Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions

Bruce M. McLaren, *Carnegie Mellon University*

This article reviews progress in developing computational models of ethical reasoning, looks in detail at two seminal machine-ethics systems—Truth-Teller and Sirocco—and discusses promising future directions.

How can machines support or, even more significantly, replace humans in performing ethical reasoning? This question greatly interests machine ethics researchers. Imbuing a computer with the ability to reason about ethical problems and dilemmas is as difficult a task as there is for AI scientists and engineers. First, ethical reasoning is based on abstract principles that you can't easily apply in a formal, deductive fashion. So, the favorite tools of logicians and mathematicians, such as first-order logic, aren't applicable. Second, throughout intellectual history, philosophers have proposed many theoretical frameworks, such as Aristotelian virtue theory,[1] the ethics of respect for persons,[2] act utilitarianism,[3] utilitarianism,[4] and prima facie duties,[5] and no universal agreement exists on which ethical theory or approach is the best. Furthermore, any of these theories or approaches could be the focus of inquiry, but all are difficult to make computational without relying on simplifying assumptions and subjective interpretation. Finally, ethical issues touch human beings profoundly and fundamentally. The premises, beliefs, and principles that humans use to make ethical decisions are quite varied, not fully understood, and often inextricably intertwined with religious beliefs.

How do you take such uniquely human characteristics and distill them into a computer program? Undaunted by the challenge, scientists and engineers have, over the past 15 years, developed several computer programs that take initial steps in addressing these difficult problems. Here, I briefly describe a few of these programs and discuss in detail two programs that I developed, both of which employ techniques from the area of AI known as *case-based reasoning* and implement aspects of the ethical approach known as *casuistry*. One of these programs, Truth-Teller, accepts a pair of ethical dilemmas and describes the salient similarities and differences between them, from both an ethical and a pragmatic perspective. The other program, Sirocco, accepts a single ethical dilemma and retrieves other cases and ethical principles that might be relevant.

Neither program was designed to reach an ethical decision. I believe that reaching an ethical conclusion is, in the end, a human decision maker's obligation. Even if I believed these computational models were up to the task of autonomously reaching correct conclusions to ethical dilemmas, having a computer program propose decisions oversimplifies the obligations of human beings and makes assumptions about the "best" form of ethical reasoning. Rather, the aim of my work has been to develop programs that produce relevant information that can help humans as they struggle with difficult ethical decisions, as opposed to providing fully supported ethical arguments and conclusions. In other words, the programs are intended to stimulate the "moral imagination"[6] and help humans reach decisions. The difficulties in developing machines that can reason ethically present an intellectual and engineering challenge of the first order to the field of machine ethics. The long history of science and technology is rife with problems that have excited the innovative spirit of scientists, philosophers, and engineers,

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

and the nascent field of machine ethics presents many new challenges. Even if I achieve my goal of creating a reliable "ethical assistant" but don't develop a fully autonomous ethical reasoner, it will be a significant step forward for machine ethics.

Attempts to build programs that support or implement ethical reasoning

Two of the earliest programs aimed at ethical reasoning, the Ethos System[7] and the Dax Cowart program,[8] were designed to help students work their way through thorny practical-ethics problems. Neither is an AI program, but each models aspects of ethical reasoning and acts as a pedagogical resource. Both programs feature an open, exploratory environment complete with video clips to provide a visceral experience of ethical-problem solving.

Donald Searing developed Ethos to accompany the engineering ethics textbook by Charles Harris, Michael Pritchard, and Michael Rabins.[6] Ethos provides a few prepackaged example dilemmas, including video clips and interviews, to help students explore real ethical dilemmas that arise in engineering. Ethos encourages rational, consistent ethical-problem solving in two ways. First, it provides a framework in which you can rationally apply moral beliefs. Second, it records the step-by-step decisions that an ethical decision maker takes when resolving a dilemma, so that those steps can later be reflected upon. The program decomposes moral decision making into three major steps:

1. framing the problem,
2. outlining the alternatives, and
3. evaluating those alternatives.

The Dax Cowart program is an interactive multimedia program designed to explore the practical-ethics issue of a person's right to die. The program focuses on the real case of Dax Cowart, a victim of severe burns, crippling injuries, and blindness who insists on his right to die throughout enforced treatment for his condition. The central question is whether he should be allowed to die. The program presents video clips of interviews with Dax's doctor, lawyer, mother, nurses, and Dax himself to let the user experience the issue from different viewpoints. The program also presents clips of Dax's painful burn treatment to provide an intimate sense of his predicament. The program periodically asks the user to make judgments on whether Dax's request to die should be granted. Depending on how the user answers, the program branches to present information and viewpoints that might cause reconsideration of that judgment.

Both Ethos and the Dax Cowart program are intended to instill a deep appreciation of the complexities of ethical decision making by letting the user interactively and iteratively engage with the various resources they provide. However, neither program involves any intelligent processing. All the steps and displays are effectively "canned," with deterministic feedback based on the user's actions.

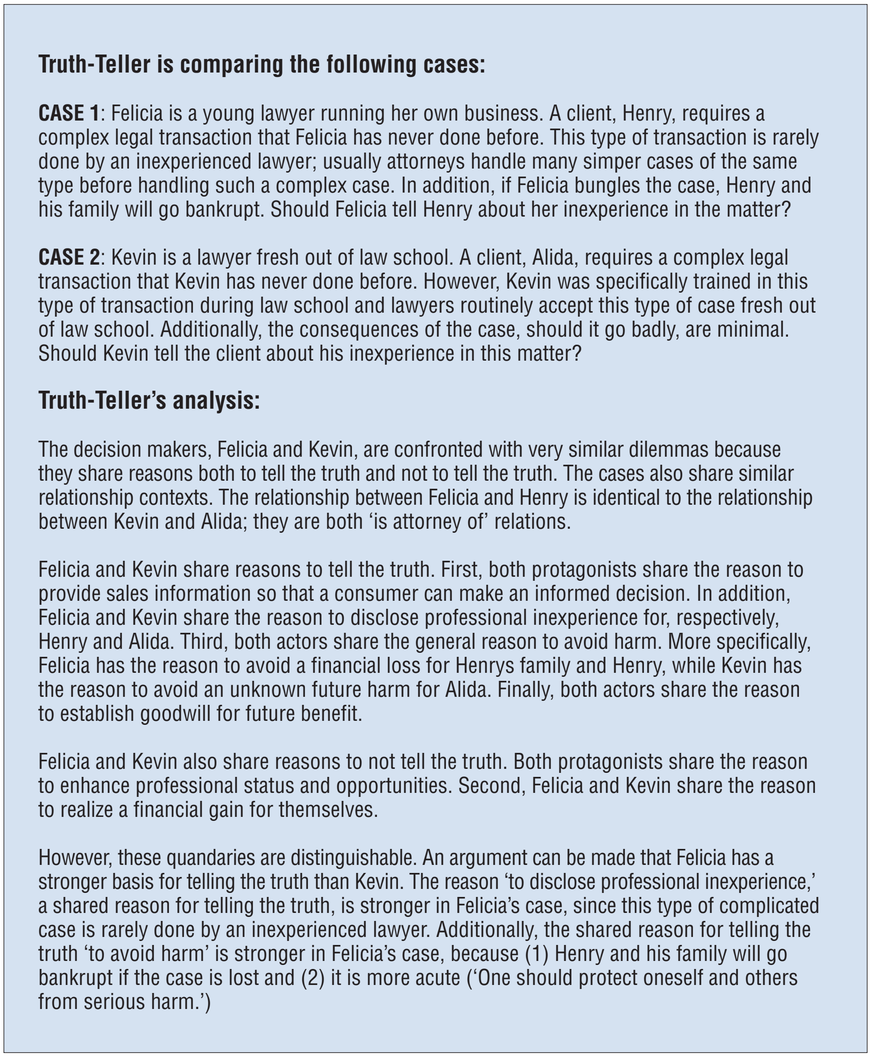
Research that has focused more specifically on the computational modeling of ethical reasoning includes that of Russell Robbins and William Wallace.[9] Their proposed computational model combines collaborative problem solving (that is, multiple human subjects discussing an ethical issue), the psychological theory of planned behavior, and the belief-desire-intention model of agency. As a decision aid, this computational model is intended to take on multiple roles, including advisor, group facilitator, interaction coach, and forecaster for human subjects as they discuss and try to resolve ethical dilemmas. This system has only been conceptually designed, not implemented, and the authors might have overreached, in a practical sense, by trying to combine such a wide range of theories and technologies in a single model. But their ideas could serve as the foundation for future computational models of ethical reasoning.

> The difficulties in developing machines that can reason ethically present an intellectual and engineering challenge of the first order to the field of machine ethics.

Earlier, Robbins, Wallace, and Bill Puka[10] implemented and experimented with a more modest system for supporting ethical-problem solving. They implemented this system as a series of Web pages, containing links to relevant ethical theories and principles, and a simple ethics "coach." The three researchers performed an empirical study in which the system's users were able to identify, for instance, more alternative ways to address a given ethical problem than subjects who used Web pages that didn't have the links or coaching. Their research is an excellent illustration of the difficulties confronting those who wish to build computational models of ethical reasoning: developing a relatively straightforward model, one that doesn't use AI or other advanced techniques, is within reach but is also limited in depth and fidelity to actual ethical reasoning. The more complex—yet more realistic—computational model that Robbins and Wallace conceived (but haven't implemented) will take considerable work to advance from concept to reality.

Unlike my research and the research I've just described, Michael Anderson, Susan Anderson, and Chris Armen aim to develop programs that reason ethically and come to their *own* ethical conclusions. Susan Anderson states this goal most clearly.[11] Anderson, Anderson, and Armen have developed prototype computational models of ethical reasoning based on well-known theoretical frameworks. The first prototype they implemented was Jeremy,[12] based on Jeremy Bentham's theory of hedonistic act utilitarianism.[3] Bentham's utilitarianism proposes a "moral arithmetic" in which one calculates the pleasure and displeasure of those affected by every possible outcome in an ethical dilemma. The Jeremy program operationalizes moral arithmetic by computing "total net pleasure" for each alternative action, using this simple formula: Total Net Pleasure = Sum-Of (Intensity * Duration * Probability) for all affected individuals. Jeremy chooses as the correct action the action with the highest total net pleasure. Rough estimates must be provided of the intensity, duration, and probability, given a small set of possible values (for example, 0.8, 0.5, and 0.2 for probability estimates), for each action per individual. The three researchers claim that Jeremy is impartial and considers all actions.

Anderson, Anderson, and Armen built a second prototype, W.D.,[12] based on William D. Ross's seven prima facie duties[5] and John Rawls' theory of reflective equilibrium.[13] The general idea behind W.D. is that Ross's theory provides a comprehensive set of duties or principles relevant to ethical cases, such as justice, beneficence, and nonmaleficence, while Rawls' approach provides the founda-



Truth-Teller is comparing the following cases:

CASE 1: Felicia is a young lawyer running her own business. A client, Henry, requires a complex legal transaction that Felicia has never done before. This type of transaction is rarely done by an inexperienced lawyer; usually attorneys handle many simper cases of the same type before handling such a complex case. In addition, if Felicia bungles the case, Henry and his family will go bankrupt. Should Felicia tell Henry about her inexperience in the matter?

CASE 2: Kevin is a lawyer fresh out of law school. A client, Alida, requires a complex legal transaction that Kevin has never done before. However, Kevin was specifically trained in this type of transaction during law school and lawyers routinely accept this type of case fresh out of law school. Additionally, the consequences of the case, should it go badly, are minimal. Should Kevin tell the client about his inexperience in this matter?

Truth-Teller's analysis:

The decision makers, Felicia and Kevin, are confronted with very similar dilemmas because they share reasons both to tell the truth and not to tell the truth. The cases also share similar relationship contexts. The relationship between Felicia and Henry is identical to the relationship between Kevin and Alida; they are both 'is attorney of' relations.

Felicia and Kevin share reasons to tell the truth. First, both protagonists share the reason to provide sales information so that a consumer can make an informed decision. In addition, Felicia and Kevin share the reason to disclose professional inexperience for, respectively, Henry and Alida. Third, both actors share the general reason to avoid harm. More specifically, Felicia has the reason to avoid a financial loss for Henrys family and Henry, while Kevin has the reason to avoid an unknown future harm for Alida. Finally, both actors share the reason to establish goodwill for future benefit.

Felicia and Kevin also share reasons to not tell the truth. Both protagonists share the reason to enhance professional status and opportunities. Second, Felicia and Kevin share the reason to realize a financial gain for themselves.

However, these quandaries are distinguishable. An argument can be made that Felicia has a stronger basis for telling the truth than Kevin. The reason 'to disclose professional inexperience,' a shared reason for telling the truth, is stronger in Felicia's case, since this type of complicated case is rarely done by an inexperienced lawyer. Additionally, the shared reason for telling the truth 'to avoid harm' is stronger in Felicia's case, because (1) Henry and his family will go bankrupt if the case is lost and (2) it is more acute ('One should protect oneself and others from serious harm.')

Figure 1. Truth-Teller's output comparing two ethical dilemmas.

tion for a procedure to make ethical decisions given those duties.

In particular, Rawls' approach inspired Anderson, Anderson, and Armen to imbue W.D. with a decision procedure that generalizes rules (or principles) from cases and tests the generalizations on further cases, with further iteration until the generated rules match ethical intuition. This procedure defines cases simply as an evaluation of a set of duties using integer estimates (ranging from –2 to 2) regarding how strongly each duty was violated or satisfied (for example, –2 represents a serious violation, and +2 is a maximal satisfaction of duty). Rawls' approach lends itself well to an AI machine learning algorithm. In particular, W.D. uses inductive-logic programming to learn Horn clause rules from each case, until the rules reach a "steady state" and can process subsequent cases without further learning.

A third program that Anderson, Anderson, and Armen developed, *MedEthEx*,[14] is very similar to W.D., except that it's specific to medical ethics and uses Tom Beauchamp and James Childress's *Principles of Biomedical Ethics*[15] in place of Ross's prima facie duties. Like W.D., MedEthEx relies on reflective equilibrium and employs integer evaluation of principles, and it uses the same machine learning technique.

This use of machine learning to support ethical reasoning is novel and quite promising. The natural fit between reflective equilibrium and inductive-logic programming is especially striking. On the other hand, their research might oversimplify the task of interpreting and evaluating ethical principles and duties. Reducing each principle and duty to an integer value on a five-point scale renders it almost trivial to apply machine learning to the resulting data, because the search space becomes drastically reduced. But can you really reduce principles such as beneficence or nonmaleficence to single values? Wouldn't people likely disagree on such simple dispositions of duties and principles?

In my experience, and exemplified by the two computational models that I later discuss, perhaps the toughest problem in ethical reasoning is understanding and interpreting the subtleties and application of principles. Very-high-level principles such as beneficence and nonmaleficence, if applied to specific situations, naturally involve bridging a huge gap between the abstract and the specific. One potential way to bridge the gap is to use cases as exemplars and explanations of "open textured" principles,[16] not simply as a means to generalize rules and principles. (Open-textured terms are conditions, premises, or clauses that aren't precise, cover a wide range of specific facts, or are highly subject to interpretation and might even have different meanings in different contexts.) This is the tack that a different group of philosophers, the casuists, take.

Truth-Teller

I intended Truth-Teller as a first step in implementing a computational model of casuistic reasoning, in which someone makes a decision by comparing a problem to paradigmatic, real, or hypothetical cases.[17] Casuistry long ago fell out of favor with many philosophers and ethicists because they believed it to be too imprecise and based on moral intuitions. However, medical ethicists have recently been using it to help solve practical dilemmas.[18,19] Unlike the approach that W.D. and MedEthEx use, casuistry (and hence Truth-Teller) focuses on the power of specific cases and case comparison, not on rules generalized from case evaluations.

Truth-Teller compares pairs of given cases presenting ethical dilemmas about whether to tell the truth.[20,21] The program marshals ethically relevant similarities and differences between the two cases from the perspective of the "truth teller" (the person facing the dilemma) and reports them to the user. In particular, it points out reasons for telling the truth (or not) that

- apply to both cases,
- apply more strongly in one case than another, or
- apply to only one case.

I adapted the dilemmas for Truth-Teller from Scruples, a party game in which participants challenge one another to resolve everyday ethical dilemmas.

Figure 1 shows Truth-Teller's output in comparing two dilemmas. As you can see, these cases share similar themes, relationships, and structure. Truth-Teller recognizes the similarity and points this out in the first paragraph of its comparison text. The truth tellers in the two scenarios, Felicia and Kevin, essentially share the same reasons for telling the truth or not, and the second and third paragraphs of Truth-Teller's output detail this. There are no reasons for telling the truth (or not) that exist in one case but not the other, so Truth-Teller makes no comment on this. Finally, the last paragraph of Truth-Teller's comparison text points out each case's distinguishing features. Felicia has a greater obligation than Kevin to reveal her inexperience owing to established custom (that is, inexperienced lawyers rarely perform this transaction) and more severe consequences (that is, Henry and his family will go bankrupt if she fails).

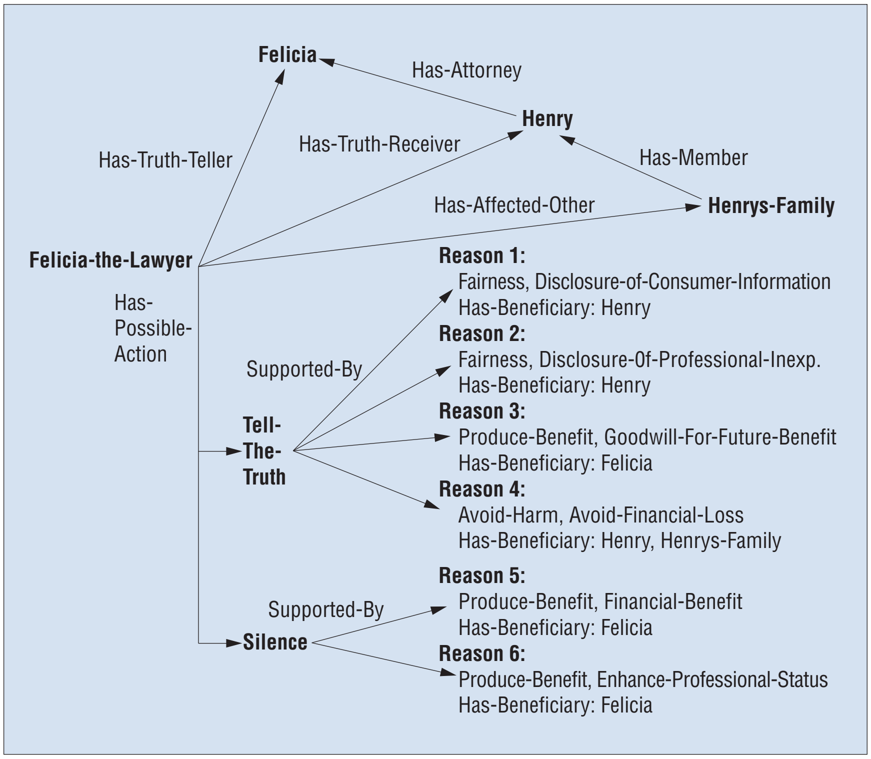


Figure 2. Truth-Teller's case representation for Felicia's case in figure 1.

Figure 2 depicts Truth-Teller's semantic representation of Felicia's case in figure 1. This representation served as input to the program to perform its reasoning. (That is, the Truth-Teller program doesn't "read" natural language input.) In this case, Felicia (the truth teller) can take one of two possible actions: tell Henry (the "truth receiver") the truth or remain silent about her inexperience. The truth teller might be able to take other actions in a scenario, such as trying to resolve a situation through a third party. Each possible action a protagonist can take has supporting reasons. For instance, two reasons for Felicia to tell the truth are

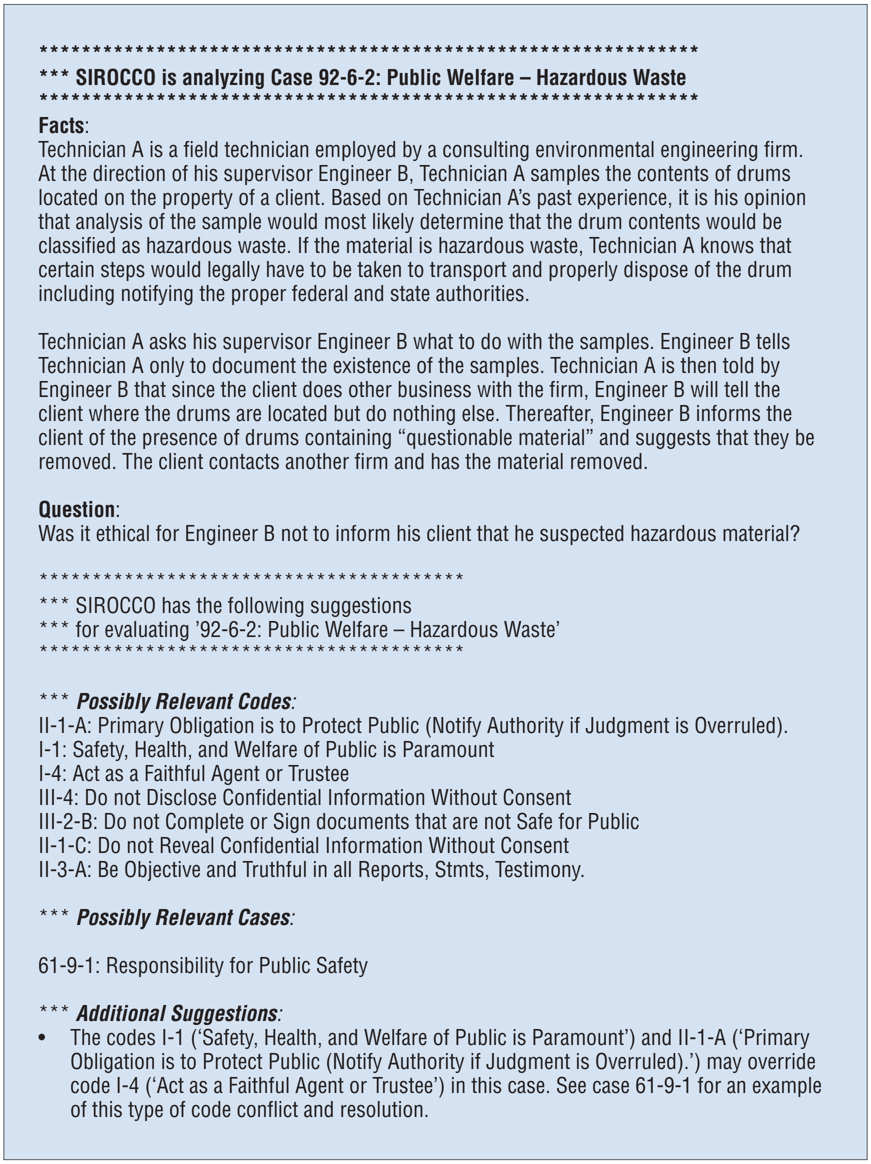
- fairness (reason 2)—Felicia has an obligation to fairly disclose her inexperience—and
- avoiding harm (reason 4)—Felicia might avoid financial harm to Henry and his family by telling the truth.

Truth-Teller compares pairs of cases by aligning and comparing the reasons that support telling the truth or not in each case. More specifically, Truth-Teller's comparison method comprises four analysis phases:

1. *Alignment* builds a mapping between the reasons in the two cases. That is, it indicates the reasons that are the same and different across the two representations.
2. *Qualification* identifies special relationships among actors, actions, and reasons that augment or diminish the reasons' importance. For example, telling the truth to a family member is typically more important than telling the truth to a stranger.
3. *Marshaling* selects particular similar or differentiating reasons for arguing that (a) one case is as strong as or stronger than the other with respect to a conclusion, (b) the cases are only weakly comparable, or (c) the cases aren't comparable.
4. *Interpretation* generates prose that accurately presents the marshalled information so that a nontechnical human user can understand it. This phase generated the text labeled "Truth-Teller's Analysis" in figure 1.

To test Truth-Teller's ability to compare cases, I performed an evaluation in which five professional ethicists were asked to grade the program's output. The experts assessed its reasonableness (R), completeness (C), and context sensitivity (CS), on a scale of 1 (low) to 10 (high), for 20 case comparisons similar to the comparison in figure 1. The experts assigned mean scores of R = 6.3, C = 6.2, and CS = 6.1 across these comparisons. The evaluation also included two comparisons written by graduate students; not surprisingly, the ethicists graded these comparisons somewhat higher, at mean scores of R = 8.2, C = 7.7, and CS = 7.8. On the other hand, two of Truth-Teller's comparisons received higher scores than one of the human evaluations.

These results indicate that Truth-Teller is moderately successful at comparing truth-telling dilemmas. Because I instructed the expert ethicists to "evaluate comparisons as you would evaluate short answers written by college undergraduates," it's quite encouraging that Truth-Teller performed as well as it did. However, two questions naturally arise: Why did the ethicists view Truth-Teller's comparisons as somewhat inferior to the humans'? How could Truth-Teller be brought closer to human performance? Several evaluators questioned Truth-Teller's lack of hypothetical analysis; the program makes fixed assumptions about the facts (that is, reasons, actions, and actors). One way to counter this might be to develop techniques that let Truth-Teller suggest hypothetical variations to problems, along the lines of the legal-reasoning program HYPO.[22] For instance, in the comparison of figure 1, Truth-Teller might suggest that, if an



```
*********************************************************************
*** SIROCCO is analyzing Case 92-6-2: Public Welfare – Hazardous Waste
*********************************************************************
Facts:
Technician A is a field technician employed by a consulting environmental engineering firm.
At the direction of his supervisor Engineer B, Technician A samples the contents of drums
located on the property of a client. Based on Technician A's past experience, it is his opinion
that analysis of the sample would most likely determine that the drum contents would be
classified as hazardous waste. If the material is hazardous waste, Technician A knows that
certain steps would legally have to be taken to transport and properly dispose of the drum
including notifying the proper federal and state authorities.

Technician A asks his supervisor Engineer B what to do with the samples. Engineer B tells
Technician A only to document the existence of the samples. Technician A is then told by
Engineer B that since the client does other business with the firm, Engineer B will tell the
client where the drums are located but do nothing else. Thereafter, Engineer B informs the
client of the presence of drums containing "questionable material" and suggests that they be
removed. The client contacts another firm and has the material removed.

Question:
Was it ethical for Engineer B not to inform his client that he suspected hazardous material?

*******************************************
*** SIROCCO has the following suggestions
*** for evaluating '92-6-2: Public Welfare – Hazardous Waste'
*******************************************

*** Possibly Relevant Codes:
II-1-A: Primary Obligation is to Protect Public (Notify Authority if Judgment is Overruled).
I-1: Safety, Health, and Welfare of Public is Paramount
I-4: Act as a Faithful Agent or Trustee
III-4: Do not Disclose Confidential Information Without Consent
III-2-B: Do not Complete or Sign documents that are not Safe for Public
II-1-C: Do not Reveal Confidential Information Without Consent
II-3-A: Be Objective and Truthful in all Reports, Stmts, Testimony.

*** Possibly Relevant Cases:

61-9-1: Responsibility for Public Safety

*** Additional Suggestions:
•  The codes I-1 ('Safety, Health, and Welfare of Public is Paramount') and II-1-A ('Primary
   Obligation is to Protect Public (Notify Authority if Judgment is Overruled).') may override
   code I-4 ('Act as a Faithful Agent or Trustee') in this case. See case 61-9-1 for an example
   of this type of code conflict and resolution.
```

Figure 3. Sirocco's output for an engineering ethical dilemma.

(unstated and thus hypothetical) long-standing relationship between Felicia and Henry exists, there is additional onus on Felicia to reveal her inexperience. Another criticism of Truth-Teller involved the program's somewhat rigid approach of enumerating supporting, individual reasons and not relating one reason to another. Some form of reason aggregation might address this issue, by discussing the overall import of supporting reasons rather than focusing on individual reasons.

Sirocco

I developed Sirocco as a second step in exploring casuistry and how to realize it in a computational model. In particular, Sirocco attempts to bridge the gap between general principles and concrete facts of cases. The program emulates how an ethical review board within the National Society of Professional Engineers decides cases by referring to, and balancing between, ethical codes and past cases.[23]

Engineering-ethics principles, while more specific than general ethical duties, such as Ross's prima facie duties (for example, justice, beneficence, and nonmaleficence), still tend to be too general to decide cases. So, the NSPE review board often uses past cases to illuminate the reasoning behind principles and as precedent in deciding new cases. Consider, for example, the following ethical code from the NSPE:

> Code II.5.a. Engineers shall not falsify or permit misrepresentation of their ... academic or professional qualifications. They shall not misrepresent or exaggerate their degree of responsibility in or for the subject matter of prior assignments. Brochures or other presentations incident to the solicitation of employment shall not misrepresent pertinent facts concerning employers, employees, associates, joint ventures or past accomplishments with the intent and purpose of enhancing their qualifications and their work.

This code specializes the more general principle of "honesty" in an engineering context. Each sentence deals with a different aspect of "misrepresentation of an engineer" and covers a wide range of possible circumstances. However, the code does not specifically state the precise circumstances that support application. Knowing whether this code applies to a particular fact situation requires that you recognize the applicability of and interpret open-textured terms and phrases in the code, such as "misrepresentation" and "intent and purpose of enhancing their qualifications." While these engineering-ethics codes are an example of abstract codes, they are by no means distinctive. Many principles and codes, generally applicable or domain specific, are abstract. Principles also typically conflict with one another in specific circumstances, with no clear resolution to that conflict. In the NSPE's analyses of over 500 engineering cases, it interprets principles such as II.5.a in the context of the facts of real cases, decides when one principle takes precedence over another, and provides a rich, extensional representation of principles.

Sirocco's goal, given a new case to analyze, is to provide the basic information with which a human reasoner—for instance, a member of the NSPE review board—could answer an ethical question and then build an argument or rationale for that conclusion.[24] Figure 3 shows an example of Sirocco's output. First, the output displays the input case's facts and the question that the case raises. This particular case involves an engineering technician who discovers what he believes to be hazardous waste, suggesting a need to notify federal authorities. However, when the technician asks his boss, Engineer B, what to do with his finding, he's told not to mention his suspicions of hazardous waste to this important client, who might face cleanup expenses and legal ramifications from the finding. The question is whether it was ethical for Engineer B to give preference to his duty to his client over public safety. Sirocco's analysis of the case consists of

- a list of possibly relevant codes,
- a list of possibly relevant past cases, and
- a list of additional suggestions.

You can run Sirocco on over 200 ethical dilemmas and view analysis such as that in figure 3 by visiting http://sirocco.lrdc.pitt.edu/sirocco/index.html.

Sirocco accepts input, or *target*, cases in a detailed case-representation language called the *Engineering Transcription Language*. ETL represents a scenario's actions and events as a *Fact Chronology* of individual sentences (that is, *Facts*). The representation uses a predefined ontology of Actor, Object, Fact Primitive, and Time Qualifier types. At least one Fact in the Fact Chronology is designated as the Questioned Fact—the action or event corresponding to the ethical question raised in the scenario. The entire ontology, a detailed description of how ETL represents cases, and over 50 examples of Fact Chronologies are at www.pitt.edu/~bmclaren/ethics/index.html.

Sirocco utilizes knowledge of past case analyses, including past retrieval of ethical codes and cases and the way past analyses utilized these knowledge elements, to support its retrieval and analysis in the target case. The program employs a two-stage graph-mapping algorithm to retrieve codes and cases. Stage 1 performs a "surface match" by retrieving all source cases from the program's database that share any fact with the target case. The cases in the database, which number over 400, are represented in EETL, an extended version of ETL. This stage then computes a score for all retrieved cases, based on fact matching between the target case and each source case, and outputs a list of ranked candidate source cases. Using A* search (an AI technique), stage 2 attempts a structural mapping between the target case and each of the *N* top-ranking candidate source cases from stage 1. This search takes into account temporal relations and abstract matches. Sirocco's Analyzer module organizes and displays the top-rated structural mappings that the A* search uncovered. The ethical codes cited in these top-rated source cases are also evaluated by the Analyzer for how well they apply to the target case. Figure 3 is an example of the Analyzer's output.

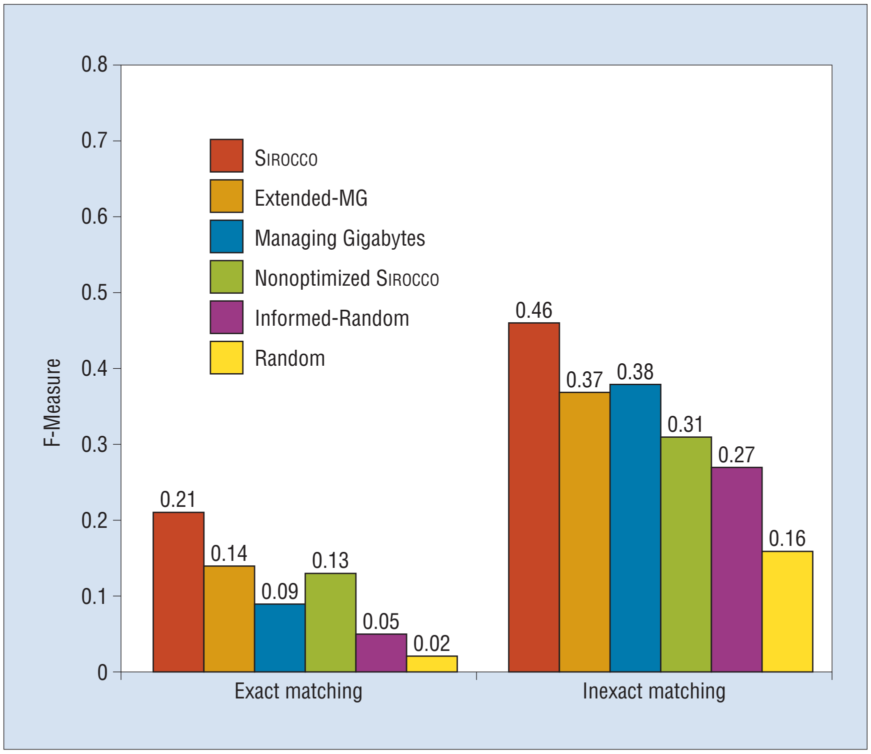


Figure 4. A comparison of Sirocco with five other methods for retrieving ethical codes and cases. The F-Measure indicates how well each method performed in comparison to humans. In exact matching, the methods and humans retrieved precisely the same codes and cases. In inexact matching, the methods and humans retrieved closely related codes and cases.

I performed a formal experiment with Sirocco to test how well it retrieved codes and cases compared to several other retrieval methods, including two full-text retrieval systems, MG (Managing Gigabytes) and Extended-MG. I also compared Sirocco to three other methods: one that randomly selects codes and cases (Random), one that randomly selects codes and cases from the most frequently cited codes and cases (Informed-Random), and an ablated version of Sirocco with operationalization functionality excised (Non-Op Sirocco). Using the *F-Measure* metric, I scored each method on the basis of how well its retrieved codes and cases overlapped with the humans' (that is, the NSPE review board's) retrieved codes and cases in evaluating the same cases. The F-measure is an information retrieval metric that combines precision and recall.[25,26] I compared the methods on two dimensions: exact matching (the method and the humans retrieved precisely the same codes and cases) and inexact matching (the method and the humans retrieved closely related codes and cases). Figure 4 summarizes the results.

The results showed that Sirocco was significantly more accurate at retrieving relevant codes and cases than all the other methods except Extended-MG, to which it came close to being significantly more accurate ($p = 0.057$). Because these automated methods, particularly the information retrieval approaches MG and Extended-MG, are arguably the most competitive with Sirocco, this experiment shows that Sirocco is an able ethics-reasoning companion. On the other hand, as figure 4 shows, Sirocco performed worse than the ethical review board (0.21 and 0.46 can be roughly interpreted as 21 percent and 46 percent overlapping with the board selections). At least some, if not most, of this discrepancy is because the inexact-matching metric doesn't fully capture correct selections. For example, in many instances Sirocco actually selected a code or case that was arguably applicable to a case but that the board didn't select. In other words, using the review board as the "gold standard" has its flaws. Nevertheless, it's fair to say that although Sirocco performs well, it doesn't perform quite at the level of an expert human reasoner.

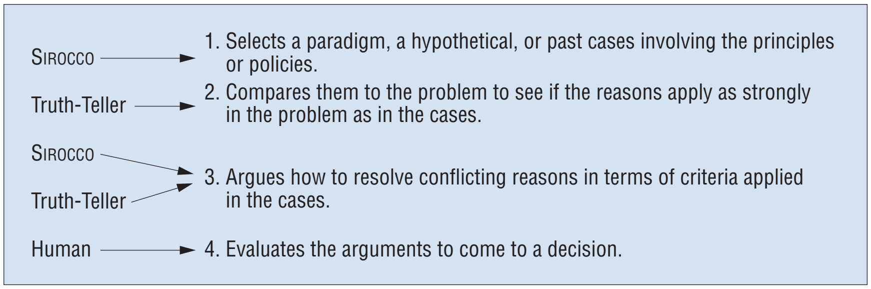

Sirocco
Truth-Teller
Sirocco
Truth-Teller
Human
1. Selects a paradigm, a hypothetical, or past cases involving the principles or policies.
2. Compares them to the problem to see if the reasons apply as strongly in the problem as in the cases.
3. Argues how to resolve conflicting reasons in terms of criteria applied in the cases.
4. Evaluates the arguments to come to a decision.

Figure 5. Truth-Teller's, Sirocco's, and a human's potential role in casuistic problem solving.

The relationship between Truth-Teller and Sirocco

Fundamentally, Truth-Teller and Sirocco have different purposes. Truth-Teller is more useful in helping users compare cases. Although Sirocco also compares cases, its results don't focus on case comparisons and presenting those comparisons to the user. Rather, Sirocco is more useful for collecting a variety of relevant information that a user should consider in evaluating a new ethical dilemma. While Truth-Teller has a clear advantage in comparing cases and explaining those comparisons, it ignores the problem of how to identify potentially comparable cases in the first place. The program compares any pair of cases it's provided, no matter how different they might be. Sirocco, on the other hand, determines which cases will most likely be relevant to a given target case and thus worth comparing.

An interesting synthesis of the two programs would be to have Sirocco retrieve comparable cases and have Truth-Teller compare them. For instance, the "algorithm" in figure 5, adapted from Albert Jonsen and Stephen Toulmin,[17] represents the general approach a casuist would take in solving an ethical dilemma.

First, given a new case, the casuistic reasoner would find cases (paradigms, hypotheticals, or real cases) that test the principles or policies in play in the new case. That is, the casuist would reach into its knowledge base of cases to find the past cases that might provide guidance in the new case. In effect, this is what Sirocco does.

Second, the reasoner would compare the new cases to the cases it retrieved. While Sirocco does this to a limited extent, this is where Truth-Teller's capability to compare and contrast given cases, at a reasonably fine level of detail, would come in.

Third, the casuist would argue how to resolve conflicting reasons. Both Truth-Teller and Sirocco have at least a limited capability to perform this step. This is illustrated, for example, in Truth-Teller's example output, at the bottom of figure 1, in which the program distinguishes the two cases by stating the reasons that apply more strongly in Felicia's case. Sirocco does this by suggesting that one principle might override another in these particular circumstances (see the "Additional Suggestions" at the bottom of figure 3).

Finally, in keeping with my vision of how to apply computational models to ethical decision making, a human would make the final decision about this ethical dilemma.

To fully realize the casuistic problem-solving approach of figure 5 and combine the complementary capabilities of Truth-Teller and Sirocco, the two programs would need common representational elements. Sirocco uses primitives that closely model some of a fact situation's actions and events to represent cases as complex narratives. In this sense, Sirocco's representational approach is more sophisticated and general than Truth-Teller's. On the other hand, Sirocco's case comparisons aren't nearly as precise and issue-oriented as Truth-Teller's.

Both Truth-Teller and Sirocco focus and rely heavily on a knowledge representation of ethics, unlike, for instance, the programs of Anderson, Anderson, and Armen, which have little reliance on representation. The knowledge representation approach to building computational models of ethical reasoning has strengths and weaknesses. Its strength is its ability to represent cases and principles at a rather fine level of detail. For instance, a detailed engineering-ethics ontology supports Sirocco, and a representation of reasons underlies Truth-Teller, as figure 2 shows. Such representation not only supports each model's reasoning approaches but also lets the models provide relatively rich explanations of their reasoning, as exemplified by the programs' output in figures 1 and 3.

On the other hand, each model's representation is necessarily specific to its task and domain. So, Truth-Teller has a rich representation of truth-telling dilemmas—but not much else. Sirocco has a deep representation of engineering-ethics principles and engineering scenarios, but no knowledge of more general ethical problem solving, such as the reasoning model embodied in W.D. and MedEthEx. So, another step required for unifying Truth-Teller and Sirocco and implementing the casuistic approach of figure 5 would be a synthesis and generalization of their respective representational models.

Lessons learned

The primary lesson I've learned from the Truth-Teller and Sirocco projects is that ethical reasoning is fundamentally different from reasoning in more formalized domains. In ethical reasoning, "inference rules" are available almost exclusively at an abstract level, in the form of principles. The difficulty of using formal logic to address and form arguments in such domains has long been recognized,[27] and some AI practitioners, particularly those interested in legal reasoning, have grappled with this issue. As Kevin Ashley pointed out, "The legal domain is harder to model than mathematical or scientific domains because deductive logic, one of the computer scientist's primary tools, does not work in it."[22]

The ethical-reasoning domain, like the legal domain, can be viewed as a weak analytic domain where the given "rules" (that is, laws, codes, or principles) are available almost exclusively at a highly abstract, conceptual level. This means that the rules might contain open-textured terms. Also, in a weak analytic domain, abstract rules often conflict with one another in particular situations with no deductive or formal means of arbitrating such conflicts. That is, more than one rule might appear to apply to a given fact situation, but neither the abstract rules nor the general knowledge of the domain provides clear resolution.

Another important lesson from the two projects is the sheer difficulty of imbuing a computer program with the sort of flexible intelligence required to perform ethical analysis. While both programs performed reasonably well in the studies I mentioned earlier, neither performed at an expert human's level. While the goal wasn't to emulate human ability, taking the task of ethical decision making away from humans, it's important that computational artifacts that purport to support ethical reasoning at least perform well enough to encourage humans to use the programs as aids in their own reasoning. As of this writing, only the Truth-

Teller and Sirocco computational models (and, perhaps to a lesser extent, Robbins, Wallace, and Puka's Web-based system) have been empirically tested in a way that might inspire faith in their performance.

My contention that computer programs should act only as aids in ethical reasoning isn't due to a high regard for human ethical decision making. Of course, humans often make errors in ethical reasoning. Rather, my position is based, as I suggested earlier, on the existence of so many plausible, competing approaches to ethical problem solving. Which philosophical method can claim to be the "correct" approach to ethical reasoning, in the same sense that we accept calculus as a means of solving engineering problems and accept first-order logic for solving syllogisms? It's difficult to imagine that a single ethical-reasoning approach embodied in a single computer program could deliver even close to a definitive approach to ethical reasoning. Of course, many approaches might be considered "good enough" without being definitive. But the bar will likely be much higher for autonomous computer-based systems making decisions in an area as sensitive and personal to humans as ethical reasoning.

Also, it's presumptuous to think that you could fully implement the subtleties of any of the well-known philosophical systems of ethics in a computer program. Any implementation of one of these theories is necessarily based on simplifying assumptions and subjective interpretation of that theory. For instance, as I mentioned before, W.D. simplifies the evaluation of Ross's prima facie duties by assigning each a score on a five-point scale. Both Truth-Teller and Sirocco also make simplifying assumptions, such as Truth-Teller representing only reasons that support telling the truth or not, and not the circumstances leading to these reasons. Of course, making simplifying assumptions is a necessary starting point for gaining traction in the difficult area of ethical reasoning.

My final reason for using computational models as only aids in ethical reasoning is my belief that humans simply won't accept autonomous computer agents making such decisions for them. But they might accept programs as advisors.

Given my view of the role of computational models and how they could (and should) support humans, a natural and fruitful next step is to use computational models of ethical reasoning as teaching aids. Ilya Goldin, Kevin Ashley, and Rosa Pinkus have taken steps in this direction. PETE (Professional Ethics Tutoring Environment) is a software tutor that leads a student step-by-step in preparing cases for class discussion. It encourages students to compare their answers to those of other students.[28]

My most recent work and interest has also involved intelligent tutoring systems.[29,30] As part of this focus, I've started to investigate whether case comparisons, such as those that Truth-Teller produces, could be the basis for an intelligent tutor. The idea is to explore whether Truth-Teller's comparison rules and procedures can

- be improved and extended to cover the kinds of reasons involved in comparing more technically complex cases, such as those tackled by Sirocco, and
- serve as the basis of a Cognitive Tutor to help a student understand and perform Truth-Teller's phases.

Another important lesson from the two projects is the sheer difficulty of imbuing a computer program with the sort of flexible intelligence required to perform ethical analysis.

Cognitive Tutors are based on John Anderson's ACT-R (adaptive control of thought-rational) theory, according to which humans use production rules, modular If-Then constructs, to perform problem-solving steps in a variety of domains.[31] Key concepts underlying Cognitive Tutors are *learn by doing*, helping students learn by engaging them in actual problem solving, and *immediate feedback*, providing guidance when students request a hint or make a mistake. For domains such as algebra, the production rules in a cognitive model indicate not only correct problem-solving steps a student might take but also plausible incorrect steps. The model provides feedback in the form of error messages, when the student takes a step anticipated by a "buggy rule," and hints, when the student asks for help.

Developing a Cognitive Tutor for case comparison presents some stiff challenges. One particular challenge is that, unlike previous domains in which Cognitive Tutors have been used, such as algebra and programming, in practical ethics answers aren't always and easily identified as correct or incorrect. Also, the rules, as I discussed earlier, are more abstract and ill defined. As a result, while learn-by-doing fits ethics case comparison very well, the concept of immediate feedback needs adaptation. Unlike feedback in more technical domains, ethics feedback might be nuanced rather than simply right or wrong, so the Cognitive Tutor approach must be adapted to this.

The rules employed in Truth-Teller's first three phases, particularly the qualification phase, provide a core set of rules that can be improved and recast as a set of rules for comparing cases in a Cognitive Tutor framework. An empirical study of case comparisons, involving more technically complex ethics cases, will enable refinement and augmentation of these comparison rules. At the same time, the empirical study of humans comparing cases might reveal plausible misconceptions about the comparison process that can serve as buggy rules that present opportunities to correct the student.

A related direction is exploring whether the priority rules of Ross's theory of prima facie duties, such as nonmaleficence normally overriding other duties and fidelity normally overriding beneficence, might benefit the Truth-Teller comparison method. At the very least it would ground Truth-Teller's approach in a more established philosophical theory (currently, priority rules are based loosely on Sissela Bok's research[32]). Such an extension to Truth-Teller would also benefit the planned Cognitive Tutor, in that it could refer to Ross's theory to support explanations to students.

Acknowledgments

Kevin Ashley contributed greatly to the ideas behind both Truth-Teller and Sirocco. US National Science Foundation Learning and Intelligent Systems grant 9720341 partly supported this research.

References

1. Aristotle, *Nicomachean Ethics*, W.D. Ross,

The Author

Bruce M. McLaren is a systems scientist at Carnegie Mellon University and the Pittsburgh Science of Learning Center. He has research interests in intelligent tutoring, collaborative learning, case-based reasoning, computational models of ethical reasoning, and Internet technologies. He also has an extensive background in practical applications, with over 20 years' experience in the commercial sector. He received his PhD in intelligent systems from the University of Pittsburgh. He's a member of the AAAI and holds one patent and one patent pending. Contact him at the Human Computer Interaction Inst.. Carnegie Mellon Univ., 2617 Newell-Simon Hall, 5000 Forbes Ave., Pittsburgh, PA 15213-3891; bmclaren@cs.cmu.edu; www.pitt.edu/~bmclaren.

ed., Oxford Univ. Press, 1924

2. I. Kant, "Groundwork of the Metaphysic of Morals," *Practical Philosophy*, translated by M.J. Gregor, Cambridge Univ. Press, 1996.
3. J. Bentham, *Introduction to the Principles of Morals and Legislation*, W. Harrison, ed., Hafner Press, 1948.
4. J.S. Mill, *Utilitarianism*, George Sher, ed., Hackett, 1979.
5. W.D. Ross, *The Right and the Good*, Oxford Univ. Press, 1930.
6. C.E. Harris, M.S. Pritchard, and M.J. Rabins, *Engineering Ethics: Concepts and Cases*, 1st ed., Wadsworth, 1995.
7. D.R. Searing, *HARPS Ethical Analysis Methodology, Method Description, Version 2.0.0*, Taknosys Software Corp., 1998.
8. R. Cavalier and P.K. Covey, *A Right to Die? The Dax Cowart Case CD-ROM Teacher's Guide, Version 1.0*, Center for the Advancement of Applied Ethics, 1996.
9. R.W. Robbins and W.A. Wallace, "Decision Support for Ethical Problem Solving: A Multi-agent Approach," to be published in *Decision Support Systems*, 2006 (available online 25 Apr. 2006 at www.sciencedirect.com).
10. R.W. Robbins, W.A. Wallace, and B. Puka, "Supporting Ethical Problem Solving: An Exploratory Investigation," *Proc. 2004 ACM SIGMIS Conf. Computer Personnel Research*, ACM Press, 2004, pp. 22–24.
11. S.L. Anderson, "Asimov's 'Three Laws of Robotics' and Machine Metaethics," *Proc. AAAI 2005 Fall Symp. Machine Ethics*, tech. report FS-05-06, AAAI Press, 2005, pp. 1–7.
12. M. Anderson, S.L. Anderson, and C. Armen, "Towards Machine Ethics: Implementing Two Action-Based Ethical Theories," *Proc. AAAI 2005 Fall Symp. Machine Ethics*, tech. report FS-05-06, AAAI Press, 2005, pp. 1–7.
13. J. Rawls, *A Theory of Justice*, 2nd ed., Harvard Univ. Press, 1999.
14. M. Anderson, S.L. Anderson, and C. Armen. "MedEthEx: Toward a Medical Ethics Advisor," *Proc. AAAI 2005 Fall Symp. Caring Machines: AI in Elder Care*, tech. report FS-05-02, AAAI Press, 2005, pp. 9–16.
15. T.L. Beauchamp and J.F. Childress, *Principles of Biomedical Ethics*, Oxford Univ. Press, 1979.
16. A. Gardner, *An Artificial Intelligence Approach to Legal Reasoning*, MIT Press, 1987.
17. A.R. Jonsen and S. Toulmin, *The Abuse of Casuistry: A History of Moral Reasoning*, Univ. of California Press, 1988.
18. C. Strong, "Justification in Ethics," *Moral Theory and Moral Judgments in Medical Ethics*, B.A. Brody, ed., Kluwer Academic, 1988, pp. 193–211.
19. B. Brody, *Taking Issue: Pluralism and Casuistry in Bioethics*, Georgetown Univ. Press, 2003.
20. K.D. Ashley and B.M. McLaren, "Reasoning with Reasons in Case-Based Comparisons," *Proc. 1st Int'l Conf. Case-Based Reasoning*, Springer, 1995, pp. 133–144.
21. B.M. McLaren and K.D. Ashley, "Case-Based Comparative Evaluation in TruthTeller," *Proc. 17th Ann. Conf. Cognitive Science Soc.*, Lawrence Erlbaum, 1995, pp. 72–77.
22. K.D. Ashley, *Modeling Legal Argument: Reasoning with Cases and Hypotheticals*, MIT Press, 1990.
23. *The NSPE Ethics Reference Guide*, Nat'l Soc. of Professional Engineers, 1996.
24. B.M. McLaren, "Extensionally Defining Principles and Cases in Ethics: An AI Model," *Artificial Intelligence J.*, vol. 150, nos. 1–2, 2003, pp. 145–181.
25. D. Lewis et al., "Training Algorithms for Linear Text Classifiers," *Proc. 19th Ann. Int'l ACM-SIGIR Conf. Research and Development in Information Retrieval*, ACM Press, 1996, pp. 298–306.
26. C.J. van Rijsbergen, *Information Retrieval*, 2nd ed., Butterworths, 1979.
27. S.E. Toulmin, *The Uses of Argument*, Cambridge Univ. Press, 1958.
28. I.M. Goldin, K.D. Ashley, and R.L. Pinkus, "Introducing PETE: Computer Support for Teaching Ethics," *Proc. 8th Int'l Conf. Artificial Intelligence & Law* (ICAIL 01), ACM Press, 2001, pp. 94–98.
29. B.M. McLaren, S. Lim, F. Gagnon, D. Yaron, and K.R. Koedinger, "Studying the Effects of Personalized Language and Worked Examples in the Context of a Web-Based Intelligent Tutor," presented at the *8th Int'l Conf. Intelligent Tutoring Systems*, 2006.
30. B.M. McLaren, L. Bollen, E. Walker, A. Harrier, and J. Sewall, "Cognitive Tutoring of Collaboration: Developmental and Empirical Steps toward Realization," *Proc. Conf. Computer Supported Collaborative Learning* (CSCL 05), Lawrence Erlbaum Associates, 2005; www.pitt.edu/~bmclaren/CSCL-2005-CameraReady-Final.pdf.
31. J.R. Anderson, *Rules of the Mind*, Lawrence Erlbaum, 1993.
32. S. Bok, *Lying: Moral Choice in Public and Private Life*, Vintage Books, 1989.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.