light of their past behavior, that are the focus of deontological theories, like Kant’s Categorical Imperative, which have been accused of ignoring consequences. The ideal ethical theory, we believe, is one that combines elements of both approaches.

The *prima facie* duty approach to ethical theory, advocated by W.D. Ross (1930), maintains that there isn’t a single absolute duty to which we must adhere, as is the case with the two theories mentioned above, but rather a number of duties that we should try to follow (some

teleological and others deontological), each of which could be overridden on occasion by one of the other duties. We have a *prima facie* duty, for instance, to follow through with a promise that we have made; but if it causes great harm to do so, it may be overridden by another *prima facie* duty not to cause harm. And the duty not to cause harm could be overridden, on occasion, by the duty to create good, if the harm is small and the good to be achieved is great. According to Ross, the fact that we have to consider a number of ethical duties, none of which is absolute, is the reason why ethical decision-making is so complicated. But how do we know which duty should be paramount in ethical dilemmas where the *prima facie* duties pull in different directions?

Ross himself had no solution to the problem of determining which duty should prevail when the *prima facie* duties give conflicting advice that is acceptable for our purposes. He was content with allowing the agent to use his or her intuition to decide which *prima facie* duty should prevail in particular situations. That would not be very helpful for a machine attempting to adopt this approach. [It doesn’t seem entirely satisfactory for human beings either. People may not have an intuition, or may have different intuitions, about which duty should be paramount in a particular situation and they are likely to emphasize the duty that permits them to rationalize doing what serves their own self-interest.] A machine needs to be given a decision principle, or a procedure for discovering a decision principle, that enables it to determine the correct action when *prima facie* duties give conflicting advice in an ethical dilemma.

Discovering a Decision Principle

The next project we tackled in our attempt to make ethics computable, therefore, was to take a *prima facie* duty theory and harness machine capabilities in order to find a way to discover a decision principle that could be used to determine the correct action when the *prima facie* duties give conflicting advice, since there was no decision principle given that we could use. John Rawls’ “reflective

equilibrium” approach (Rawls 1951) to creating and refining ethical principles seems reasonable and has inspired our solution to the 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 principle that agrees with intuition that can be used to determine the correct action when prima facie duties give conflicting advice.

Since we wanted to focus on the critical problem of discovering a decision principle required for a machine to implement a prima facie duty ethical theory, in the process establishing a prototype solution to the problem, we constrained the task as much as possible. We used a well-known prima facie duty theory in the domain of biomedicine that has fewer duties than Ross’ more general theory and applied it to a common, but narrow, type of ethical dilemma in that domain to develop and test our solution to the problem. We chose the domain of biomedicine, in part, because the field of biomedical ethics is well developed with much agreement among ethicists as to what is and is not ethically acceptable in particular cases.

The prima facie duty theory that we used is Beauchamp and Childress’ Principles (Duties) of Biomedical Ethics. (Beauchamp and Childress 1979) The type of dilemma that we considered (Anderson et al. 2006a) involved three of their four duties: Respect for the Autonomy of the patient as long as the patient sufficiently understands his/her condition and decisions are made free of external and internal constraints, Nonmaleficence (not causing harm to the patient) and Beneficence (promoting patient welfare).

The general type of ethical dilemma that we considered was one that many health care professionals have faced: *A health care professional 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 professional shouldn’t challenge the patient’s autonomy unnecessarily. On the other hand, the health care professional might have concerns about why the patient is refusing treatment – that is, whether the decision is fully autonomous. Besides the duty to respect patient autonomy, this type of dilemma involves the duty not to cause harm to the patient (nonmaleficence) and/or the duty to promote patient welfare (beneficence), since the recommended treatment is designed to prevent harm to, and/or benefit, the patient.

In this type of dilemma, the options for the health care professional are just two – either to accept the patient’s decision or not – and there are a finite number of specific types of cases using the representation scheme we adopted for possible cases. Our representation scheme consisted of an ordered set of values for each of the possible actions that could be performed, where those values reflected whether the particular prima facie duties were satisfied or violated (if they were involved) and, if so, to which of two

possible degrees. We learned from Bentham, in our earlier work, that the degree of satisfaction or violation of a duty can be very important. To test our approach, we used -2 to represent a strong violation of a particular duty, -1 to represent a weaker violation, 0 when the duty is not involved, +1 for some affirmation and +2 for a strong affirmation of the duty.

Consider the following example of a specific ethical dilemma of the type previously described and how it was represented numerically: *A 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 permit him to take medications. The patient understands the consequences of this refusal. Should the health care professional accept his decision or try again to convince him to take the antibiotic?* In this case, accepting the patient’s decision involves a +2 for respect for the autonomy of the patient, since it’s a fully autonomous decision, a -1 for nonmaleficence since it will lead to some harm for the patient that could have been prevented, and -1 for beneficence since the patient will lose some benefit that he could have received from taking the antibiotic. Questioning the patient’s decision, on the other hand, would involve a -1 for respecting patient autonomy (the patient’s autonomy is being challenged, but he is not being forced to do something against his will), a +1 for nonmaleficence and a +1 for beneficence, since taking the antibiotic would lead to the patient avoiding some harm as well as benefitting him to some degree. From this we generated a case profile: Accept: +2, -1, -1; Try Again: -1, +1, +1.

We used ethicists’ intuitions to tell us the degree of satisfaction/violation of the assumed duties within the range stipulated, and which actions would be preferable, in enough specific cases from which a machine learning procedure arrived at a general principle (confirmed by ethicists) that resolved all cases of the type of dilemmas it would face. [In this and other cases of dilemmas of this type, more specifically, we abstracted the correct answers from a discussion of similar types of cases given by Buchanan and Brock in their article “Deciding for Others: The Ethics of Surrogate Decision Making” (Buchanan and Brock 1989).] We believe that there is a consensus among bioethicists that these are the correct answers. Medical ethicists would say, in the present case, that one should accept the patient’s decision.

It turns out that, with our allowable range of values for the three possible duties that could be at stake, there are 18 possible case profiles (considering that there are only three possible values for autonomy, since we never force treatment on a patient: +2, +1 and -1) and that given the correct answer to just 4 of these profiles enabled the computer to abstract a principle [using *inductive logic programming* (ILP)] that gave the correct answer for the remaining 14 cases. The principle learned was the following: A health care professional should challenge a patient’s decision if it isn’t fully autonomous and there is

either any violation of nonmaleficence or a severe violation of beneficence.

Of course, the principle was implicit in the judgment of ethicists, but we don't believe that it had ever been explicitly stated before. It gives us hope that not only can ethics help to guide machine behavior, but that machines can help us to discover the ethics needed to guide such behavior. Furthermore, we developed a way of representing the needed data and a system architecture for implementing the principle.

Applying the Decision Principle

We have developed three applications of the principle: (1) MedEthEx (Anderson et al. 2006b), a medical ethics advisor system for dilemmas of the type that we considered. (2) A medication reminder system, EthEl, for the elderly that not only issues reminders at appropriate times, but also determines when an overseer (health care provider or family member) should be notified if the patient refuses to take the medication. (Anderson and Anderson 2008) (3) An instantiation of EthEl in a Nao robot, which we believe is the first example of a robot that follows an ethical principle in determining which actions it will take. (Anderson and Anderson 2010)

MedEthEx is 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 use by someone unfamiliar with the representation details required by the decision procedure, a user interface was developed that (1) asks ethically relevant questions of the user regarding the particular case at hand, (2) transforms the answers to these questions into the appropriate representations, (3) sends these representations to a decision procedure, (4) presents the answer provided by the decision procedure, and (5) provides a justification for this answer.

EthEl is faced with an ethical dilemma that is analogous to that from which the principle was learned, in that the same duties are involved and "try again" corresponds to notifying an overseer when a patient refuses to take a prescribed medication and "accept" corresponds to not notifying the overseer when the patient refuses to take it. EthEl receives input from an overseer (most likely a doctor), including: the prescribed time to take a medication, the maximum amount of harm that could occur if this medication is not taken (for example, 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 and violation levels over time, a function of the maximum amount of harm or good and the number of hours for this effect to take place. This value is used to increment duty satisfaction and violation levels for the remind action and, when a patient disregards a reminder, the notify action. It is used to decrement don't

remind and don't notify actions as well. A reminder is issued when, according to the principle, the duty satisfaction or violation levels 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 duty satisfaction or violation levels have reached the point where notifying the overseer is ethically preferable to not notifying the overseer.

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 that never does so. In between, a system such as EthEl takes into account ethical considerations in deciding when to contact an overseer. 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.

We have embodied this software prototype in Aldebaran Robotic's Nao robot, a platform that provides out-of-the-box capabilities sufficient to serve as the foundation for implementation of principle-driven higher-level behaviors. These capabilities include walking, speech recognition/generation, gripping, touch-sensitivity, wifi internet access, face and mark recognition, infra-red capabilities, sonar, sound localization, and telepresence. These, combined with wifi RFID tagging of Nao's charges for identification and location purposes, permit Nao to assume obligations towards users such as promising to remind them of when to take medications, etc. and seeking them out when it is time to do so. Notice that full language understanding, full vision, and other complex behaviors are not necessary to produce a useful robotic assistant that can accomplish these tasks in an ethical manner. For instance, communication for the tasks described can be achieved through simple spoken or touch input and output; navigation of a common room can be achieved through a combination of limited vision, sonar, and touch; location and identification of people can be accomplished with sound localization, face and mark recognition, and wifi RFID tagging.

In our current implementation, Nao is capable of finding and walking towards a patient who needs to be reminded to take a medication, bringing the medication to the patient, engaging in a natural language exchange, and notifying an overseer by e-mail when necessary. To our knowledge, Nao is the first robot whose behavior is guided by an ethical principle.

Generalizing the Approach

Having had success in developing a method for discovering a decision principle needed to resolve ethical

dilemmas when prima facie duties give conflicting advice, we next wanted to find a method for generating the ethics needed for a machine to function in a particular domain from scratch, without making the assumptions used in our prototype. We previously made assumptions about the prima facie duties in the type of dilemmas it would face, as well as the range of possible satisfaction/violation of the duties.

These assumptions were based on well-established ideas in ethical theory, but we want now to make the fewest assumptions possible. Some of the assumptions we list below have been implicit in the work that we have done so far, and we believe that they are necessary if ethical judgments are to have validity at all or to make sense of their application to machines. Some have come from a realization that there is something more basic to ethics than duties. The others come from insights of three great theorists in the history of Ethics.

In our current approach to discovering and implementing ethics for a machine, we make the following assumptions:

(1) We are concerned with the *behavior* of machines – their *actions* rather than their status – so we have adopted the action-based approach to ethical theory rather than the virtue-based approach.

(2) There is at least one ethically significant *feature* of dilemmas that are classified as being ethical that needs to be considered in determining the right action.

(3) There is at least one *duty* incumbent upon the agent/machine in an ethical dilemma, either to maximize or minimize the ethical feature(s).

(4) We accept Bentham’s insight (1781) that ethical features may be present to a greater or lesser degree in ethical dilemmas (e.g. more or less pleasure may result from performing the possible actions) and this affects how strong the corresponding duties are in that dilemma.

(5) If there is more than one duty, corresponding to more than one ethically significant feature of ethical dilemmas (which we think is likely in true ethical *dilemmas*), then since the duties may conflict with one another, we should consider them to be *prima facie* duties, requiring a decision principle to give us the correct answer in cases of conflict.

(6) John Rawls’ “reflective equilibrium” approach (Rawls 1951) to creating and refining ethical principles seems reasonable and can be used to solve to the problem of coming up with a decision principle/principles when there are several prima facie duties that give conflicting advice in ethical dilemmas. 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 principle that agrees with intuition that can be used to determine the correct action when prima facie duties give conflicting advice.

(7) It is the intuitions of ethicists that should be used in adopting the reflective equilibrium approach to determining decision principles. We believe that there is an expertise that comes from thinking long and deeply about

ethical matters. Ordinary human beings are not likely to be the best judges of how one should behave in ethical dilemmas. We are not, therefore, adopting a sociological approach to capturing ethics, since we are concerned with *ideal* behavior rather than what most people happen to think is acceptable behavior. [Also, ethicists tend to reject ethical relativism, which is typically not the case with sociologists; and it is essential in order to give meaning and weight to ethical judgments that they not just be matters of opinion.]

(8) Finally, we accept the Kantian insight (Kant 1785) that, to be rational, like cases must be treated in the same fashion. What’s right for one must be right for another (others). We cannot accept contradictions in the ethics we embody in machines. [We believe that humans should not accept contradictions in their own, or others’, ethical beliefs either.] With two ethically identical cases – i.e. cases with the same ethically relevant feature(s) to the same degree – an action cannot be right in one of the cases, while the comparable action in the other case is considered to be wrong. Formal representation of ethical dilemmas and their solutions make it possible for machines to spot contradictions that need to be resolved.

Believing that it is unacceptable to hold contradictory views in ethics has led us to the conclusion that if we encounter two cases that appear to be identical ethically, but it is believed that they should be treated differently, then there must be an ethically relevant difference between them. If the judgments are correct, then there must either be a *qualitative* distinction between them that must be revealed, or else there must be a *quantitative* difference between them. This can be translated into either a difference in the ethically relevant features between the two cases, i.e. a feature which appears in the one but not in the other case, or else a wider range of satisfaction or violation of existing features must be considered which would reveal a difference between the cases, i.e. there is a greater satisfaction or violation of existing features in the one, but not the other, case. [These options, by the way, get at the bone of contention between Mill (1863) and Bentham in developing Hedonistic Utilitarianism. Bentham thought that one only needs to consider different quantities of pleasure/displeasure to differentiate between cases, whereas Mill was convinced that there were higher and lower pleasures to be taken into account as well, i.e. a qualitative distinction between the cases.] Can there be any other way of rationally defending our treating one case differently from another? It would seem not.

We now envision, when developing a machine that will function more or less autonomously in a particular domain, that there will be a dialogue with ethicists to determine the ethically relevant features of possible dilemmas that such a machine may encounter, and correlative duties, plus the correct behavior when faced with those dilemmas. From this information the machine should be able to come up with a principle, or principles, to resolve dilemmas that it may encounter, even those that have not been anticipated. The principle, or principles, it comes up with may be

implicit in the judgments of ethicists, but to date has/have not as yet been explicitly stated. In this way, work in machine ethics may help to advance the study of ethics in general.

We are now working on generating from scratch, in an automated fashion, the ethically relevant features, correlative duties, and the range of intensities required, as well as discovering a decision principle(s) for resolving conflicts for the types of dilemmas our autonomous medication reminder system might face, hoping to devise a model for creating an ethic that can be used for autonomous systems in other domains as well.

Imagining a dialogue between the learning system and an applied ethicist, using our medication reminder system as an example, we can see that (in principle) we can hone down what is required to enable the ethicist to begin to teach the system the ethically relevant features, correlative duties and eventually the range of intensities required, from which decision principles can be discovered. The system prompts the ethicist to give an example of an ethical dilemma that a medication reminder system might face, asking the ethicist to state the possible actions that could be performed, which one is preferable, and what feature is present in one of the actions, but not in the other. From this information, a duty that is at least *prima facie* can be inferred, either to maximize or minimize the feature, depending upon whether the action that has the feature is preferable or not. Information is stored in the system, including a representation of a *positive* case (that one action is preferable to the other) and a *negative* one (that the opposite action is not preferable).

The system might then prompt the ethicist to give an example of a new ethical dilemma where the judgment of the ethicist would be the reverse of the first case (i.e. instead of notifying the overseer as being correct, one should not notify the overseer). Prompting the ethicist, the system determines whether in this case a *second* feature is present, which should be maximized or minimized, or whether the difference between the two cases amounts to a difference in the *degree* to which the original feature is present. As new features are introduced (often as a result of resolving apparent contradictions within the existing representation scheme), with corresponding *prima facie* duties, the system begins to formulate and then refine a decision principle to resolve cases where the *prima facie* duties pull in different directions.

We envision the system prompting the ethicist to enter in just the types of cases that will enable it to obtain the data it needs to learn a decision principle as efficiently as possible, i.e. to infer an ethically acceptable decision principle with the fewest number of cases.

There are two advantages to discovering ethically relevant features/duties, and an appropriate range of intensities, with this approach to learning what is needed to resolve ethical dilemmas. First, it can be tailored to the domain with which one is concerned. Different sets of ethically relevant features/*prima facie* duties can be discovered, through considering examples of dilemmas in

the different domains in which machines will operate. A second advantage is that features/duties can be added or removed, if it becomes clear that they are needed or redundant.

In addition, we believe that there is hope for discovering decision principles that, at best, have only been implicit in the judgments of ethicists and may lead to surprising new insights, and therefore breakthroughs, in ethical theory. This can happen as a result of the computational power of today's machines that can keep track of more information than a human mind and require consistency. Inconsistencies that are revealed will force ethicists to try to resolve those inconsistencies through the sharpening of distinctions between ethical dilemmas that appear to be similar at first glance, but which we want to treat differently. There is, of course, always the possibility that genuine disagreement between ethicists will be revealed concerning what is correct behavior in ethical dilemmas in certain domains. If so, the nature of the disagreement should be sharpened as a result of this procedure; and we should not permit machines to make decisions in these domains.

Future Work

While we believe that the type of representation scheme that we have been developing will be helpful in categorizing and resolving ethical dilemmas in a manner that permits machines to behave more ethically, we envision an extension and an even more subtle representation of ethical dilemmas in future research. We need to consider more possible actions available to the agent, where there is not necessarily a symmetry between actions (i.e. where the degree of satisfaction/violation of a duty in one is mirrored by the opposite in the other). Also, ideally, one should not only consider present options, but possible actions that could be taken in the future. It might be the case, for instance, that one present option, which in and of itself appears to be more ethically correct than another option, could be postponed and performed at some time in the future, whereas the other one cannot, and this should affect the assessment of the actions.

Consider the following ethical dilemma: You had promised your elderly parents that you would help them by cleaning out the overflowing gutters on their house this afternoon. Just as you are about to leave, a friend calls to say that her car has broken down some distance from your apartment and she needs a ride. She reminds you that you owe her a favor; but helping her would take the rest of the afternoon and, as a result, you would not be able to keep your promise to your parents. What should you do? Let's assume that the benefit for each party is the same and that honoring a promise is a stronger obligation than returning a favor, so it would appear that the right action is to clean out your parents' gutters this afternoon. But it might also be the case that you could clean out your parents' gutters tomorrow afternoon without any substantial loss of benefit or harm resulting from the postponement -- the weather is

expected to be clear for at least the next day -- whereas your friend must have assistance today. You can't postpone helping your friend until another day. Shouldn't this information factor into the assessment of the ethical dilemma? Projecting into the future will complicate things, but it will yield a more ethically correct assessment and should eventually be incorporated into the process.

Acknowledgement

We would like to acknowledge Mathieu Rodrigue of the University of Hartford for his efforts in implementing the algorithm used to derive the results in this paper.

References

Anderson, M., Anderson, S. & Armen, C. (2004), "Toward Machine Ethics" in Proceedings of AAAI Workshop on Agent Organizations: Theory and Practice, San Jose, CA, July.

Anderson, M., Anderson, S., and Armen, C. (2006a), "An Approach to Computing Ethics," *IEEE Intelligent Systems*, Vol. 21, no. 4.

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

Anderson, M. and Anderson, S. (2008), "EthEl: Toward a Principled Ethical Eldercare Robot" in *Proceedings of the AAAI Fall 2008 Symposium on AI in Eldercare: New Solutions to Old Problems*, Arlington, Virginia, November.

Anderson, M. and Anderson, S. (2010), "An Ethical Robot", *Scientific American*, October.

Beauchamp and Childress (1979), *Principles of Biomedical Ethics*. Oxford, UK: Oxford University Press.

Bentham, J. (1781), *An Introduction to the Principles of Morals and Legislation*, Clarendon Press, Oxford.

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

Kant, I. (1785), *The Groundwork of the Metaphysic of Morals*, trans. by H. J. Paton (1964). New York: Harper & Row.

Mill, J.S. (1863), *Utilitarianism*, Parker, Son and Bourn, London

Rawls, J. (1951), "Outline for a Decision Procedure for Ethics", *The Philosophical Review* 60(2): 177-197.

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