17

Computational Models of Ethical Reasoning

Challenges, Initial Steps, and Future Directions

Bruce M. Mclaren

Introduction

How can machines support, or even more significantly replace, humans in performing ethical reasoning? This is a question of great interest to those engaged in Machine Ethics research. Imbuing a computer with the ability to reason about ethical problems and dilemmas is as difficult a task as there is for Artificial Intelligence (AI) scientists and engineers. First, ethical reasoning is based on abstract principles that cannot be easily applied in formal, deductive fashion. Thus the favorite tools of logicians and mathematicians, such as first-order logic, are not applicable. Second, although there have been many theoretical frameworks proposed by philosophers throughout intellectual history, such as Aristotelian virtue theory (Aristotle, edited and published in 1924), the ethics of respect for persons (Kant 1785), Act Utilitarianism (Bentham 1789), Utilitarianism (Mill 1863), and prima facie duties (Ross 1930), there is no universal agreement 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 in a profound and fundamental way. The premises, beliefs, and principles employed by humans as they make ethical decisions are quite varied, not fully understood, and often inextricably intertwined with religious beliefs. How does one take such uniquely human characteristics and distil them into a computer program?

Undaunted by the challenge, scientists and engineers have over the past fifteen years developed several computer programs that take initial steps in addressing these difficult problems. This paper provides a brief overview of a few of these programs and discusses two in more detail, both focused on reasoning from cases, implementing aspects of the ethical approach known as casuistry, and developed by the author of this paper. One of the programs developed by the author, Truth-Teller, is designed to accept a pair of ethical dilemmas and describe the salient similarities and differences between the cases from both an ethical and pragmatic perspective. The other program, SIROCCO, is constructed to accept a single ethical dilemma and retrieve other cases and ethical principles that may be relevant to the new case.

Neither program was designed to reach an ethical decision. The view that runs throughout the author's work is that reaching an ethical conclusion is, in the end, the obligation of a *human* decision maker. Even if the author believed the computational models presented in this paper 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 in this 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" (Harris, Pritchard, and Rabins, 1995) and help humans reach decisions.

Despite the difficulties in developing machines that can reason ethically, the field of machine ethics presents an intellectual and engineering challenge of the first order. The long history of science and technology is ripe with problems that excite the innovative spirit of scientists, philosophers, and engineers. Even if the author's goal of creating a reliable "ethical assistant" is achieved short of developing a fully autonomous ethical reasoner, a significant achievement will be realized.

Efforts to Build Computer Programs that Support or Model Ethical Reasoning

Two of the earliest programs aimed at ethical reasoning, Ethos and the Dax Cowart program, were designed to assist students in working their own way through thorny problems of practical ethics. 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.

The Ethos System was developed by Searing (1998) to accompany the engineering ethics textbook written by Harris and colleagues (1995). Ethos provides a few prepackaged example dilemmas, including video clips and interviews, to help students explore real ethical dilemmas that arise in the engineering profession. Ethos encourages rational and consistent ethical problem solving in two ways: first, by providing a framework in which one can rationally apply moral beliefs; and second, by recording the step-by-step decisions taken by an ethical decision maker in 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 (Cavalier and Covey 1996). The program focuses on the single, 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 of the case is whether Dax should be allowed to die. The program presents actual video clips of interviews with Dax's doctor, lawyer, mother, nurses, and Dax himself to allow the user to 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 user is periodically asked to make judgments on whether Dax's request to die should be granted, and, dependent on how one answers, the program branches to present information and viewpoints that may cause reconsideration of that judgment.

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

Work that has focused more specifically on the computational modeling of ethical reasoning includes that of Robbins and Wallace (2007). Their proposed computational model combines collaborative problem solving (i.e., multiple human subjects discussing an ethical issue), the psychological Theory of Planned Behavior, and the Belief-Desire-Intention (BDI) 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 subjects as they discuss and try to resolve ethical dilemmas. This system has only been conceptually designed, not implemented, and the authors may have overreached in a practical sense by trying to combine such a wide range of theories and technologies in a single computational model. However, the ideas in the paper could serve as the foundation for future computational models of ethical reasoning. Earlier, Robbins, Wallace, and Puka (2004) did implement and experiment with a more modest Web-based system designed to support ethical problem solving. This system was implemented as a series of Web pages, containing links to relevant ethical theories and principles and a simple ethics "coach." Robbins and his colleagues performed an empirical study in which users of this system were able to identify, for instance, more alternative ways to address a given ethical problem than subjects who used Web pages that did not have the links or coaching. The Robbins and colleagues work 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 does not 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 conceived by Robbins and colleagues has not been implemented and will take considerable work to advance from concept to reality.

Unlike the other work just cited, as well as the work of this author – which purports to support humans in ethical reasoning rather than to perform autonomous ethical reasoning – Anderson, Anderson, and Armen have as a goal developing programs that reason ethically and come to their *own* ethical conclusions (Anderson 2005, p. 10). They have developed prototype computational models of ethical reasoning based on well-known theoretical frameworks. The first prototype they implemented was called *Jeremy* (Anderson, Anderson, and Armen 2005a), based on Jeremy Bentham's theory of Hedonistic Act Utilitarianism (Bentham 1789). 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 the following simple formula: Total Net Pleasure = Sum-Of (Intensity * Duration * Probability) for all affected individuals. The action with the highest Total Net Pleasure is then chosen as the correct action. Rough estimates of the intensity, duration, and probability, given a small set of possible values (e.g., 0.8, 0.5, and 0.2 for probability estimates), for each action per individual must be provided. Anderson et al. claim that *Jeremy* has the advantage of being impartial and considering all actions.

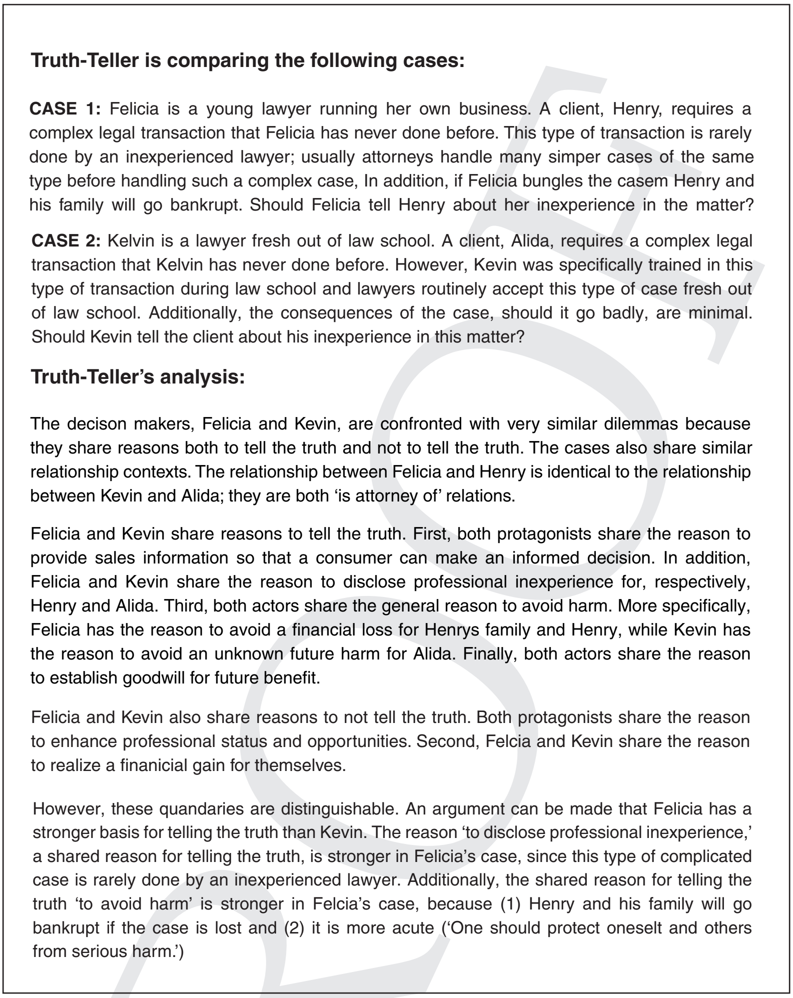
Anderson et al. built a second prototype, W. D. (2005a), based on W. D. Ross's seven prima facie duties (Ross 1930) and reflective equilibrium (Rawls 1971). The general idea behind W. D. is that Ross's theory provides a comprehensive set of duties/principles relevant to ethical cases, such as justice, beneficence, and nonmaleficence, whereas Rawls's approach provides the foundation for a "decision procedure" to make ethical decisions given those duties. In particular, the Rawls' approach inspired a decision procedure in which rules (or principles) are generalized from cases and the generalizations are tested on further cases, with further iteration until the generated rules match ethical intuition. Cases are defined simply as an evaluation of a set of duties using integer estimates (ranging from –2 to 2) regarding how severely each duty was violated (e.g., –2 represents a serious violation of the duty, +2 is a maximal satisfaction of duty). The Rawls approach lends itself well to an AI machine-learning algorithm and, in fact, is the approach adopted by Anderson et al. 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 the need for further learning. A third program developed by Anderson et al. (2005b), MedEthEx, is very similar to W. D., except that it is specific to medical ethics and uses Beauchamp and Childress's Principles of Biomedical Ethics (1979) in place of Ross's prima facie duties. MedEthEx also relies on reflective equilibrium and employs the same notion of integer evaluations of principles and the machine-learning technique of W. D.

Anderson and colleagues' idea to use machine-learning techniques to support ethical reasoning is novel and quite promising. The natural fit between Rawls's reflective equilibrium process and inductive logic programming is especially striking. On the other hand, the work of Anderson et al. may oversimplify the task of interpreting and evaluating ethical principles and duties. Reducing each principle and/or duty to an integer value on a scale of five values renders it almost trivial to apply a machine-learning technique to the resulting data, because the search space becomes drastically reduced. Yet is it really possible to reduce principles such as beneficence or nonmaleficence to single values? Wouldn't people likely disagree on such simple dispositions of duties and principles? In this author's experience, and exemplified by the two computational models discussed in the following sections, 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 (Gardner 1987), not just as a means to generalize rules and principles. This is the tack taken by a different group of philosophers, the casuists, and is the general approach the ethical reasoning systems discussed in the following sections employ.

Truth-Teller

Truth-Teller, the first program implemented by the author to perform ethical reasoning, compares pairs of cases presenting ethical dilemmas about whether or not to tell the truth (Ashley and McLaren1995; McLaren and Ashley 1995). The program was intended as a first step in implementing a computational model of casuistic reasoning, a form of ethical reasoning in which decisions are made by comparing a problem to paradigmatic, real, or hypothetical cases (Jonsen and Toulmin 1988). Casuistry long ago fell out of favor with many philosophers and ethicists because they believe it to be too imprecise and based on moral intuitions, but in recent times, casuistry has been employed as a technique to help solve practical dilemmas by medical ethicists (Strong 1988; Brody 2003). In contrast to the approach embodied in W. D. and MedEthEx just described, casuistry (and hence Truth-Teller) is focused on the power of specific cases and case comparison, not on the rules that are generalized from the evaluation of cases.

The Truth-Teller program marshals ethically relevant similarities and differences between two given cases from the perspective of the "truth teller" (i.e., the person faced with the dilemma) and reports them to the user. In particular, it points out reasons for telling the truth (or not) that (1) apply to both cases, (2) apply more strongly in one case than another, or (3) apply to only one case.



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 casem Henry and his family will go bankrupt. Should Felicia tell Henry about her inexperience in the matter?

CASE 2: Kelvin is a lawyer fresh out of law school. A client, Alida, requires a complex legal transaction that Kelvin 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 decison 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, Felcia and Kevin share the reason to realize a finanicial 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 Felcia'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 oneselt and others from serious harm.')

Figure 17.1. Truth-Teller's output comparing Felicia's and Kevin's cases.

The dilemmas addressed by the Truth-Teller program were adapted from the game of Scruples™, a party game in which participants challenge one another to resolve everyday ethical dilemmas.

Figure 17.1 shows Truth-Teller's output in comparing two dilemmas adapted from the Scruples game. As can be seen, these cases share very 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 this is detailed by Truth-Teller in the second and third

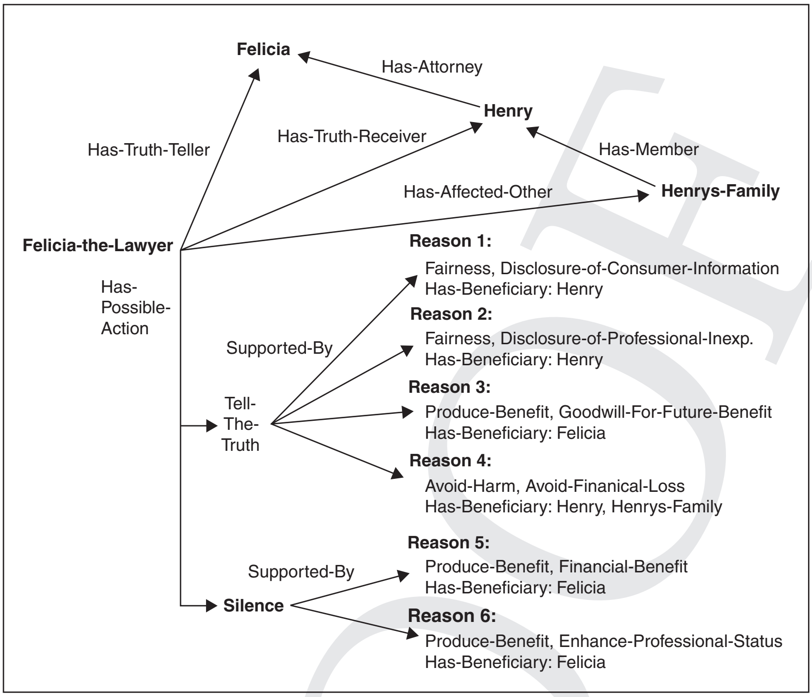


Figure 17.2. An example of Truth-Teller's case representation.

paragraphs of its output. 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, Truth-Teller points out the distinguishing features of the two cases in the last paragraph of its comparison text. Felicia has a greater obligation than Kevin to reveal her inexperience due to established custom (i.e., inexperienced lawyers rarely perform this transaction) and more severe consequences (i.e., Henry and his family will go bankrupt if she fails).

Figure 17.2 depicts Truth-Teller's semantic representation of the Felicia case of Figure 17.1. This is the representation that is provided as input to the program to perform its reasoning. In this case, Felicia is the "truth teller," and the actor who may receive the truth, or the "truth receiver," is Henry. Felicia can take one of two possible actions: tell Henry the truth or remain silent about her inexperience. It is also possible that the truth teller may have other actions he or she can take in a scenario, such as trying to resolve a situation through a third party. Each of the possible actions a protagonist can take has reasons that support it. For instance, two of the reasons for Felicia to tell the truth are (Reason 2) fairness – Felicia has an obligation to fairly disclose her inexperience – and (Reason 4) avoiding harm – Felicia might avoid financial harm to Henry and his family by telling the truth.

Truth-Teller compares pairs of cases given to it as input 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 phases of analysis:

(1) **Alignment**: build a mapping between the reasons in the two cases, that is, indicate the reasons that are the same and different across the two representations
(2) **Qualification**: identify special relationships among actors, actions, and reasons that augment or diminish the importance of the reasons, for example, telling the truth to a family member is typically more important than telling the truth to a stranger
(3) **Marshaling**: select particular similar or differentiating reasons to emphasize in presenting an argument that (1) one case is as strong as or stronger than the other with respect to a conclusion, (2) the cases are only weakly comparable, or (3) the cases are not comparable at all
(4) **Interpretation**: generate prose that accurately presents the marshaled information so that a nontechnical human user can understand it.

To test Truth-Teller's ability to compare cases, an evaluation was performed in which professional ethicists were asked to grade the program's output. The goal was to test whether expert ethicists would regard Truth-Teller's case comparisons as high quality. Five professional ethicists were asked to assess Truth-Teller as to the reasonableness (R), completeness (C), and context sensitivity (CS) on a scale of 1 (low) to 10 (high) of twenty of Truth-Teller's case comparisons, similar to the comparison in Figure 17.1. The mean scores assigned by the five experts across the twenty comparisons were R=6.3, C=6.2, and CS=6.1. Two human comparisons, written by graduate students, were also included in the evaluation and, not surprisingly, these comparisons were graded somewhat higher by the ethicists, at mean scores of R=8.2, C=7.7, and CS=7.8. On the other hand, two of Truth-Teller's comparisons graded higher than one of the human evaluations.

These results indicate that Truth-Teller is moderately successful at comparing truth-telling dilemmas. Because the expert ethicists were given the instruction to "evaluate comparisons as you would evaluate short answers written by college undergraduates," it is quite encouraging that Truth-Teller performed as well as it did. However, the following two questions naturally arise: Why were Truth-Teller's comparisons viewed as somewhat inferior to the human's and 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 (i.e., reasons, actions, and actors). One possible way to counter this would be develop techniques that allow Truth-Teller to suggest hypothetical variations to problems along the lines of the legal-reasoning program HYPO (Ashley 1990). For instance, in the comparison of Figure 17.1, Truth-Teller might suggest that, if an (unstated and thus hypothetical) longstanding relationship between Felicia and Henry exists, there is additional onus on Felicia to reveal her inexperience. Another criticism of Truth-Teller by the evaluators involved the program's somewhat rigid approach of enumerating individual supporting reasons, which does not relate 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

SIROCCO, the second ethical reasoning program created by the author, was developed as a second step in exploring casuistry and how it might be realized in a computational model. In particular, SIROCCO was implemented as an attempt to bridge the gap between general principles and concrete facts of cases. The program emulates the way an ethical review board within a professional engineering organization (the National Society of Professional Engineers – NSPE) decides cases by referring to, and balancing between, ethical codes and past cases (NSPE 1996).

The principles in engineering ethics, although more specific than general ethical duties such as Ross's prima facie duties (e.g., justice, beneficence, and nonmaleficence), still tend to be too general to decide cases. Thus, 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 code from the NSPE:

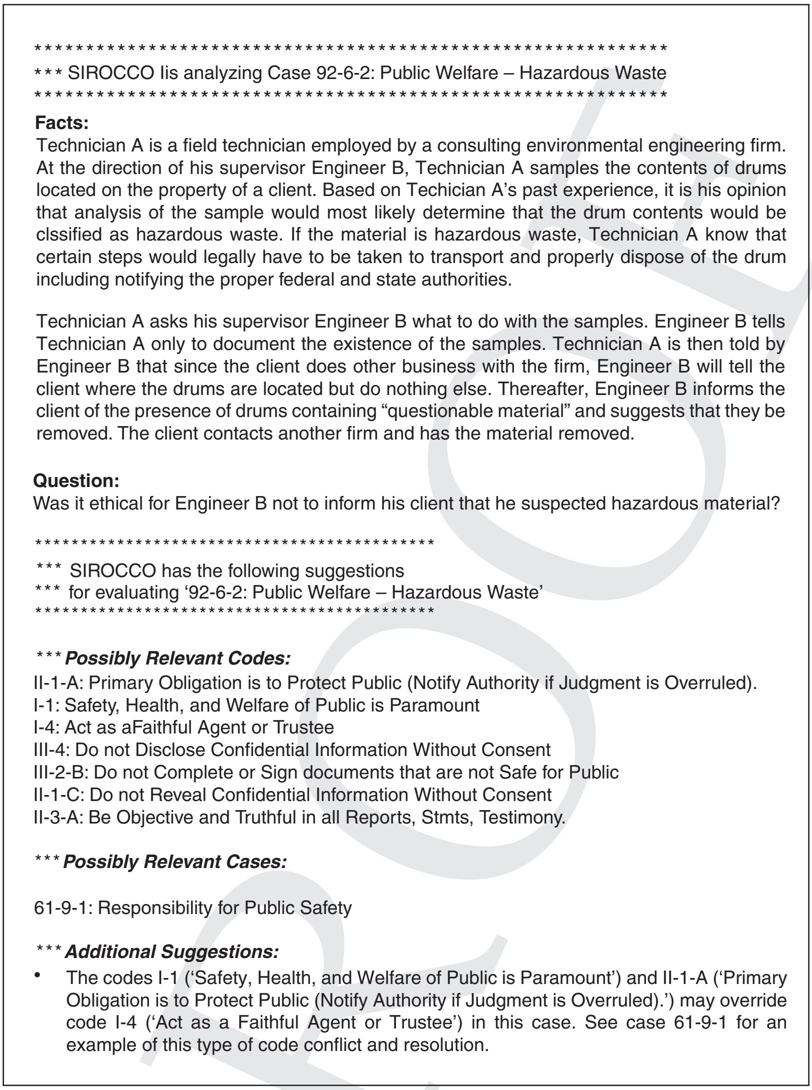
> *Code II.5.a*. Engineers shall not falsify or permit misrepresentation of their … academic or professional qualifications. They shall not misrepresent or exaggerate their degree of responsibility in or for the subject matter of prior assignments. Brochures or other presentations incident to the solicitation 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 ethical code specializes the more general principle of "honesty" in an engineering context. Each of the three sentences in the code deals with a different aspect of "misrepresentation of an engineer," and each sentence covers a wide range of possible circumstances. The precise circumstances that support application, however, are not specifically stated. Knowing whether this code applies to a particular fact-situation requires that one 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." Note that although these engineering ethics codes are an example of abstract codes, they are by no means exceptional. Many principles and codes, generally applicable or domain-specific, share the characteristic of being abstract. It is also typical for principles to conflict with one another in specific circumstances, with no clear resolution to that conflict. In their analyses of over five hundred engineering cases, the NSPE 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 and *extensional* representation of principles such as II.5.a.

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 (McLaren 2003). An example of SIROCCO's output is shown in Figure 17.3. The facts of the input case and the question raised by the case are first displayed. 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 is told not to mention his suspicions of hazardous waste to this important client, who might face clean-up expenses and legal ramifications from the finding. The question raised 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: (1) a list of possibly relevant codes, (2) a list of possibly relevant past cases, and (3) a list of additional suggestions. The interested reader can run the SIROCCO program on more than two hundred ethical dilemmas and view analysis such as that shown in Figure 17.3 by going to the following Web page: 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). SIROCCO's language represents the actions and events of a scenario as a Fact Chronology of individual sentences (i.e., Facts). A predefined ontology of Actor, Object, Fact Primitive, and Time Qualifier types are used in the representation. At least one Fact in the Fact Chronology is designated as the Questioned Fact; this is the action or event corresponding to the ethical question raised in the scenario. The entire ontology, a detailed description of how cases are represented, and more than fifty examples of Fact Chronologies can be found at: http://www.pitt.edu/~bmclaren/ethics/index.html.

SIROCCO utilizes knowledge of past case analyses, including past retrieval of principles and cases, and the way these knowledge elements were utilized in the past analyses to support its retrieval and analysis in the new (target) case. The program employs a two-stage graph-mapping algorithm to retrieve cases and codes. Stage 1 performs a "surface match" by retrieving all *source* cases – the cases in the program's database, represented in an extended version of ETL (EETL), totaling more than four hundred – that share any fact with the target case. It computes a score for all retrieved cases based on fact matching between the target case and each source case, and outputs a list of candidate source cases ranked by scores. Using an AI search technique known as A* search, Stage 2 attempts a structural mapping between the target case and each of the N top-ranking candidate source cases from Stage 1. SIROCCO takes temporal relations and abstract matches into account in this search. The top-rated structural mappings uncovered by the



```
***************************************************************
*** SIROCCO Iis 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 Techician A's past experience, it is his opinion
that analysis of the sample would most likely determine that the drum contents would be
clssified as hazardous waste. If the material is hazardous waste, Technician A know 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 aFaithful 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 17.3. SIROCCO's output for case 92–6–2.

A* search are organized and displayed by a module called the Analyzer. The output of Figure 17.3 is an example of what is produced by the Analyzer.

A formal experiment was performed with SIROCCO to test how well it retrieved principles and cases in comparison to several other retrieval techniques, including two full-text retrieval systems (Managing Gigabytes and Extended-MG). Each

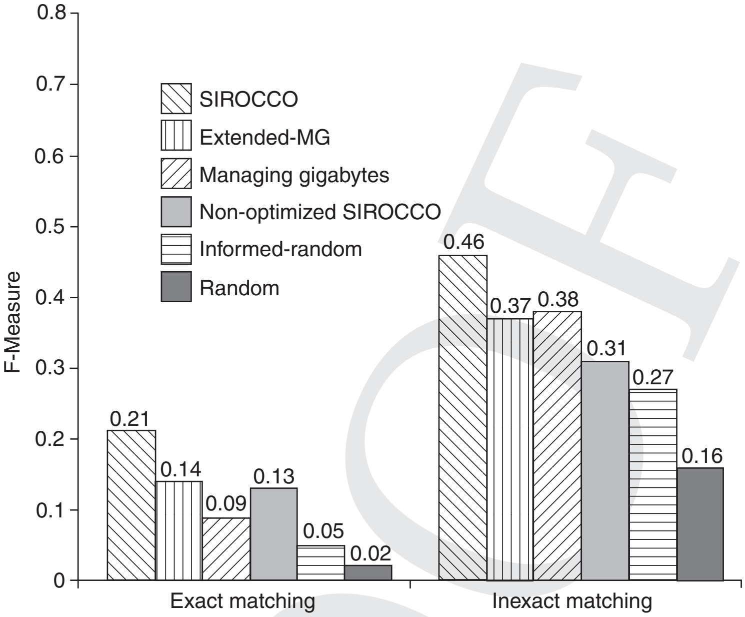


Figure 17.4. Mean F-Measures for all methods over all of the trial cases.

method was scored based on how well its retrieved cases and codes overlapped with that of the humans' (i.e., the NSPE review board) retrieved cases and codes in evaluating the same cases, using a metric called the *F-Measure*. The methods were compared on two dimensions: exact matching (defined as the method and humans retrieving precisely the same codes and cases) and inexact matching (defined as the method and humans retrieving closely related codes and cases). A summary of the results is shown in Figure 17.4.

In summary, SIROCCO was found to be significantly more accurate at retrieving relevant codes and cases than the other methods, with the exception of EXTENDED-MG, for which it was very close to being significantly more accurate (p = 0.057). Because these methods are arguably the most competitive automated methods with SIROCCO, this experiment shows that SIROCCO is an able ethics-reasoning companion. On the other hand, as can be seen in Figure 17.4, SIROCCO performed beneath the level of the ethical review board (0.21 and 0.46 can be roughly interpreted as being, respectively, 21 percent and 46 percent overlapping with the board selections). At least some, if not most, of this discrepancy can be accounted for by the fact that the inexact matching metric does not fully capture correct selections. For instance, there were many instances in which SIROCCO actually selected a code or case that was arguably applicable to a case, but the board did not select it. In other words, using the review board as the "gold standard" has its flaws. Nevertheless, it can be fairly stated that although

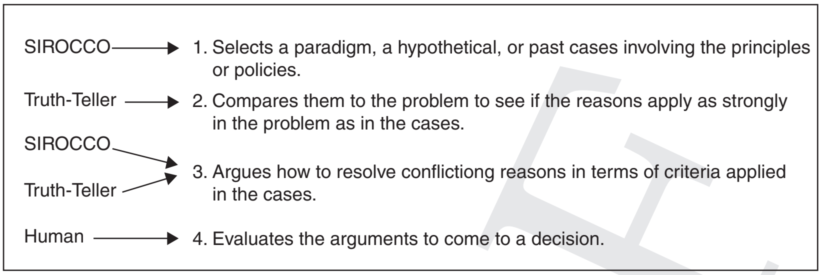
SIROCCO performs well, it does not perform quite at the level of an expert human reasoner at the same task.

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 and recognize important similarities and differences between the cases. Although SIROCCO also compares cases, its results are not focused on case comparisons and presenting those comparisons to the user. Rather, SIROCCO is more useful for collecting a variety of relevant information, principles, cases, and additional information that a user should consider in evaluating a new ethical dilemma. Whereas Truth-Teller has a clear advantage in comparing cases and explaining those comparisons, it ignores the problem of how potentially "comparable" cases are identified in the first place. The program compares any pair of cases it is provided, no matter how different they may be. SIROCCO, on the other hand, uses a retrieval algorithm to determine which cases are most likely to 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 cases. For instance, see the casuistic "algorithm" depicted in Figure 17.5. This "algorithm," adapted from the proposed casuistic approach of Jonsen and Toulmin (1988), 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. The casuist reaches 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 compares the new cases to the cases it retrieves. Although 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 argues 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 17.1, 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 may override another in these particular circumstances (see the "Additional Suggestions" at the bottom of Figure 17.3). Finally, a decision is made about this ethical dilemma. In keeping with the author's vision of how computational models should be applied to ethical decision making, neither Truth-Teller nor SIROCCO provides assistance on this step. This is the province of the human decision maker alone.

To fully realize the casuistic problem-solving approach of Figure 17.5 and combine the complementary capabilities of Truth-Teller and SIROCCO, the two



SIROCCO → 1. Selects a paradigm, a hypothetical, or past cases involving the principles or policies.

Truth-Teller → 2. Compares them to the problem to see if the reasons apply as strongly in the problem as in the cases.

SIROCCO, Truth-Teller → 3. Argues how to resolve conflictiong reasons in terms of criteria applied in the cases.

Human → 4. Evaluates the arguments to come to a decision.

Figure 17.5. Casuistic problem solving – and Truth-Teller's, SIROCCO's, and a human's potential role in the approach.

programs would need common representational elements. In SIROCCO, primitives that closely model some of the actions and events of a fact-situation are used 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 are not nearly as precise and issue-oriented as Truth-Teller's.

Both the Truth-Teller and SIROCCO projects are focused and rely heavily on a knowledge representation of ethics, in contrast to, for instance, the programs of Anderson et al., which have little reliance on representation. The knowledge-representation approach to building computational models of ethical reasoning has both strengths and weaknesses. The strength of the approach is the ability to represent cases and principles at a rather fine level of detail. For instance, a detailed ontology of engineering ethics is used to support the SIROCCO program, and a representation of reasons underlies Truth-Teller, as shown in Figure 17.2. Not only does such representation support the reasoning approaches of each model, but it also allows the models to provide relatively rich explanations of their reasoning, as exemplified by the output of the programs shown in Figures 17.1 and 17.3. On the other hand, the respective representations of the two models are necessarily specific to their tasks and domains. Thus, 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 model of reasoning that is embodied in the W. D. and MedEthEx programs of Anderson et al. So, another step that would be required to unify Truth-Teller and SIROCCO and implement the casuistic approach of Figure 17.5 would be a synthesis and generalization of their respective representational models.

Lessons Learned

The primary lesson learned from the Truth-Teller and SIROCCO projects is that ethical reasoning has a fundamentally different character than 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 in addressing and forming arguments in such domains using formal logic has long been recognized (Toulmin 1958), and some practitioners in AI, particularly those interested in legal reasoning, have also grappled with this issue. As pointed out by Ashley, "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" (1990, p. 2).

The domain of ethical reasoning, like the legal domain, can be viewed as a *weak analytic domain* characterized in which the given "rules" (i.e., laws, codes, or principles) are available almost exclusively at a highly abstract, conceptual level. This means that the rules may contain open-textured terms. That is, conditions, premises, or clauses that are not precise or that cover a wide range of specific facts, or are highly subject to interpretation and may even have different meanings in different contexts. 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 may appear to apply to a given fact-situation, but neither the abstract rules nor the general knowledge of the domain provide clear resolution.

Another important lesson from the Truth-Teller and SIROCCO projects is the sheer difficulty in imbuing a computer program with the sort of flexible intelligence required to perform ethical analysis. Although both programs performed reasonably well in the aforementioned studies, neither could be said to have performed at the level of an expert human at the same task. Although the goal was not to emulate human ability, taking the task of ethical decision making away from humans, it is 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, the Web-based system of Robbins et al., 2004) have been empirically tested in a way that might inspire faith in their performance.

It is important to make clear that the author's contention that computer programs should only act as aids in ethical reasoning is not due to a high regard for human ethical decision making. Of course, humans often make errors in ethical reasoning. Rather, the author's position is based, as suggested earlier, on the existence of so many plausible competing approaches to ethical problem solving. Which philosophical method can be claimed to be the "correct" approach to ethical reasoning in the same sense that calculus is accepted as a means of solving engineering problems or first-order logic is used to solve syllogisms? It is 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 there are lots of approaches that might be considered "good enough" without being definitive. However, the bar is likely to be held much higher for autonomous machine-based systems making decisions in an area as sensitive and personal to humans as ethical reasoning. Second, it is presumptuous to think that the subtleties of any of the well-known philosophical systems of ethics could be fully implemented 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, the W. D. program simplifies the evaluation of Ross's prima facie duties by assigning each a score on a five-point scale. Both the Truth-Teller and SIROCCO programs also make simplifying assumptions, such as Truth-Teller representing only reasons that support telling the truth or not, and not the circumstances that lead to these reasons. Of course, making simplifying assumptions is a necessary starting point for gaining traction in the difficult area of ethical reasoning. The third and final reason the author advocates for computational models being used only aids in ethical reasoning is the belief that humans simply won't accept autonomous computer agents making such decisions for them. They may, however, accept programs as advisors.

Future Directions

Given the author's 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. Goldin, Ashley, and Pinkus (2001) have taken steps in this direction. PETE 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 the answers of other students.

The author's most recent work and interest has also been in the area of intelligent tutoring systems (McLaren, DeLeeuw, and Mayer, in press; McLaren et al. 2009). As such, the author has started to investigate whether case comparisons, such as those produced by Truth-Teller, could be used as 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 the phases taken by the Truth-Teller program.

Cognitive Tutors are based on Anderson's ACT-R theory (Anderson 1993), according to which humans use production rules, modular IF-THEN constructs, to perform problem-solving steps in a wide variety of domains. Key concepts underlying Cognitive Tutors are "learn by doing," which helps students learn by engaging them in actual problem solving, and immediate feedback, which provides guidance to students at the time they request a hint or make a mistake. For domains like algebra, the production rules in a cognitive model indicate 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, not the least of which is that, unlike previous domains in which Cognitive Tutors have been used, such as algebra and programming, in practical ethics answers are not always and easily identified as correct or incorrect, and the rules, as explained earlier, are more abstract and ill-defined. As a result, although learning by doing fits ethics case comparison very well, the concept of immediate feedback needs to be adapted. Unlike more technical domains, ethics feedback may be nuanced rather than simply right or wrong, and the Cognitive Tutor approach must accordingly 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 within 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 subjects' comparing cases may reveal plausible misconceptions about the comparison process that can serve as buggy rules or faulty production 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 (1930), 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 Bok (1989). Such an extension to Truth-Teller would also benefit the planned Cognitive Tutor, as explanations to students could be supported with reference to Ross's theory.

Acknowledgments

This chapter was originally published as a journal article in IEEE Intelligent Systems (McLaren, 2006). Kevin Ashley contributed greatly to the ideas behind both Truth-Teller and SIROCCO. This work was supported in part by NSF-LIS grant No. 9720341.

References

Anderson, J. R. (1993). *Rules of the Mind*. Mahwah, NJ: Lawrence Erlbaum.

Anderson, S. L. (2005). Asimov's "Three Laws of Robotics" and Machine Metaethics. *Proceedings of the AAAI 2005 Fall Symposium on Machine Ethics*, Crystal City, VA. Technical Report FS-05–06, 1–7.

Anderson, M., Anderson, S. L., and Armen, C. (2005a). Towards Machine Ethics: Implementing Two Action-Based Ethical Theories. *Proceedings of the AAAI 2005 Fall Symposium on Machine Ethics*, Crystal City, VA. Technical Report FS-05–06, 1–7.

Anderson, M., Anderson, S. L., and Armen, C. (2005b). MedEthEx: Toward a Medical Ethics Advisor. *Proceedings of the AAAI 2005 Fall Symposium on Caring Machines: AI in Elder Care*, Crystal City, VA.

Aristotle, (edited and published in 1924) *Nicomachean Ethics*. W. D. Ross, editor, Oxford, 1924.

Ashley, K. D. (1990). *Modeling Legal Argument: Reasoning with Cases and Hypotheticals*. Cambridge: MIT Press, 1990.

Ashley, K. D. and McLaren, B. M. (1995). Reasoning with Reasons in Case-Based Comparisons. In the *Proceedings of the First International Conference on Case-Based Reasoning*, Sesimbra, Portugal.

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

Bentham, J. (1789). *Introduction to the Principles of Morals and Legislation*. In W. Harrison (ed.), Oxford: Hafner Press, 1948.

Bok, S. (1989). *Lying: Moral Choice in Public and Private Life*. New York: Random House, Inc. Vintage Books.

Brody, B. (2003). *Taking Issue: Pluralism and Casuistry in Bioethics*. Georgetown University Press.

Cavalier, R. and Covey, P. K. (1996). *A Right to Die? The Dax Cowart Case CD-ROM Teacher's Guide, Version 1.0*, Center for the Advancement of Applied Ethics, Carnegie Mellon University, Pittsburgh, PA.

Gardner, A. (1987). *An Artificial Intelligence Approach to Legal Reasoning*. Cambridge, MA: MIT Press.

Goldin, I. M., Ashley, K. D., and Pinkus, R. L. (2001). Introducing PETE: Computer Support for Teaching Ethics. *Proceedings of the Eighth International Conference on Artificial Intelligence & Law* (ICAIL-2001). Eds. Henry Prakken and Ronald P. Loui. Association of Computing Machinery, New York.

Harris, C. E., Pritchard, M. S., and Rabins, M. J. (1995). *Engineering Ethics: Concepts and Cases*. 1st edition. Belmont, CA: Wadsworth Publishing Company.

Jonsen, A. R. and Toulmin, S. (1988). *The Abuse of Casuistry: A History of Moral Reasoning*. Berkeley, CA: University of California Press.

Kant, I. (1785). Groundwork of the Metaphysic of Morals, in *Practical Philosophy*, translated by M. J. Gregor, Cambridge: Cambridge University Press, 1996.

McLaren, B. M. and Ashley, K. D. (1995). Case-Based Comparative Evaluation in Truth-Teller. In the *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*. Pittsburgh, PA.

McLaren, B. M. (1999). *Assessing the Relevance of Cases and Principles Using Operationalization Techniques*. Ph.D. Dissertation, University of Pittsburgh

McLaren, B. M. (2003). Extensionally Defining Principles and Cases in Ethics: an AI Model; *Artificial Intelligence Journal*, Volume 150, November 2003, pp. 145–181.

McLaren, B. M. (2006). Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions. *IEEE Intelligent Systems*, Published by the IEEE Computer Society. July/August 2006. 29–37.

McLaren, B. M., DeLeeuw, K. E., & Mayer, R. E. (2011). Polite web-based intelligent tutors: Can they improve learning in classrooms? *Computers & Education*, 56, 574–584. doi: 10.1016/j.compedu.2010.09.019.

McLaren, B. M., Wegerif, R., Mikšátko, J., Scheuer, O., Chamrada, M., & Mansour, N. (2009). Are your students working creatively together? Automatically recognizing creative turns in student e-Discussions. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. Graesser (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (AIED-09), Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling. (pp. 317–324). IOS Press.

Mill, J. S. *Utilitarianism*. (1863). In George Sher, (Ed.) Indianapolis, Indiana, USA: Hackett Publishing Company, 1979.

National Society of Professional Engineers (1996). *The NSPE Ethics Reference Guide*. Alexandria, VA: the National Society of Professional Engineers.

Rawls, J. (1971). *A Theory of Justice*, 2nd Edition 1999, Cambridge, MA: Harvard University Press.

Robbins, R. W. and Wallace, W. A. (2007). A Decision Aid for Ethical Problem Solving: A Multi-Agent Approach. *Decision Support Systems*, 43(4): 1571–1587.

Robbins, R. W., Wallace, W. A., and Puka, B. (2004). Supporting Ethical Problem Solving: An Exploratory Investigation. In the *Proceedings of the 2004 ACM Special Interest Group on Management Information Systems and Computer Personnel Research*, 22–24.

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

Searing, D. R. (1998). *HARPS Ethical Analysis Methodology, Method Description. Version 2.0.0.*, Lake Zurich, IL: Taknosys Software Corporation, 1998.

Strong, C. (1988). Justification in Ethics. In Baruch A. Brody, editor, *Moral Theory and Moral Judgments in Medical Ethics*, 193–211. Dordrecht: Kluwer Academic Publishers.

Toulmin, S. E. (1958). *The Uses of Argument*. Cambridge, England: Cambridge University Press.