Developing a General, Interactive Approach to Codifying Ethical Principles

Michael Anderson
Department of Computer Science
University of Hartford
West Hartford, CT 06776
anderson@hartford.edu

Susan Leigh Anderson
Department of Philosophy
University of Connecticut
Stamford, CT 06901
Susan.Anderson@UConn.edu

Abstract

Building on our previous achievements in machine ethics (Anderson *et al.* 2006a-b, 2007, 2008), we are developing and implementing a general interactive approach to analyzing ethical dilemmas with the goal to apply it toward the end of codifying the ethical principles that will help resolve ethical dilemmas that intelligent systems will encounter in their interactions with human beings. Making a minimal epistemological commitment that there is at least one ethical duty and at least two possible actions that could be performed, the general system will: 1) incrementally construct, through an interactive exchange with experts in ethics, a representation scheme needed to handle the dilemmas with which it is presented, and 2) discover principles implicit in the judgments of these ethicists in particular cases that lead to their resolution. The system will commit only to the assumption that any ethically relevant features of a dilemma can be represented as the degree of satisfaction or violation of one or more duties that an agent must take into account to determine which of the actions that are possible in that dilemma is ethically preferable.

Introduction

In broadest terms, we are trying to determine to what extent ethical decision-making is computable. There are domains where intelligent machines could play a significant role in improving the quality of life of human beings as long as ethical concerns about their behavior can be overcome by incorporating ethical principles into them. If ethical decision-making can be computed to an acceptable degree (at least in domains where machines might impact human lives) and incorporated into machines, then society should feel comfortable in allowing the continued development of increasingly autonomous machines that interact with humans. An important bi-product of work on "machine ethics" is that new insights in ethical theory are likely to result. We believe that machines, through determining what consistently follows from accepting particular ethical judgments, may be able to discover new ethical principles even before ethical theorists are able to do so themselves.

We assert that ethical decision-making is, to a degree, computable. At this time there is no universally agreed upon ethical theory, yet ethicists are generally in agreement about the right course of action in many particular cases and about the ethically relevant features of those cases. We maintain that much can be learned from those features and judgments. Following W.D. Ross (1930), who argued that simple, single-principle ethical theories do not capture the complexities of ethical decision-making, they believe that acknowledging the relevant features of ethical dilemmas may lead to determining several *prima facie* duties the agent should attempt to follow. The major challenges in Ethical Theory are determining these duties and the principles needed to decide the ethically preferable action when the *prima facie* duties give conflicting advice, as often happens. Computers can help us to abstract these principles from particular cases of ethical dilemmas involving multiple *prima facie* duties where experts in ethics have clear intuitions about the ethically relevant features of those cases and the correct course of action.

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 machines are likely to interact with human beings, there is likely to be a consensus that machines should defer to the best interests of the humans affected. If this were not the case, then it would be ill-advised to create machines that would interact with humans at all.

Related Research

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. The general machine ethics research agenda involves testing the feasibility of a variety of approaches to capturing ethical reasoning, with differing ethical bases and implementation formalisms, and applying this reasoning in systems engaged in ethically sensitive activities. Researchers must investigate how to determine and represent ethical principles, incorporate ethical principles into a system's decision procedure, make ethical decisions with incomplete and uncertain knowledge, provide explanations for decisions made using ethical principles, and evaluate systems that act upon ethical principles.

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 implementation: Christopher Grau (2006) considers whether the ethical theory that best lends itself to implementation in a machine, Utilitarianism, should be used as the basis of machine ethics; Tom Powers (2006) assesses the viability of using deontic and default logics to implement Kant's categorical imperative. Jim Moor (2006), on the other hand, investigates the nature of machine ethics in general, making distinctions between different types of normative agents, and its importance.

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 (2003), 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. Colin Allen, Gary Varner, and Jason Zinser (2000) have suggested that a "moral Turing test" be used to evaluate systems that incorporate an ethical dimension.

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 (2006-7). These workshops of 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 (Turkle 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).

Our Previous Research

In our previous work, we have developed a proof-of-concept system that discovered a novel ethical principle that governs decisions in a particular type of dilemma that involves three *prima facie* duties and applied this principle in making ethical decisions in two proof of concept systems (Anderson *et al.* 2004, 2005a-c, 2006a-d). The principle was discovered using machine-learning techniques to abstract relationships between the three duties from cases of the particular type of ethical dilemma where ethicists are in agreement as to the correct action.

In this work, we adopted the *prima facie* duty approach to ethical theory which, as they believe, better reveals the complexity of ethical decision-making than single, absolute duty theories like Hedonistic Act Utilitarianism. 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, and journalistic ethics, for example.

There are two well-known *prima facie* duty theories: Ross' theory (1930), dealing with general ethical dilemmas, that has seven duties; and Beauchamp's and Childress' four Principles of Biomedical Ethics (1979) (three of which are derived from Ross' theory) that is intended to cover ethical dilemmas specific to the field of biomedicine. Because there is more agreement between ethicists working on biomedical ethics than in other areas, and because there are fewer duties, we began development of their *prima facie* duty approach to computing ethics using Beauchamp's and Childress' Principles of Biomedical Ethics.

Beauchamp's and Childress' Principles of Biomedical Ethics include: 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.

The domain of our previous work was medical ethics, consistent with their choice of *prima facie* duties, and, in particular, a representative type of ethical dilemma that involves three of the four Principles of Biomedical Ethics: Respect for Autonomy, Nonmaleficence and Beneficence. The type of dilemma that the PIs considered was one that health care workers often face: A health care worker has recommended a particular treatment for her competent adult patient and the patient has rejected that treatment option. Should the health care worker try again to change the patient's mind or accept the patient's decision as final? The dilemma arises because, on the one hand, the 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.

In this type of dilemma, the options for the health care worker are just two, either to accept the patient's decision or not, by trying again to change the patient's mind. In this attempt to make a *prima facie* duty ethical theory computable, our single type of dilemma encompassed a finite number of specific cases, just three duties, and only two possible actions in each case. We abstracted, from a discussion of similar types of cases given by Buchanan and Brock (1989), the correct answers to the specific cases of the type of dilemma they considered. We believe that there is a consensus among bioethicists that these are the correct answers.

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. Implementing the theory required that we first discover a principle to determine the correct action when the duties give conflicting advice and then fashion a decision procedure using this principle. John Rawls' "reflective equilibrium" (1951) approach to creating and refining ethical principles inspired their 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 principle that agrees with intuition. The PIs abstracted a principle from representations of specific cases of ethical dilemmas where experts in ethics have clear intuitions about the correct action. Their representation of the ethical dilemmas consisted of an ordered set of values for each of the possible actions that could be performed, where these values reflected whether the particular *prima facie* duties are satisfied or violated and, if so, to what degree.

We developed a system (Anderson et al. 2006b) that uses machine learning techniques to abstract relationships between the *prima facie* duties from particular ethical dilemmas where there is an agreed upon correct action. Their 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.

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 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 so abstracted supports an insight of Ross' that violations of the duty of nonmaleficence should carry more weight than violations of the duty of beneficence. We offer this principle as evidence that making the 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.

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.

We next explored two applications of the discovered principle that governs three duties of Beauchamp's and Childress' Principles of Biomedical Ethics. In both applications, they developed a program where a machine could use the principle to determine the correct answer in specific ethical dilemmas. The first, MedEthEx (Anderson et al. 2006a) is a proof of concept medical ethical advisor system; the second, EthEl, is a proof of concept system in the domain of eldercare that determines when a patient should be reminded to take medication and when a refusal to do so is serious enough to contact an overseer. EthEl exhibits more autonomy than MedEthEx in that, whereas MedEthEx gives the ethically correct answer to a human user who will act on it or not, EthEl herself acts on what she determines to be the ethically correct action.

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 a user unfamiliar with the representation details required by the decision procedure to use the principle, 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 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) 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.

In our second application, EthEl (Anderson & Anderson *in submission*) makes decisions about when and how often to remind a competent patient to take medications and when to notify an overseer if the patient refuses to do so. The previously discovered ethical principle is pertinent to this dilemma as it is analogous to the original dilemma — 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. In this case, the principle balances the importance of not allowing the patient to be harmed or lose substantial benefits from not taking the medication against the duty to respect the autonomy of the patient.

Our Current Research

We are currently developing and implementing a general interactive approach to analyzing ethical dilemmas and applying it towards the resolution of ethical dilemmas that will be faced by intelligent systems in their interactions with human beings. Our approach leverages our previous proof-of-concept system that discovered an ethical principle used by MedEthEx and EthEl, expanding it to include discovering ethical features and duties and generalizing it so that it can be applied to other domains. As this system will work in conjunction with an ethicist, the tasks required to fulfill the goal must be partitioned into those best handled by the ethicist and those best handled by the system. These tasks include:

1. providing examples of ethical dilemmas
2. establishing ethically relevant features in a dilemma
3. identifying duties that correspond to features in a dilemma
4. determining the satisfaction/violation levels of duties pertinent to a dilemma
5. specifying actions that are possible in a dilemma
6. indicating the ethically preferable action in a dilemma, if possible

7. constructing a succinct, expressive representation of a dilemma
8. discovering and resolving contradictory example dilemmas
9. abstracting general principles from example dilemmas
10. discerning example dilemmas that are needed to better delineate abstracted principles
11. applying general principles to new example dilemmas
12. explaining the rationale of the general principle as applied to dilemmas

Although it may be the case that these tasks need further decomposition, roughly the first half of the above tasks will be the primary responsibility of the ethicist who will train the system and the second half will be assigned to the system itself. The ethicist will need to give examples of ethical dilemmas, indicate which actions are possible in these dilemmas, and give the generally agreed upon solutions to these dilemmas. Further, the ethicist must establish the ethically relevant features of the dilemmas as well as their corresponding duty satisfaction/violation levels. The system will use this information to incrementally construct an increasingly more expressive representation for ethical dilemmas and discover ethical principles that are consistent with the correct actions in the examples with which it is presented. It will prompt the ethicist for further information leading to the resolution of contradictions as well as for example dilemmas that are needed to better delineate the abstracted principles. Further, it will use the abstracted principles to provide solutions to examples that are not part of its training and, along with stored examples, to provide explanations for the decisions it makes.

In our previous 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 principle to resolve the conflict.

The system requires a dynamic representation scheme, i.e. one that can change over time. Having 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. The understanding of ethical duties, and their relationships, evolves over time.

We start with a representation where there is one, unnamed ethical duty and a principle that states that an action that satisfies this duty is ethically preferable to an action that does not. As examples of particular dilemmas are presented to the system with an indication of the ethically preferable action, the number of duties and the range of their satisfaction/violation levels are likely to increase, and, further, the principle will evolve to incorporate them. Alan Bundy and Fiona McNeill's work in dynamic representations (2006) motivates a number of dimensions along which the representation scheme for ethical dilemmas that evolves in our system might change, including the addition of new features, deletion of superfluous features, merging of multiple features into one, splitting of a single feature into many, and repair of existing data.

Addition of new features

To maintain a consistent principle, contradictions are not permitted. If the system is given two training cases that have the same representations for the possible actions in the dilemmas, but different actions are said to be the correct one, then ethically distinguishable features of the cases must not be expressible in the existing representation scheme. This would demonstrate the need for an additional duty or finer partitioning of the range of satisfaction/violation values of one or more existing duties.

	Benefit
A:	Yes
B:	No

A is preferable to B.
(a)

	Benefit
A:	Yes
B:	No

B is preferable to A.
(b)

	Benefit	Harm
A:	Yes	Yes
B:	No	No

B is preferable to A
(c)

	Benefit	Harm
A:	Yes	No
B:	No	No

A is preferable to B
(d)

Figures 1 a-d. Generation of a new feature.

For example, suppose the following dilemma, and solution, had been given by the ethicist at the beginning of the training process:
Dr. X is faced with the choice of advising patient Y to take medication Z (option A) or not (option B). If the patient takes Z it will benefit the patient, a benefit that will be lost if the patient does not take Z. Should X advise the patient to take Z (option A) or not (option B)? The correct answer is that option A is preferable to option B.
At this time the only ethical feature that was needed to represent the dilemma was a *prima facie* duty to benefit someone, if at all possible. Thus, the dilemma could be represented as in Fig. 1a.

Next, suppose the following dilemma and solution were provided to the system by the ethicist:
Dr. A is faced with the choice of advising patient Y to take medication Z (option A) or not (option B). Taking Z would provide some benefit, but would also cause serious side effects for Y. Y would lose the benefit, but also avoid the serious side effects, if Z is not taken. In this case B is preferable to A.
If the second dilemma is summarized using the simple representation scheme developed to this point, it will be represented as in Fig. 1b. Since this contradicts the earlier training example, the representation scheme must be changed. Clearly, ethically distinct dilemmas cannot be correctly captured by this simplistic representation scheme, so it must be modified. Questioning the ethicist to learn what is different about the two dilemmas — that the second one involves the possibility of the patient being harmed whereas the first one does not — could elicit the need for a second duty: a *prima facie* duty to cause the least harm. If the new dilemma is represented as in Fig. 1c and the earlier one as in Fig. 1d, there is no longer a contradiction. (A further training example involving only these two duties, creating a contradiction with the latest example, would generate a need for distinguishing between different levels of harm and benefit.)

Deletion of superfluous features

That a *prima facie* duty can be removed from the representation scheme can be deduced from the fact that the duty is found to be superfluous in determining the decision principle, e.g. what was once thought to be a duty is no longer recognized as being one, or realizing that a new duty always has the same value, across actions, as one that is already established. It could also be determined that a wider range of duty satisfaction/violation values was postulated than is necessary by establishing that two different duty satisfaction/violation values always generate the same result in terms of the ethical principle.

For example, suppose that society evolves from having individual ownership of material objects to the point where everything is communally owned. (Even personal hygiene objects like toothbrushes are communally owned, becoming single use goods that the community supplies to anyone needing them.) Before the full transition to the communal society, it might have been thought important to recognize individual ownership of goods in settling ethical disputes, whereas that feature later drops out, becoming irrelevant. This sort of transition is in fact happening in communities designed for retirees who are looking for support in their declining years.

Merging features

As an example of merging features, consider a particular ethical dilemma that might have suggested a *prima facie* duty to provide benefit to someone, if at all possible. Later, another ethical dilemma might have suggested a *prima facie* duty to help someone, if at all possible. That a *prima facie* duty is duplicated might happen because the machine will adopt the language the ethicist uses in each case, not recognizing that "benefiting" and "helping" are synonyms. Once this is recognized, the two duties can be merged into a single duty.

As another example, in a dialogue between the system and ethicist concerning particular ethical dilemmas, it might have been thought to be important to distinguish between three levels of harm: minor, moderate and maximal. Yet it could turn out that the last two categories could be collapsed into one as far as determining an ethical principle is concerned. Perhaps it's sufficient that an action is likely to lead to at least moderate harm in order for an action to be considered ethically unacceptable.

Splitting features

Consider a variation on the previous example, in a dialogue between the system and ethicist concerning particular ethical dilemmas, where it might have been thought that there was a need to distinguish between only two levels of harm: minor and maximal. Yet it could turn out that a need arises for a level in between the two categories for determining an ethical principle concerned. Such an addition might be considered as splitting one of the categories into two separate categories.

Repair of the representation of existing data

The evolution of a representation scheme over time also entails the repair of data expressed in an earlier representation scheme. This will require, in the cases of adding, merging, and splitting duties, the addition of new duties and, in the case of deleting, merging, and splitting, the deletion of old duties. New duties can be added to existing representations with a zero degree satisfaction/violation (as in figure 5d) with the realization that future inconsistency may signal the

need for updating this value. No longer needed duties can simply be deleted.

An illustration

To illustrate this general approach to discovering the features of ethical dilemmas and their corresponding duty satisfaction/violation levels, in a more formal way, as well as principles needed to resolve the dilemmas, consider the hypothetical dialogue between a system such as we envision and ethicist in Table 1. We have begun with the same type of dilemma that we considered in our previous work concerning MedEthEx, attempting to see how one might generate the features of specific ethical dilemmas, corresponding duties and principles that resolve these dilemmas.

References

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

Anderson, M., Anderson, S. & Armen, C. 2006b. An Approach to Computing Ethics. Special Issue of IEEE Intelligent Systems on Machine Ethics, August.

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

Anderson, M. & Anderson, S. EthEl: Toward a Principled Ethical Eldercare Robot. 2008. Workshop on Assistive Robots at the Conference on Human-Robot Interaction, March.

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

Boden, M. 2006. Robots and Anthropomorphism. In (Metzler 2006).

Bringsjord, S., Arkoudas, K. & 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, pp48-57, Cambridge University Press.

Bundy, A. and McNeill, F. 2006. Representation as a Fluent: An AI Challenge for the Next Half Century. IEEE Intelligent Systems ,vol. 21, no. 3, pp. 85-87, May/June.

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.

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

McLaren, B. M. 2003. Extensionally Defining Principles and Cases in Ethics: an AI Model, Artificial Intelligence 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 (Metzler 2007).

Turkle, S. 2006. Robot as Rorschach: New Complicities for Companionship. In (Metzler 2006).

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. 260-277.)

Table 1: Generating duties, satisfaction levels, and principles

System (S): Give me a case of the type of dilemma we are considering.	[∂(0,1)] prefer A1 over A2 if $\Delta\partial \geq 1$	*Single, unnamed feature with 2 values*
Ethicist (E): **Case 1:** A patient refuses to have simple and safe cosmetic surgery to remove a very large non-cancerous growth on his face that has led to repeated rejection from others. His depression from his condition has lead to his being resigned to be an outcast. Should the doctor try again (T) to convince the patient to have the surgery or accept (A) his decision as final?	A1 = T and A2 = A or A1 = A and A2 = T	*There are two possible actions in the given dilemma: T and A*
S: What ethically relevant feature is present in one of the possible actions and not in the other?		
E: T could benefit the patient, while the patient would lose this benefit with A.	∂ = cause_benefit or ∂ = eliminate_benefit	*Ethically relevant feature is "benefit"*
S: Which action is ethically preferable in this case?		
E: T is ethically preferable to A.	A1 = T and A2 = A ∂ = cause_benefit $A1_{cause\_benefit} = 1$, $A2_{cause\_benefit} = 0$	*Ethical duty is "cause benefit"; T satisfies duty, A does not*
S: So, all other things being equal, one has a duty to benefit the patient rather than not?		
E: Yes.	[cause_benefit(0,1)], prefer A1 over A2 if Δcause_benefit ≥ 1 Case1(+,T=1,A=0)	*Instantiation of ∂ to "cause benefit"* *Store positive Case 1*
S: Can you think of a case of this type of dilemma that involves a possible benefit for the patient, which would be lost if the patient's decision is accepted, in which one shouldn't try again to change the patient's mind?		*Is there a counter-example to "cause_benefit"?*
E: Yes. **Case 2:** A patient refuses to take a medication that could only help to alleviate some symptoms of a virus that can't be cured by the medication but must run its course. One needn't try again to change the patient's mind in this case.	Case2(-,T=1,A=0) A is preferred so, A1 = T and A2 = A, $A1_{cause\_benefit} = 1$, $A2_{cause\_benefit} = 0$ and Δcause_benefit ≥ 1 is false	*Store negative Case 2* *A contradiction has been established that needs to be resolved*
S: What is a difference between Case 1 and Case 2?		
E: There is greater benefit for the patient at stake in trying again in Case 1 than in Case 2.	$T1_{cause\_benefit} > T2_{cause\_benefit}$	*The value of "cause benefit" for T is not 1 in both cases*
S: So we need to distinguish between levels of benefit?		
E: Yes.	[cause_benefit(0,1,2)], prefer A1 over A2 if Δcause_benefit ≥ 2 Case1(+,T=2,A=0)	*Increase range of cause_benefit and minimum satisfactory differential* *Repair Case 1*

S: Can you think of a case where there was the greater benefit at stake for the patient, where it would be wrong to try to change the patient's mind about accepting treatment that would benefit him in this way?		
E: Yes. Case 3: A patient refuses to have a life-saving blood transfusion because of long standing religious beliefs.	Case3(-,T=2, A=0) A is preferred so, A1 = T and A2 = A, $A1_{benefit}$ = 2, $A2_{benefit}$ = 0 and Δcause_benefit≥ 2 is false	*Store negative Case 3* *A contradiction has been established that needs to be resolved*
S: What is different between case 1 and case 3?		
E: In case 3 the patient's decision to forego treatment was a fully autonomous one, whereas in case 1 that was not true.	∂ = cause_autonomy or ∂ = eliminate_autonomy	*Ethically relevant feature is "autonomy"*
S: So there is a duty to respect a fully autonomous decision of a patient that is distinct from a duty to benefit the patient?		
E: Yes.	A1 = T and A2 = A ∂ = cause_autonomy $A1_{autonomy}$ = 0, $A2_{autonomy}$ = 1	*Ethical duty is "cause autonomy"; A satisfies duty, T does not*
S: And the duty to respect a fully autonomous decision can outweigh the duty to provide considerable benefit?		
E: Yes.	[autonomy(0,1)], [benefit(0,1, 2)], prefer A1 over A2 if (Δcause_benefit≥ 2 and ~Δcause_autonomy≥ 1) Case1(+,$T_{benefit}$ =2,$A_{benefit}$ =0, $T_{autonomy}$ = 0, ,$A_{autonomy}$ = 0) Case3(-,$T_{benefit}$ =2,$A_{benefit}$ =0, $T_{autonomy}$ = 0, ,$A_{autonomy}$ = 1)	*Add new feature and principle clause* *Repair Cases*