An Approach to Computing Ethics

Michael Anderson, *University of Hartford*

Susan Leigh Anderson, *University of Connecticut*

Chris Armen, *Amherst College*

This approach employs machine learning to help determine the ethically correct action when different ethical duties give conflicting advice.

It might seem impossible to "compute" ideas that humans feel most passionately about and have such difficulty codifying: their ethical beliefs. Despite this, our interdisciplinary team of an ethicist and computer scientists believe that it's essential that we try, since there will be benefits not only for the field of artificial intelligence, but ethics as well. We've been attempting to make ethics computable for three reasons. First, to avert possible harmful behavior from increasingly autonomous machines, we want to determine whether one can add an ethical dimension to them. Second, we want to advance the study of ethical theory by making it more precise. Finally, we want to solve a particular problem in ethical theory—namely, to develop a decision procedure for an ethical theory that involves multiple, potentially competing, duties.

We've adopted the action-based approach to ethical theory, where the theory tells us how we should act in ethical dilemmas. This approach lends itself to machine implementation by giving the agent either a single principle or several principles to guide its actions, unlike other approaches that don't clearly specify the correct action in an ethical dilemma. A good action-based ethical theory should have these qualities:[1]

- *Consistency.* The theory shouldn't contradict itself by saying that a single action in a given set of circumstances is simultaneously right and wrong.
- *Completeness.* It should tell us how to act in any ethical dilemma in which we might find ourselves.
- *Practicality.* We should be able to follow it.
- *Agreement with intuition.* The actions it requires and forbids should agree with expert ethicists' intuition.

The approach to computing ethics that we describe in this article is illustrated by MedEthEx,[2] a system that uses machine learning to resolve a biomedical ethical dilemma. As you'll see, for a dilemma involving three ethical duties and 18 possible cases, MedEthEx needed only four training sets to create an ethically significant decision principle that covered the remaining cases. This, we believe, lends support for our approach to computing ethics.

One approach to computing ethics

We started our project by programming the one theory that clearly attempts to make ethics computable: Hedonistic Act Utilitarianism. According to one of its creators, Jeremy Bentham, HAU simply involves doing "moral arithmetic."[3] HAU maintains that an action is right when, of all the possible actions open to the agent, it will likely result in the greatest net pleasure, or happiness, taking all those affected by the action equally into account. HAU involves first calculating the units of pleasure and displeasure that each person affected will likely receive from each possible action. It then subtracts the total units of displeasure from the total units of pleasure for each of those actions to get the total net pleasure. The action likely to produce the greatest net pleasure is the correct one. If the calculations end in a tie, where two or more actions are likely to result in the greatest net pleasure, the theory considers these actions equally correct.

The program Jeremy[4] is our implementation of HAU with simplified input requirements. Jeremy presents the user with an input screen that prompts for an action's description and the name of a person that action affects. It also requests a rough estimate of the amount (very pleasurable, somewhat pleasur-

1541-1672/06/$20.00 © 2006 IEEE
Published by the IEEE Computer Society

able, not pleasurable or displeasurable, somewhat displeasurable, or very displeasurable) and likelihood (very likely, somewhat likely, or not very likely) of pleasure or displeasure that the person would experience from this action. The user enters this data for each person affected by the action and for each action under consideration. When data entry is complete, JEREMY calculates the amount of net pleasure each action achieves. (It assigns 2, 1, 0, –1, or –2 to pleasure estimates and 0.8, 0.5, or 0.2 to likelihood estimates, and sums their product for each individual affected by each action.) It then presents the user with the action or actions achieving the greatest net pleasure.

An ideal version of a system such as JEREMY might well have an advantage over a human being in following HAU because you can program it to do the arithmetic strictly (rather than simply estimate), be impartial, and consider all possible actions. We conclude, then, that machines can follow HAU at least as well as human beings and perhaps even better, given the same data that human beings would need to follow the theory.

Even though HAU is consistent, complete, and can be made practical, most ethicists believe that it fails the test of agreement with intuition. Despite John Stuart Mill's heroic attempt in chapter five of *Utilitarianism* to show that considerations of justice can be subsumed under the utilitarian principle,[5] ethicists generally believe that HAU can allow for the violation of individual rights if this will likely result in the greatest net good consequences, taking everyone affected into account. One could, for instance, construct a case to show that HAU permits killing one unimportant person to save the lives of five important persons. This violates the intuition that it's wrong to kill one person to save several persons.

We have, however, adopted an aspect of HAU in our current approach to ethical decision making. When applying an ethical duty to a particular dilemma, we consider such factors as the duty's intensity and duration and the number of persons affected—which we have initially combined as the level of satisfaction or violation of the duty involved.

A more comprehensive ethical theory

In agreement with W.D. Ross,[6] we believe that all single-principle, absolute-duty ethical theories (such as HAU and Kant's Categorical Imperative, a principle that requires you to act in a way that can be universalized) are unacceptable because they don't appreciate the complexity of ethical decision making and the tensions that arise from different ethical obligations pulling us in different directions.

Ross's theory consists of seven *prima facie duties*. A prima facie duty is an obligation that we should try to satisfy but that can be overridden on occasion by another, stronger duty. Ross's suggested list of prima facie duties (which he says can be altered) captures the best of several single-principle ethical theories, while eliminating defects by allowing for exceptions. His suggested duties are those of

- *fidelity*—you should honor promises and live up to agreements that you've voluntarily made,

> In a given ethical dilemma, one of Ross's duties could tell us that a particular action is right, while another could tell us that the same action is wrong, making the theory inconsistent.

- *reparation*—you should make amends for wrongs you've done,
- *gratitude*—you should return favors,
- *justice*—you should treat people as they deserve to be treated, in light of their past behavior and rights they might have,
- *beneficence*—you should act so as to bring about the most amount of good,
- *nonmaleficence*—you should act so as to cause the least harm, and
- *self-improvement*—you should develop your talents and abilities to the fullest.

The first four duties are Kantian in spirit. The next two duties—beneficence and nonmaleficence—derive from the single utilitarian principle. However, they reflect Ross's insight that you must separate the possible good consequences and the likely harm that can be caused. The duty of nonmaleficence is stronger than that of beneficence, to account for our intuition that it's wrong to kill one person to save five. Finally, the last duty, that of self-improvement, captures the best of "ethical egoism" by acknowledging that we have a special duty to ourselves that we don't have to others.

While everyone agrees that Ross's duties seem intuitively plausible, he doesn't tell us how to determine the ethically correct action when the duties give conflicting advice, beyond saying that you should use your intuition to resolve the conflict. Unfortunately, this would let you rationalize doing whatever you feel like doing, by maintaining that a duty that supported that action is the most important one in the dilemma.

Without an objective decision procedure, furthermore, the theory can fail all the requirements of an acceptable action-based ethical theory. In a given ethical dilemma, one of Ross's duties could tell us that a particular action is right, while another could tell us that the same action is wrong, making the theory inconsistent. By not giving us a single ethically correct action in that dilemma, so that we don't know what we ought to do, the theory could also be considered incomplete and impractical. Finally, because you could rationalize doing an action that an ethical expert, and most of us, would consider wrong, the theory could fail the test of agreement with intuition.

We've concluded that the ideal ethical theory incorporates multiple prima facie duties, like Ross's theory, with some sort of a decision procedure to determine the ethically correct action in cases where the duties give conflicting advice.

A decision procedure for competing duties

We've formulated a method that could help make a multiple prima facie duty theory, like Ross's, workable. Our method essentially adopts John Rawls' *reflective equilibrium* approach to creating and refining ethical principles, which goes back and forth between particular cases and principles.[7] First, we find or create ethical dilemmas where tension exists between the prima facie duties and where ethicists have reached a consensus as to the correct action. (The system can learn a decision principle only to the extent that ethical experts agree on the answers to particular dilemmas.) We then use machine learning to abstract a general decision principle from those cases. Finally, we test this principle on further cases and refine it as needed to reflect ethicists' intuitions about the correct action in these other cases.

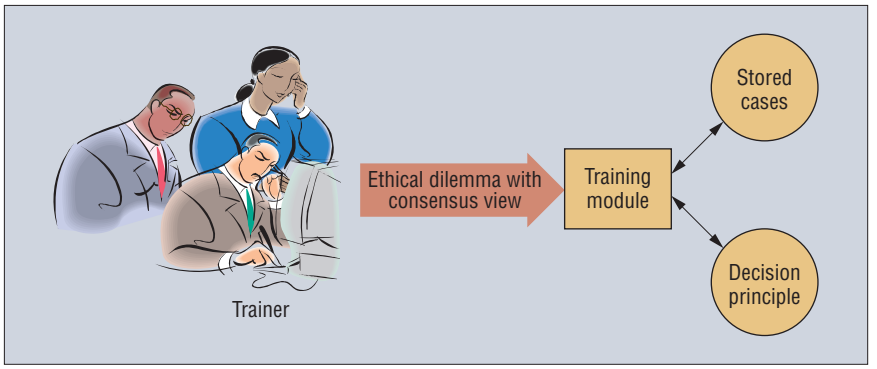


Figure 1. Developing a decision principle.

Our method uses a trainer (see figure 1) to develop the decision principle. It prompts the expert ethicist for an action's description and an estimate of each of the affected duties' satisfaction or violation level (very satisfied, somewhat satisfied, not involved, somewhat violated, or very violated). The expert enters this data for each action under consideration. When data entry is complete, the trainer seeks the intuitively correct action from the expert. It combines this information with the input case to form a new training example that it stores and uses to refine the decision principle. After such training, the decision principle can be used to provide the correct action for this case, should it arise in the future, as well as for all previous cases encountered. Furthermore, because the decision principle learned is the least-specific one required to satisfy cases seen so far, it might be general enough to be used to determine correct actions in previously unseen cases as well.

To capture expert ethical opinion, we use machine learning, currently *inductive-logic programming*, to learn the relationships between the duties involved in a particular dilemma.

ILP is a machine learning technique that inductively learns relations represented as first-order Horn clauses, classifying positive and negative examples of a relation.[8] To train a system using ILP, you present it with examples of the target relation, indicating whether they're positive (true) or negative (false). The object of training is for the system to learn a new hypothesis that, in relation to all input cases, is complete (covers all positive cases) and consistent (covers no negative cases).

We chose this machine learning technique for a number of reasons. First, the properties of the relationships between a set of prima facie duties aren't clear. For instance, do they form a partial order? Are they transitive? Do subsets of duties have different properties than other subsets? Simply assigning linear weights to the duties isn't sufficiently expressive to capture the relationships between those duties.[2] ILP provides a rich representation language that's better able to express these potentially nonclassical relationships. Furthermore, representing the relationships as Horn clauses lets us automatically confirm a decision principle's consistency regarding the relationships between duties across all cases. Finally, ILP's declarative representation language lets us more readily express, consult, and update commonsense background knowledge regarding duty relationships.

The decision principle learned is based on a predicate, *supersedes*(*Action1*, *Action2*), that's true when the first of the two actions that it is given is ethically preferable to the second. We represent each action as an ordered collection of values specifying the level of satisfaction or violation for each duty involved. We use this range of values: –2 represents a serious violation, –1 represents a less serious violation, 0 indicates that the duty is neither satisfied nor violated, +1 indicates minimal satisfaction, and +2 indicates maximal satisfaction.

A *decision procedure* determines the correct action for a particular case. The satisfaction or violation values for the duties of each possible action in a given case are provided to the decision procedure, which then determines the ethically preferable action (if any) according to the learned decision principle. The decision procedure uses resolution, an automatic method of theorem proving, to test the *supersedes* predicate. For example, if *supersedes*(*Action1*, *Action2*) is true, *Action1* supersedes *Action2* according to the decision principle and will be output by the decision procedure as the ethically preferable action.

Ethical advisor systems

A good first step toward the eventual goal of developing machines that can follow ethical principles is creating programs that enable machines to act as ethical advisors to human beings.[2] We begin this way for four pragmatic reasons.

First, one could start by designing an advisor that gives guidance to a select group of persons in a finite number of circumstances, thus reducing the assignment's scope.

Second, the general public will probably more easily accept machines that just advise human beings than machines that try to behave ethically themselves. In the first case, it's human beings who will make ethical decisions by deciding whether to follow the machine's recommendations, preserving the idea that only human beings will be moral agents. The next step in the Machine Ethics project is likely to be more contentious: creating machines that are autonomous moral agents.

Third, a problem for AI in general, and so for this project too, is how to get needed data—in this case, the information from which to make ethical judgments. With an ethical advisor, human beings can be prompted to supply the needed data.

Finally, ethical theory hasn't advanced to the point where there's agreement, even by ethical experts, on the correct answer for all ethical dilemmas. An advisor can recognize this fact, passing difficult decisions that must be made in order to act onto the human user.

Figure 2 depicts a general architecture for an ethical advisor system. A user inputs details of a particular case into the system and is presented with the ethically preferable action in accordance with the decision principle. A *knowledge-based interface* provides guidance in selecting the duties involved and their satisfaction or violation levels for the case. The interface uses knowledge derived from ethicists concerning the dimensions and duties of the particular ethical dilemma. Knowledge is represented as finite-state automata (FSA) for each duty entailed. Questions pertinent to the dilemma serve as start and intermediate states, and intensities of duty satisfaction or violation levels (as well as requests for more information) are final states. The input to the interface is the user's responses to the questions posed; the output is a case with duty satisfaction or violation levels corresponding to these responses. This interface provides the experienced guidance necessary to navigate the subtleties of deter-

mining satisfaction or violation levels of duties in particular cases.

Given the details of the case from the knowledge-based interface, the decision procedure consults the decision principle and determines whether one action supersedes all others in the current case. If it discovers such an action, it outputs that action as the ethically correct action in this case—that is, the action that's consistent with the system's training.

An ethical advisor: MedEthEx

We decided to create an ethical advisor, applying our method of creating a decision principle, using a theory that has only four prima facie duties, a constrained domain, and lacks a decision procedure: Tom Beauchamp and James Childress's Principles of Biomedical Ethics.[9] PBE uses Ross's duties of beneficence, nonmaleficence, and justice and adds the principle of *respect for autonomy*. This last principle reflects the recent shift from a paternalistic model of the healthcare worker-patient relationship to one giving the patient a more active role. For a patient's healthcare decision to be fully autonomous,[10] it must meet three criteria. First, it must be based on sufficient understanding of his or her medical situation and the likely consequences of forgoing treatment. Second, it must be sufficiently free of external constraints (for example, pressure by others or external circumstances, such as a lack of funds). Finally, it must be sufficiently free of internal constraints (for example, pain or discomfort, the effects of medication, or irrational fears or values that will likely change over time).

We chose PBE and biomedical ethical dilemmas for five reasons. First, PBE uses a more manageable total of four duties, instead of Ross's seven. Second, one member of our research team has a biomedical-ethics background. Third, healthcare workers will likely have the information needed to judge whether a particular duty is involved in an ethical dilemma and to judge that duty's intensity. Fourth, more agreement exists among biomedical ethicists as to the ethically preferable action than in other areas of applied ethics. Finally, there's a pressing need for ethical advice in this area, as biomedical research introduces new, challenging ethical dilemmas and as baby boomers begin to age (many ethical dilemmas involve end-of-life care).

MedEthEx offers guidance on the following type of biomedical ethical dilemma:

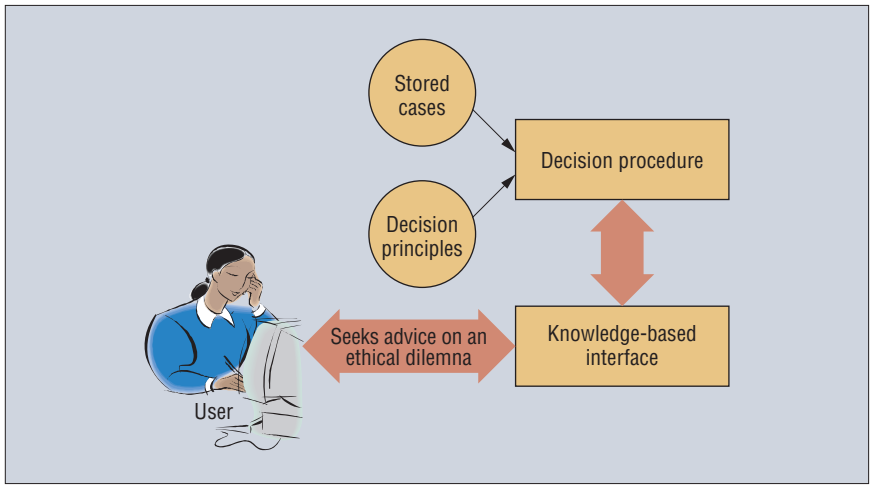


Figure 2. An architecture for an ethical advisor system.

> Given the details of the case from the knowledge-based interface, the decision procedure consults the decision principle and determines whether one action supersedes all others in the current case.

A healthcare professional has recommended a particular treatment for her competent adult patient, but the patient has rejected it. Should the healthcare professional try 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 shouldn't challenge the patient's autonomy unnecessarily. On the other hand, the healthcare professional might have concerns about why the patient is refusing the treatment—that is, whether the decision is fully autonomous. This dilemma is constrained to three of the four duties of PBE (nonmaleficence, beneficence, and respect for autonomy) and involves only two possible actions in each case. We've drawn on the intuitions of Allen Buchanan and Dan Brock[11] and our project's ethicist (whose views reflect a general consensus) to determine the correct actions in particular cases of this type of dilemma.

In the type of dilemma we consider, we can assign specific meanings to each duty's possible values. For nonmaleficence,

- –2 means that this action will likely cause severe harm to the patient that could have been prevented,
- –1 means that this action will likely cause some harm to the patient that could have been prevented,
- 0 means that this action isn't likely to cause or prevent harm to the patient,
- +1 means that this action will likely prevent harm to the patient to some degree, and
- +2 means that this action will likely prevent severe harm to the patient.

For beneficence,

- –2 means that the other action would likely have improved the patient's quality of life significantly,
- –1 means that the other action would likely have improved the patient's quality of life somewhat,
- 0 means that neither action is likely to improve the patient's quality of life,
- +1 means that this action will likely improve the patient's quality of life somewhat, and
- +2 means that this action will likely improve the patient's quality of life significantly.

For respect for autonomy,

- –1 means not immediately acquiescing to the patient's wishes but trying again to change the patient's mind,

Table 1. The levels of duty satisfaction or violation for the two possible actions in four MedEthEx training cases. A check mark indicates the ethically correct action in each case.

Training case no. and action	Nonmaleficence value	Beneficence value	Autonomy value
Case 1			
Try Again	+2	+2	–1
✓ Accept	–2	–2	+2
Case 2			
Try Again	0	+1	–1
✓ Accept	0	–1	+1
Case 3			
✓ Try Again	+1	+1	–1
Accept	–1	–1	+1
Case 4			
✓ Try Again	0	+2	–1
Accept	0	–2	+1

- +1 means that the healthcare worker acts according to the patient's wishes but believes that the patient's decision isn't fully autonomous, and
- +2 means that the healthcare worker acts according to the patient's wishes and believes that the patient's decision is fully autonomous.

(Because this dilemma always involves autonomy, but never to the extent of forcing a treatment on the patient, 0 and –2 aren't options.)

As an example, consider a specific case of the type of dilemma we're considering. A patient refuses to take an antibiotic that's almost certain to cure an infection that would otherwise likely lead to his death. He decides this on the grounds of long-standing religious beliefs that forbid him to take medications. The correct action in this case is for the healthcare worker to accept the patient's decision as final because, although severe harm (his death) will likely result, his decision can be seen as being fully autonomous. The healthcare worker must respect a fully autonomous decision of a competent adult patient, even if he or she disagrees with it, because the decision concerns the patient's body and a patient should have control over what is done to his or her body. This dilemma appears as training case 1 in table 1. In this case, the predicate *supersedes*(Accept, Try Again) would be true and *supersedes*(Try Again, Accept) would be false.

Training MedEthEx

We presented MedEthEx with four positive training cases, drawn from the 18 possible cases, and four negative cases derived by simply exchanging the actions of the positive training examples. Training case 1 was just described in the previous paragraph.

In training case 2, a patient won't consider taking medication that could only help alleviate some symptoms of a virus that must run its course. He refuses the medication because he has heard untrue rumors that the medication is unsafe. Even though the decision is less than fully autonomous, because it's based on false information, the little good that could come from taking the medication doesn't justify trying to change his mind. So, the doctor should accept his decision.

In training case 3, a patient with incurable cancer refuses further chemotherapy that will let him live a few months longer, relatively pain free. He refuses the treatment because, ignoring the clear evidence to the contrary, he's convinced himself that he's cancer-free and doesn't need chemotherapy. The ethically preferable answer is to try again. The patient's less than fully autonomous decision will lead to some harm (dying sooner) and deny him the chance of a somewhat longer life (a violation of the duty of beneficence), which he might later regret.

In training case 4, a patient, who has suffered repeated rejection from others due to a very large noncancerous abnormal growth on his face, refuses to have simple and safe cosmetic surgery to remove the growth. Even though this has negatively affected his career and social life, he's resigned himself to being an outcast, convinced that this is his lot in life. The doctor is convinced that his rejection of the surgery stems from depression due to his abnormality and that having the surgery could vastly improve his entire life and outlook. The doctor should try again to convince him because so much of an improvement is at stake and his decision is less than fully autonomous.

Table 1 summarizes the levels of duty satisfaction or violation for both of the possible actions in all four training cases and indicates the correct action in each case.

We can more succinctly characterize the cases using the difference between the values for duties in the ethically preferable action and the values for corresponding duties in the less preferable action. For example, in training case 1, the differences between the duties of the Accept and the Try Again actions are –4, –4, 3. Positive differences signify duties that are favored in the ethically preferable action (respect for autonomy in this example); negative differences signify duties that are favored in the less preferable action (nonmaleficence and beneficence in this example).

Figure 3 denotes each possible case in the dilemma under consideration. It represents each case as a triple denoting the differences in corresponding duties, depicted as a point in 3-space. The space as a whole represents all 729 discrete, three-duty cases possible given the range of values permitted for each duty's level of satisfaction or violation. The highest, right-most point (4, 4, 4) represents the case in which each duty is maximally in favor of the ethically preferable action. The lowest, left-most point (–4, –4, –4) represents the case in which each duty is maximally in favor of the less ethically preferable action. Blue points signify positive training cases, red points signify negative training cases, and white points signify the remaining 14 possible positive cases.

The learning task is to find a set of clauses that covers all the positive training examples while not covering any negative training examples. Figure 4 illustrates the set of clauses defining the *supersedes*(*Action1*, *Action2*) predicate learned by MedEthEx, where Δ<*duty*> denotes the difference between *Action1*'s <*duty*> value and *Action2*'s <*duty*> value.

Each clause specifies a lower bound for each of the three duty differentials that must hold for that clause to be true. As each clause is joined to the others disjunctively, any one true clause will cause the *supersedes* predicate to be true. For example, the third clause states that in order for it to consider *Action1* ethically preferable to *Action2*,

- the value for nonmaleficence must be 1 or more in favor of *Action1*,
- the value for beneficence can be any value (as –4 is the lowest possible bound), and

- the value for respect for autonomy can be in favor of *Action2* by no more than 2.

This set of clauses, in relation to the type of dilemma under consideration, represents a decision principle that states that a healthcare worker should challenge a patient's decision if it isn't fully autonomous and there's either any violation of nonmaleficence or a severe violation of beneficence. This philosophically interesting result gives credence to Rawls' method of reflective equilibrium. Through abstracting a principle from intuitions about particular cases and then testing that principle on further cases, we've produced a plausible principle that tells us which action is ethically preferable when specific duties pull in different directions in a particular ethical dilemma. Furthermore, this abstracted principle supports Ross's insight that violations of nonmaleficence should carry more weight than violations of beneficence.

Visualizing the learned decision principle

Figure 5 graphically represents the semantics of each clause that MedEthEx learned. Each blue point represents the training case for which the clause was generated, white points represent the remaining positive cases that the clause covers, red points represent the negative training cases, and the volume represents the space assumed to contain only positive cases.

Figure 5a represents the first clause, generated to cover training case 1. It covers eight of the remaining positive cases and shows how the learning algorithm greedily occupies space surrounding a training example that doesn't contain a negative example. In this case, the resulting volume simply represents the commonsense inference that if any duty differential for *Action1* in a case is equal to or greater than a positive training case's duty differential, that case will also be positive.

Figure 5b represents the second clause, for training case 2. It covers one of the remaining positive cases and shows how the space occupied by the learning algorithm can be somewhat speculative. Although this space contains the same commonsense inference as that for training case 1, it extends its boundaries to lower values than those of its training case. The intuition is that, because this space contains no negative cases, we can safely assume that the cases this space covers are positive until we receive evidence to the contrary. If and when the training module finds a negative case in this space, it will modify the clause to account for it. This is in the spirit of reflective equilibrium, which advocates abstracting principles from cases while permitting modification of these principles when new cases require it.

Figure 5c represents the third clause, for training case 3. It covers five of the remaining positive cases and, like the space generated for training case 2, extends its boundaries lower than those of its training case. Finally, figure 5d represents the fourth clause, for training case 4. It covers the last remaining positive cases and extends its boundaries lower than its training case as well.

Clearly, other partitions of this space can cover all positive cases. Figures 5a, 5c, and 5d reflect this particular implementation's bias toward constraining the fewest number of duty differentials. Figure 5a shows that only the duty differential for respect for autonomy has any bearing on the first clause's truth value, permitting two degrees of freedom in the graph—the duty differentials of beneficence and nonmaleficence. Likewise, figures 5c and 5d show the degrees of freedom permitted in the last two clauses—the duty differentials of beneficence and nonmaleficence, respectively.

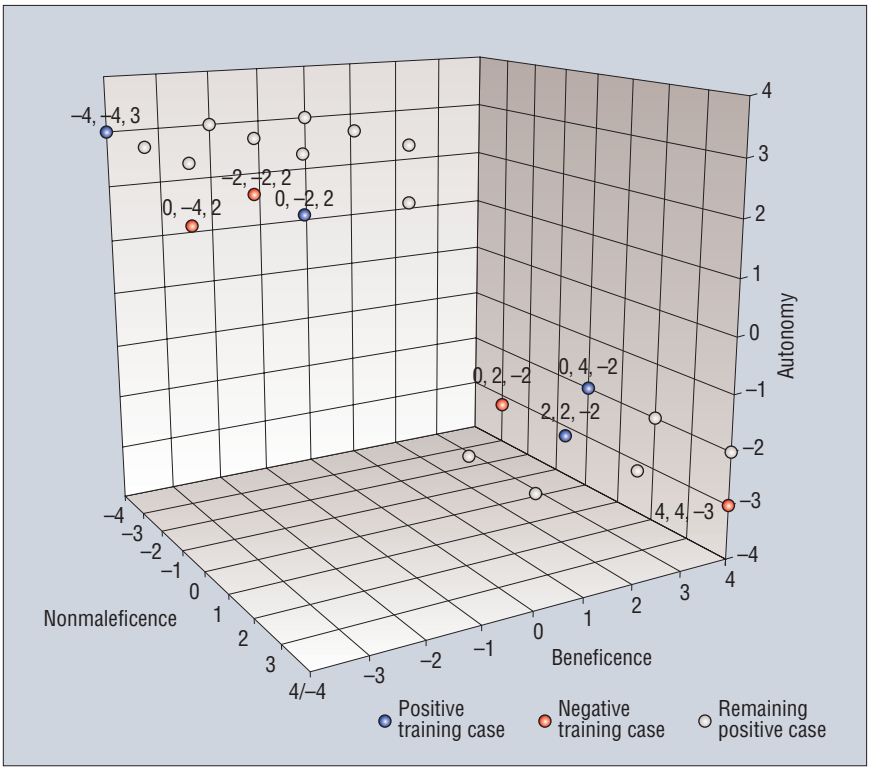


Figure 3. All 18 possible cases of the biomedical ethical dilemma. We represent each case as a triple denoting the differences in corresponding duties, depicted as a point in 3-space.

$$\Delta nonmaleficence \geq -4 \wedge \Delta beneficence \geq -4 \wedge \Delta autonomy \geq 3$$
$$\vee$$
$$\Delta nonmaleficence \geq -1 \wedge \Delta beneficence \geq -3 \wedge \Delta autonomy \geq -1$$
$$\vee$$
$$\Delta nonmaleficence \geq 1 \wedge \Delta beneficence \geq -4 \wedge \Delta autonomy \geq -2$$
$$\vee$$
$$\Delta nonmaleficence \geq -4 \wedge \Delta beneficence \geq 3 \wedge \Delta autonomy \geq -2$$

Figure 4. The set of clauses defining the *supersedes*(*Action1*, *Action2*) predicate learned by MedEthEx.

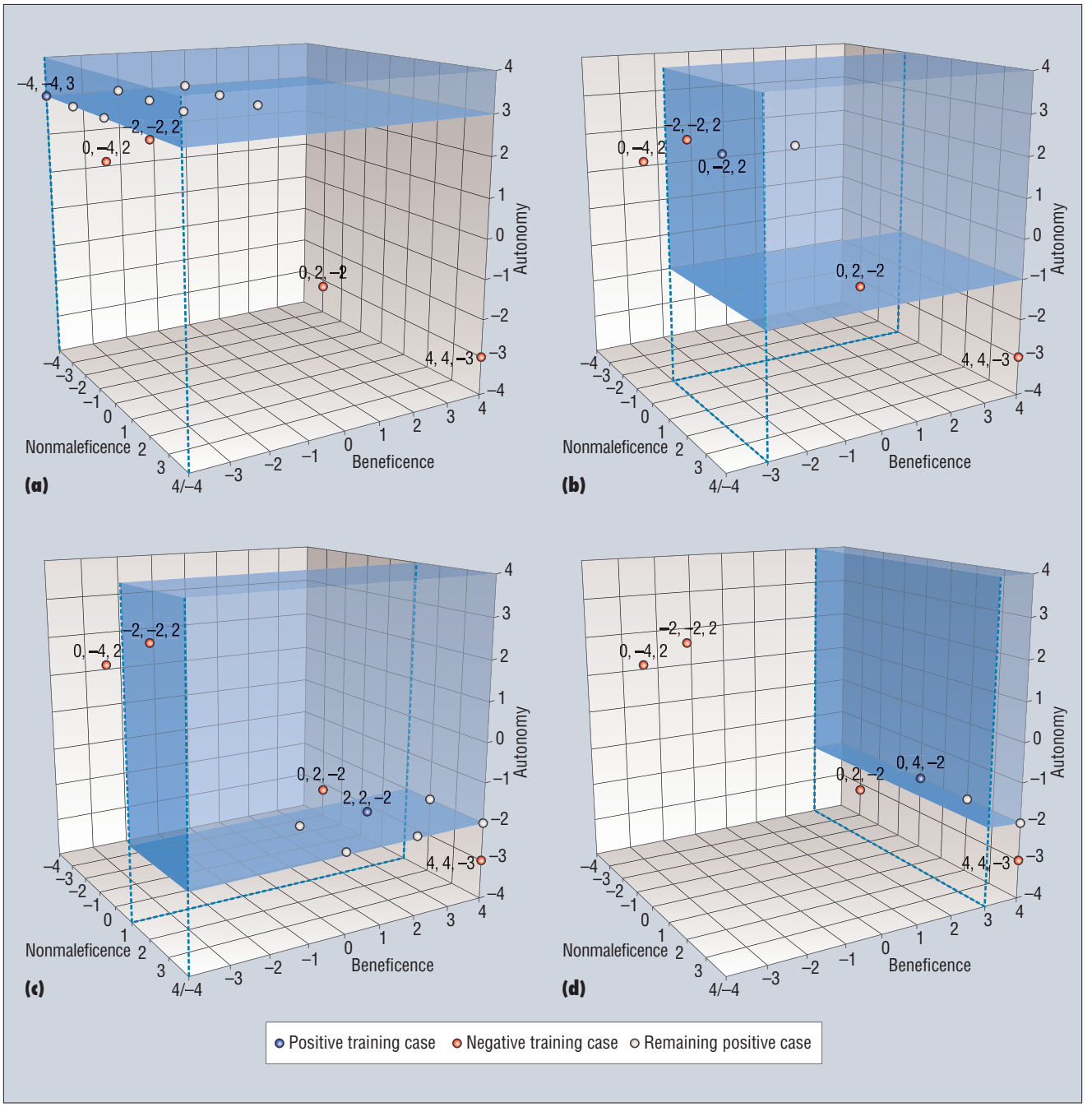


Figure 5. The duty differential space covered by (a) the first clause of the decision principle that MedEthEx learned, (b) the second clause, (c) the third clause, and (d) the fourth clause.

Figure 6 represents the union of figures 5a through 5d. As such, it represents the total space of duty differentials in the dilemma under consideration where MedEthEx learned *supersedes*(*Action1*, *Action2*) as true. We can further verify (or refine) this space by developing other positive and negative cases and determining where they fall within it. If a positive case falls outside the space or a negative case falls within it, we must refine the space to account for it. If we can find no such examples, the space has been further verified.

Future research

We plan to develop MedEthEx further to see whether the learned decision principle works in other dilemmas involving the same three duties. It will be interesting to add the fourth duty, justice, to see to what extent there's a consensus among bioethicists in cases where this duty is involved from which we can abstract a decision principle. There's disagreement about what is just among those working in ethics in other domains, but there might not be disagreement among bioethicists. Furthermore, we would like to see if our approach to learning decision principles

will prove viable for other sets of duties, including sets of higher cardinality, and in other domains. It's reasonable to believe that each specific applied ethics domain (legal ethics, business ethics, journalistic ethics, and so on) involves juggling a set of prima facie duties that's specific to that domain. In each case, there will be the problem of abstracting a decision principle to determine the correct action when the duties conflict. We plan, therefore, to look at other domains to see whether our approach to creating an ethical-advisor system might be helpful in solving ethical dilemmas for those who work in those domains.

Our long-term goal, of course, is to have a machine follow ethical principles itself, rather than simply advising human beings as to how to behave ethically. We believe, though, that the first step in the development of machine ethics must be to work on making ethics computable. If that task can't be accomplished, at least to the extent to which ethics experts are in agreement as to what's ethically right, then creating a machine that behaves ethically will be impossible. Creating ethical-advisor systems lets us explore the extent to which ethics can be computed in specific domains. Once ethics experts are comfortable with the results, then an ethical dimension can, at least in principle, be incorporated into machines that function in those domains. This should not only avert unethical behavior on the part of machines, but also allow them to do tasks that we would have previously thought only human beings should do.

The process of making an ethical theory precise enough to be computed will likely sharpen and revise the theory itself. This research provides an opportunity for applying AI techniques in a new domain and developing new areas of applied ethics, as well as making a contribution to ethical theory itself.

Our results demonstrate that a problem in ethical theory—devising a decision procedure for an ethical theory involving multiple prima facie duties—can be solved at least in a constrained domain and that AI techniques can help solve it. So, we believe that not only can you train a machine to make ethical decisions but also that machines can help human beings codify the principles that should guide them in ethical decision making. ■

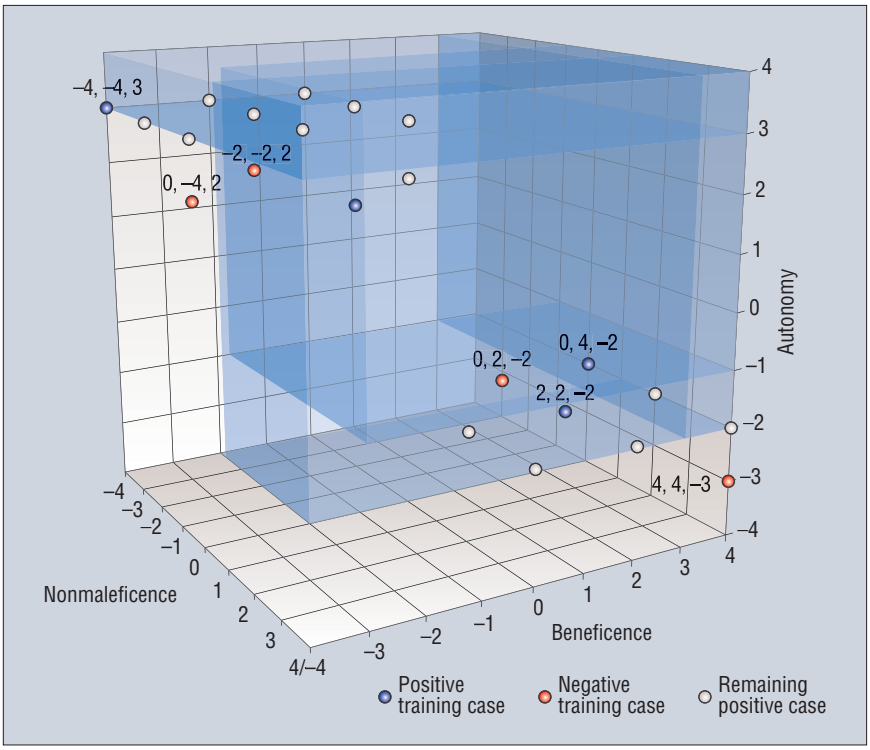


Figure 6. The duty differential space covered by the combined clauses.

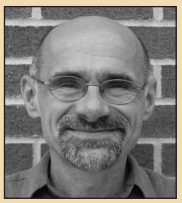
References

1. S. Anderson, "We Are Our Values," *Questioning Matters*, D. Kolak, ed., Mayfield, 1999, p. 599.
2. M. Anderson, S. Anderson, and C. Armen, "Toward Machine Ethics: Implementing Two Action-Based Ethical Theories," *Proc. AAAI 2005 Fall Symp. Machine Ethics*, AAAI Press, 2005, pp. 1–16.
3. J. Bentham, *An Introduction to the Principles and Morals of Legislation*, Oxford Univ. Press, 1799.
4. M. Anderson, S. Anderson, and C. Armen, "MedEthEx: Toward a Medical Ethics Advisor," *Proc. AAAI 2005 Fall Symp. Caring Machines: AI in Eldercare,* AAAI Press, 2005, pp. 9–16.
5. J.S. Mill, "Utilitarianism," *Utilitarianism and Other Writings*, M. Warnock, ed., New American Library, 1974, chapter 5.
6. W.D. Ross, *The Right and the Good*, Clarendon Press, 1930.
7. J. Rawls, "Outline for a Decision Procedure for Ethics," *Philosophical Rev.*, vol. 60, 1951.
8. N. Lavrac and S. Dzeroski, *Inductive Logic Programming: Techniques and Applications*, Ellis Harwood, 1997.
9. T.L. Beauchamp and J.F. Childress, *Principles of Biomedical Ethics*, Oxford Univ. Press, 1979.
10. T.A. Mappes and D. DeGrazia, *Biomedical Ethics*, 5th ed., McGraw-Hill, 2001, pp. 39–42.
11. A.E. Buchanan and D.W. Brock, *Deciding for Others: The Ethics of Surrogate Decision Making*, Cambridge Univ. Press, 1989, pp. 48–57.

The Authors

Michael Anderson is an associate professor of computer science at the University of Hartford. Contact him at anderson@hartford.edu.

Susan Leigh Anderson is a professor of philosophy at the University of Connecticut, Stamford. Contact her at susan.anderson@uconn.edu.



Chris Armen is a visiting assistant professor at Amherst College's Department of Mathematics and Computer Science. Contact him at carmen@amherst.edu.