

Assessing Case Analyses in Bioengineering Ethics Education: Reliability and Training

Ilya M. Goldin¹, Kevin D. Ashley², Rosa L. Pinkus³

Abstract - We describe a novel Assessment Instrument in the domain of bioengineering ethics which can be applied to a wide variety of students' ethics case analyses. It is sensitive to knowledge gain and it can be applied reliably. In a new study, we reevaluate its reliability and find that coder agreement is higher on concepts that are especially relevant to the case at hand. Additionally, we describe how we train coders, and how we use Open Source software to compile and examine coder annotations.

Index Terms - ethics assessment, case analysis, reliability, student-authored cases

INTRODUCTION

Engineering ethics is an area of interest for professionals and for policy makers. The Accreditation Board for Engineering and Technology (ABET) requires that “by the time of their graduation [students are expected to have] an understanding of professional and ethical responsibility.” [1] The National Science Foundation has repeatedly solicited proposals under the program “Ethics Education in Science and Engineering.”

While educators in any discipline require valid, reliable, and objective tools for assessment, assessment in engineering ethics, as in any applied ethics domain, is problematic, as ethics is “ill-defined.” In an ill-defined domain, problems rarely have a definitive answer as the problem solver needs to frame the problem prior to analyzing it. As such, students may produce a wide range of acceptable responses to a given case; thus the challenge for assessment.

We propose a novel Assessment Instrument that can accommodate a variety of ways students analyze a given case. First, we describe the challenges of assessment in engineering ethics, focusing on a particular class in Bioengineering Ethics, which shaped our Instrument. We then evaluate validity and reliability of the Instrument. We conclude by describing what training is necessary to apply it.

ADDRESSING THE DIFFICULTIES OF ASSESSMENT

In professional ethics, except for paradigm cases, problems rarely have a definitive answer, and the problem solver often needs to define the problem better through *framing*. That is, she must add constraints and these in turn affect how ethics principles or professional ethics codes apply to the problem at hand. Mapping the principles to the situation and weighing the effects of alternative actions or creative compromises in order

to resolve conflicts are skills that must be learned. A given problem, as posed, can become a different problem through framing. The correctness of a particular resolution to an ethics case depends on how one frames the case. [2]

One's methodological approach to the domain also affects assessment. For example, approaches based on theory or principles, which are “top down,” may facilitate assessment because both suggest a “ready-made” frame. Assessing whether or not a student uses the principles in a theory should be comparatively straightforward. For example, an analysis that uses Utilitarianism correctly must include the notion of promoting the greatest good for the greatest number. All other qualifiers should also conform to this. The “Four Principles Approach” [3] shifts the analysis from abstract theories closer to the facts of the case. The principles complicate assessment however, because the student needs not only to invoke the principles but to prioritize them. Casuistry [4], i.e., reasoning with cases, does help map the principles to the problem, but adds yet another level of complexity. Given that this is a “bottom-up” approach to analysis, the student has to assess morally relevant facts and concepts, and select appropriate analogous cases for comparison. Here, “framing” the case and the “ill-defined” nature of ethics become paramount. [5, 6]

As one can imagine, creating an assessment instrument to evaluate if a student has correctly learned to apply a method of moral reasoning presents specific challenges. Given the wide range of ways one can teach applied ethics and even define ethics itself, we began by identifying 1) what is important to teach, 2) to assess, and 3) how to assess it. As the central goal in the course was to teach a set of analytical skills and moral problem solving methods, we focused on assessing students' mastery of these. Previous work has focused on instruments that assess a student's general level of moral development, reasoning or judgment [7-11] and those that survey moral values (reviews in [12, 13]).

Specifically, we sought to calibrate the Instrument so that it would reflect learning within a semester-long class: a required graduate and undergraduate course, Bioengineering Ethics, taught by author Pinkus. The course is “designed to instruct...bioengineering students to identify the values embedded in their practice. It stresses the acquisition of methods of moral problem solving to be used to identify, analyze and resolve dilemmas posed when these values conflict.” [14] The methods match the course textbook. [15]

As the course emphasizes the skills of ethics case analysis, the capstone assignment is an extended analysis of a

¹ Ilya M. Goldin, Intelligent Systems Program, Learning Research and Development Center, University of Pittsburgh, goldin@pitt.edu

² Kevin D. Ashley, Intelligent Systems Program, Learning Research and Development Center, University of Pittsburgh, ashley@pitt.edu

³ Rosa L. Pinkus, Neurosurgery/Medicine, School of Medicine, University of Pittsburgh, pinkus@pitt.edu

complex case. In an effort to stimulate students to frame problems, this assignment asks them to devise and analyze their own engineering ethics dilemmas, first in a class presentation and then in writing.

This approach, where students create cases close to their professional expertise and interests, has been shown to be a positive factor in student learning. [14] But this pedagogical technique constitutes yet another assessment challenge. In previous work [16, 17] it has been proposed that a student analysis of a case be compared to experts' "gold standard" analyses. However, the variety of cases that students create themselves combined with the variety of ways in which each case can be framed means that it is often impossible to stipulate a gold standard.

In lieu of a gold standard, our Assessment Instrument [14, 18] looks for evidence that a student has grasped analytical tools we call "higher-level moral reasoning skills" (HLMRS), which, we induced, were representative of the learning goals of the Bioengineering Ethics course. The five measures are whether a student:

- Employs professional engineering knowledge to frame issues.
- Views the problem from multiple levels.
- Flexibly moves among multiple levels.
- Identifies analogous cases and articulates ways the cases were analogous.
- Employs a method of moral reasoning in conducting the analysis.

The Instrument enables a "coder" to indicate in free-form text whether a student's analysis demonstrates skills a through d. With respect to the fifth HLMRS, however, "Employs a method of moral reasoning in conducting the ethical analysis," we have attempted to create a standardized method by which one can code for evidence of the skill.

We operationalized this HLMRS in terms of a list of over 40 pedagogically important concepts associated with methods of moral reasoning that has been compiled from case books [6, 15, 19] and capstone assignments from several semesters of the Bioengineering Ethics course. The list includes concepts like common morality, utilitarianism, safety, and conflict of interest. Although this list is not exhaustive, it is as comprehensive as possible for the domain of bioengineering ethics so as to enhance the content validity of the Instrument.

A student *de facto* frames a case when she writes it and, in analyzing it, she introduces a set of moral reasoning concepts in a particular way. The Instrument recognizes that a student may *label*, *define*, or *apply* a concept in a case analysis. A concept is said to be *labeled* as such if the term for the concept is present; *defined* if a dictionary-like definition of the concept is present; and *applied* if the concept has been brought to bear appropriately in the context of the particular case. The coders then annotate a case analysis for each concept that they believe the student invokes, such as in this excerpt from an actual student case analysis, where utilitarianism is the concept of interest. We use the tags <lab> and </lab> for examples of labeling, and <app> for applying: "<app>From a <lab>utilitarian</lab> standpoint, if the

transplant can save Mr. Creighton from certain death, then it should be done.</app>" Unfortunately, the student does not define utilitarianism here. Coders using the Instrument can find a definition of this and other concepts in the provided Glossary. (Additionally, coders receive written instructions.) Here is the Glossary entry for utilitarianism:

Utilitarianism: "Utilitarian thinking favors bringing about the greatest total amount of good that we can."

[15] There are three ways to formulate this total good: the cost/benefit approach, the act utilitarian approach and the rule utilitarian approach. Each approach must be "universal" or put to the test of promoting good for all, not just a single person. In calculating the greatest good, however, there is a danger of harming the individual. (See act utilitarianism, rule utilitarianism, risk/benefit analysis, cost/benefit analysis, universalizability.)

Example: In the Tuskegee Study, researchers decided to lie to the prospective subjects and told them that if they signed up for the study, they would get up to date treatment for their "bad blood." Without this lie, the men would not have participated in the study. The researchers justified their actions (wrongly) by looking to the long-term benefits of the study rather than the short-term unethical conduct. In the end, this utilitarian approach tainted the entire study and some now question whether using the results is ethical.

Throughout, the Glossary cross-references related concepts and cites authoritative sources [3, 15], creating a network of concepts, definitions, and examples that serves as an annotating aid. It should be especially useful where a student applies a concept without explicitly labeling it, or if the coder suspects an error in how the student defined or applied a concept.

The assessment instrument accommodates the variety of cases created by students, and the variety of ways students frame them. Given this diversity, not all of the concepts taught in the course and listed in the instrument will be relevant to every case. Which concepts are relevant depends on how the student frames the problem. Since the instrument comprehensively lists concepts covered in the course, the concepts relevant to a particular case will generally be a subset of those listed. If the coder finds that the student labels, defines, or applies a concept that is missing from the instrument, he may write it in the spaces provided.

We are aware of one other instrument for annotating case analyses of cases authored by the students themselves. [20] That instrument asks coders to tally principles and theories invoked by the students, where the master list is "utilitarianism, consequentialism, deontology, virtue, autonomy, justice, beneficence and nonmaleficence." Our Instrument in effect subsumes that approach and expands it. The additional concepts and HLMRS in our Instrument make it useful not only for teachers focused on theory or principles, but also for those who use a case-based approach.

EMPIRICAL EVALUATION OF THE ASSESSMENT INSTRUMENT

In [18], we demonstrated validity and reliability of the Instrument. Below, we summarize those validity findings, and then take a deeper look at the dataset.

Validity describes how well the Instrument measures that which is important, and intended, to observe; we showed that our Instrument is a valid measure, because it is sensitive to relevant skills of students in performing ethics case analysis. We investigated whether our instrument is sensitive to student learning gains during a semester-long course. We compared student skills at ethics case analysis, as measured by our instrument, at pre- and posttest times. We argue that if one assumes that students actually learn in the class, then their learning can only be reflected in posttest scores if our instrument is sensitive to changes in learning.

A graduate bioengineering ethics class (excluding two dropouts) participated in the study. On the first day of the semester, the students analyzed an ethics case, *The Artificial Heart*. For the posttest, after their final exam, the students analyzed one of two cases, *The Price is Right or Trees* (Table 1). Students did not choose which case to analyze, and due to time constraints, their responses were all one or two pages long. The resulting dataset was annotated by three coders: M, K, and A. M and K were independent coders who had been blinded to the fact that we were measuring validity, and to the pretest vs. posttest experimental design. Coder A was also independent; although she was not blinded, we know of no bias on her part. As the coders applied the Instrument, they did not connect their LDA annotations to specific utterances; as a result, we look at what they recorded for entire papers.

Briefly, the coders found evidence of higher-level moral reasoning skills significantly more often at posttest than at pretest, and also found evidence of concepts significantly more at posttest than at pretest. Further, at pretest, coders found that students mostly apply concepts, while on the posttest they label and define them in addition to applying.

Dataset	Students	"Target" Concepts
pre ("Artificial Heart")	13	25
post case 1 ("The Price is Right")	6	19
post case 2 ("Trees")	7	15

TABLE 1: "TARGET" CONCEPTS WERE POSSIBLY RELEVANT TO CASES IN THE SENSITIVITY STUDY (THE INSTRUMENT LISTS 41 CONCEPTS).

Because this experiment used well-studied cases, in addition to the full set of concepts, we also looked at a subset of "target" concepts that could be possibly relevant to each of the cases (Table 1). The "target" concepts were either considered relevant by an expert (she had an MA in Ethics, and was a teaching assistant for several courses that covered these cases), or they were annotated as occurring in actual student analyses.

We found that the focus on target concepts magnified the students' tendency to invoke more concepts on the posttest than on the pretest, and to label and define concepts on the posttest in addition to applying them. While this calculation is confounded by the fact that the number of target concepts at pretest is higher than at posttest, the confound also implies a higher probability that a student could invoke a concept on the

pretest than on the posttest, which emphasizes the fact that students fail to invoke concepts at pretest.

The study shows that the Instrument is sensitive to student learning over the course of the semester, and is thus a valid measure of learning; it captures at least one important aspect of the quality of case analyses. In another result of the study, although all coders found a significant difference in student use of HLMRS and concepts between pretest and posttest, there were also significant differences among the coders themselves. In a new study, we evaluate their agreement.

I. Inter-rater Reliability: Methodology

Reliability means that the results obtained from applying the Instrument are not a one-time occurrence, but a reproducible event. To evaluate reliability, we look at whether independent coders agree when they apply our Instrument.

Following [18], we calculate one agreement score for the four HLMRS, and another for the concepts operationalizing the HLMRS "Employed a method of moral reasoning".

We treat the HLMRS questions as binary variables. Each question can then be viewed as a two-by-two contingency table, a standard data representation for calculating agreement. We calculate agreement as Cohen's κ (Kappa). [21-25]

We calculate agreement about concepts with and without the assumption that labeling, defining, and applying are independent actions. The assumption of independence shows whether coders agree more about, say, labeling than defining. However, the assumption is sometimes inappropriate. If we do not assume that labeling, defining, and applying a concept are independent actions, then there are eight possible "LDA combinations." The independence assumption can be in error, because not all LDA combinations are equally probable. For example, defining a concept without labeling it is unlikely.

The independence issue necessitates further technical considerations. With the assumption, we can represent data on labeling, defining, and applying with two-by-two contingency tables, and compute agreement as κ . When we relax the assumption, with eight LDA combinations per coder, we have an eight-by-eight contingency table (64 comparisons between two coders). This leads to a sparse data distribution and tends to decrease κ . "The magnitude of κ is influenced by...the number of categories in the measurement scale... The larger the number of scale categories, the greater the potential for disagreement, with the result that unweighted κ will be lower with many categories than with few." [25]

The κ metric computes the probability of actual (observed) agreement corrected by the expected probability of agreement; κ ranges from 1 (perfect agreement) to -1 ("perfect" disagreement). While κ is widely recommended for reporting agreement, it is also susceptible to the problems of bias and prevalence, which can obscure the true levels of agreement in a dataset.

Bias means that κ can be artificially inflated because of differences in the coders' beliefs about the actual distribution of LDA combinations in our dataset. Prevalence means that κ can be artificially depressed because of imbalance in the "true", coder-independent underlying distribution of LDA. In

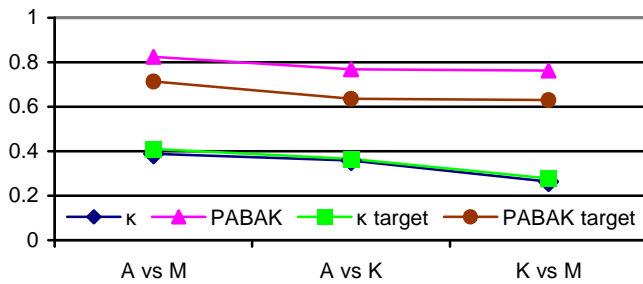


Figure 2: κ and PABAK on target and unfiltered concepts

fact, we know that the actual distribution is unbalanced: first, there must always be many more instances of the absence of labels, definitions, or applications than of their presence (recall that not all of the concepts taught in the course and listed in the assessment instrument will be relevant to every case); second, it is to be expected that coders agree more on these absences. Following [23-25], we computed bias-adjusted κ (BAK) and prevalence-adjusted bias-adjusted κ (PABAK).

Since BAK and PABAK complement κ , it is customary to interpret all three on the same scale. We found that κ and BAK diverge little for our data, which means that there is little difference in coder bias; thus, we omit BAK. However, κ and PABAK do differ, which implies that prevalence of some LDA combinations over others depresses our κ values. We therefore focus on PABAK, and report κ for completeness.

As a guide to interpreting the metrics, [25] cite an interpretation of κ from [26]: $\kappa \leq 0$ is poor, .01-.20 is slight, .21-.40 is fair, .41-.60 is moderate, .61-.80 is substantial, and .81-1 is almost perfect. We are aware of other interpretations; “the choice of such benchmarks, however, is inevitably arbitrary, and the effects of prevalence and bias on κ must be considered when judging its magnitude.” [25]

II. Inter-rater Reliability: Results

We begin by treating LDA as not independent (i.e., we use the eight-by-eight contingency table for agreement calculations). We see substantial (Figure 2, K vs. M PABAK=0.763) to excellent (A vs. M PABAK=0.824) agreement on labeling, defining, and applying concepts. PABAK is higher than uncorrected κ due to the prevalence of some annotations, i.e., κ obscures the actual level of agreement of our coders.

Arguably, if we restrict the inter-rater reliability calculation to the “target” concepts, the agreement scores would be more meaningful, because they would reflect agreement on what is really germane to each case. (For expository purposes, let us now call the full sensitivity dataset “unfiltered”.) On target concepts, we find substantial agreement (from A vs. K PABAK=0.630 to A vs. M PABAK=0.714) on labeling, defining, and applying concepts. PABAK is still higher than uncorrected κ due to the prevalence of some annotations. Compared to the unfiltered sensitivity dataset, even though PABAK scores have dropped, κ scores have increased. The decrease in PABAK scores is caused by the removal of a large number of *none* instances, which reduces the prevalence imbalance, but also reduces the number of instances of agreement. The κ increase at the same

San Juan, PR

time is due to removal of instances of disagreement (i.e., off-diagonal instances); this shows that untrained coders agree more on concepts that are relevant than on all concepts.

We now make the assumption that labeling, defining and applying are independent of each other so that we can understand the dataset in greater detail. Looking at all concepts, we find substantial to excellent agreement (Figure 1), from PABAK=0.787 for K vs. M for applying concepts to PABAK=0.982 (A vs. M for defining). PABAK is much higher than uncorrected κ .

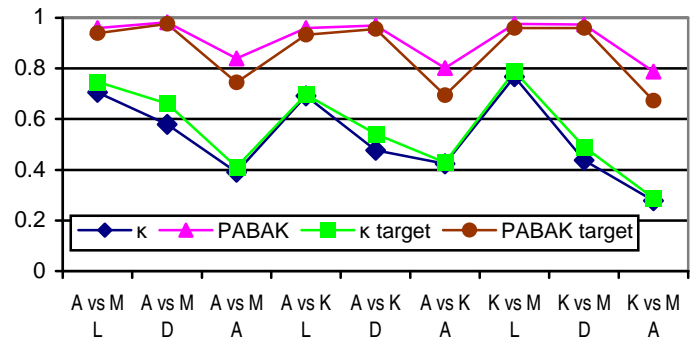


Figure 1: κ and PABAK assuming LDA independence

We can now also look more closely at the drop in PABAK and increase in κ on the target concepts. The drop in PABAK holds across all three ways of invoking a concept, but is most prominent for applying (mean change over all coder pairs is -0.106). Similarly, κ also increases for all three, but most dramatically for defining (mean change over the three coder pairs is 0.066). These results suggest that when we focus on “target” concepts, there is a decrease in agreement about the *lack* of concepts (especially concept applications), and there is a complementary increase in agreement about the *presence* of concepts (especially concept definitions).

When we consider the work of coders on the sensitivity dataset, we see that they agree the most on whether a student has labeled a concept, less so on defining, and the least on whether a concept has been applied. This result corresponds to our intuitions about the relative difficulty of annotating labels, definitions, and applications. Interestingly, PABAK scores on labeling decrease on target concepts. Clearly, the task of detecting whether a concept has been labeled as such is simple—one need only look for the right word. It is apparent from this PABAK decrease that the task still elicits poor coder performance, perhaps because it is easy to miss a concept name in the text, or perhaps because coders are overwhelmed by looking for so many concepts at once. At the same time, an automated computer “coder” might detect labels well.

By contrast, the coders’ agreement on annotating the four HLMRS is low (Figure 3) despite the fact that in [18] all coders found significantly more evidence of HLMRS at posttest than at pretest. For the first time, we lack an effect due to prevalence (PABAK is lower than κ for two of three coder pairs), and strong differences in coder bias (BAK is much lower than κ for two of three coder pairs), but regardless, scores are at best fair. This is not surprising: the HLMRS are complex skills, and it takes an experienced domain expert to

July 23 – 28, 2006

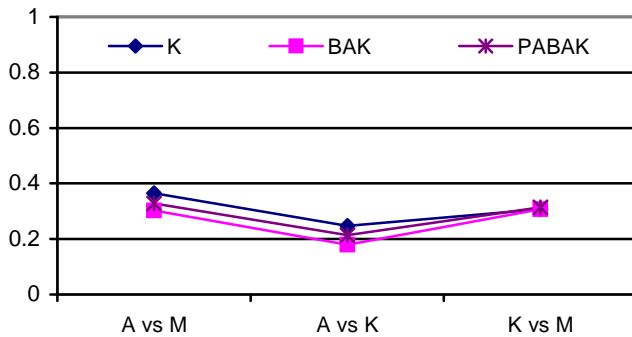


Figure 3: Average HLMRS agreement

be able to detect them in student essays. While our coders thought they found significantly more evidence of the HLMRS at posttest than at pretest, the HLMRS are not operationalized in sufficient objective detail for the coders to agree. These findings highlight the value of operationalizing the HLMRS “Employing a method of moral reasoning” via LDA. We believe that work like [27] will aid in operationalizing other HLMRS.

Results here support our findings in [18] that coders agree on whether students label, define and apply ethics concepts, but not on whether students employ higher-level moral reasoning skills. In [18], we measured agreement on the capstone assignments described above, in which students analyze at length the cases that they themselves create. Indeed, the capstones are the authentic context for the Instrument, the reason it was devised. (Capstones ought to be even more difficult to code than the essays in the sensitivity study, because this study used standardized cases, which had comparatively “pre-framed” ethical dilemmas, whereas in capstones, the students must do the hard work of framing, and the coders need to confirm whether the frame is well-constructed.) We conclude that even though the question of whether a student has labeled, defined or applied a concept is subject to the interpretation of coders, the LDA operationalization (coupled with coder training) makes it easier to approach consensus.

ANNOTATION AND TRAINING PROCEDURES

We have developed a training procedure for our coders. The objective is that trained coders be able to annotate independently of any “experts.” The procedure makes use of publicly available Open Source software that allows us to compile and examine coder annotations. We hope that the training procedure will aid others in adopting the Instrument in new engineering education contexts, such as professional ethics beyond bioengineering.

Both graduate and undergraduate students have performed coding for us. We consider it a prerequisite that the student have some experience in the relevant field of engineering ethics, such as having taken a semester-long Bioengineering Ethics class. Additional background knowledge in ethics and/or engineering is obviously helpful, as is experience with ethics case analysis. For example, one coder we trained

recently was a graduate engineering student with professional experience. He had taken six weeks of the Bioengineering Ethics course when we began training. To introduce him to case analysis and concepts of moral reasoning, we asked him to read and outline [15, Ch. 1-4], to read the Glossary, and to practice annotation on two term papers.

We begin by asking a new coder to annotate a case analysis using the paper form of the instrument. This form contains four open-ended questions about higher-level moral reasoning skills a) through d), and it operationalizes HLMRS e) via labeling, defining, and applying as discussed above.

The instrument is accompanied by written instructions about the logistics of coding, and the Glossary. The instructions distinguish among the label, definition, and application of a particular concept. Significantly, they alert the coder to the chance that a student “may apply a concept without explicitly labeling it” or that a definition or application of a concept may contain an error that indicates misunderstanding of the concept. The instructions urge the coder to refer to the Glossary on these occasions, and to use appropriate notation in the margins of the case analysis. Finally, the instructions provide brief explanations of HLMRS a) through d). For example, “*Identify different perspectives* means that the respondent has analyzed the case from different points of view.”

The first case analysis that the coder annotates should be representative of the best student performance. Ideally, it should have examples of several HLMRS, and it should have instances of comprehensive framing of moral reasoning concepts, including their labels, definitions, and applications. We believe that it is important to show the coder at this juncture what students (ought to) aim to achieve in their works. By implication, it becomes apparent when that achievement is lacking. To foster independence, we ask the coder to work through an entire case analysis as best she can.

Once one case analysis is annotated, the coder reviews her annotations together with an experienced coder, who points out any omissions and misclassifications, and clarifies misconceptions about the HLMRS and concepts. At the same time, new coders are particularly adept at pointing out weaknesses in the coding scheme in general, in the list of concepts in the Instrument, or in the definitions and examples in the Glossary. Consequently, when adapting the Instrument to a new context, it is very helpful to consider the observations of new coders. For example, one coder in training recently pointed out that it was unclear how he should code a negative example of a concept. The Instrument includes the moral reasoning concepts responsibility of bioengineer, and honesty. But the coder argued that a student’s paper analyzed instances of “irresponsibility” of bioengineer, and “dishonesty.” The ensuing discussion clarified the meaning of the concepts both for the coder and for the person “in charge” of the training.

Annotation with the paper version of the Instrument is convenient for the coder when the case analysis is in hard copy itself. However, annotation on computer can provide significant advantages. Chief among them is the ability to run reports against the annotated corpus of documents, such as to

retrieve all definitions or applications of any concept. For example, one may ask for all definitions of “common morality.” Just as easily, one may request tallies of the relevant annotations. If any documents are annotated by multiple coders, computer software can make it easy to compare their annotations and to calculate inter-rater reliability over an entire corpus.

Our annotation tool is an open-source software package called GATE. [28] Beyond supporting annotation, GATE provides useful Natural Language Processing functionality that supports data analysis. For example, one can use the built-in NLP technology to make queries like “what is the head noun in all definitions of the concept ‘informed consent’?”

The chief disadvantage of annotation with GATE is the relative difficulty of including free-form comments in ones annotations, i.e., the equivalent of scribbling in the margins. However, we feel that this kind of data gathering is more appropriate in the exploratory phase of developing an assessment instrument. Once development is complete, and one wishes to apply the instrument to a new corpus, free-form data are more difficult to organize than “closed” form data, and hence, more difficult to use.

We encourage adoption of the Instrument to engineering ethics domains beyond bioengineering. The HLMRS are abstract and characterize reasoning in moral dilemmas in general. The LDA operationalization should stand as well. Even the list of concepts may require little modification, as bioengineering ethics includes many conceptual issues from across the spectrum of engineering ethics. One may wish to add entries for any new concepts to the Glossary and to customize those examples in the Glossary that focus on bioengineering issues.

CONCLUSION

The Assessment Instrument is an innovative approach to assessment in the ill-defined domain of bioengineering ethics. The findings here support the conclusion [18] that the Instrument can be applied reliably. An effective coder training procedure and computer-aided annotation improve the usefulness of the Instrument.

REFERENCES

[1] Accreditation Board of Engineering and Technology, *Criteria for Accrediting Engineering Programs*. 2004: Baltimore, MD.

[2] Goldin, I.M., K.D. Ashley, and R.L. Pinkus. *Teaching Case Analysis through Framing: Prospects for an ITS in an ill-defined domain*. in *Workshop on Intelligent Tutoring Systems for Ill-Defined Domains, 8th International Conference on Intelligent Tutoring Systems*. 2006. Jhongli, Taiwan.

[3] Beauchamp, T.L. and J.F. Childress, *Principles of biomedical ethics*. 5 ed. 2001, New York, N.Y.: Oxford University Press. 454.

[4] Jonsen, A.R. and S.E. Toulmin, *The abuse of casuistry: a history of moral reasoning*. 1988, Berkeley: University of California Press. 420.

[5] Kuczewski, M.G., *Fragmentation and consensus: communitarian and casuist bioethics*. 1997, Washington, D.C.: Georgetown University Press. 177.

[6] Pinkus, R.L., *Engineering ethics: balancing cost, schedule, and risk--lessons learned from the space shuttle*. 1997, New York: Cambridge University Press. 379.

[7] Kohlberg, L., *The Philosophy of Moral Development. Essays on Moral Development*. Vol. 1. 1981, San Francisco: Harper & Row.

[8] Gibbs, J.C. and K. Widaman, *Social Intelligence: Measuring the Development of Sociomoral Reflection*. 1982, Prentice-Hall: New Jersey. p. 191-211.

[9] Rest, J.R., D. Narvaez, M.J. Bebeau, and S.J. Thoma, *Postconventional Moral Thinking: a Neo-Kohlbergian Approach*. 1999, New Jersey: Lawrence Erlbaum Associates.

[10] Lind, G., *Moral Regression in Medical Students and Their Learning Environment*. *Revista Brasileira de Educacao Médica*, 2000. 24(3): p. 24-33.

[11] Comunian, A.L. *Structure of the Padua Moral Judgment Scale: A Study of Young Adults in Seven Countries*. in *110th Annual Conference of the American Psychological Association*. 2002. Chicago, IL.

[12] Rudnicka, E. *A review of instruments for measuring moral reasoning/values*. in *10th International Conference on Industry, Engineering, and Management Systems*. 2004. Cocoa Beach, FL.

[13] Lynch, D.C., P.M. Surdyk, and A.R. Eiser, *Assessing professionalism: a review of the literature*. *Medical Teacher*, 2004. 26(4): p. 366-373.

[14] Pinkus, R.L., C. Gloeckner, and A. Fortunato, *Cognitive Science Meets Applied Ethics: Lessons Learned for Teaching*. in preparation.

[15] Harris, C.E., Jr., M.S. Pritchard, and M.J. Rabins, *Engineering Ethics: Concepts and Cases*. 2nd ed. 2000, Belmont, CA: Wadsworth.

[16] Hébert, P.C., E.M. Meslin, and E.V. Dunn, *Measuring the Ethical Sensitivity of Medical Students: A Study at the University of Toronto*. *Journal of Medical Ethics*, 1992. 18(3): p. 142-147.

[17] Savulescu, J., R. Crisp, K.W.M. Fulford, and T. Hope, *Evaluating ethics competence in medical education*. *Journal of Medical Ethics*, 1999. 25(5): p. 367-374.

[18] Goldin, I.M., R.L. Pinkus, and K.D. Ashley, *Assessing Case Analyses in Bioengineering Ethics Education*. in preparation.

[19] Pence, G.E., *Classic cases in medical ethics: accounts of cases that have shaped medical ethics, with philosophical, legal, and historical backgrounds*. 3rd ed. 2000, Boston: McGraw-Hill. 509.

[20] Shapiro, J. and R. Miller, *How medical students think about ethical issues*. *Acad Med*, 1994. 69(7): p. 591-3.

[21] Cohen, J., *A coefficient of agreement for nominal scales*. *Educational and Psychological Measurement*, 1960. 20: p. 37-46.

[22] Carletta, J., *Assessing agreement on classification tasks: the kappa statistic*. *Computational Linguistics*, 1996. 22(2): p. 249-254.

[23] Byrt, T., J. Bishop, and J.B. Carlin, *Bias, prevalence and kappa*. *J Clin Epidemiol*, 1993. 46(5): p. 423-9.

[24] Di Eugenio, B. and M. Glass, *The Kappa statistic: a second look*. *Computational Linguistics*, 2004. 30(1).

[25] Sim, J. and C.C. Wright, *The kappa statistic in reliability studies: use, interpretation, and sample size requirements*. *Phys Ther*, 2005. 85(3): p. 257-68.

[26] Landis, J.R. and G.G. Koch, *The measurement of observer agreement for categorical data*. *Biometrics*, 1977. 33(1): p. 159-74.

[27] Martin, T., K. Rayne, N.J. Kemp, J. Hart, and K. Diller, *Teaching Adaptive Expertise in Biomedical Engineering Ethics*. *Science and Engineering Ethics*, 2005. 11: p. 257-276.

[28] Cunningham, H., D. Maynard, K. Bontcheva, and V. Tablan. *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications in 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. 2002. Philadelphia.