

Teaching Case Analysis through Framing: Prospects for an ITS in an ill-defined domain

Ilya M. Goldin

Intelligent Systems Program
Learning Research &
Development Center
University of Pittsburgh
goldin@pitt.edu

Kevin D. Ashley

Intelligent Systems Program
Learning Research &
Development Center
University of Pittsburgh
ashley@pitt.edu

Rosa L. Pinkus

Neurosurgery/Medicine
School of Medicine
University of Pittsburgh
pinkus@pitt.edu

Abstract. Intelligent Tutoring Systems research has made assumptions that may be violated in ill-defined tasks. We describe ethics case analysis, an important educational yet ill-defined task in the domain of bioengineering ethics. We discuss an ITS approach for this task that involves structuring the learning experience and using AI to help guide student peer reviewers.

Keywords: intelligent tutoring system, ill-defined domain, bioengineering ethics, case analysis, framing

INTRODUCTION

A basic assumption in the design of Intelligent Tutoring Systems is that an ITS needs to interpret student output. This is necessary so that the ITS can provide appropriate feedback, or to assess knowledge for student modeling, or to measure performance. In domains such as geometry or algebra, it is often possible to design the problems that students solve so that the result is numeric or at least symbolic and constrained in a way that an ITS can interpret. In other words, these problems are well-defined.

An accepted approach to teaching bioengineering ethics is to use textual cases to illustrate important domain concepts and methods of moral reasoning. The role that cases play in the classroom is to permit students to immerse themselves—to situate their problem solving—in the complexity and variety of authentic ethical dilemmas. This methodology, a kind of problem-based learning (PBL), is distinct from and often complements alternative teaching practices, which may emphasize philosophical moral theories. The PBL setting requires that students learn to analyze cases. For example, the case analysis method taught in the popular textbook (Harris et al., 2000) asks students to consider the morally relevant facts of the case, both known and unknown; to structure the analysis via the conceptual issues that can relate the facts to each other; to use their moral imagination to propose and compare alternative resolutions to the dilemma; and finally to justify a particular resolution.

The fact that students may produce a wide range of acceptable responses marks the case analysis task as ill-defined. An ethical dilemma may have no good resolutions at all, or it may have multiple ones. In fact, usually only “paradigm” cases will have clear-cut definitive answers. Realistic problems will be more complex and more equivocal as one considers alternative frameworks for resolving the problem and justifying the resolutions. Not only do ethical problems rarely have definitive answers, the answers depend in part on how the problem is conceived or “framed”, as well as on the justifications. (Pinkus et al., in preparation) A large part of problem solving in this ill-defined domain involves constructing (i.e., framing) a representation of the problem, which may include additional constraints and possible actions. As a result, a given problem as posed can become a different problem depending on the constraints and conditions that a solver adds in order to better define the problem or to suggest alternative actions. In addition, for many ethics problems, the issue is not only identifying relevant moral principles to justify an answer, but also mapping the concepts in the principles to the situation at hand, which may not be clear-cut and may require a good deal of subjective interpretation.

One consequence of the ill-defined nature of ethics case analysis is that the most appropriate representation for student case analyses is free-form text. Only natural language enables describing the problem scenarios in sufficient detail and for considering the implications of particular details on alternative resolutions. Another consequence is that even if an ITS could understand the text, the possibility of a wide range of acceptable answers makes providing feedback, modeling student knowledge, or measuring performance correspondingly difficult for a human tutor, much less a machine.

We have studied how one educator addresses the ill-defined nature of ethics case analysis by encouraging students to frame the cases they analyze and by gauging their skills with a manually administered specially-

designed Assessment Instrument. We have demonstrated the Instrument's validity and reliability in assessing some important moral reasoning skills. (Goldin et al., 2006; Goldin et al., in preparation) Our objective here is to propose an ITS that helps to teach analysis of textual ethics cases. Our design adapts a computer-supported collaborative learning program (SWoRD or Scaffolded Writing and Rewriting in the Discipline) to support peer review of the case analyses and leverages a database of student-authored analyses manually annotated with the Assessment Instrument. First, we describe the pedagogical technique for dealing with this ill-defined task and the Instrument, and summarize our evaluation of the Instrument's validity and reliability. We then outline our proposed system, suggest how it will incorporate AI techniques in helping students learn to analyze ethics cases, and how to evaluate it. To conclude, we review related work, and consider links to other ill-defined domains.

ETHICS CASE ANALYSIS THROUGH FRAMING: ASSESSMENT CHALLENGES

Educators have developed some ingenious pedagogical strategies using PBL to deal with the ill-defined nature of ethics case analysis. Author Pinkus, a professional ethicist, assigns students a capstone exercise in a required Bioengineering Ethics class: the task is to write a one or two page case study based on a student's area of engineering expertise. In these cases, the protagonist, an engineer, is faced with a dilemma, which is caused by, or will have ramifications for, the engineer's professional duties. The course of action is unclear, and requires analysis. Each student presents the case to the class for comment and then writes a paper that analyzes the case using the methods taught in the class. This approach, where students create cases close to their professional expertise and interests, has been shown to be a positive factor in student learning. (Pinkus et al., in preparation)

From the viewpoint of intelligent tutoring systems, this approach poses challenges. An ITS needs not only to be able to model a student's solution where many alternative solutions are acceptable, but also a case that the student has designed herself. Indeed, even if we simplify the task by asking students to analyze an assigned case (and forego the pedagogical benefits of student-authoring of cases), students trained to apply the Harris method will do so differently because of how they *frame* the case.

Framing Dilemmas through Labeling, Defining, Applying Concepts

Framing is one strategy for dealing with ill-defined domains. As noted, ethics problems, except in paradigm cases, rarely have a definitive answer or even a definitive description. Students must add constraints to the facts of the case and thus articulate what the ethical dilemma is. This in turn, affects how a moral principle or a professional ethics code applies to the problem. Mapping the principles to the situation and weighing the effects of alternative actions or using one's moral imagination to create compromises in order to resolve conflicts are among the skills that students must learn. The methods of moral reasoning described in the Harris text and elsewhere provide a conceptual framework for viewing the cases. The framework can be derived from moral theories, principles, the moral imagination, or from tacit rules embedded in engineering practice. Once a case is framed, the moral dilemmas that characterize it can be articulated. Then, professional knowledge and key concepts can serve to "filter" morally relevant facts and alternative resolutions can be proposed.

Thus, one approach to ill-defined aspects of ethics case analysis requires the problem-solver to define the problem better through framing. The latitude one has in framing a case, however, is not intuitive to an engineering student. Typically in engineering, one is given a problem already framed and asked to solve it, such as in textbook standardized ethics cases. Given the importance of framing in ethics problem-solving, asking students to create their own cases is an ideal pedagogical exercise; it requires them to frame the problems.

Although the facts of a case are the most obvious properties of a dilemma, their relative importance becomes apparent only after they are framed within the conceptual issues implicit in the case. Thus, even if an analysis ought to begin with identification of facts, this is insufficient for purposes of student modeling or assessment. Consider, for instance, the conceptual issue of informed consent. A typical dilemma that involves this issue is whether a person has been properly informed about the risks inherent in a medical procedure, and has granted consent to be exposed to these risks, probably with the hope of deriving some benefit. For example, a participant in evaluating a new medical treatment must be informed and grant consent before being subjected to the treatment. To analyze a case in terms of informed consent, it is important to recognize at least two protagonists: one with expert knowledge and skills; the other in need of the expert's service, but whose permission is a prerequisite for accepting those services. The expert will have to inform the person in need, especially about risks and benefits of accepting the service and of any alternatives that could be used. The person accepting the service can only grant consent if this "informing" is non-coercive, and if he demonstrates understanding.

The example of the informed consent frame comes from an awareness of professional responsibilities, which is an outcome of "role morality," i.e., the obligations inherent in one's role as a professional, as well as from the related concepts of autonomy and respect for persons. One may also frame a concept from personal or common morality. Personal morality means that one's personal values can be a legitimate factor in how one views a case, even if these values are clearly not shared by others. In the informed consent case, the physician may personally

not agree with a patient's decision to forego life-sustaining treatment for her pancreatic cancer yet legal considerations and medical ethics guide the physician to respect the patient's informed decision. Common morality means that concepts like honesty and respect for persons are universally shared, and can guide framing. Experienced ethical reasoners are "careful to identify issues and to specify conditions under which specific professional role obligations recommend particular actions, [to elaborate] conditions which would affect the moral analysis of a problem, in part through posing hypothetical variations of the problem, and [to justify] resolutions in terms of those conditions which they conclude apply in the problem." (Keefer & Ashley, 2001)

Thus, one way a student can frame a case is to claim that particular concepts, such as informed consent, constitute the frame through which the case ought to be viewed. The next step is to define the issue in a general, abstract way, like informed consent in the above paragraph; this shows an understanding of the properties of the issue removed from the details of a given case. Finally, the student must explain how the definition maps to the case at hand. We call these three steps *labeling* the concept as such; *defining* it; and *applying* it.

"Sometimes apparent moral disagreement turns out to rest on conceptual differences where no one's motives are in question. These are issues about the general definitions, or meanings, of concepts." (Harris et al., 2000, p. 46) Definitions are particularly important when open-ended terms are in play, such as "acceptable risk" posed by an engineer's creation, or "the public" whose safety the engineer ought to hold paramount. Such open-ended language figures prominently in abstract ethics principles, and even in the more detailed codes of ethics. Yet "attempts to specify the meanings of terms ahead of time can never anticipate all of the cases to which they do and do not apply. No matter how precisely one attempts to define a concept, it will always remain open-ended; that is, it will always remain insufficiently specified, so that some of its applications to particular circumstances will remain problematic." (Harris et al., 2000, p. 50) This requires a problem-solver to apply the concept to the specifics of the case, i.e., to say whether a given fact situation constitutes an occurrence of a concept. When a problem-solver labels, defines and applies an open-ended, i.e., ill-defined, term, she frames the conceptual issue.

Consider this excerpt from a good case analysis (Figure 1) showing defined or applied concepts. (Coders annotate on a computer, and the terms that are concept labels can be automatically identified; thus, we omit the labels.) Using her definition of the concept of "responsibility of a bioengineer" as a hub, the author applies the concepts of confidentiality, autonomy, and safety to consider whether to disclose the pilot's condition.

Example: <concept-applied="safety"> Jeff - the bioengineering student - is concerned for the safety of the pilot and of his future passengers since he plans to continue flying despite the doctor's advice to stop. </concept-applied> <concept-applied="responsibility-of-bioengineer"> Jeff wonders whether he is responsible for telling the airline of Joe's condition since the neurosurgeon and his advisor will not. </concept-applied> <concept-applied="confidentiality"> Would Jeff be breaching the confidentiality constraints set by the IRB form if he informed the airline? </concept-applied> Is there a solution to this issue that would serve the best interests of all parties?

<concept-defined="responsibility-of-bioengineer"> Researchers have various obligations and prerogatives associated with their profession, and these responsibilities can be referred to as their role morality. </concept-defined> [1] For example, <concept-applied="responsibility-of-bioengineer"> researchers have responsibilities to their experimental subjects mandating that the subjects' safety be of utmost importance. Furthermore, <concept-defined="autonomy"> researchers should respect their subjects' autonomy, allowing them to decide if they want to participate and allowing them to discontinue the experiment at any point. </concept-defined="autonomy"> <concept-applied="confidentiality"> Furthermore, a subject's identity should be kept as confidential as possible. </concept-applied="confidentiality"></concept-applied="responsibility-of-bioengineer"> Except under unusual circumstances, <concept-defined="confidentiality"> the only people who should have access to the subject's records are the investigators and staff who run the experiments. </concept-defined> According to the Bioethics Advisory Commission (of August 2001), "Protecting the rights and welfare of those who volunteer to participate in research is a fundamental tenet of ethical research". [2] <concept-applied="responsibility-of-bioengineer"> When deciding whether or not to inform the airline of Joe's condition, Jeff needs to be cognizant of the responsibilities he has toward his subjects, particularly his responsibility to respect their confidentiality. However, he also has to consider <concept-applied="safety"> his responsibility to protect Joe's safety, which may be in danger if he continues to fly despite his medical condition. </concept-applied="safety"> In addition to researchers' obligations to their subjects, they also have obligations to society. </concept-applied="responsibility-of-bioengineer">

Figure 1: An excerpt from a case analysis, annotated for defined or applied concepts.

Assessment Instrument for Labeling, Defining, and Applying

In this way, labeling, defining, and applying (LDA) serve as an operationalization of framing of concepts, similar to how the NSPE Board of Ethical Review fleshes out abstract ethics code provisions when they analyze exemplar cases for posterity. (Ashley & McLaren, 2001) The LDA operationalization is embedded in an Assessment Instrument for bioengineering ethics case analysis. (Pinkus et al., in preparation) The Instrument is a set of questions that invites coders to assess whether students acquire Higher-Level Moral Reasoning Skills (HLMRS). LDA operationalize one of these skills with questions about labeling, defining and applying of over 40 concepts. The list of concepts has been derived from the Harris text and from student essays, and includes consideration of moral theories, principles, codes of ethics, and common, personal, and role moralities. (Other

HLMRS recognize other ways to frame a case, e.g., from professional knowledge rather than from conceptual issues, but that requires a different operationalization of framing; for example, see (Martin et al., 2005).)

We evaluated the Assessment Instrument’s validity by measuring whether it reflects student learning. We also conducted a study to evaluate how reliably the instrument can be applied by comparing how well independent human coders agreed on their coding assignments. (Goldin et al., in preparation) The sensitivity study showed that the LDA operationalization is valid, because it is sensitive to student learning gains during a semester-long class. We compared student skills at analysis of short assigned ethics cases, as measured by the Assessment Instrument, at pre- and posttest times ($n=13$, Figure 2). Students labeled, defined, or applied very few concepts at pretest, and significantly more at posttest. At pretest, students do not label, define or apply (code “none”) 94.5% of the concepts, and they never do all three (code “LDA”). At post-test, they invoke significantly more concepts, including when they comprehensively label, define and apply the concept.

Coder Pair	Label	Define	Apply	Other 4 HLMRS
Trained vs. trained (N=12)	0.893	0.892	0.868	0.324
Naïve vs. trained (N=29)	0.626	0.472	0.577	0.158

Table 1: Agreement between coders (Cohen’s Kappa) on Assessment Instrument annotations

The reliability study showed that the Assessment Instrument can be applied reliably by trained independent coders, and fairly reliably even by untrained coders. Three pairs of coders annotated 41 student-authored term papers using the Assessment Instrument. Agreement for the LDA (Table 1) was much higher than that for the more abstract HLMRS, reflecting the value of operationalizing one of the HLMRS. Note that for measuring IRR, one usually trains coders on a range of possible answers to some particular question. In our study, each term paper contained a new, student-authored case, meaning that we could never train our coders on the particular “question” they would annotate. Consequently, the high level of agreement on LDA is especially noteworthy. The promising results of evaluating the Instrument lead us to design the system described below.

PROPOSED SYSTEM ARCHITECTURE

Given the challenges posed by the ill-defined nature of case analysis in bioengineering ethics to the task of designing an ITS, it is clear that traditional ITS technology is insufficient. The system needs to accommodate the student-authored cases and analyses. While Natural Language Understanding is an active area of research, we are a long way from computer comprehension of student essays. That means that humans need to do the bulk of understanding the texts. At the same time, we hold out the hope that gradual advances in NLU technology will some day permit the system to bear a greater load. Thus, our goal is to produce a system that

- organizes the process of case creation, analysis, and gathering feedback so that it may
- enhance this process to the extent that technology allows today, and
- collect data on this process that can be used to improve the state of the art.

Case analysis is a writing task. Traditional classroom writing instruction suffers from two basic flaws: the teacher can be overburdened by the obligation to provide feedback as class size grows, and even in small classes students may lack the opportunity to respond to feedback by submitting multiple essay drafts. One way to address this is to ask student peer reviewers to provide feedback to each other. This fits with the requirement that humans need to do the bulk of understanding the texts. By staging peer review online with the help of a system like SWoRD (Cho & Schunn, 2005), we tackle goals (a) and (c). Ideally, peer feedback helps drive home the significance of having framed the case in one way rather than another, and the act of peer reviewing constitutes a learning opportunity in itself. Furthermore, peer review assures the students that their classroom work is not an abstract exercise, but has real consequences—lessons professional ethics courses seek to teach. We intend to compare the learning effects of SWoRD-aided peer review versus traditional classroom practice.

Finally, we aim to enhance the peer review process by asking how the system could encourage students to

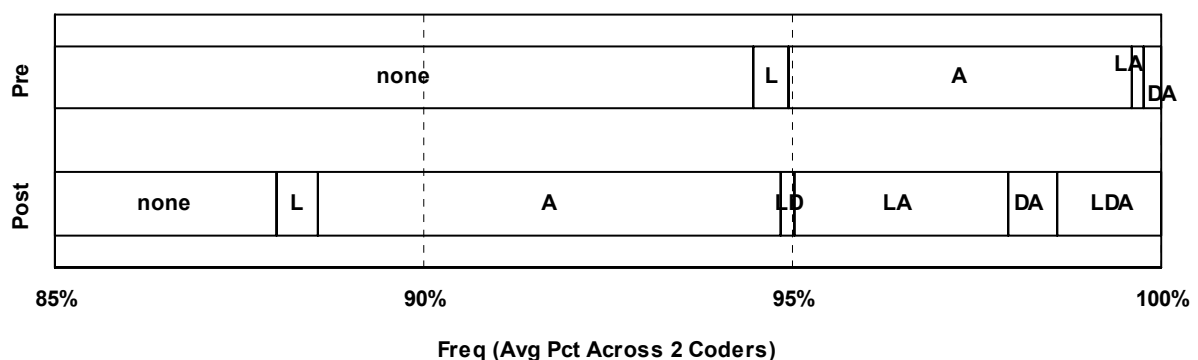


Figure 2: Change in LDA scores from pre- to posttest; LA = Labeled and Applied, not Defined, etc.

4. Evidence of moral reasoning skills

This dimension is about the evidence of moral reasoning skills in the author's case analysis. Did the author appear to: (1) employ professional engineering knowledge to frame the issues, (2) view the problem from multiple levels (e.g., that of the engineer, the employer, the client, the patient, the public, regulators, etc.), (3) flexibly move among the multiple levels in his/her analysis, (4) identify analogous cases and explain the analogies, and (5) employ a method of moral reasoning in conducting the analysis? In connection with (5), did the author identify moral reasoning concepts relevant to analyzing the case? Did the author label, define, and apply these concepts?

Your Comments: Provide specific comments about the paper's evidence of moral reasoning skills. If the author *employed relevant professional knowledge to frame the issues, or viewed the problem from multiple levels or moved flexibly among those levels, or identified relevant analogous cases and adequately explained the analogies, or used a moral reasoning method in conducting the analysis*, point that out and congratulate them! If the author did not do so, try to suggest potential fixes to these problems. In looking for evidence of a method of moral reasoning, look to see if the author identified relevant moral reasoning concepts in analyzing the case, or labeled, defined, and applied such concepts. Suggest relevant concepts and how to label, define, and apply them in the case. (The **GLOSSARY** defines and provides examples of many concepts from this course.)

<Peer reviewer writes free-form comments here>

Your Rating: Based on your comments above, how would you rate the evidence of moral reasoning skills in the author's case analysis?

- | | | |
|--------------------------|---------------|---|
| <input type="checkbox"/> | 7. Excellent | The paper shows strong evidence of all five moral reasoning skills. The author labels, defines, and applies relevant concepts in his/her case analysis. |
| <input type="checkbox"/> | 6. Very good | The paper shows strong evidence of all but one of the first four moral reasoning skills. Regarding the fifth, the author labels, defines, and applies most of the relevant concepts in his/her case analysis. |
| <input type="checkbox"/> | 5. Good | The paper shows some evidence of two of the first four moral reasoning skills. Regarding the fifth, the author labels, defines, and applies some of the relevant concepts in his/her case analysis. |
| <input type="checkbox"/> | 4. Average | The paper shows some evidence of only one of the first four moral reasoning skills. Regarding the fifth, the author either labels or applies some of the relevant concepts in his/her case analysis but does not always label, define and apply each concept. |
| <input type="checkbox"/> | 3. Poor | The paper shows almost no evidence of any of the first four moral reasoning skills. Regarding the fifth, the author alludes to some of the relevant concepts in his/her case analysis but does not always label, define, and apply each concept. |
| <input type="checkbox"/> | 2. Very poor | The paper shows no evidence of any of the first four moral reasoning skills. Regarding the fifth, the author alludes to some of the relevant concepts in his/her case analysis but does not label, define, and apply each concept. |
| <input type="checkbox"/> | 1. Disastrous | The paper shows no evidence of any of the first four moral reasoning skills. Regarding the fifth, the author does not label, define, or apply any relevant concepts in his/her case analysis. |

Figure 3: Moral reasoning skills criterion for peer reviewers

frame cases. We examine two strategies: first, adapt SWoRD to the domain of case analysis; second, provide reviewers domain-specific feedback that builds on existing case analyses already annotated for framing.

Adapting SWoRD to Peer Reviewing of Ethics Case Analyses

SWoRD is a web-based instructional system that supports reciprocal student authoring and student peer reviewing. Its aim so far has been to improve writing by focusing reviewers on prose flow, logical argument, and insight. Of course, these are aspects of writing that an ethics case analysis should also include. Prose flow concerns how well the author identifies the main points and transitions from one point to the next. Logical argument is “the extent to which each paper is logically coherent in terms of text structure that organizes various facts and arguments,” and “how well the main arguments are supported with evidence.” Insight involves “the extent to which each paper contributes new knowledge and insight to the reader. In classes, this is operationally defined as new knowledge or insight beyond required class texts and materials.” (Cho & Schunn, 2005).

Our goal for SWoRD is to help students learn not only writing skills, but domain reasoning skills: how to analyze ethics cases by framing. We hope that if we reach the peer reviewers, not only will they learn, but they will in turn reach the student authors. SWoRD has been deployed and evaluated in many classrooms, including domains as varied as psychology and physics, but with the focus on writing quality described above. It has not been applied to improve domain reasoning skills, nor in the context of engineering ethics.

We will adapt SWoRD to engineering ethics by defining a criterion that focuses student peer reviewers on the higher-level moral reasoning skills, and, in particular, on the skill “using a method of moral reasoning,” operationalized in terms of whether or not authors label, define, and apply ethical concepts. In essence, the criterion teaches the reviewers to apply the Assessment Instrument, a valid and reliable method of assessment. The goal is to teach the reviewers to critique the case analyses in terms of the HLMRS, and relevant ethical concepts in particular. This, in turn, will encourage student authors to do the same. A new page of the form that

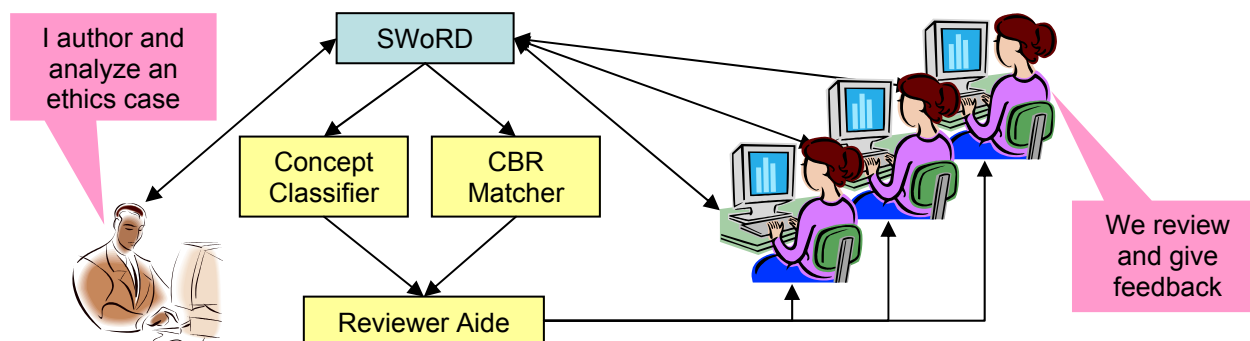


Figure 4: System architecture: SWoRD plus Reviewer Aide with Concept Classifier and CBR Matcher

reviewers fill out (Figure 3) comprises a detailed description of the new criterion, followed by instructions, a space for comments, and a seven point scale along which the reviewer rates the paper. In addition to focusing the reviewers and authors on the HLMRS, the system will facilitate access to an already developed Glossary of the concepts, their definitions and application examples, like the entry for confidentiality:

Confidentiality requires that all professionals hold information about a client/patient/subject “safe” i.e. undisclosed because it was given to the professional with the idea that it would not be disclosed to others. This may include information given by the client or information gained by the professional in work paid for by the client. (Harris et al., 2000, p. 132)

Example: Jeff is an airplane pilot who has been discovered to have vertigo in a research study in which he is enrolled. When the researchers involved report this finding to Jeff’s medical physician without first asking Jeff, Jeff becomes furious because his confidentiality has been breached and because he could lose his pilot’s license.

The question remains whether the peers can provide good feedback. As the SWoRD creators note, “One of the fundamental challenges is that peer reviewers are novices in their disciplines.” (Cho & Schunn, 2005) Presumably, novice reviewers face most difficulty with domain-specific reasoning skills. Novices will lack the expertise to make informed judgments about higher level moral reasoning skills in the domain of engineering ethics. Apart from the new domain-specific criterion we will introduce, SWoRD itself addresses this challenge with the distributed expertise of multiple reviewers, statistical checks on review accuracy, and authors’ back-reviews. Evaluations of SWoRD have shown that student authors improve their writing more from feedback of multiple peers than from single-peer or single-expert feedback (Cho & Schunn, 2005), even if the expert’s feedback is of higher quality, according to a second, blinded expert.

Using AI to Enhance Peer Reviewing

Aside from adapting SWoRD, we are exploring how we can enhance peer review through system-generated feedback. The goal here is to ease the job of the reviewers who are faced not only with providing feedback to their peers, but also with learning difficult new skills themselves. While the new domain-specific criterion asks the reviewers to note what concepts are relevant to the case at hand, we hope to be able to show the reviewers what concepts have already been defined or applied by the author, and what concepts were relevant in similar cases. We will test out two new approaches: a *Concept Classifier* to detect what concepts have been defined or applied in a new case analysis; and a *CBR Matcher* to locate existing case analyses similar to the new one, and to report what concepts were defined or applied in them. The results of the processing performed by the *Concept Classifier* and *CBR Matcher* will serve as input to a new *Reviewer Aide* component. The *Reviewer Aide* will use a model of peer review to determine what feedback to display to the peer reviewers and in what form.

Both the *Concept Classifier* and the *CBR Matcher* tailor their feedback to the case analysis under review, and focus the feedback on framing through reporting on concept definitions and applications. We will train the *Concept Classifier* and the *CBR Matcher* on an already collected corpus of case analyses. It contains approximately 150 term papers written by graduate and undergraduate students as capstone exercises in Pinkus’s Bioengineering Ethics course. We are in the process of completing manual annotation of the corpus using the Assessment Instrument to indicate where students define and apply any moral reasoning concepts, as in Figure 1. The annotation is being performed at the sentence level with the GATE natural language engineering software (Cunningham et al., 2002); approximately two thirds of the corpus has been annotated already. The papers represent the “multi-disciplinary” knowledge domain that comprises bioengineering ethics.

The database can be considered a “casuistry” of cases that provides examples of how a set number of ethical concepts and principles are used to frame and define issues that students have identified. The sample excerpt from a case analysis illustrates the coders’ annotations. All these case analyses have been authored by students taught with the Harris text as part of their final class projects. They cover many of the topics addressed in the

Your Comments: Provide specific comments about the paper’s evidence of moral reasoning skills. Suggest relevant concepts and how they apply. (The **GLOSSARY** has definitions and examples of many concepts from this course.)

Notes from the Reviewer Aide:

- This paper refers to ‘safety’ on lines 220 and 225, but there doesn’t seem to be a definition. Here is a definition and example of safety from the **GLOSSARY**.
- This paper defines ‘responsibility of a bioengineer’. Here are some sample definitions of responsibility of a bioengineer from similar case analyses. Does your author get it right?

Figure 5: Reviewer Aide prompt to peer reviewers

Harris text, such as honesty and the obligation to disclose information, in a bioengineering context. Students created and analyzed their own bioengineering ethics fact situations dealing with such questions as, “Should a graduate researcher report a defect in the cusp of a tissue-engineered heart valve that he is evaluating for another purpose?” and “How can an informed consent be written to enable a study of unexpected slips?” In connection with our experiments testing the sensitivity of the Assessment Instrument, we also collected 28 short student analyses of standardized bioengineering ethics cases such as one adapted from the first artificial heart transplant, which have also been annotated using the Assessment Instrument

Our hope is that the *Concept Classifier* can learn to detect concept definitions and applications in new case analyses. We will try to train a Naïve Bayes bag-of-words classifier (McCallum & Nigam, 1998) on definitions and applications in our corpus, starting with the most frequently occurring concepts (the full Instrument has over 40 concepts). We will then augment the term vectors with positional and natural language features like distance to beginning of document and part-of-speech tags. If this machine learning experiment is successful, then the *Concept Classifier* could make a probabilistic judgment whether a student author has:

1. Labeled a moral reasoning concept using proper terminology, and defined or applied it.
2. Labeled a moral reasoning concept, but failed to define or apply it.
3. Defined or applied a moral reasoning concept, but failed to label it as such.
4. Failed to label, define or apply a moral reasoning concept that is salient for an assigned case.

Another way to enhance peer reviewing with the help of AI is to show reviewers examples of how cases similar to the one at hand have been analyzed. We will attempt to create a *CBR Matcher*, which will use a Nearest Neighbor algorithm to make a probabilistic judgment whether a new essay is similar to existing analyses in the corpus. If so, then moral reasoning concepts discussed in existing essays may also be relevant to the new one; it would be easy to report what concepts were discussed in existing essays thanks to manual annotation by human coders. Similarity between case analyses can be determined by any shared concepts, or by whether the essays belong to the same curricular bioengineering ‘track’. We will categorize our corpus according to six “specialty tracks” defined by the Bioengineering Department at the University of Pittsburgh within its graduate degree program: Cellular and Organ Engineering; Biomechanics of Organs, Tissues, and Cells; Biosignals and Imaging; Physiology and Biophysics; Neural Engineering; Rehabilitation Engineering and Human Movement. The results from this search for similar cases can be used to aid the peer reviewers in providing feedback to authors. Of course, the *CBR Matcher* presupposes that similar cases will have concepts framed in similar ways, and that reviewers and authors will benefit from such information.

The *Reviewer Aide* determines what feedback to display to the reviewers (Figure 5) based on input from the *Concept Classifier* and the *CBR Matcher*. It will make decisions about relevance, timeliness, informativeness, and accuracy of the feedback using criteria like “use only the search results from *Concept Classifier* that exceed a relevance threshold,” “do not overwhelm reviewer with too many search results,” “always refer to specific textual passages in feedback to reviewer,” and “do not draw reviewer’s attention to concepts that she already discussed in her comments.” (We will refine criteria and thresholds after usability evaluations with reviewers.)

The feedback from the *Reviewer Aide* depends on the input from the *Concept Classifier* and *CBR Matcher*. While the criteria outlined above will moderate the quality of the feedback, the feedback might still be inaccurate, untimely, or irrelevant. Ultimately, the reviewer has to assess the *Reviewer Aide*’s feedback, and choose to incorporate it (or not) in her own feedback to the author. Thus, the quality of the feedback the author sees should improve even if the reviewer disagrees with the *Reviewer Aide*: first, the reviewer will filter low-quality feedback, and second, the reviewer will modify mediocre feedback for the author’s benefit. Since even in rejecting the *Reviewer Aide*’s comments, the reviewer has been prompted to consider their relevance, we can use rejected or reviewer-modified feedback to evaluate and improve the system’s performance. We can maximize this effect by having the *Reviewer Aide* present different feedback to different reviewers.

While we have high hopes for the system proposed here, we will also measure its contributions empirically. We will compare three conditions: traditional classroom teaching vs. SWoRD augmented with the domain-specific moral reasoning criterion vs. SWoRD augmented with the new criterion as well as with the *Reviewer Aide*, *Concept Classifier*, and *CBR Matcher*. To measure student learning, we will compare the capstone exercises described above across the conditions, i.e., the authentic task in this context. The measures for this task must include both traditional holistic grading and the Assessment Instrument, so as not to favor any condition

with a tailored learning measure. Our use of SWoRD with and without the *Reviewer Aide* facilitates additional comparisons across the two conditions: the value of the *Reviewer Aide* will be apparent first, if student authors give higher ratings to those reviewers who use the *Aide*, and second, if the student essays improve more between drafts when reviewers use the *Aide*. The development of the *Concept Classifier* and *CBR Matcher* will prompt another set of evaluations. As both are machine learning techniques, the appropriate comparisons will be on measures like precision and recall. Furthermore, data on the accuracy of these components will be necessary in tuning the *Reviewer Aide*'s internal thresholds for presenting feedback to the reviewers.

RELATED WORK

Our focus on objectively measuring whether students learn moral reasoning skills provides ethics pedagogy with new empirical methods. Significantly, our system does not require a special case representation, it can handle never-before-seen cases within the domain of bioengineering ethics, and it orchestrates a useful collaboration through a reciprocal student authoring and reviewing. Thus, the system is likely to engage students more actively in ethical reasoning over a wider range of cases than "textbook-on-computer" resources like (Madsen; , "Online Ethics Center for Engineering and Science", 2006). Software like (Andersen et al., 1996; Searing, 2000) and web-based systems like (Goldin et al., 2001; Keefer & Ashley, 2001; McLaren, 2003; McLaren & Ashley, 1999; Robbins, 2005) support interactive case analysis, but unlike our approach, they lack instructional feedback and opportunity for collaborative learning.

Our focus on detecting LDA of key concepts complements research in automated essay scoring (AES) on higher-level features that indirectly relate to the quality of a written analysis and that improve perceived validity of the scoring model. In detecting instances of defining and applying, our *Concept Classifier* would deal directly with student texts as in (Landauer et al., 2003a, 2003b; Larkey, 1998), but does so by finding "proxes" or stand-ins for good case analyses as in (Burststein et al., 2001; Burststein et al., 2003; Page, 1966, 2003).

We seek to advance ITSs for writing by applying an ITS in a domain where it is the norm to analyze ill-defined problems in natural language. ITS that work with essays include Select-a-Kibitzer (Wiemer-Hastings & Graesser, 2000), Summary Street (Steinhart, 2001), AutoTutor (Graesser et al., 2000), Apex (Lemaire & Dessus, 2001), and Criterion (Higgins et al., 2004). These systems not only evaluate student essays on the fly, but they also provide feedback and encourage students to correct and rewrite their essays and resubmit them for new feedback. So far, however, ITS for writing can detect only fairly general features. For instance, Criterion and e-rater learn to detect 'discourse segments' like thesis, main idea, supporting idea, and conclusion, but they work on essays that have a rigid structure (an introduction, three supporting paragraphs, and a conclusion, about 300 words long). Alternatively, systems like the Intelligent Essay Assessor (Landauer et al., 2003a) that can address term-paper length essays detect general features, like coverage or absence of broad topics based on comparisons to past graded papers. An ability to detect finer-grained features such as examples of defined and applied concepts by the *Concept Classifier* would enable a writing ITS to give more detailed feedback.

CONCLUSION

In designing an ITS for engineering ethics, our objective is to extend the SWoRD approach to a technical domain where ill-defined problem solving and conceptual framing of cases are important. Our system will direct the feedback on HLMRS and on labeling, defining, and applying concepts through the student peer reviewers, who will decide whether to pass it along. This may help filter out any inapplicable feedback. Deciding which feedback to pass along is also a learning opportunity for the peer reviewers, who are students in the same class. Our approach offers a way to leverage the domain expertise embodied in a corpus of student papers from past offerings of the Bioengineering Ethics class.

The ill-defined properties of bioengineering ethics case analysis discussed here are not uncharacteristic of case analysis in other domains. The underlying task of case analysis requires the kind of argumentation that one finds in rhetoric, and, by extension, other humanities disciplines, and especially other problem-based learning scenarios. We believe that the Assessment Instrument can help engineering educators in engineering ethics domains beyond bioengineering (Goldin et al., 2006), and we hope that the combination of the Instrument with SWoRD and AI techniques will also generalize to those settings. While intelligent tutoring technology has been successful at well-defined tasks, its role in ill-defined tasks is less clear, and may require a more cautious approach, such as using AI in a support of human peer reviewing.

ACKNOWLEDGMENTS

This work has been supported in part by National Science Foundation Engineering and Computing Education grant #0203307. We thank our collaborators Christian Schunn and Janyce Wiebe for help writing an NSF proposal to continue this work.

REFERENCES

- Andersen, D., Cavalier, R., & Covey, P. (1996). *A Right to Die? The Dax Cowart Case*: Routledge.
- Ashley, K. D., & McLaren, B. M. (2001). *An AI Investigation of Citation's Epistemological Role*. Proceedings of Eighth International Conference on Artificial Intelligence & Law (ICAIL-01).
- Burstein, J., Marcu, D., Andreyev, S., et al. (2001). *Towards Automatic Classification of Discourse Elements in Essays*. Proceedings of Meeting of the Association for Computational Linguistics.
- Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. In *IEEE Intelligent Systems*.
- Cho, K., & Schunn, C. D. (2005). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers and Education, in press*.
- Cunningham, H., Maynard, D., Bontcheva, K., et al. (2002, July, 2002). *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*. Proceedings of 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia.
- Goldin, I. M., Ashley, K. D., & Pinkus, R. L. (2001). *Introducing PETE: Computer Support for Teaching Ethics*. Proceedings of International Conference on Artificial Intelligence & Law (ICAIL-2001), St. Louis, MO.
- Goldin, I. M., Ashley, K. D., & Pinkus, R. L. (2006). *Assessing Case Analyses in Bioengineering Ethics Education: Reliability and Training*. Proceedings of International Conference on Engineering Education, San Juan, Puerto Rico.
- Goldin, I. M., Pinkus, R. L., & Ashley, K. D. (in preparation). Sensitivity and Reliability of an Instrument for Assessing Case Analyses in Bioengineering Ethics Education.
- Graesser, A., Wiemer-Hastings, P., Wiemer-Hastings, K., et al. (2000). Using Latent Semantic Analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments, 8*, 149-69.
- Harris, C. E., Jr., Pritchard, M. S., & Rabins, M. J. (2000). *Engineering Ethics: Concepts and Cases* (2nd ed.). Belmont, CA: Wadsworth.
- Higgins, D., Burstein, J. C., Marcu, D., et al. (2004). Evaluating Multiple Aspects of Coherence in Student Essays. In *Proceedings of the Annual Meeting of HLT/NAACL*.
- Keefer, M. W., & Ashley, K. D. (2001). Case-based Approaches to Professional Ethics: a systematic comparison of students' and ethicists' moral reasoning. *Journal of Moral Education, 30*(4), 377-98.
- Landauer, T. K., Laham, D., & Foltz, P. W. (2003a). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. a. B. Shermis, Jill (Ed.), *Automated Essay Scoring* (pp. 87-112).
- Landauer, T. K., Laham, D., & Foltz, P. W. (2003b). Automatic essay assessment. *Assessment in Education, 10*(3).
- Larkey, L. S. (1998). *Automatic Essay Grading Using Text Categorization Techniques*. Proceedings of 21st Int'l Conference on Research and Development in Information Retrieval (SIGIR-1998), Melbourne.
- Lemaire, B., & Dessus, P. (2001). A System to Assess the Semantic Content of Student Essays. *Journal of Educational Computing Research, 24*(3), 305-20.
- Madsen, P. Ethical Judgments in Professional Life. Retrieved April 8, 2006, from <http://www.andrew.cmu.edu/course/80-241/>, login: guest, password: guest
- Martin, T., Rayne, K., Kemp, N. J., et al. (2005). Teaching Adaptive Expertise in Biomedical Engineering Ethics. *Science and Engineering Ethics, 11*, 257-76.
- McCallum, A. K., & Nigam, K. (1998). *A comparison of event models for naive Bayes text classification*. Proceedings of 1st AAAI workshop on learning for text categorization, Madison, WI.
- McLaren, B. M. (2003). Extensionally Defining Principles and Cases in Ethics: an AI Model. *Artificial Intelligence Journal, 150*, 145-81.
- McLaren, B. M., & Ashley, K. D. (1999). *Case Representation, Acquisition, and Retrieval in SIROCCO*. Proceedings of Third International Conference on Case-Based Reasoning, Munich, Germany.
- Online Ethics Center for Engineering and Science. (2006). Retrieved April 8, 2006, from <http://onlineethics.org/>
- Page, E. B. (1966). *The imminence of grading essays by computer*. Proceedings of Phi Delta Kappan.
- Page, E. B. (2003). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 62*(2), 243-54.
- Pinkus, R. L., Gloeckner, C., & Fortunato, A. (in preparation). Cognitive Science Meets Applied Ethics: Lessons Learned for Teaching.
- Robbins, R. (2005). The Ethical Assistant.
- Searing, D. R. (2000). Ethos System: Taknosys Software Corporation.
- Steinhart, D. J. (2001). *Summary Street: An Intelligent Tutoring System for Improving Student Writing through the use of Latent Semantic Analysis*. University of Colorado, Boulder, Colorado.
- Wiemer-Hastings, P., & Graesser, A. (2000). Select-a-Kibitzer: A Computer Tool that Gives Meaningful Feedback on Student Compositions. *Interactive Learning Environments, 8*(2), 149-69.