# GENETH: A General Ethical Dilemma Analyzer

**Michael Anderson**
University of Hartford
West Hartford, CT, USA
anderson@hartford.edu

**Susan Leigh Anderson**
University of Connecticut
Storrs, CT, USA
susan.anderson@uconn.edu

## Abstract

We contend that all behavior of autonomous systems should be guided by explicit ethical principles determined through a consensus of ethicists. Such principles ensure the ethical behavior of complex and dynamic systems and further serve as a basis for justification of their actions as well as a control abstraction for managing unanticipated behavior. To provide assistance in developing ethical principles, in particular those pertinent to the behavior of autonmous systems, we have developed GENETH, a general ethical dilemma analyzer that, through a dialog with ethicists, codifies ethical principles in any given domain.

## 1 Introduction

Autonomous systems not only produce change in the environment but can monitor this environment to determine the effects of their actions and decide which action to take next. Ethical issues concerning the behavior of such complex and dynamic systems are likely to exceed the grasp of their designers and elude simple, static solutions. We assert that the behavior of such systems should be guided by explicit ethical principles determined through a revisable consensus of ethicists.

We believe it has been shown that ethical decision-making is, to a degree, computable. Utilitarianism, which advocates that determining the ethically correct action is a matter of performing "moral arithmetic" considering the likely future consequences of actions, has many advocates [ Bentham, 1799; Singer, 1979]. Even critics admit that utilitarian reasoning should be at least part of ethical decision-making [Ross, 1930]. [Anderson and Anderson, 2007] maintains that additional factors, such as respect for patient autonomy, can also be represented numerically.

To ensure ethical behavior, a system's possible actions should be weighed against each other to determine which is the most ethically preferable at any given moment. It is likely that ethical action preference of a large set of actions will be difficult or impossible to define extensionally as an exhaustive list of instances and instead will need to be defined intensionally in the form of rules. This more concise definition is possible since action preference is only dependent upon a likely smaller set of *ethically relevant features* that actions involve. Given this, action preference can be more succinctly stated in terms of satisfaction or violation of *duties* to either minimize or maximize (as appropriate) each feature. We refer to intensionally defined action preference as a *principle.*

A principle defines a binary relation over a set of actions that partitions it into subsets ordered by ethical preference with actions within the same partition having equal preference. As this relation is transitive, it can be used to sort a list of possible actions and find the most ethically preferable action(s) of that list. This relation could form the basis of *principle-based behavior*: a system decides its next action by using its principle to determine the most ethically preferable one(s). If such principles are explicitly represented, they have the further benefit of helping justify a system's actions as they can provide pointed, logical explanations as to why one action was chosen over another.

Although it may be fruitful to develop ethical principles for the guidance of autonomous machine behavior, it is a complex process that involves determining what the ethical dilemmas are in terms of ethically relevant features, which duties need to be considered, and how to weigh them when they pull in different directions. To help contend with this complexity, we have developed GENETH, a *general ethical dilemma analyzer* that, through a dialog with ethicists, helps codify ethical principles in any given domain including those pertinent to the behavior of autonmous systems.

## 2 GENETH

As it is likely that in many particular cases of ethical dilemmas ethicists agree on the ethically relevant features and the right course of action, generalization of such cases can be used to help discover principles needed for ethical guidance of the behavior of autonomous systems. A principle abstracted from cases that is no more specific than needed to make determinations complete and consistent with its training can be useful in making provisional determinations about untested cases. Cases can also provide a further means of justification for a system's actions: as an action is chosen for execution by a system, clauses of the principle that were instrumental in its selection can be

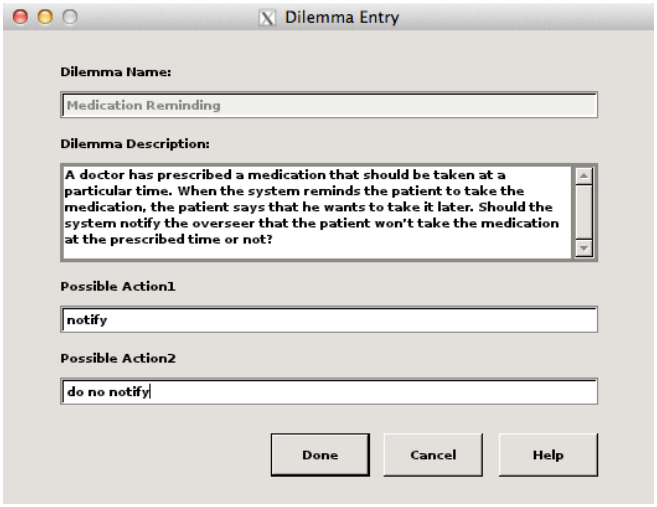


Figure 1 *Dilemma Entry*

determined and, as clauses of principles can be traced to the cases from which they were abstracted, these cases and their origin can be ascertained and used as justification for a system's action.

GENETH uses *inductive concept learning* [Lavrač and Džeroski, 1997] to infer a *principle of ethical action preference* from cases that is complete and consistent in relation to these cases. That is, a definition of a predicate $p$ is discovered such that $p(a_1,a_2)$ returns *true* if action $a_1$ is ethically preferable to action $a_2$. The principles discovered are *most general specializations,* covering more cases than those used in their specialization and, therefore, can be used to make and justify provisional determinations about untested cases. To minimize bias, GENETH is commited only to a knowledge representation scheme based on the concepts of ethically relevant features with corresponding degrees of presence/absence from which duties to minimize/maximize these features with corresponding degrees of satisfaction/violation of those duties are inferred. The system has no a priori knowledge regarding what these features, degrees, and duties might be but determines them as it is presented with example cases. Besides minimizing bias, there are two other advantages to this approach. Firstly, the principle in question can be tailored to the domain with which one is concerned. Different sets of ethically relevant features and duties can be discovered, through consideration of examples of dilemmas in the different domains in which machines will operate. Secondly, features and duties can be added or removed if it becomes clear that they are needed or redundant.

GENETH starts without any knowledge concerning particular features, degrees, or duties and a most general principle that simply states that all actions are equally ethically preferable (that is $p(a_1,a_2)$ returns true for all pairs of actions). An ethical dilemma and its two possible actions are input (Figure 1), defining the domain of the current cases and principle. The system then accepts example cases of this dilemma. A case is represented by the ethically relevant features it exhibits, as well as the determination as to which is the correct action given these features. Features are further delineated by the degree to which they are present or absent in one of the actions in question. From this information, duties are inferred either to maximize that feature (when it is present in the ethically preferable action or absent in the non-ethically preferable action) or minimize that feature (when it is absent in the ethically preferable action or present in the non-ethically preferable action). As features are presented to the system, the representation of cases is modified to include these inferred duties and the degree to which the cases satisfy or violate each one. Figure 2 shows a confirmation dialog for a case in which two features were input (of which only one is showing as features are tabbed in the interface) and two corresponding duties, as well as their degree of satisfaction/violation for each action in this case, were inferred.

As new cases of a given ethical dilemma are presented to the system, new duties and wider ranges of degrees are generated in GENETH through resolution of contradictions that arise. With two ethically identical cases – i.e. cases with the same ethically relevant feature(s) to the same degree of satisfaction or violation – an action cannot be right in one of these 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 detect such contradictions as they arise. If the determinations are correct, then there must either be a *qualitative* distinction or a *quantitative* difference between them that must be revealed. This can be translated into a difference in the ethically relevant features between the two cases, that is, a feature that appears in one but not in the other case; or a wider range of the degree of presence or absence of existing features must be considered that would reveal a difference between the cases, that is, there is a greater degree of presence or absence of existing features in one but not in the other case. In this fashion, GENETH systematically helps construct a concrete representation language that makes explicit features, their possible degrees of presence or absence, duties to maximize or minimize them, and their possible degrees of satisfaction or violation.

Ethical preference is determined from differentials of satisfaction/violation values of corresponding duties of two actions. Given two actions $a_1$ and $a_2$ and duty $d$, this differential can be notated as $da_1 - da_2$ or simply $\Delta d$. If an action $a_1$ satisfies a duty $d$ more (or violates it less) than another action $a_2$, then $a_1$ is ethically preferable to $a_2$ with respect to that duty. For example, given a duty with the possible values of $+1$ (for satisfied), $-1$ (for violated) and $0$ (for not involved), the possible range of the differential between the corresponding duty values is $-2$ to $+2$. That is, if this duty was satisfied in $a_1$ and violated in $a_2$, the differential for this duty in these actions would be $1--1$ or $+2$. On the other hand, if the this duty was violated in $a_1$ and satisfied in $a_2$, the differential for this duty in these actions would be $-1-1$ or $-2$. Although a principle can be defined that captures the notion of ethical preference in these cases simply as $p(a_1,a_2) \rightarrow \Delta d = 2$, such a definition overfits the given cases leaving no room for it to make determinations concerning untested cases. To overcome

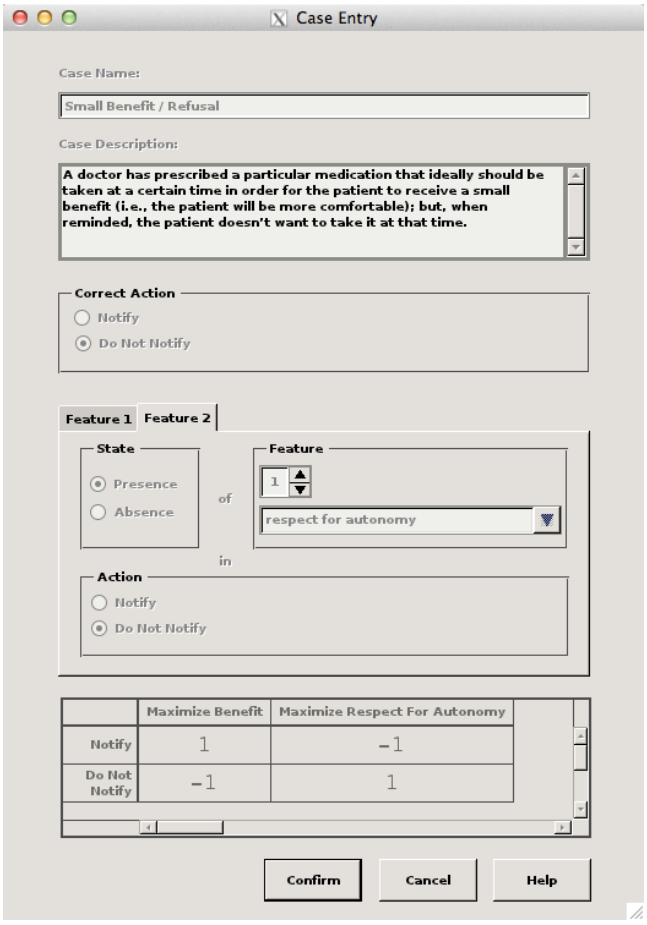


Figure 2 *Case Entry*

this limitation, what is required is a less specific principle that still covers (i.e. returns true for) positive cases (those where the first action is ethically preferable to the second) and does not cover negative cases (those where the first action is not ethically preferable to the second).

GENETH's approach is to generate a principle that is a *most general specification* by starting with the most general principle (i.e. one that covers all cases, positive and negative) and incrementally specialize it so that it no longer covers any negative cases while still covering all positive ones. These conditions correspond to the logical properties of consistency and completeness, respectively. In the single duty example above, the most general principle can be defined as $p(a_1,a_2) \rightarrow \Delta d \geq -2$ as the duty differentials in both the positive and negative cases satisfy the inequality. The specialization that the system employs is to incrementally raise the lower bounds of duties. In the example, the lower bound is raised by 1 resulting in the principle $p(a_1,a_2) \rightarrow \Delta d \geq -1$ which is true for the positive case (where $\Delta d = +2$) and false for the negative one (where $\Delta d = -2$). Unlike the earlier overfitted principle, this principle covers a positive case not in its training set. Consider when duty $d$ is neither satisfied or violated in $a_2$ (denoted by a 0 value for that duty). In this case, given a value of 1, $a_1$ is ethically preferable than $a_2$ since it satisfies $d$ more. This untested case is correctly covered by the principle as $\Delta d = 1$ satifies its inequality.

This simple example also shows why determinations on untested cases must be considered provisional. Consider when duty $d$ has the same value in both actions. These cases are negative examples (neither action is ethically preferable to the other in any of them) but all are still covered by the principle as $\Delta d = 0$ satifies its inequality. The solution to this inconsistency in this case is to specialize the principle even further to avoid covering these negative cases resulting in the final consistent and complete principle $p(a_1,a_2) \rightarrow \Delta d \geq 1$. This simply means that, to be considered ethically preferable, an action has to satisfy duty $d$ by at least 1 more than the other action in question (or violate it less by at least that amount) .

The system helps create a complete and consistent principle in a number of ways. It generates negative cases from positive ones entered (simply reversing the duty values for the actions in question) and presents them to the concept learner as cases that should not be covered. Determinations of cases are checked for plausibility by ensuring that the action deemed ethically preferable satisfies at least one duty more than the less ethically preferable action (or at least violates it less). As a contradiction indicates inconsistency, the system also checks for these between newly entered cases and previous cases, prompting the user for their resolution by a change in the determination, a new feature, or a new degree range for an existing feature in the cases. The system can provide guidance that leads to a more complete principle. It *seeks* cases from the user that either specify the opposite action of that of an existing case as ethically preferable or contradicts previous cases (i.e. cases that have the same features to the same degree but different determinations as to the correct action in that case). The system also seeks cases that involve duties and combinations of duties that are not yet represented in the principle. In doing so, new features, degree ranges, and duties are discovered that extend the principle, permitting it to cover more cases correctly. Lastly, incorrect system choice of minimization or maximization of a newly inferred duty signals that further delination of the case in question is needed.

To increase transparency, inferred principles are translated into more readily understandable textual representations and divided into disjuncts, each displayed on its own tab on the interface. Figure 3 shows one such translated disjunct (of the three that make up the complete principle; see Figure 4 for this) that entails the duty to minimize harm given the actions (*notify* and *do not notify*) for a medication reminding dilemma. The formal representation of this disjunct has the same structure as that generated in the previous simple example with its actions and duty instantiated:

$$p(\text{notify}, \text{do not notify}) \rightarrow \Delta\text{min harm} \geq 1$$

To further help mitigate the complexity of principle development, GENETH saves a dilemma's cases, features,

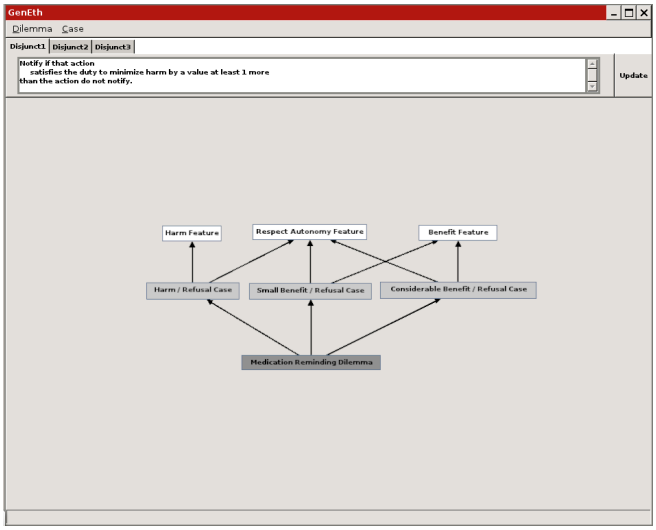


Figure 3 *Principle Display and Triplestore Interface*

and duties in a triplestore that maintains the relationships between these components. A graph is displayed that permits point and click inspection/editing of these components and relationships. Figure 3 shows the relationships hasCase and hasFeature for a medication reminding dilemma. The system also permits logical entailment between clauses of principles and the cases used to derive them to also be explored. Support for a clause can be generated by displaying all cases for which that clause returns true.

## 3 System Validation

As a first validation of GENETH, it was used to *re*discover representations and principles necessary to represent and resolve a general type of ethical dilemma in the domain of medical ethics previously discovered in [Anderson *et al.*, 2006]. In that work, an ethical dilemma was considered that involved the duties of beneficence, nonmaleficence, and respect for autonomy and a principle discovered that correctly (as per a consensus of ethicists) balanced these principles in all cases represented. The principle discovered was then used as a basis for an expert system and to guide the behavior of an autonomous robot [Anderson and Anderson, 2007; Anderson and Anderson, 2010]. This previous work assumed specific duties and ranges of satisfaction/violation degrees for these duties thus baising the learning algorithm toward them. GENETH lifts these assumptions, only assuming that such duties and ranges exist without specifying what they are.

To begin, GENETH was given the name and description of the previous ethical dilemma (one that a medication reminding system might face) and its two possible actions, *notify* and *do not notify*, specified (Figure 1). Next, the first case of this dilemma was input into the system where *notify* is the preferable action:

*A doctor has prescribed a medication that needs to be taken at a particular time or the patient will be harmed by not taking it at that time. When reminded, the patient refuses to take it at that time.*

The fact that *notify* is the ethically preferable action is because there is a particular ethically relevant feature in *do not notify*, namely, the presence of harm. Eliciting this information from the user, the system infers that there is a *duty to minimize harm* and asks for confirmation of that fact from the user. Given this confirmation, the system then infers the minimal range of degree of satisfaction/violation for this duty ($-1$ to $+1$) and assigns $+1$ for this duty to the *notify* action and $-1$ to the *do not notify* action. The system then generates the most general principle for the given duty, creates the opposite negative case for the given positive one by negating the values of the feature in each action, and then specializes the principle to uncover the negative one producing:

$$p(\text{notify}, \text{do not notify}) \rightarrow \Delta\text{min harm} \geq -1$$

In this instance, when requested, the system seeks a case where the ethically preferable action is *do not notify* as no case has yet been presented to the system in which that action is ethically preferable. The case entered is

*A doctor has prescribed a particular medication that ideally should be taken at a certain time in order for the patient to receive a small benefit (i.e., the patient will be more comfortable); but, when reminded, the patient doesn't want to take it at that time.*

The fact that *do not notify* is the ethically preferable action in this case is because there is a new ethically relevant feature involved, namely, benefit. What is unusual in this case is that the ethically preferable action displays an absence of this feature. Eliciting this information from the user, the system infers that there is a *duty to minimize benefit* and asks for confirmation of that fact from the user. In this case, no such confirmation is given and from this the system determines that there is instead a *duty to maximize benefit*, infers the minimal range of degree of satisfaction/violation for this duty ($-1$ to $+1$) and assigns $+1$ for this duty to the *notify* action and $-1$ to the *do not notify* action. The system then notes that it is not possible for an action to be ethically preferable and not satisfy at least one duty more (or violate it less) than the non-ethically preferable one. To resolve this there must either be a new ethically relevant feature or a wider range of the degree of presence or absence of existing feature. Since only one duty is involved in this case, widening the range of the possible degrees of that duty will not resolve the problem. Determining this, the system instead prompts the user for a new ethically relevant feature that is present in the *do not notify* action that makes it ethically preferable. The user then inputs the fact that not notifying respects the autonomy of the patient and confirms the inference that there is a *duty to maximize respect for autonomy*. Given this confirmation, the system then infers the minimal range of degree of satisfaction/violation for this new duty ($-1$ to $+1$) and assigns $+1$ for this duty to the *do not notify* action and $-1$ to the *notify* action.

At this point, the three duties that were assumed in the previous work have been inferred by the system from ethically relevant features input by the user. The system then asks the user to reconsider the previously entered case in light of the newly inferred duties and the user determines that the duty to maximize respect for autonomy is involved in the first case as well and assigns it a value of $-1$ for this duty to the *notify* action and $+1$ to the *do not notify* action in this case. The system then generates the most general principle for the given duties, creates the opposite negative cases for the given positive ones, and then specializes the principle to uncover the negative cases while still covering the positive ones producing:

$$p(\text{notify}, \text{do not notify}) \rightarrow$$
$$\Delta\text{min harm} \geq 1$$
$$\vee$$
$$\Delta\text{min harm} \geq -1 \wedge \Delta\text{max autonomy} \geq -1$$

Noting that the inferred duty to maximize benefit does not figure into the current principle, the system seeks a case from the user that, this time, has a greater value in the ethically prerferable action for that duty than in the other action. The case offered in response was:

*A doctor has prescribed a particular medication that would provide considerable benefit for the patient (e.g. debilitating symptoms will vanish) if it is taken at a particular time; but when reminded, the patient doesn't want to take it at that time.*

The ethically preferable action in this case is to *notify* and that action satisfies the duty to maximize benefit more than *do not notify*. As this case also involves the duty to maximize respect for autonomy and does so with exactly the same values as the previous case, the system notes that a contradiction exists. Given this, the user is asked to revisit the cases and either revise the determination of one of them or find a qualitative or quantitative difference between them. The user decides that a wider range of satisfaction/violation for benefit is required and the new case is differentiated from the previous case by assigning $+2$ for the duty to maximize benefit for the notify action and a $-2$ for the do not notify action. The system then generates the most general principle for the given duties, creates the opposite negative cases for the given positive ones, and then specializes the principle to uncover the negative cases while still covering the positive ones producing:

$$p(\text{notify}, \text{do not notify}) \rightarrow$$
$$\Delta\text{min harm} \geq 1$$
$$\vee$$
$$\Delta\text{max benefit} \geq 3$$
$$\vee$$
$$\Delta\text{min harm} \geq -1 \wedge \Delta\text{max benefit} \geq -3 \wedge \Delta\text{max autonomy} \geq -1$$

As this principle gives equivalent responses for the current dilemma to that given by the principle discovered in the previous research[1], GENETH has been shown able, in its interaction with an ethicist, to not only discover this principle but also to determine the knowledge representation scheme required to do so while making minimal assumptions.

The next step in system validation is to introduce a case not used in the previous research and show how GENETH can leverage its power to extend this principle. This new case is:

*A doctor has prescribed a particular medication that ideally should be taken at a particular time in order for the patient to receive a small benefit; but, when reminded, the patient refuses to respond, one way or the other.*

The ethically preferable action in this case is *notify* (the overseer needs to be informed that the patient is not responding) but, when given values for its features, the system determines that it contradicts a previous case in which the same values and features call for *do not notify*. Given this, the user is asked to revisit the cases and decides that the new case involves the absence of the ethically relevant feature of interaction. From this, the system infers a new duty to maximize interaction that, when the user supplies values for it in the contradicting cases, resolves the contradiction. The system then generates the most general principle for the given duties, creates the opposite negative cases for the given positive ones, and then specializes the principle to uncover the negative cases while still covering the positive ones and produces the final principle:

$$p(\text{notify}, \text{do not notify}) \rightarrow$$
$$\Delta\text{min harm} \geq 1$$
$$\vee$$
$$\Delta\text{max interaction} \geq 1$$
$$\vee$$
$$\Delta\text{max benefit} \geq 3$$
$$\vee$$
$$\Delta\text{min harm} \geq -1 \wedge \Delta\text{max benefit} \geq -3 \wedge$$
$$\Delta\text{max autonomy} \geq -1 \wedge \Delta\text{max interaction} \geq -1$$

The system, in conjunction with an ethicist, has instantiated its knowledge representation scheme to include: the ethically relevant features of *harm*, *interaction*, *benefit*, and *respect for autonomy* and the corresponding duties (and the specific degrees to which these duties can be satisfied or violated) to *minimize harm* ($-1$ to $+1$), *maximize interaction* ($-1$ to $+1$), *maximize benefit* ($-2$ to $+2$), and *maximize respect for autonomy* ($-1$ to $+1$). Actions are represented as tuples of values for these duties and cases are represented as tuples derived from the corresponding differential of these values, i.e. the difference of the value of the more ethically

[1] The current dilemma differs slightly from the previous one as it is more in service of autonomous systems and does not require as wide a range of values for the duty to maximize respect for autonomy.

preferable action's duty value and the value of the less ethically preferable action's duty value. As an example, consider the last case. *Notify* would be represented as the tuple (0, 1, -1, 1) (for the duties involving harm, benefit, autonomy, and interaction, respectively) and *do not notify* as the tuple (0, -1, 1, -1). The case would therefore be represented by the tuple comprised of the differences of these values: (0, 2, -2, 2). The fact that *notify* is ethically preferable to *do not notify* is supported by the second clause of the principle: the difference in the values for the duty to maximize interaction is 2 (greater than 1).

The representation scheme captures the complexity of ethical decision-making: that there are prima facie obligations that are utilitarian (maximizing benefit and minimizing harm), as well as those that are deontological (such as maximizing respect for autonomy), all based on ethically relevant features of ethical dilemmas. Specifying these elements sufficiently captures the ethical content of these dilemmas. Furthermore, the fact that there can be tensions between the prima facie duties necessitates that a principle for mediating between them must be discovered.

The discovered principle is complete and consistent with respect to its training cases and is general enough to cover cases not in this set. Given appropriate values for its pair of action's satisfaction/violation of the current duties, this principle can be used to determine, with respect to these duties, when notifying an overseer is ethically preferable to not notifying an overseer when a patient refuses to take his/her medication both directly and indirectly through inaction.

We believe that these results are promising and are encouraged to be optimistic regarding the possibility that GENETH will be instrumental in discovering principles that will permit machines to behave in a more ethical manner. We envision an extension and an even more subtle representation of ethical dilemmas in future research. The system will 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, as one should not only consider present options, the set of possible actions should include those 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.

## 4 Related Research

Although many have voiced concern over the impending need for machine ethics for decades (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.

Efforts by others that do attempt implementation have largely 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 without making judgements.

## 5 Conclusion

It can be argued that *machine ethics* ought to be the driving force in determining the extent to which autonomous systems should be permitted to interact with human beings. Autonomous systems that behave in a less than ethically acceptable manner towards human beings will not, and should not, be tolerated. Thus, it becomes paramount that we demonstrate that these systems will not violate the rights of human beings and will perform only those actions that follow acceptable ethical principles. Principles offer the further benefits of serving as a basis for justification of actions taken by a system as well as for an overarching control mechanism to manage unanticipated behavior of such systems. Developing principles for this use is a complex process and new tools and methodologies will be needed to help contend with this complexity. We offer GENETH as one such tool and have shown how it can help mitigate this complexity.

## Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grant Numbers IIS-0500133 and IIS-1151305. We would also like to acknowledge Mathieu Rodrigue of the University of California at Santa Barbara for his efforts in implementing the algorithm used to derive the results in this paper.

## References


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

[Anderson and Anderson, 2007] Anderson, M. and Anderson, S. L., Machine Ethics: Creating an Ethical Intelligent Agent, Artificial Intelligence Magazine, 28:4, Winter 2007.

[Anderson and Anderson, 2010] Anderson, M. and Anderson, S. L., "Robot be Good", Scientific American Magazine, October 2010.

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

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

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

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

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

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

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

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

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

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

[Singer, 1979] Singer, P. *Practical Ethics*, Cambridge University Press, Cambridge, 1979.

[Waldrop, 1987] Waldrop, M. M. 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.)