Supporting Students in the Analysis of Case Studies for Professional Ethics Education

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Doctor of Philosophy
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Mayya Sharipova

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9

Abstract

Intelligent tutoring systems and computer-supported collaborative environments have been designed to enhance human learning in various domains. While a number of solid techniques have been developed in the Artificial Intelligence in Education (AIED) field to foster human learning in fundamental science domains, there is still a lack of evidence about how to support learning in so-called ill-defined domains that are characterized by the absence of formal domain theories, uncertainty about best solution strategies and teaching practices, and learners' answers represented through text and argumentation.

This dissertation investigates how to support students' learning in the ill-defined domain of professional ethics through a computer-based learning system. More specifically, it examines how to support students in the analysis of case studies, which is a common pedagogical practice in the ethics domain.

This dissertation describes our design considerations and a resulting system called Umka. In Umka learners analyze case studies individually and collaboratively that pose some ethical or professional dilemmas. Umka provides various types of support to learners in the analysis task. In the individual analysis it provides various kinds of feedback to arguments of learners based on predefined system knowledge. In the collaborative analysis Umka fosters learners' interactions and self-reflection through system suggestions and a specifically designed visualization. The system suggestions offer learners the chance to consider certain helpful arguments of their peers, or to interact with certain helpful peers. The visualization highlights similarities and differences between the learners' positions, and illustrates the learners' level of acceptance of each other's positions.

This dissertation reports on a series of experiments in which we evaluated the effectiveness of Umka's support features, and suggests several research contributions.

Through this work, it is shown that despite the ill-definedness of the ethics domain, and the consequent complications of text processing and domain modelling, it is possible to build effective tutoring systems for supporting students' learning in this domain. Moreover, the techniques developed through this research for the ethics domain can be readily expanded to other ill-defined domains, where argument, qualitative analysis, metacognition and interaction over case studies are key pedagogical practices.

Acknowledgements

The dissertation would not exist without the help and support of others. First and foremost, I would like to thank my supervisor Prof. Gordon McCalla for inspirational, financial, and intellectual support related to the academics, but also support much beyond academics: related to a new for me life in Canada, and life after graduating university. I am grateful to Gord for opening up a new research area for me — intelligent tutoring systems, and for giving me freedom and opportunities to explore this area beginning with choosing my own research topic, and letting me to decide which direction I will take this topic. I am thankful to him for giving me a chance to teach his class, the first class I taught in Canada. Gord once commented that educating someone is to make him or her progressively more independent, and I believe Gord was successful in this.

I would like to thank Dr. Julita Vassileva for being beyond an outstanding committee member, for giving me an opportunity to run two experiments in her class, for her expert prospective on the research topic, for sharing her own experiences on conducting research, and for her inspirational talks on the role of women in computer science.

I am thankful to the committee member Dr. Chris Dutchyn for his thoughtful remarks on my research, and for the best computer science lectures I have ever had chance to attend.

I would like to thank the committee member Dr. Jim Greer for his extensive knowledge and expertise in the field which he brought in judging my thesis.

I am thankful to the committee member Dr. Emer O'Hagan for her unique, outside of the computer science field, perspective on the research topic, that provoked my deeper thinking.

I would like to thank an external examiner Dr. Roger Nkambou for bringing his expertise in ill-defined domains for judging my thesis.

I'm thankful to other faculty members for their involvement in the experimental part of my research, to Dr. Diane Martz for giving me a chance to run an experiment in her class at the University of Saskatchewan, to Bev Demmert for an experiment in her class at the SIAST, and to Dr. Susan Elliott Sim for an experiment in her class at the University of Toronto.

I am also grateful to all my fellow students who contributed to this work in various ways. Special thanks to Carrie Demmans Epp for helping me to organize a experiment at the University of Toronto, to Terry Peckham for helping to organize an experiment in the SIAST, to Jennifer Seaton for helping to organize an experiment at the University of Saskatchewan, to Elena Dergacheva for helping to evaluate experimental results, and all other SIAST, University of Saskatchewan and University of Toronto students who participated in the research experiments.

Lastly, I would like to express thanks to my family, my fiance Mohammad Hashemian for his tremendous support and encouragement throughout all these years, and my mother Nina Sharipova and brother Timur Sharipov for their constant unconditional love and support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Since the proliferation of computers, the impact of technology on education is increasingly growing. Extensive research to enhance human learning with the help of computers resulted in the formation of a new research field — Artificial Intelligence in Education (AIED). Breakthroughs in this research field triggered the development and adoption of numerous Intelligent Tutoring Systems (ITSs) and Computer-Supported Collaborative Environments that are helping hundreds of thousands of learners of various age.

A limitation of AIED research and techniques is their predominant concentration on supporting learning for well-defined problems in fundamental science domains. Many so-called ill-defined domains and problems have received relatively little attention in the AIED community, and there is a lack of evidence how to support learning in these domains through technology. This situation is not acceptable considering that the skills and the training needed to enhance these skills is different in ill-defined domains than in well-defined domains. Ill-defined domains require learners to gain "soft" skills such as finding divergent creative solutions, evaluating them, expressing personal opinions and beliefs, making judgements and defending them, etc. [39]. These skills help learners to tackle many real-life and real-world problems that are ill-defined by nature, and learning these skills constitutes an important part of education.

Recently the AIED/ITS community has realized the need to expand its arena to ill-defined domains. The community grew interested in computer techniques for teaching ill-defined domains, initiating a series of workshops for ill-defined domains[1]. Several techniques for building ITSs for ill-defined domains have been proposed in these workshops, but there are still many open research questions to be answered and challenges to be addressed. According to the workshops' organizers some of these challenges include: defining computational models for open-ended exploration with appropriate meta-cognitive scaffolding; constraining students to productive behaviors in otherwise underspecified domains; providing effective feedback when the problem-solving model is underspecified; structuring of learning experiences in the absence of a clear problem, strategy, and answer; developing user models that accommodate the uncertainty, dynamicity, and multiple perspectives of ill-defined domains; designing interfaces that can guide learners to productive interactions without artificially constraining their work [55].

One of the ill-defined domains that has received particularly little attention in AIED research is professional ethics. And this comes as no surprise considering that there is a little consensus how to effectively

[1]Intelligent Tutoring Systems for Ill-Defined Domains Workshops 2006, 2007, 2008, 2010.

support learning in professional ethics [2, 40], and even less consensus and understanding as to how to support it with technology. This research addresses this gap, and investigates techniques for designing an intelligent tutoring system for professional ethics domain.

1.1 Problem

The question we address in this dissertation is how computer systems can facilitate professional ethics learning, and specifically, how computer systems can support students in the analysis of case studies for professional ethics education.

The past two decades witnessed a significant increase in professional ethics courses [40]. A popular teaching pedagogy in those courses is the analysis of case studies that represent key professional issues and concepts. Analysis of case studies is a common teaching practice in many other domains, and outside of the ethics domain has been well practised in history, philosophy, business, public policy, etc. In the professional ethics domain case studies are represented as fictional or real-life situations that illustrate conflicts and dilemmas that can arise in professional life. Examples of case studies could be professional situations where a main protagonist experiences conflict of interest, or is requested by his employer to do something not completely legal such as installation of unlicensed software, or is in the position of having to trade one safety concern for another, etc. Learners are asked to analyse a case study by identifying professional issues involved in the case study, proposing a resolution for the case study dilemma that best addresses these issues, and by justifying the proposed resolution with argumentation.

Analysis of a case study is a difficult task: it involves problem investigation, looking for relevant information, finding and weighing various alternative resolutions, considering their consequences, providing argumentation for a chosen resolution. Many students, especially novice learners, provide a naive resolution to a given case and simple justification of their resolution, they don't consider and assess alternatives, their justification is largely based on general moral principles or simple consequences rather than on role-specific obligations [41]. Common ways to assist these kind of learners are through peer interactions and training of metacognitive strategies that will develop the learners' ethical reasoning and broaden their perspectives [41, 68, 69]. Well-designed computer-learning environments have a potential to play a crucial rule in this assistance: not only can they help students to structure their case analysis, but they can also provide personalized suggestions, foster the development of students' metacognitive skills, and organize collaboration among students that will challenge and develop their reasoning.

Analysis of realistic ethical problems is a classic example of an ill-defined problem. In ill-defined problems there is no complete formal domain theory [20] or it is ambiguous and subject to debate; solving ill-defined problems can require metacognitive skills [53]; there is no single solution to a problem, and solutions cannot be automatically verified [47]; there is no agreement on solution strategies, and hence there is no consensus on instructional strategies. New kind of skills are needed to solve ill-defined problems, and new kinds of tutoring systems are needed to help learners acquire these skills.

The special characteristics of ill-defined problems make them difficult for computer processing. Therefore, building tutoring systems for ill-defined problems calls for the adoption of alternative non-standard techniques. The process of building any ITS requires addressing four components of the classic ITS architecture: domain model, student model, tutoring model, and the user interface [10]. In the context of developing an ITS for the professional ethics domain important questions are how to organize these ITS components considering the ill-defined nature of the case analysis task:

- Domain model: What constitutes the domain model for professional ethics and for a particular case study? What is the domain knowledge and what skill set needs to be mastered by students? How can this domain model be represented in a computer system?
- Student model: What are the measures of students' performance? What is deemed as successful performance? And if it is impossible for a tutoring system to provide definite evaluation of students' answers, what could be the proxies for this evaluation that are congruent with the students' learning and development? What are the ways to represent an open student model?
- Tutoring model: What are effective pedagogical interventions to support students in the analysis of case studies? How can a collaborative learning environment be built to provide students with opportunities to have their thinking challenged by other students, to stimulate students' interaction over a given case, and to broaden their perspectives?
- Interface: How should a system interface be organized that both allows students to freely express their argumentation in natural language, and also is constrained enough for machine processing?

1.2 Approach

This research is based on a computer learning environment called Umka that we developed to support students in the analysis of case studies for professional ethics education. In this environment students first work on the case analysis individually and supply possible resolutions for the given case, arguments or issues for and against resolutions, choose the best resolution and provide justification for it. After that students have an opportunity to work collaboratively seeing and commenting on each others' reasoning. Different case studies can be supplied by instructors or the researchers for the students. Over the course of this research we designed, experimented and evaluated different features of Umka with the goal of identifying features that offer the most effective support.

While investigating the domain modelling issues, we did an extensive literature review on ethics education, metacognition and ill-defined domains to understand what skills students need to acquire through professional ethics education. We experimented with different ways to build a domain model for a particular case study: i) assembling a system predefined domain model that includes potential arguments that students can come up with for the case, possible feedback for these arguments, possible case study resolutions etc., or ii) organizing a domain model on the fly based on the answers of all students. We represented a domain model in natural English language as a set of isolated arguments tagged for or against a given case study.

Regarding the student model, instead of designing the system to evaluate students' textual analyses, which would be an impossible task for computer processing, we identified a number of metrics which could serve as proxies for this evaluation of students' performance. These metrics are: the number of arguments in a student's analysis, the resolution chosen by the student for the case study dilemma, the mark given to the student by her peers for her analysis and for the feedback she provided to her peers. These metrics were the basis of building an open group learner model. We represented the open group learner model through a circle visualization. We experimented with different features of the circle visualization. In its final incarnation students' positions about a given case study are depicted as circles of different sizes, colors and intensities representing the number of arguments in a student's position, chosen resolutions and the level of the acceptance by peers, respectively. The students' circle positions are placed at certain distance from each other, proportional to the difference between arguments in their positions. For calculating the difference between arguments of students we experimented with various ways to compare students' arguments, including calculating arguments' similarity score through Latent Semantic Analysis and Weighted Textual Matrix Factorization algorithms.

When developing the tutoring model, we explored multiple support types for stimulating learners' interactions, and developing their metacognitive skills. These support types include various system feedback on a learner's arguments, hints on arguments the learner hasn't yet considered, suggestions to consider certain arguments and positions of the learner's peers, different visualization formats. We evaluated the efficiency of the proposed support types through the educational impact they produced: the increase in the students' productive behaviours and changes in the students' analyses. We also measured the students' attitudes towards the proposed support types through the post-study questionnaires.

When considering the best arrangement for the system interface, we decided to organize the interface as a grid. The grid interface has different sections for different parts of the case analysis: arguments for and against a certain resolution, a chosen resolution, conclusions etc. This organization allowed the system to easily dissect for computer processing different parts of the case analysis including separate arguments tagged for or against a certain resolution, but at the same time this organisation gave students freedom to express their arguments in natural language.

This dissertation considers three main areas concerning the development of a tutoring system for professional ethics. The areas involve how to organize tutoring support for students' individual analysis, how to build a collaborative learning environment where students can interact with each other about their analyses, and ways for the system (Umka) to effectively diagnose students' arguments. Hypotheses examining the impact of Umka's support are presented, and their evaluation is provided based on the series of experimental studies.

1.3 Contributions

The goal of this research project is the development of techniques for organizing tutoring systems to support students in the analysis of case studies for professional ethics education. The contributions of this research are multifold. The primary set of contributions is to the field of the Artificial Intelligence in Education, and these particular subfields of it: ill-defined domains, metacognition, open group learner modelling, visualization, collaborative environments, domains requiring natural language interaction.

In this research we also surveyed literature on pedagogy in professional ethics education, and best teaching practices both in traditional classroom settings and through computer-based learning environments. Based on the reviewed literature and our own experimental results, we have formulated techniques for building computer-based learning systems for the development of skills necessary for the professional ethics domain. Thus, the secondary set of contributions is concerned with professional ethics education.

Specific contributions of the research are the following:

- Development of an effective interface design for an intelligent tutoring system for the analysis of case studies. This kind of interface gives learners the freedom to express their argumentation about a case study in natural human language, but has enough constraints for the computer processing of their argumentation.

- Development of techniques for an ITS to support learners in the individual analysis of case studies. This dissertation presents tutoring techniques that can effectively support a learner while she works individually in the case analysis. The dissertation also sheds light on techniques that didn't work that well, and didn't find substantial usage among the students.

- Development of techniques to organize collaborative learning environments that stimulate productive student interactions, and hone students' metacognitive skills such as reflection and open-mindedness to new perspectives. This dissertation provides experimental evidence as to the ways to organize collaborative environments and support in them.

- Development of an open group learner model through the visualization of students' positions on a given case. This dissertation proposes a novel way to illustrate learners' positions on a given case study in the class based on a circle visualization. The dissertation also demonstrates the effect of the visualisation on stimulating students' productive interactions with each other, and on fostering students' metacognitive processes of reconsidering and expanding their initial positions.

- Identification of effective methods for "diagnosing" students' arguments represented in a textual form. This dissertation presents the evaluation of a number of text comparison algorithms. These algorithms have been evaluated for the task of finding how similar to each other or different from each other specific students' arguments are.

- Discovery of correlations between specific students' collaborative behaviours and students' reconsidering and broadening their initial positions. This dissertation demonstrates what specific student interactions with each other in the collaboration stage are beneficial for stimulating students' metacognitive processes of reconsidering and broadening their initial perspectives on a given case study.

- Development of alternative metrics to measure students' learning in ill-defined domains, where the evaluation of the students' performance is impeded by the absence of a single right solution to a given problem, existence of multiple solutions that are subject to debate through textual argumentation, and the students' answers not rendered well for computer processing. This dissertation outlines metrics that can be used to measure students' learning in the analysis of case studies in the professional ethics domain, but also can be utilized in other domains that employ analysis of case studies as a pedagogical practice.

All these contributions have a goal to expand AIED's repertoire of techniques for supporting learning in ill-defined domains. Another goal of the research is to provide concrete approaches for building practical tutoring systems for the analysis of case studies.

1.4 Dissertation Outline

This dissertation is separated into seven chapters.

The related work will be reviewed in Chapter 2. In this chapter we will first get a bird's eye view of ill-defined domains, and various tutoring systems developed to support learning in this kind of domain. We will then narrow our focus to the ethics domain, outlining the specifics of this domain, its pedagogical goals and practices, and discussing techniques for building computer-based learning environments that address the needs of this domain.

Chapter 3 will present the main thesis statement. This will include three major research questions, and the methods to answer the research questions considering the ill-defined nature of the ethics domain. The chapter will also introduce the proposed approach: the Umka tutoring system and its key features.

Chapters 4, 5 and 6 are allocated for the in-depth discussions of each separate version of Umka: versions 1, 2, and 3.

Chapter 4 will present the initial version of the Umka tutoring system, and its components: a domain model, a tutoring model, a student model, and an interface. The chapter will justify the rationale behind choosing particular support types, and algorithms for the analysis of students' answers. It will also demonstrate a scenario of a student's workflow in Umka. Quantitative experimental results of a human subject experiment using this version of Umka, and qualitative analysis of the questionnaire answers will be presented.

Chapter 5 will introduce version 2 of Umka. This version incorporates visualization of students' positions as a way to support students in a case analysis, and as a way to demonstrate their open group learner model. A major part of the chapter will be devoted to the discussion of the results of several other experimental studies including the testing of quantitative hypotheses, analysis of qualitative questionnaire results, evaluating the accuracy of the text similarity algorithms, and analysis of the effectiveness of students' interactions.

Chapter 6 will describe the final modifications to Umka's interface, visualization and support types. It will present a comprehensive experimental evaluation of these modifications in new empirical studies.

The thesis will conclude with Chapter 7 which will summarize findings in the form of answers to the initial research questions, state the research contributions and broader implications, and offer possible avenues for further investigation.

Chapter 2

Related Work

This chapter will firstly review a broad literature on ill-defined domains: their characteristics, and work that has been done in the field of intelligent tutoring systems about how to support learning in them. The second part of the chapter will focus more specifically on the ethics domain. Instructional principles and pedagogy in ethics education, and computer learning systems developed to support learning in the ethics domain will be discussed.

2.1 Ill-defined Problems

Ill-defined problems have received relatively little attention in the literature of ITS and AIED, but these kinds of problems are encountered in every domain and in everyday life. This section will outline characteristics of ill-defined problems, and will discuss techniques for designing domain, student and tutoring models for building ITSs for ill-defined problems.

2.1.1 Characteristics of Ill-defined Problems

From the point of view of the ITS and AIED fields domains or problems can be divided into two categories: well-defined (well-structured) and ill-defined (ill-structured). A problem is called well-structured to the extent it has the following characteristics [94]:

1. There is a definite criterion to test a solution, and a process to apply this criterion.

2. The initial problem state, the goal state, all other states reachable during the problem solution, the changes between states, possible effects of them on the external world, and knowledge that the problem solver can acquire about the problem, can be represented in a problem space with a practicable amount of computation and memory.

Examples of well-defined problems are solving an algebra equation or a quantitative physics problem. In these problems there is only one right solution and there is a predefined known number of steps to reach it. Since the number of states of a well-defined problem, and the algorithms for solving such problems are well known, and the solution algorithms are frequently mathematical or logic-based [105], the techniques for building ITSs for well-defined problems are more straightforward, and hence they are more well-studied.

In contrast, ill-defined problems have different characteristics. Examples of ill-defined problems are designing a house, composing a musical piece, deciding a policy in politics and international relations, finding a way to improve the education system, etc. The notion of ill-defined or ill-structured domain is in itself ill-defined; there are no clear boundaries between ill-defined and well-defined domains, but a continuum ranging from well-defined to ill-defined [20]. The AIED and education literature suggest that problems are ill-defined to the extent they have the following characteristics [20, 53, 95]:

- *An incomplete, indefinite start state of the problem.* Problem definition is ambiguous and not clear [67]. One or more of the problem elements are unknown [39, 53], vaguely defined, have unclear goals [39], have very few initial constraints (Reitman 1973 as cited in [94]). Concepts and relations of the problem are un- or under-specified, open textured, or intractable with a lack of absolute definition [52, 53, 54]. The description of the problem requirements or all relevant concepts and terms and problem components is absent or underspecified (Reitman 1964, 1965 as cited in [53]). For example, if we have a problem as to how to improve the education system, the definition of "improvement" is ambiguous, not definite, and could mean different things for different people.

- *Unclear, indefinite strategies for finding solutions.* There is no well-defined algorithm to apply to the initial state to derive the goal state [67]. A common strategy for well-defined domains — decomposing a problem into subproblems — often doesn't work, because problems in ill-defined domains don't often decompose into subproblems [52]. It is uncertain which concepts, rules, and principles are necessary for the solution and how they are organized [39].

  Solving the problem begins with the problem recharacterization, refinement, generating more constraints in order to define the particular issues clearly, and this recharacterization could be the subject of debate [53, 105], (Reitman 1973 as cited in [94]).

  A solution often relies on reasoning analogically with cases and examples [53], but relationships between concepts, rules, and principles are inconsistent between cases, or there could be no prototypic cases because case elements are differentially important in different contexts and, because they interact, there are no general rules or principles for describing or predicting most of the cases [39].

  An ill-defined problem is often not considered to be solved when one solution is presented but may be readdressed by multiple, often distinct solutions [53].

  Solving an ill-defined problem often involves uniquely human interpersonal activities such as the expression of personal opinions or beliefs, making and defending judgements about the problem [39, 92], design, novelty [52]. As people approach ill-defined problems with their personal knowledge, beliefs, attitudes, etc., there is always disagreement among problem solvers about the solutions [53]. Thus, a solution often includes arguments about why a particular solution could be a good solution [105].

  Ill-defined problems have a large or complex solution space that prohibits one from enumerating all possible characterizations or solutions [53]. The size of the the knowledge database for ill-structured problems makes simulation difficult or even impossible [105].

- *The indefinite, unclear goal state of the problem.* The form of the final correct solution is not clear, and it is not clear whether the correct solution even exists [67], or there exist multiple solutions [53]. Either there exist no systematic well-accepted methods to validate correct solutions (John McCarthy (1956) as cited in [53]), or there are multiple ambiguous criteria for evaluating solutions [39, 52]. There are no uniquely specifying means to validate problem solutions or cases [53]. For example, there are no standard ways to verify legal arguments, pieces of music or architecture [52]. Hence solutions to ill-defined problems are judged not as right or wrong, but rather "regarded in terms of some level of plausibility or acceptability" [105, p. 305], agreement or disagreement, or preferability [20].

2.1.2 Skills For Solving Ill-defined Problems

The dissimilar nature of ill-defined and well-defined problems presupposes different skill sets for solving them. Kitchener [43] was the first who proposed a three-level model of problem-solving skills: cognitive, metacognitive, and epistemic. Epistemic skills refer to a person's ability to assess the limits and certainty of his or her knowledge, the legitimacy of solutions [43]. Metacognition is defined as thinking about thinking, "one's knowledge concerning one's own cognitive processes and products or anything related to them" [17, p. 232]. Metacognition consists of two processes: i) metacognitive knowledge — knowledge about oneself as a learner and factors affecting cognition, knowledge of strategies to achieve learning goals, etc., and ii) metacognitive regulation — self-management of one's own thinking such as evaluation, planning, regulation [11, 18].

Kitchener argued that well-defined problems can be solved using cognitive and metacognitive skills, while ill-defined problems in addition require epistemic skills [43].

Jonassen [39], Shin [91], Shin et al. [92], and Shraw et al. [82] supported Kitchener's idea that well-defined and ill-defined problems require separate cognitive processes. Jonassen [39] argued that the metacognitive activities of composing and carrying out a problem-solving plan are necessary for all kinds of problems. However, the nature of these metacognitive processes is different: "... ill-structured problem solving should engage meta-metacognitive processes whereby individuals monitor the epistemic nature of the problems they are solving and the truth value of alternative solutions, not just the comprehension-monitoring metacognitive strategies that serve well-structured problem solving" [39, p. 81]. The metacognitive process for ill-defined domains include individuals' reflection on what they know, the certainty of their knowledge, their potential biases, consideration of others' beliefs and alternative solutions, selection or synthesis of a unique solution. Shin in his Ph.D thesis [91] demonstrated that cognition (domain specific knowledge and structural knowledge) is necessary to successfully solve well-structured problems, while metacognition as well as cognition are essential for solving ill-defined problems.

Some domains like law, ethics, history, public policy, and architecture are inherently ill-defined, as are most of the problems in them [53]. This means that the very subject matter to be learned in these domains is metacognitive and epistemic, and ITSs supporting learning in these domains should focus on enhancing these skills in learners.

2.1.3 Techniques for Building ITSs for Ill-defined Domains: Domain Model

In ill-defined domains there is no complete formal domain theory [20], or at least there is lack of widely accepted domain theories uniquely specifying relevant concepts, relations, and procedures [52, 53], or there are competing domain principles that are subject to debate [53]. For example, music composition and architecture have incomplete domain theories [20]; ethics has competing theories that are subject to debate. Ill-defined domains are abundant with open-textured concepts that are abstract, have several interpretations, and rely on natural language [20]. All this makes the modelling of ill-defined domains a challenging task, necessitating modification of the traditional approaches and adoption of new specific approaches. Several approaches presented below are either new specific approaches developed for ill-defined domains, or an application of traditional approaches of domain modelling to ill-defined domains (these are the expert system approach, the cognitive approach, and the constraint-based modelling approach).

The Expert System Approach

McLaren proposed that his expert system for ethical reasoning Truth-Teller can be used to build an intelligent tutor that aids students in case comparison [59]. Truth-Teller compares a problem case to some paradigm cases, and identifies similarities and differences between them, and presents this analysis to the user. All cases are on the same topic whether a protagonist should tell the truth, and are represented semantically with identified protagonists, reasons to tell the truth, and reasons for silence.

The expert system approach to domain modelling requires a significant prior effort of constructing an expert system, and hence can be applied for only a small part of a given domain. For example, McLaren's system worked only with cases on the topic of truth telling.

The Cognitive Approach

Ill-defined tasks don't have definite strategies for finding solutions, and hence cannot be modelled with complete cognitive task models. However, several attempts have been made to build partial, incomplete cognitive models of ill-defined tasks, that still turned out to be useful.

One such attempt is in the RomanTutor, an ITS for learning to operate Canadarm2, a robotic arm deployed on the international space station which has seven degrees of freedom [19]. Users' plans consisting of multiple actions were recorded by the tutoring system, and the plans were annotated as successful or buggy depending on whether they reached the goal or not. Then a data mining algorithm was applied to learn successful sequences of actions frequently repeated among several users. These mined sequences were used to build a partial domain model.

Another attempt to approach ill-defined domain modelling from the cognitive perspective was made in a French language tutor [71]. The goal of this tutor is to teach learners feature discrimination between the Passe Compose and Imparfait verb tenses, so students will know which tense to use in a particular situation. Researchers first developed a theory driven model specifying rules that can provide evidence for the usage of a particular verb tense, then they augmented their model with heuristics obtained from expert and novice users. Developers of the French tutor acknowledge the incompleteness of their model.

The tasks of operating a robotic arm or using a correct tense of a french verb are not that far removed from well-defined tasks, given that both these tasks have a specific goal whose success and failure is definitive (i.e. the arm moves where it should, the tense is used correctly). This may explain why it was possible to have cognitive domain modelling for these tasks.

The Constraint-Based Modelling Approach

The constraint-based modelling approach could be a tractable way to represent a domain for ill-defined problems. It doesn't require listing the correct solutions; neither does it require providing the solution procedure. An example of a constraint-based tutor for an ill-defined task is SQL-Tutor [66]. SQL-Tutor supports students in learning how to write SQL queries that address given problem statements.

Although the judgement about the correctness of a particular SQL query is relatively straightforward, the authors of SQL-Tutor classify the query writing as an ill-defined problem because a problem statement can be ambiguous, the transformation of the problem statement into an SQL query is underspecified, and there could be multiple correct solutions [67]. SQL-Tutor has around 700 constraints that all correct solutions should satisfy. Some constraints specify syntactic properties of queries, and check, for example, that the SELECT clause of a SQL query is not empty, or all the names in the FROM clause are valid. Other constraints specify semantic properties of queries, whether a student's solution is a correct solution for a given problem by comparing it to the ideal solution [66].

Weak Theory Modelling: Modelling Law Concepts

Weak theory modelling attempts both to model what is known about the domain, as well as to leave room for ambiguity, ill-definedness, and personal interpretation.

Aleven's CATO system [1] uses factors to model ill-defined law concepts. In the legal domain even experts disagree on the definitions of concepts. For example, there is no exact definition of a trade secret, but rather there is a list of various factors to be considered in determining whether information is a trade secret. Thus, the CATO system domain model is represented as a hierarchy of factors. All factors are linked to more abstract legal issues (e.g. a "trade secret" issue) with strong or weak, positive or negative links. A positive link in the hierarchy indicates that the factor supports an issue, a negative link indicates that the factor favours an opposite conclusion on an issue. All cases in the CATO system are also annotated with factors, and the system using the hierarchy of factors can reason about how strongly a particular case represents an issue. What is important in this model is that the factor hierarchy is not the standard AND/OR goal tree often used in well-defined domains: the links represent support for making a case stronger or weaker with respect to an issue. This allows the modelling of ill-defined law issues by demonstrating how strong factors in the cases represent them.

Weak Theory Modelling: Modelling Negotiation Skills

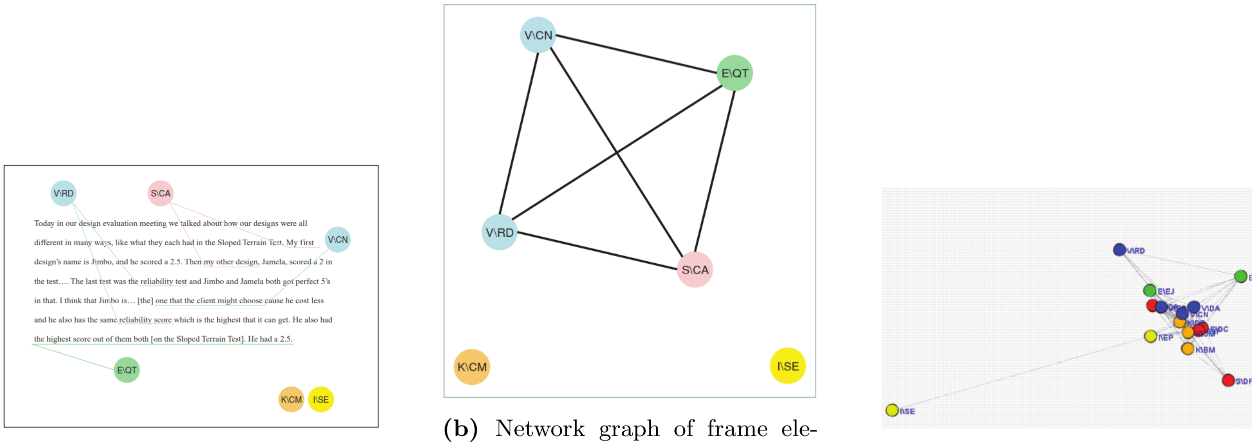
A similar approach to modelling the domain knowledge was taken in the BiLAT system [42]. BiLAT is an ITS that teaches U.S. soldiers negotiation skills. Negotiation skills consist of knowledge of negotiation principles (e.g. understanding what one really wants), and procedural steps for successful negotiation (e.g. before the meeting conduct research on your meeting partner). For domain modelling theses principles and procedures were distilled into a hierarchical list of learning objectives. Each learning objective consists of standards that if followed will lead to success in BiLAT. Some standards are required (labelled REQ), while others may not always be needed or appropriate (labelled USL for usually necessary). This allows some modelling of ambiguity, typical for a negotiation domain.

2.1.4 Techniques for Building ITSs for Ill-defined Domains: Student Model

The diagnosis of a student's answers, students' errors, student's solution strategy and behaviour constitutes the building of a student model. In ill-defined domains there are multiple solution paths to reach answers, and very often answers are presented in textual form. This section summarizes techniques for diagnosing student behaviour and building student models in ill-defined domains.

Weighted Constraint-Based Model

Le and Pinkwart [48] proposed the use of a weighted constraint-based model to detect the solution strategy used by a student for problems with multiple solutions. In their approach every constraint on correct solutions has an associated probability measuring the importance of a constraint, and is referred to as a constraint weight. Every solution strategy has a set of constraints to satisfy. The ITS, using constraint weights, calculates the plausibility of every possible solution strategy, and chooses the strategy with the highest plausibility as a diagnosed strategy pursued by the student. The experiment of Le and Pinkwart in the logic programming domain demonstrated that adding constraint weights can improve the error diagnosis of a constraint-based ITS. Their tutoring system with a weighted constraint-based model was about twice as precise in diagnosing errors as an ITS with a simple constraint model. Although Le and Pinkwart applied their weighted constraint-based model in the relatively well-defined domain of the logic programming, ill-defined domains can also make use of it. Ill-defined problems very often have several possible solution paths, and the model of Le and Pinkwart can be used to detect a solution path pursued by a student.



(a) Coded data

(b) Network graph of frame elements at one point of time

(c) Final network graph

Figure 2.1: The Epistemic Network Analysis model (reprinted from [86, p. 6] with permission from MIT press).

Sociocultural Model

One of the ways to model students' learning progression in ill-defined domains is to compare their thinking and acting with the thinking and acting of professionals from a certain community of practice. Learning to solve ill-defined problems in professional domains implies becoming part of community of practice and sharing with this community similar ways to solve problems. Learning doesn't only involve mastering skills and knowledge, but involves mastering the culture of the community which, according to Shaffer and Graesser [85], consists of the following components: 1) skills: the things that people within the community do; 2) knowledge: the understandings that people in the community share; 3) values: the beliefs that members of the community hold; 4) identity: the way that members of the community see themselves; 5) epistemology: the warrants that justify actions or claims as legitimate within the community. This collection of skills, knowledge, values, identity, and epistemology forms the epistemic frame of the community.

Shaffer and Graesser [85] developed a model that can capture the progression of students' epistemic frames. The model called Epistemic Network Analysis (ENA) measures the extent to which an individual has the ways of thinking, talking, and acting that are characteristic of a particular community of practice. Given the identified student's frame elements — skills, values, identity, etc. (Fig. 2.1 a), ENA constructs a network graph which shows the strength of association between each pair of frame elements at each particular time moment (Fig. 2.1 b), and then constructs a cumulative network graph for the whole time (Fig. 2.1 c), where frame elements (nodes) that are linked more often in the data are closer to each other than those that are linked less often in the data [86]. ENA can quantify the student's epistemic frame using concepts from social network analysis, such as density and centrality. For their experimental domain Shaffer and Graesser chose ecological thinking, where students learn to deal with land use issues in ecologically sensitive areas.

Evaluating Students' Answers in a Textual Form

Very often students' solutions to ill-defined problems are presented in the form of text. Students' answers are usually short, ambiguous, with grammatical errors. This and the imperfection of natural language processing techniques make it quite challenging to evaluate students' answers for building an accurate student model. Several research groups responded to this challenge, giving rise to interesting methods to evaluate textual answers and to constrain environments for a more accurate diagnosis. Most of these methods are based on the comparison of a student's answer with some benchmark answers prestored in the system. The following methods were proposed:

- *Content Overlap.* Content overlap, or keyword search is the most simple method for assessing a student's answer, and is based on the number of overlapping words between the student's answer and an expert answer. The higher this number, the more correct the student's answer is considered to be. Content overlap was used in a number of ITSs. For example, Walker et al. [106] used content overlap to evaluate students' posts in their intercultural competence tutor. Their method was able to generate assessment that matched human ratings in 64% of cases.

- *Combination of Content Overlap and Interface Design.* A well-designed interface can significantly aid the system's understanding of students' answers. An ITS that uses a combination of keyword search and interface tools to predict students' input is Rashi — an inquiry tutor in biology, geology, forestry, and art history [16]. In the Rashi system a student types her observations, inferences, hypotheses, arguments, evidence. The system, using keyword search, offers expert statements that are similar to the student's, and if the student chooses one of these statements, the system recognizes the student's comment as the same as the expert comment. The interface tools in Rashi also help to understand the student's input. The student collects data using the system interface tools, which means that the system can identify which data has and has not been explored. Even if the system is unable to match any student textual statements, it recognizes the data collected by the student. Using all this matched information, the system builds the student model which is a subset of the expert argument [15].

- *Latent Semantic Analysis.* Latent Semantic Analysis (LSA) is a method for extracting and representing the meaning of words from a large corpus [46]. LSA in combination with other methods has also been a popular technique for the evaluation of students' written answers in several ITSs. For example, AutoTutor (a tutor that teaches physics, computer literacy and research ethics through a conversation with a learner in natural language) compares a student's answer with prestored expectations and misconceptions, and formulates its next dialogue move as the corresponding feedback [26, 27]. Experiments with AutoTutor demonstrated that LSA performs matching of students and predefined system's answers almost as well as human judges. The correlation between AutoTutor and a human expert has been approximately $r = 0.5$, whereas two human experts correlate approximately $r = 0.6$.

A combined LSA and content overlap method proved to be the most effective method for the iSTART system [63]. iSTART, Interactive Strategy Training for Active Reading and Thinking, is a tutoring environment for improving students' self-explanations of science texts. After reading a science text, a student is asked to self-explain certain parts of the text. iSTART evaluates the student's self-explanation by comparing it with a benchmark answer. The iSTART approach correlated with a human judgement in 67% cases.

- *Textual Entailment.* Textual entailment is the task of deciding, given two text fragments, whether the meaning of one text is entailed (can be inferred) from another text. Rus et al. [80] used a textual entailment algorithm, implemented in their Entailer software, to evaluate students' answers in the iSTART system. They used their Entailer software to judge whether a student's answer elaborates, is entailed by, or paraphrases an original sentence from the text. The Entailer's results produced very high correlations with human evaluations (paraphrase: r=0.818, entailment: r=0.741, elaboration: r=0.673), and were significantly more accurate than the results from the content-overlap method (paraphrase: r=0.659, entailment: r=0.57, elaboration: r=0.515), and from the LSA method (paraphrase: r=0.574, entailment: r=0.469, elaboration: r=0.416).

- *Lexical Semantics and Word Weighting.* In this method semantic similarity between a student's answer and an expected system answer is calculated based on some semantic database. A popular database to use is WordNet that allows the assessment of the relatedness of two concepts, based on how close they are in the database, what the relationships are between them, and how similar are their definitions. Rus et al. [79] used lexical semantics to evaluate students' answers in the iSTART system, and their evaluation demonstrated the correlation of 0.606 with human judgement.

- *Support Vector Machines.* Support vector machines (SVM) are classifiers for classifying non-linearly separable classes [37]. SVMs have been successful at tackling a wide range of classification problems, including text classification. A support Vector Machine (SVM) classifier was the best performing method for the assessment of students' essays in Hughes et al.'s study [37]. In their study schoolchildren read tree related articles, and were asked to summarize the content of these articles in an integrative essay. SVMs reached an accuracy of around 0.6–0.8 in detecting in children's essays the presence of all necessary essay components: top-level claims, evidence for those claims, integration between texts, etc.

- *Combination of Content and Context.* Usually diagnosis of students' answers is based on their content, but contextual cues — characteristics of the students and how students were answering — can also significantly aid the diagnosis. Lehman et al. [50] conducted a comparison study between three models of the evaluation of students' self-explanations in the Autotutor system: the content model, the context model, and the combined model. Their context model included a student's answer's response time, performance (correct or incorrect), response time to the pretest, the order of topic presentation. Lehman et al. found out that the context model was the most successful (74%) for discriminating correct from incorrect responses. However, it was the content model that performed best for discrimination of particular error types within incorrect answers (67.6%). The Combined model did not yield any noticeable improvements over the best performing individual model for either task.

- *Using Various Linguistic Features.* To assess the quality of students' essays in the Writing-Pal intelligent tutoring system McNamara et al. [62] explored various linguistic features. The Writing-Pal system teaches high school students a number of strategies for writing high quality essays, and provides activities to practise these strategies. McNamara et al. discovered what linguistic features are correlated with the quality of students' essays. For example, in differentiating between the parts of essays (i.e. introduction, body, conclusion), the linguistic features that are the most predictive seem to be measures of length, cohesion, and word difficulty. Thus, introductory paragraphs are shorter, include few cohesion cues, and contain words that are more specific, more meaningful, and less common in English. The researchers evaluated the performance of this differentiating model for the task of identifying if a paragraph is an introductory, body or concluding in the predefined 182 paragraphs randomly chosen from students' essays. The performance of the differentiating model was quite high — 65% accuracy versus 66% accuracy for a human judgement. The grading of a whole essay was still work in progress for the McNamara's group.

- *Graphical Representations.* Instead of evaluating the content of a text, students may be asked to represent their reasoning in a graphical form, and evaluation of the produced graphical representation of the text can be made. This approach was taken in several ITSs to evaluate students' scientific reasoning, such as the Belvedere system [98], and legal arguments, such as the Largo system [73]. The analysis of graphs is more amenable to machine processing than the analysis of arguments represented in textual form, which allows more reliable diagnosis of students' knowledge. For example, Pinkwart et al. [75] demonstrated that diagrams' features made by students in their Largo system are correlated with students' number of years in law school and students' standardized test scores that assess their ability to evaluate argumentation reasoning.

- *Collaborative, Peer Assessment.* In many collaborative systems students are asked to evaluate each others' answers. A number of models have been developed to intelligently combine students' evaluations for the production of a more objective, more accurate assessment of their solutions. Cho and Schunn [8] developed a model to assess students' writing using peer reviews. Their system, called SWoRD (Scaffolded Writing and Rewriting in the Discipline), allows students to submit a paper, get it reviewed by peers, rewrite it based on the reviews, and submit it again. In SWoRD as a reviewer submits his reviews, SWoRD assigns an accuracy to the reviewer. The calculation of the accuracy of the reviewer is based on the comparison of his ratings to the mean ratings for that set of papers across (typically six) reviewers. To make a final grade for the writing, the scores of all reviewers are combined, giving less weight to scores of less accurate reviewers, and more weight to scores of more accurate reviewers. Cho and Schunn [8] didn't directly evaluate their assessment model by comparing it with the assessment of an expert. However, an indirect proof of the success of their model is the fact that students who used SWoRD improved their writing more than students who used feedback from a single expert.

Visualizing Learners' Knowledge, Interactions and Argumentation for Enhanced Performance

It has long been known in the AIED community that open learner modeling, that is presenting to a learner the system's evaluation of her performance, has positive effect on that learner's progress. An open learner model promotes the student's reflection on her knowledge and skills, encourages self-assessment, supports planning and monitoring, allows the student to take greater control and responsibility over her learning [6, 14]. An open group learner model, presenting to the students the performance of the whole class, has extra advantages of helping learners to reflect on their progress in the group context, prompting or supporting collaborative and/or competitive interactions amongst groups of learners, allowing students to recognize and realize opportunities for learning with more capable peers [6, 7].

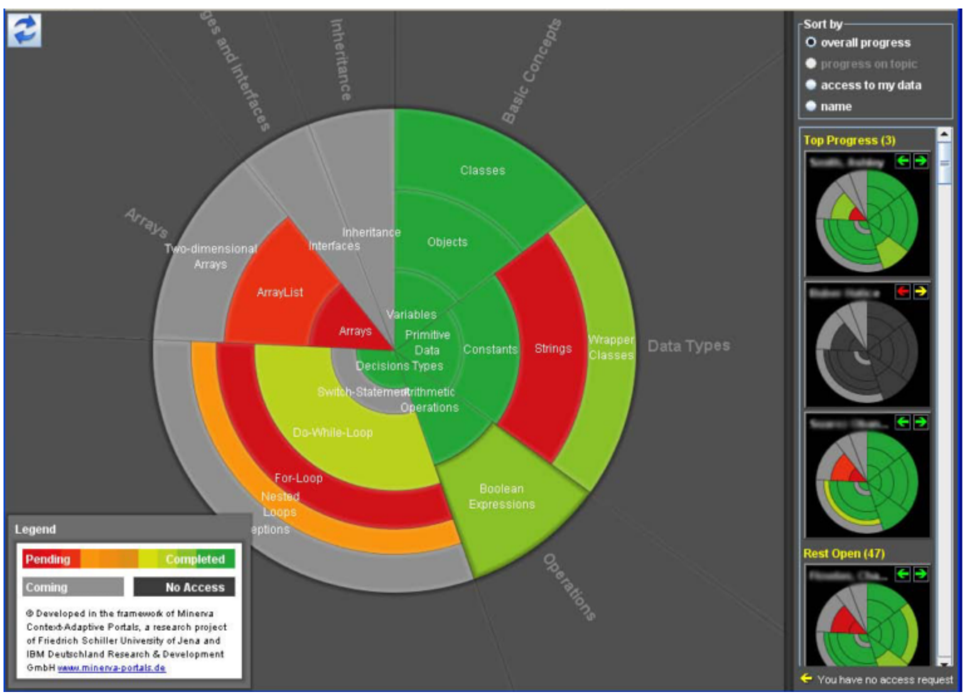


Figure 2.2: Progressor's interface. The right pane displays progress charts of peers (reprinted from [36, p. 304] with permission from ACM).

Researchers developed a number of visualizations to support collaborative learning and interaction. Brusilovsky's research group was working on developing visualizations to support social learning [5, 30, 36]. One of their systems, called Progressor, displays a student's progress as a pie chart [36]. Using different colors the pie chart demonstrates which topics have been covered by the student, and which have not been, and the level of coverage for each topic (Figure 2.2). In Progressor's visualization students can compare their

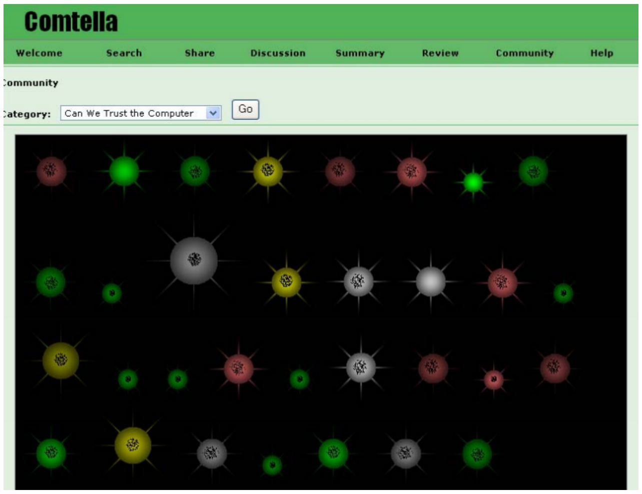


Figure 2.3: Comtella's interface: presenting each user as a star (reprinted from [97, p. 61] with permission from University of Saskatchewan).

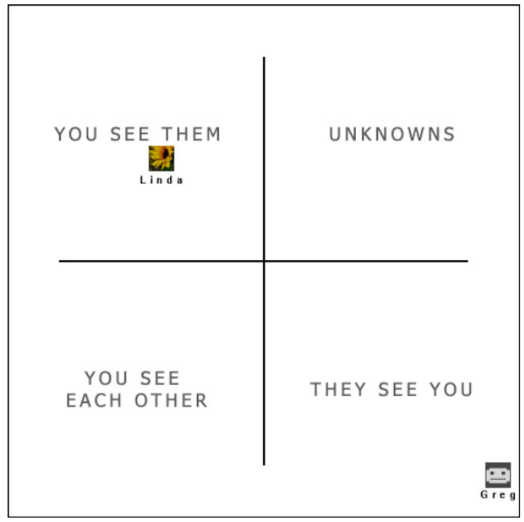


Figure 2.4: Comtella's interface: visualizing interpersonal relations among users (reprinted from [107, p. 226] with permission from Springer).

Forums	Description
Privacy	Big Brother, databases, risks, protection, awareness, philosophical views
Freedom of Speech	Censorship, anonymity, laws, offensive/dangerous speech
Intellectual Property	Fair-use, copying music/movies/software, solutions, copyrights vs patents
Wiretapping and Encryption	Role of secrecy, trust in government, cryptography
Computer Security and Crime	Hacking, hactivism, law, identity theft, privacy and civil liberties, crime fighting
Computers and Work	Changing nature of work, impact on employment, employee monitoring, teleworking
Broader social issues	Computers and community, digital divide, bad technologies, who benefits the most
Can we trust the computer?	What can go wrong, Therac-25 case study, reliability and safety, computer models
Ethics and Professionalism	Professional codes and guidelines, cases, aspects of professional ethics

Figure 2.5: Comtella's interface: visualizing the rank of contributions (reprinted from [107, p. 230] with permission from Springer).

pie charts with the pie charts of their peers (displayed in the right pane of the Figure 2.2). Progressor's visualization proved to create positive motivational impacts: the visualization of other students' progress, especially of class leaders, encouraged students to work more and ahead of the course schedule, and served as a guidance on what next relevant quizzes to attempt.

Vassileva's research group has been working on designing visualizations that provide social awareness about the other learners' existence, actions and contributions, and encourage the adoption of social norms such as social comparison and reciprocation [103]. In their Comtella system they experimented with a number of visualizations that proved to increase students' participation: representing users as stars of various sizes, colors, and brightness based on the number and quality of their contributions (Figure 2.3); visualizing interpersonal relations among participants demonstrating if they read, rank, and comment each other's contributions (Figure 2.4); using "hotter" colors and bigger font sizes for participants' contributions with higher rankings (Figure 2.5) [97, 107].

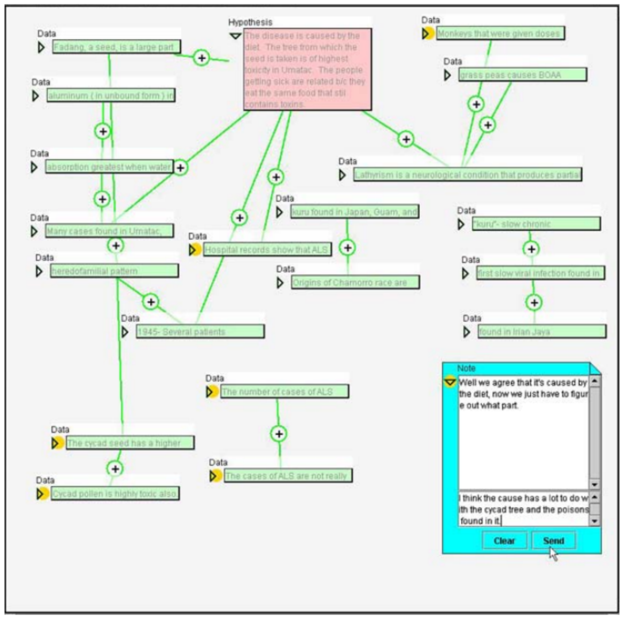


Figure 2.6: A knowledge map from the Belvedere system: visualizing students' argumentation (reprinted from [99, p. 1108] with permission from Elsevir.)

Graphical argumentation tools scaffold students' argumentation by enabling students to see the structure of their arguments, facilitating construction of better arguments and communications around them [38]. Suthers et al. showed that the visualization of arguments scaffolds collaborative knowledge construction better than threaded discussions [99]. In their study students who used visual knowledge maps (Figure 2.6), in comparison with students who used threaded discussions, created more hypotheses earlier, elaborated on them more, and had greater collaborative integration of information. Similarly, in the field of more formal, legal argumentation Pinkwart et al. [74] has demonstrated the superiority of graphical representations over a purely text-based alternative. Particularly, the lower-aptitude students who graphically construct legal arguments in their LARGO system showed higher performance than the lower-aptitude students who used a purely text-based alternative — they better identified and understood the argument components.

2.1.5 Techniques for Building ITSs for Ill-defined Domains: Tutoring Model

The tutoring techniques presented in the section were developed to address specific needs of ill-defined domains.

Teaching Through a Model and Examples

A teaching technique through a model and examples was implemented in Aleven's system CATO [1]. The CATO system helps law students to acquire case-based argumentation skills. It demonstrates to students a model as to how to legally reason about a case based on the case's similarities and differences with other cases. CATO also provides computer-generated examples. Aleven in his Ph.D. thesis [1] demonstrated that this type of support through the model and examples was enough for statistically significant improvements in students' basic argumentation skills.

Edutainment — Teaching Through Games

Edutainment — the utilization of serious games for education — seems to be a promising and popular tutoring strategy employed in many ITSs for ill-defined domains. In some ITSs games are a part of the system's teaching repertoire; in others a game is the core of an ITS. An example of the former is Writing-Pal [12], an ITS for teaching students various strategies to write high quality essays. Each strategy lesson in Writing-Pal is complimented by numerous challenges represented in the form of games.

An ITS with a game at its core uses an interactive narrative to tutor students. An interactive narrative environment for learning presents a story to a learner with the goal that the learner by immersing into the environment and interacting with the story can build the necessary domain knowledge and skills. Often the story is presented as a game; the story unfolds based on the actions of the learner in the game. One of such system BiLAT teaches cultural awareness and negotiation skills for bilateral engagements [42]. One of the scenarios in BilAT places a learner in some Iraqi village. The learner interacts with different characters or agents in the game by choosing actions from the menu. Evaluation of the BiLAT system using U.S. soldiers discovered significant situation-judgement gains by participants without prior negotiation experience. For those with prior negotiation experience results did not demonstrate significant change.

Supporting Development of Reasoning Via an Inquiry Environment

Structured reasoning, the ability to reason about cases, to postulate theories and hypotheses, to recognize whether data supports hypotheses is a necessary skill for ill-defined domains. Dragon et al. [15] showed that the development of structured reasoning can be effectively supported by a well-designed inquiry environment. Dragon's inquiry learning infrastructure called Rashi presents authentic cases to students from four domains: geology, biology, art history, and forestry. Students generate questions to investigate a case, gather and analyse data, and generate hypotheses based on inferences made from evidence. Rashi supports inquiry learning by providing tools for gathering, organizing, visualizing, and analyzing information, and by giving feedback to students on their argumentation and inquiry process. Evaluation of the Rashi system in several classrooms showed that the usage of the system leads to a greater participation, better learning outcomes – more and richer hypotheses than from the inquiry-style lectures, which confirms that Rashi's infrastructure can enhance inquiry instruction and structured reasoning [108].

Supporting Through Graphical Representations

A number of tutoring systems have been developed that support students through graphical representations. One of them, LARGO [74], supports students in the development of argumentation skills by rendering argu-

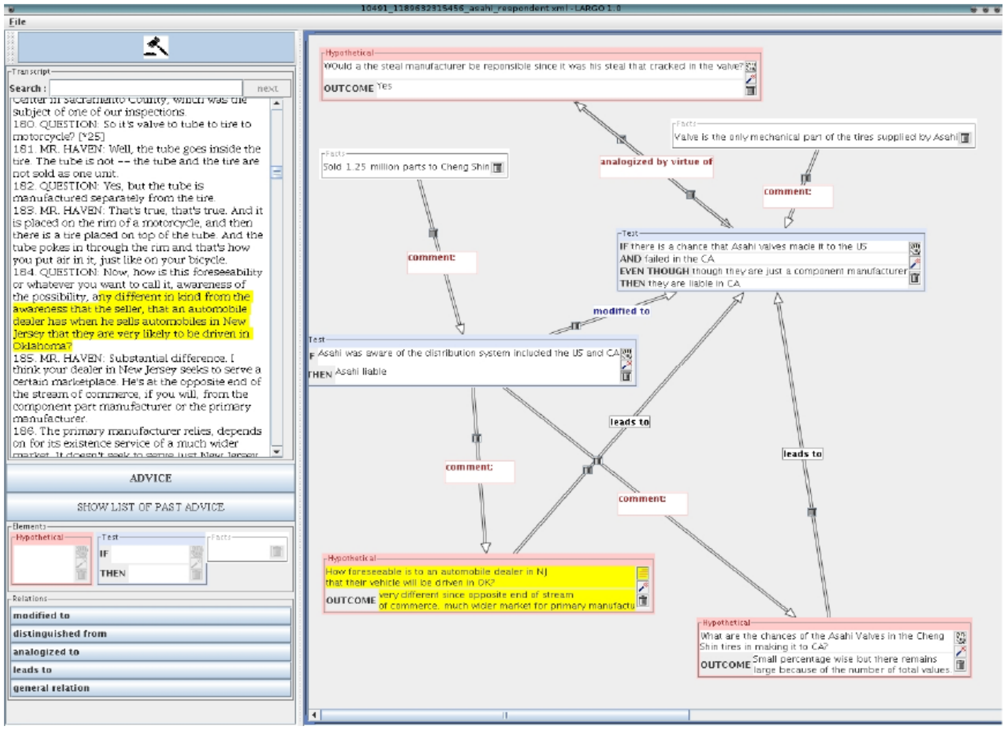


Figure 2.7: LARGO representation of an argument (reprinted from [75, p. 93] with permission from Springer).

ments graphically. In the LARGO (Legal ARgument Graph Observer) system students read texts about U.S. Supreme Court oral arguments, and then graphically represent their reasoning. Students construct graphs by dragging different elements of an argument into the workspace (e.g. proposed hypotheses, hypothetical challenges, responses, relations between them: distinction, analogy, etc.), filling in appropriate text, and linking graph elements to parts of the text (Figure 2.7). Pinkwart et al [74] discovered that LARGO was especially helpful for low performing students. The low performing students who used their tool significantly outperformed in the post-test other low performing students who didn't use their tool.

Supporting Collaboration

Collaboration and social interactions have been recognized as powerful methods to support learning in ill-defined domains. But for collaboration to be the most effective for learning it should be mediated by a computer environment. This mediation can be organized through facilitating interactions, analysing students' dialogues [61], helping the peer reviewing process [25], highlighting conflicts and disagreements [13], etc. Disagreements and conflicts can be a good learning opportunity for students, when students need to identify and resolve conflicts in their viewpoints, present alternatives, and request and give explanations [13]. Detecting these learning opportunities and correspondingly facilitating collaboration are the main functions of COLER (COllaborative Learning environment for Entity-Relationship modeling) [13]. In COLER students resolve database modelling problems, first building entity relationship diagrams individually, and then collaboratively while communicating via a chat window. A computer coach finds significant differences between individual and group diagrams by using a problem-specific glossary of terms and subgraph matching. Based on the found differences COLER generates advice in the form of suggestions or questions that try to encourage students to discuss or participate. The evaluation study conducted with COLER [9] showed that 40% of the total COLER advice instances were followed up by the students. Researchers didn't evaluate the learning outcomes of individual students, but they observed that all final group diagrams were better than the individual diagrams.

Weak Theory Scaffolding

Although ill-defined domains are characterized by the absence of well-structured theories, they still have some weak theories about how to solve problems in the domains. For example, the procedure for the analysis of a case study in the ethics domain includes the following steps: 1) identify known morally relevant facts and unknown morally relevant facts; 2) identify conceptual issues; 3) identify specific ethical issues and general ethical issues; 4) suggest possible resolutions, examine consequences, and select a resolution [23]. This guidance was used as a basis to build the Pete system [23]. Pete supports students in the analysis of case studies in two ways. Firstly, it walks students through the steps of the ethical analysis by providing instructions and asking students to fill in predefined forms. Secondly, PETE stores responses of past students, so the student after finishing a particular step can compare his or her response with responses of students from the past. The results of the evaluation of PETE are unknown.

Supporting by Pedagogically Augmented Simulations

One way to study complex ill-defined problems is to simulate them. There are many computer programs that simulate natural and urban disasters, migration of people, terrorist attacks, epidemics, criminal activities, etc. Simulation by itself is a powerful pedagogical strategy. Simulations make it possible for students to study the cause-and-effect relationships of their actions without the risks, costs or damages these actions may bring about in real environments [21]. But simulation alone is not sufficient for effective learning: often students find it difficult to understand the simulation process and results. Therefore, some researchers have tried to augment computer simulation with pedagogical models.

One such attempt was the work of Furtado and Vasconcelos [21] to support students' understanding of the simulation of resource allocation in the ExpertCop tutorial system. ExpertCop aims to train police officers in the activity of preventive policing allocation. The activity involves deciding how many police teams to allocate for what time and for what points of the city with the intention of minimizing crime. In ExpertCop students (police officers) analyse the geographic and statistical data for a selected geographical region, and perform the allocation of the police force based on such analysis. Then, ExpertCop, using the students' allocation plan, simulates how crime rates change — what crimes were prevented, and what crimes occurred. This cycle can be repeated as many times as the students finds necessary. As each new allocation is performed, the system shows the student whether or not the modification brought about an improvement in the crime rate. The evaluation of ExpertCop demonstrated that as a result of the interaction with the system the students improved their understanding of the allocation process, motives and causes for crimes, and made these concepts more specific and practical.

2.2 Ethics Domain

After discussing general approaches for building systems for a number of ill-defined domains, in this section we will focus on the ethics domain. The objectives of this section are to clarify the goals of ethics education and the main pedagogical approaches to achieve them, as well as to highlight advances in computer-based learning systems designed to support ethics learning.

2.2.1 Goals of Ethics Education

A popular resource on general ethics education, the Hasting Center's report [100], identifies the goals of ethics education as follows: "to provide students with those concepts and analytical skills that will enable them to grapple with broad ethical theory in attempting to resolve both personal and professional dilemmas, as well as to reflect on the moral issues facing the larger society" [100, p. 48], "providing them with the tools for a more articulate and consistent means of justifying their moral judgements and of describing the process of their ethical thinking" [100, p. 54].

The Hasting Center's report emphasizes that the goals of ethics education should not be to change students' behaviour or to make students passively accept a teacher's beliefs, institutional or professional convictions, but rather the goals of ethics education should be to help students establish their own convictions and to help them to develop the skills for analysis and critique of their own and others' convictions.

Talking more specifically about the goals of professional ethics education, it is common to cite Harris [32], who states these goals as follows: 1) to stimulate the ethical imagination of students; 2) to help students recognize ethical issues; 3) to help students analyze key ethical concepts and principles that are relevant to the particular profession or practice; 4) to help students deal with ethical disagreement, ambiguity, and vagueness; 5) to encourage students to take ethical responsibility seriously; 6) to increase student sensitivity to ethical issues; 7) to increase student knowledge of relevant standards; 8) to improve ethical judgement; 9) to increase students' will power.

Many skills to be learned in professional ethics classes are metacognitive in nature, because they go beyond ethics-specific domain knowledge and skills, and involve thinking about thinking. Some of these skills are the skill of describing one's own ethical thinking, critically appraising one's own and others' convictions, reflecting on conflicting arguments and making decisions based on this reflection. The foundation researcher in metacognition, Flavell, considered many of these skills to be metacognitive in nature, and important for making wise and thoughtful life decisions [18].

2.2.2 Pedagogical Framework of Ethics Education

Professional ethics education is concerned with the development of various skills in students: recognizing relevant ethical issues, moral reasoning skills, decision making skills, etc. Hence, important questions for educators and researchers are how to organize learning environments that will foster the development of these skills, and what specific instructional interventions and features of learning environments are responsible for this development.

Analysis of case studies has been recognized as one of the most effective pedagogical approaches to achieve the goals of ethics education [33, p. 18]. Through the analysis of case studies, students can practice various moral problem-solving skills: recognition of ethical issues, identification of possible solutions to resolve the dilemma of a case study, moral reasoning and resolution among others [25, 33].

One of the most effective ways to organize analysis of case studies is through peer discussions [70, 84]: "the process of participating in arguments or even of listening to others argue and justify their opinions or solutions may be enough to enhance learning, even without in-group teaching, explanation, or assessment" [96, p. 188]. "Group discussions of controversial moral dilemmas ... provide concentrated practice in moral problem solving that is stimulated by peers challenging one another's thinking, reexamining of assumptions, being exposed to different viewpoints, building argumentation, and responding to counterarguments" [51, p. 44]. "The opportunity to learn that is provided by collaboration is the opportunity to analyze multiple perspectives and to draw a conclusion from that analysis... Collaborative learning is learning from analysis of the other's perspective, and from the other's analysis of one's own perspective, and from a new synthesis of those analyses" [45, p. 179].

Researchers have further investigated what specific features of peer discussions maximize students' learning. A well-known researcher on moral development, and the inventor of the theory of stages of moral development, Lawrence Kohlberg, argued that it is the cognitive conflict and exposure to the next higher moral reasoning stage that moves children up in their moral development [44]. Kohlberg identified the following stages of moral development: 1) obedience and punishment orientation; 2) self-interest orientation; 3) interpersonal accord and conformity; 4) authority and social-order maintaining orientation; 5) social contract orientation; 6) universal ethical principles. His research group discovered that exposing students to the argumentation of the stage that was one level up from their own advanced the students' moral development, while exposure to lower stages or stages that were much higher that the students' own stages was not beneficial. This effect was named the "*Plus One*" effect.

Berkowitz and Gibbs [4] discovered other features of students' discussions that are congruent with their learning. Through their experimentation they found out that the best predictor of students' reasoning development are transitive discussions between peers. A transactive discussion is defined by the researchers as reasoning that operates on reasoning of another: "rather than merely providing consecutive assertions, discussants operate on each other's reasoning". "Transacts are characterized by listeners' efforts to integrate the speaker's statements into his own framework before generating a response. Transacts are responses that attempt to extend the logic of the speaker's argument, refute the assumptions of the speaker's argument, or provide a point of commonality or resolution between the two conflicting positions" [70, p. 90].

Nucci also highlighted conflict as a characteristic of effective moral discussions. Stage progress occurs more often in learners who disagreed about the moral solution to a dilemma: "consensus on the outcome reduced the likelihood that students would challenge or otherwise respond to one another's reasoning and thus reduced the impact of the discussion on students' existing notions of morality" [70, p. 90].

Nucci, similar to Kohlberg, also argued that stage disparity is another ingredient of effective moral discussions. But, unlike Kohlberg, he believed that the optimal distance for the development to occur should be on the order of one-half stage. And this stage disparity is a typical disparity among classrooms students. This implies that "normal heterogeneity among students is sufficient for effective moral discussion" [70, p. 90].

In summary, analysis of case studies supplemented by group discussion has been recognized as one of the most effective pedagogies to develop students' ethical problem-solving skills. The features of group discussion that promote development most are: interaction with peers who disagree about a resolution to a given dilemma, interaction with peers whose moral reasoning is of slightly higher stage, and transactive discussions with peers based on the reasoning about the reasoning of each other.

2.2.3 Evaluation of Student's Ethical problem-solving

Assessing students' ability to recognize and resolve ethical dilemmas is not trivial. As ethics specialists acknowledge, ethical dilemmas are often complex, open-ended and ill-defined, and "to date, methods to assess students' ability to resolve ethical dilemmas remain largely undeveloped" [93]. Nevertheless, several assessment schemes have been proposed for the evaluation of students' case analyses.

Lu [51] in her Ph.D thesis evaluated students' ethical reasoning based on four components: 1) identification of ethical issues (ethical sensitivity); 2) adoption of multiple viewpoints (ethical viewpoint); 3) resolution of ethical dilemmas (ethical options); 4) justification of decisions and actions (ethical justification).

Graves et al. [29] used a mixed method approach to evaluate students' ethical decision making. First, the researchers conducted a qualitative analysis of students' experiences in the domain of interest, and developed a thematic model based on this analysis. Based on the thematic model the researchers developed Ethical Perceptions Scale (EPS), which allowed them to obtain quantitative measures on individual and group perceptual/interpretive preferences for ethical decision making.

Another scheme was developed by Shuman et al. [93], and included the following components:

1. Recognition of dilemmas — a student's ability to identify and frame the key dilemmas in a case study.

2. Information — a student's ability to use facts from the case study, and bring in information from his own experiences.

3. Analysis — a student's ability to include citations to analogous cases with consideration of risk elements with respect to each alternative.

4. Perspective — a student's ability to hold a global view of the situation, considering the perspectives of the employer, the profession, and society as well as the individual who is the focus of the case.

5. Resolution — a student's ability to consider potential risks and/or public safety concerns, and their ability to propose a creative middle ground ("win-win" situation).

Shuman and colleagues proved the validity of their assessment scheme by demonstrating the high consistency between raters' evaluations of student works, and by demonstrating distinctions between pre- and post-assessment results in a semester long ethics course [24, 93].

Goldin, Ashley and Pinkus [24] further elaborated the first item of this assessment scheme — a student's ability to identify and frame the key dilemmas. They say that a student frames the case when he recognizes all ethical issues or concepts involved in this case, and when the student is able to label, define and apply all these concepts in the case analysis. "A concept is said to be labelled 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" [24, p. 15].

2.2.4 ITSs and Computer Tools to Support Students in The Analysis of Case Studies

Students often find analysis of case studies to be challenging partly due to inadequate practice and the lack of individually tailored feedback, both of which are difficult to render in traditional classroom settings. A number of computer tutoring systems and tools have been developed to address this problem by providing students with multiple opportunities to practise, to get individualized feedback, and to participate in collaborative learning.

One thing to notice is that ethical decision making is a complex ill-structured solving process. Students' ethical analysis is represented as a text, there are no constraints to solutions and arguments students can come up with, as there is no predefined algorithm of ethical decision making that students can follow. All this makes it infeasible to construct an ITS for ethics in accordance with the classic ITS model. Similarly, it is not feasible at this time to support the whole process of ethical decision making, considering the overall complexity of it. Thus, computer systems described in this section have adopted alternative techniques to support students, and are limited to support only some part of their ethical decision making. Nevertheless, the creative approaches that these systems have employed to support learning in the ethics domain make them worth studying.

Systems Supporting the Process

A number of systems have been developed to support the process of ethical decision making either by providing a scaffolding structure, guidance or some other help during this process.

Supporting Through Structure. One of the ways to help students in the analysis of case studies is to structure students' case analysis. A system can walk students through the steps of ethical analysis by providing instructions and asking students to fill in predefined forms. One of the first systems with this approach was Ethos [83]. Ethos would guide students in their ethical analysis based on the procedure developed by Harris et al. [33].

A similar approach was adopted in the PETE system developed by Goldin et al. [23]. The only enhancement of Goldin's system over the Ethos system was that the PETE system stored responses of past students, so a student after finishing a particular step can compare his or her response with responses of the students from the past.

The Agora system[1] takes a step further by giving instructors the possibility to customize the ethical analysis structure that students should follow. Agora, an electronic tool for education in ethics and technology, was developed by three Dutch universities, and was used in teaching engineering ethics to thousands of students [102]. The Agora team had a goal to develop a system that would not limit the ethical analysis process to the application of a single ethical theory or principle, and to allow more than one way to analyse a case. Thus, Agora does not offer a fixed sequence of steps, but allows instructors to use different sequences, depending on the kind of analysis to be carried out. An instructor builds her model for ethical analysis of a particular case study using a large variety of available steps (Figure 2.8). Students analysing the case study will follow the steps predefined by the instructor for this specific case study. The Agora software also has predefined driving questions for each step that can facilitate students' progress.

Supporting for Framing a Case. Analysis of a case study begins with identifying ethical issues or concepts of the case, what is called "framing the case". A person frames the concept by claiming through the labelling process that particular concepts constitute the frame through which the case is to be viewed [25]. The next step is to define these concepts in general. The final step is to explain how the definition maps to the given case. Thus, the framing process consists of labelling concepts, defining them, and applying them to the case.

Goldin et al. [25] developed an instrument that facilitates students' development of skills necessary for the framing of a case. The instrument is based on the collaborative peer reviewing system. Students are given a case to analyse, and are asked to frame the case by labelling, defining, and applying ethical concepts to the case. Then, students review each other's case analyses assessing how well authors label, define, and

[1] http://www.ethiekentechniek.nl/site/

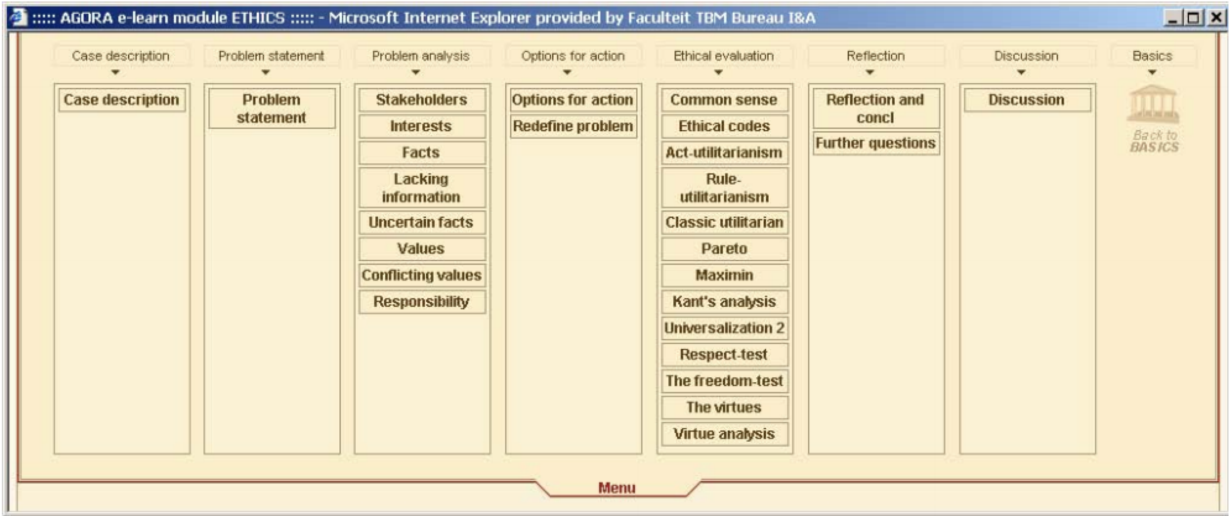


Figure 2.8: Agora interface: the set of available steps an instructor can choose from (reprinted from [102, p. 285] with permission from Springer).

apply ethical concepts. The system facilitates the assessment by offering an assessment form to fill out, and providing examples of the definition and application of concepts.

The authors' plan was to enhance the system by providing a Reviewer Aide functionality. The Reviewer Aide would use machine intelligence techniques to detect what concepts have been defined or applied in the student's analysis, and to locate existing case analyses similar to the new one. The results of the realization of this plan, and the results of any formal assessment of the Goldin et al. instrument are unknown.

Providing Content Help. Another way to support students in a case-based problem-solving is to provide them with various content help. The content help may have a form of summary of guidelines, standards, codes of professional ethics, or traditional ethical principles. As Robbins, Wallace, and Puka have demonstrated [78], even this simple type of support can improve students' ethical analysis. These researchers summarized and simplified traditional ethical principles of Virtue Theory, Utilitarianism, Kantian deontology, and the Ethic of Care, and made this information available online at any time through their web-based decision aid tool. The students who used their tool produced better solutions to a given IT ethical dilemma: they were able to identify the case's main issue more often, they discussed principles of the provided ethical theories more often than participants who did not use their tool, and they put more effort as measured by word count into their analysis.

Retrieving similar cases. Studying similar cases can facilitate students in making more informed ethical decisions. The SIROCCO system developed by McLaren for his Ph.D dissertation [58] retrieves facts, principles, and past cases that are relevant for the analysis of new case studies. SIROCCO can also explain the reasons for its predictions. This retrieval in SIROCCO is based on a computational model that utilizes analyzed past cases. SIROCCO accepts new cases in a special format as an ordered list of facts with identified actors, objects, actions and time. Already analyzed cases in SIROCCO are also stored in this format, but in addition include components of the analysis: the protagonist whose action is questioned, relevant ethical codes, past cases supporting or conflicting with the case, a conclusion for the case (ethical, unethical, undecided), and other details [60]. Given a new case, SIROCCO retrieves relevant analyzed case studies and codes cited by those cases based on the matching of the facts between a new case and analyzed cases. McLaren planned to integrate SIROCCO into an ITS for engineering students as a retrieval tool to support their reasoning and arguing about a case. The results of this integration is unknown.

Systems Supporting Argumentation

Argumentation is an integral part of ethical decision making, and assessing students' arguments and providing feedback on them is an important feature of an ethics ITS. Methods to enhance students' argumentation skills were extensively investigated in the ITSs for the legal domain. While the goals of the legal and ethics argumentations are different (in legal reasoning a student needs to argue a given position, in ethical learning the goal is to integrate as many positions as possible and to develop your own position), methods developed for the legal domain can be utilized in the ethics domain as well, and hence are worth reporting here.

Scheuer et al. [81] surveyed various techniques that computer argumentation tools use to analyse students' arguments and provide feedback on them. The most common technique of argument analysis is to ask students to represent their arguments using diagrams. The computer assessment of students' arguments is done using the arguments' structure, their syntactical and graphical representation, but without considering their content. Accordingly, corrective feedback is provided if the structure of an argument is unacceptable (e.g circular arguments, invalidly connected contribution types) or incomplete (e.g. a required contribution type is still missing in the argument diagram). This assessment technique is employed in the systems Belvedere and LARGO [81].

A purely syntactic assessment is easy to implement, but this assessment is limited to problems of a general type, and cannot be used to provide problem-specific hints because it lacks the necessary domain knowledge. Thus, some systems (Belvedere, Rashi, CATO, LARGO, etc.) attempt to represent this domain knowledge by having an expert argumentative model for certain problems. A student's solution to a given problem is compared with a predefined expert model, and if significant discrepancies are found, feedback is provided. The matching between the expert model and a student model as in purely syntactic systems is mostly based on the similarity of argumentative diagrams for a specific problem.

A few systems attempt to assess the content of arguments as well, and this is usually done using collaborative filtering. Students are asked to provide quality ratings for arguments of their peers, and these ratings are collected and numerically combined using collaborative filtering resulting in an overall quality rating. An example of an argumentation system with collaborative content assessment is LARGO [81].

Figure 2.9: A screenshot of the Conundrum system (reprinted from [56, p. 23] with permission from University of Saskatchewan).

Serious Games

McKenzie and McCalla [57] took a different approach to enhance students' ethics education — by creating a case study in the form of a serious game, where a student is embedded in a case and can reflect on the case as he makes the decisions affecting it. They developed a game called Conundrum (Figure 2.9). The Conundrum system presents the case as various acts, each of which consists of scenarios with issues and dilemmas that may arise in the information technology field. Students interact with these scenarios, making various decisions, and observing the consequences of these decisions: good or bad, anticipated or not. Conundrum promotes students' reflection, by allowing them after their initial exploration of their decisions to go back to any part of the played scenario and try different decisions. A goal is that the students by taking various roles, making different decisions, and seeing the consequences of these decisions in the game, will gain realistic experience and a deeper appreciation of the ethical issues that can arise in real life. Conundrum also provides a toolkit that students can use to develop their own acts, thus allowing them to appreciate deeper subtleties of ethics through synthesis of their own cases.

Systems Supporting Learning through a Case Adaptation

Computer systems can support students in learning ethics by adapting a case for a personalized learning experience. Adaptive Educational Interactive Narrative System (AEINS) [35] teaches ethics to 8-12 year old children in a personalized way by changing the storyline in the case. AEINS presents a story to a student containing different teaching moments (e.g. “do not lie”, “do not cheat”, “be sincere”, etc.), and asks the student to take actions and decisions in the story. The selection of the teaching moments depends on a current student model which shows what skills the student has already acquired. The selection is done by the pedagogical model which is represented as a set of production rules, describing how a student ideally would use the system and how the system reacts to his/her actions. Analysis of the log files of the AEINS system [34] demonstrated the effectiveness of this adaptation, since the majority of the learners were presented with the appropriate teaching moments, in which they were at first making wrong decisions.

AEINS also provides explicit and implicit feedback to kids as they make ethical decisions. An example of an explicit feedback when a student seemed to be unaware about the badness of stealing is: “don’t you know that stealing is wrong and you can go to jail”. An example of an implicit feedback when a student was following a good model of behaviour is: “your mum is proud of you!!”. Since this system is designed for kids, and the decisions kids are supposed to make are relatively straightforward and unambiguous, it was possible to organize this type of feedback in the system.

Organizing Collaborative Environments to Support Case-Based Learning

Smartly organized collaborative environments can be of tremendous help to students in the analysis of case studies, and in the development of their reflection and argumentation skills. Even as simple a feature as providing the possibility to see other people’s ideas can bring benefit to student. Lees [49] studied the benefits for students in seeing each other ideas in the Value Exchange software. The Value Exchange software[2] is used for teaching of ethics to undergraduate students in health-related degree programs. Students are given case studies, and asked to express their perspectives, thoughts, ideas, arguments, and values. As Lees demonstrated in her Masters thesis [49] seeing other people’s reports impacted the majority of learners. Similar thinking let learners expand their thinking, because they could see different reasons that other people used to justify their position. Opposite thinking widened a horizon, gave new insight into the case scenario, and seemed to be an incentive to strengthen one’s own arguments. For some participants the reading of other people’s analysis led to new understanding about themselves.

Even more benefits are achieved when students are given an opportunity to discuss a case study with each other. The study of Benbunan-Fich and Hiltz [3] demonstrated that collaboration through an asynchronous learning network enhances the quantity and quality of the solutions to an ethical case scenario, and increases students’ perception of learning in comparison with the individual learning.

Similarly, Voigt in his Ph.D. thesis [104] was investigating how to design electronic case-based learning that can bring benefits for students in terms of skills development, motivation and learning efficacy. One of his findings was that real time interactions contribute positively to group work.

Collaborations can be also organized around peer-reviewing, as it was done in Goldin et al.’s system [25],

[2] http://aut.values-exchange.co.nz/

where students provide reviews to each other's framings of a given ethical case. This peer-reviewing also opens up the possibly for students to practice and improve their assessment skills.

2.3 Summary

This chapter provided an overview of the literature and research on ill-defined domains devoted to defining ill-defined problem spaces, distilling their distinguishing characteristics, formulating the skills necessary for solving ill-defined problems, and developing domain, student, and tutoring models of ITSs that adequately address the needs of ill-defined domains. While past research has produced many insights and interesting approaches, the work in the area of ill-defined domains is still quite preliminary. Many proposed approaches were either limited to a very specific domain, and could not be easily generalized to other ill-defined domains, or were not thoroughly tested, and even for some approaches the evaluation was completely unknown.

This chapter also summarized more specific research on ethics and professional ethics domains: the goals of ethics education, effective pedagogical practices to achieve them, computer systems and tools developed to support the analysis of case studies which is a common pedagogical practice in ethics education. Considering the sheer complexity of the ethics problem-solving process, and its difficult rendering for computer processing, it came as no surprise that the computer systems developed here were limited to support only some small parts of ethical decision making, had very constrained interfaces for students' input, used other human raters for the assessment of students' input, and proposed alternative tutoring techniques affordable for implementation.

The reviewed literature demonstrated that building a tutoring system for an ill-defined domain, and specifically for professional ethics, is an interesting and difficult challenge. The rest of the thesis is devoted to taking on this challenge, where we will propose and evaluate a tutoring system for supporting students in the analysis of case studies.

Chapter 3

Thesis Statement

The goal of this dissertation is to develop techniques for building an ITS for supporting students in the analysis of case studies for professional ethics education. The terms "ITS", "intelligent tutoring system", "tutoring system", and "computer-based learning environment" will be used in this dissertation interchangeably.

3.1 Research Questions

A general research question is:

> How to design a computer-based learning environment to support ethical problem-solving? In particular, what features and interventions of a computer-based learning environment can effectively support students in the analysis of case studies?

This major research question can be divided into three more specific research questions:

- RQ1: *What processes and interventions of a tutoring system can effectively support students in their individual ethical problem-solving?* What are the best ways for a tutoring system to support a learner while he or she works alone on a case analysis?
- RQ2: *How should collaborative learning environments be organized to support ethical problem-solving?* What features of collaborative environments trigger productive interactions among students for developing their analytical and metacognitive skills and for broadening their perspectives?
- RQ3: *What are ways to diagnose students' arguments presented in a textual form?* All system interventions begin with the understanding and diagnosing a student's input. The question is how can we get away from the need for full natural language understanding, while at the same time making the diagnosis of students' textual arguments accurate enough for effective interventions by a learning environment to occur?

3.2 How to Answer the Research Questions

In this section we will supply and justify the criteria for determining how the above-stated research questions can be answered. These criteria will also determine how deeply these questions must be answered to fulfil the goals of this research project.

RQ1 and RQ2 will be answered if we identify system interventions and ways to organize collaborative learning environments that effectively support students in ethical problem-solving.

RQ3 will be answered when we identify the most accurate way(s) to diagnose students' arguments that allow the effective interventions to occur. For our proposed interventions (see the next Section 3.3), the diagnosis of students' arguments is equivalent to the comparison of students' arguments with each other and with the system's arguments to determine if the given arguments are focused on the same or different issues. We are seeking ways that result in the highest accuracy in determining if the two given arguments are on the same or different issues as judged by a human expert. One thing to note here is that we do not intend to develop a new algorithm for diagnosing students' arguments. Our intention is to use already developed text comparison algorithms, such as word-based search, Latent Semantic Analysis, etc., and identify the most effective among them for our needs.

How can we measure the effectiveness of proposed interventions for RQ1 and RQ2? In measuring the effect of the interventions of a new computer-based learning environment, it is common in the AIED/ITS community to talk about learning gains. Students take a pretest, use the new computer environment, and then take a posttest. The difference between students' results in the pretest and posttest are called learning gains, and attributed to the effectiveness of the new learning environment. This works well for problems with single definite solutions, such as many maths or quantitative physics problems.

However, for ill-defined domains including the ethics domain, this approach for measuring the system's impact is not possible, because of the absence of formal or well-accepted methods to verify solutions, a lack of criteria by which solutions are judged, and disagreement even among domain experts regarding the adequacy of solutions [53]. Specifically for the ethics domain, methods to assess students' ability to resolve ethical dilemmas remain largely undeveloped; already existing assessment schemes require trained coders, and have been tested to be sensitive only for measuring learning gains across a whole semester of study [93]. All this makes it impractical to use the standard pretest/posttest procedure to objectively evaluate the learning gains of students for our intended short-run interventions in Umka. Thus, instead we adopted different metrics for the evaluation of the effectiveness of the system interventions:

1. *Measuring productive student behaviors.* Rather than measuring the students' learning, we will measure the extent to which the system was able to create necessary students' experiences, and foster students' behaviors that are associated with higher learning. Section 2.2.2 on pedagogy in ethics and moral education listed the behaviors that are associated with students' moral development: participating in arguments, justifying opinions, responding to counterarguments, re-examining assumptions, analyzing multiple perspectives, etc. Thus, the more the system was able to stimulate students' productive behaviors, the more effective are the system's interventions.

   These productive behaviors of students are, in fact, metacognitive processes. To analyze a case study, students need to be aware of and regulate their ethical thinking: to question their own assumptions, analyse their own arguments and motivations, make sure they have covered all the facts, have not factored in their own beliefs or prejudices too strongly, have uncovered all the possible directions for analyzing the case, have taken measures to improve their analysis through interaction with other students. In section 2.1.2 we demonstrated that metacognitive processes are important for solving ill-defined problems. Thus, measuring how strong students' metacognitive skills are, we indirectly measure the students' ability to solve ill-defined problems in ethics.

2. *Measuring students' attitudes towards different system features.* This is done by asking students to evaluate helpfulness of various system support features, and tracking students' usage of them. The students' frequent usage of certain support types most likely indicates their effectiveness in helping students. A high evaluation of the helpfulness of the support types by the learners can be an additional indicator of their helpfulness.

3. *Measuring the changes in a student's analysis as the result of the system interventions.* This can be accomplished by comparing the student's initial individual analysis before any system interventions with the student's final work after the interventions. Based on the discussion in Section 2.2.3 on the evaluation of an ethical analysis, this comparison involves identifying differences in the number of issues and arguments presented, differences in the quality of the discussed issues and arguments, and differences in the justification of the final resolution.

4. *Comparing with other systems, or the same system with different types of interventions.* Contrasting students' recorded behaviors, their attitudes from questionnaires, and resulting changes in their analyses across different systems or across different interventions of a given system allows the relative assessment of the effectiveness of given interventions in comparison with other interventions, and allows the identification of the most effective system interventions.

3.3 Proposed Approach

The research will be approached from the social constructivist perspective, that states that individuals create their own new understandings through meaningful social interactions with others [72]. Particularly, peer interactions with sociocognitive conflicts lead to higher levels of reasoning and learning [72].

The solution that we propose to the problem of how to support students in the analysis of case studies is to design and build a computer-based learning environment called Umka where students can work on a case analysis both individually and collaboratively. Allowing an individual analysis of a student lets him to represent his own point of view about a given case study, and to track the progression of his point of view. Allowing students to collaborate on a case brings extra benefits for students of engaging in healthy competition, seeing alternative points of view and different arguments, sharpening analytical skills by giving feedback to peers, and honing metacognitive skills of reconsidering and broadening the students' own points of view in light of the views of the peers.

Considering the scarcity of established effective techniques for tutoring and building tutoring systems in ill-defined domains in general, and in the ethics domain in particular, we adopted an exploratory approach in our research study. Drawing from the AIED and pedagogical literature, and techniques explored in other tutoring systems for ill-defined domains, and imitating human tutoring interventions in real classroom settings in ethics classes, we identify, propose, implement and test various support techniques in Umka. These techniques are to be further refined and tested in each subsequent version of Umka.

To answer RQ#1 on the ways to support a learner in individual case analysis we implemented a number of tutoring support types in Umka version 1 (Chapter 4). These support types include guidance on the steps of the ethical analysis, various kinds of system feedback on students' arguments, and hints on important case arguments that the student missed. We then measured the impact of the proposed support types using the metrics discussed in the previous section (Section 3.2), and identified the most effective support types for individual case analysis.

To answer RQ#2 on the ways to organize effective collaborative environments for ethical problem-solving, we also explored a number of approaches. The initial Umka's collaborative support to students was in the form of suggestions to consider certain "helpful" arguments of their peers (Chapters 4 and 5). This support type was later augmented in version 2 of Umka with a visualization of students' positions on a given case study (Chapter 5). In the last, third version of Umka we also examined the effectiveness of the system suggesting to students certain other students they can interact with, interaction with whom was deemed to be "helpful" by the system (Chapter 6).

The implementation of the proposed support types is mainly based on correctly diagnosing students' arguments. By diagnosing students' arguments we mean comparison of students' arguments with each other and with the system's arguments. We proposed to diagnose students' arguments based on the interworking of an interface design that allows positive and negative arguments to be indicated by the student, and a text comparison algorithm that allows the content of the arguments to be compared (Chapter 4). To answer RQ#3 on the ways to diagnose students' arguments, we evaluated several text comparison algorithms for judging the semantic similarity between short student arguments. These algorithms included one that is already known in the AIED community — Latent Semantic Analysis, and a relatively new algorithm called Weighted Textual Matrix Factorization.

Each of the next three chapters will discuss in depth the three versions of the Umka learning environment. They will demonstrate the progression of Umka from version 1 to version 3, along with the features that were being augmented with every version. The chapters will also present the experiments, and the results specific for each version.

Chapter 4

Version 1 of Umka

This chapter presents the first version of Umka with its initial design considerations in all essential components of a tutoring system: an interface, a domain model, a tutoring model, and a student model. The chapter also presents the evaluation of Umka version 1 carried out in our first experiment in the fall of 2011.

4.1 Interface

4.1.1 Justification of The Chosen Interface

One of the challenges in designing an interface for a tutoring system in an ill-defined domain is to make the interface restrictive enough to enable a necessary level of the computer processing of students' input, but at the same time flexible enough to enable productive natural students' work and interactions.

Analysis of a case study is carried out in written text. A case study analysis encompasses possible resolutions to a given case study dilemma, arguments for and against a particular resolution, and a conclusion that integrates all possible resolutions and arguments, and explains why a chosen resolution is the best.

There were two main requirements for the design of Umka's interface. On one hand, for the computer processing of a student's input, the interface should provide a way to easily dissect constituent parts of a student's analysis: separate arguments, a chosen resolution, a conclusion, etc. On the other hand, it should allow students to elaborate their ethical analysis in a natural human language. To address these two interface requirements, we came up with an interface representation as a grid (Figure 4.1). The interface grid had predefined parts for writing arguments, choosing a resolution, writing a conclusion. Hence, these parts could be easily separated. Moreover, this structure of the interface gave students freedom to write their analysis in a natural language; the students just needed to put parts of the analysis into the appropriate predefined interface slots.

Moreover, to further distinguish whether a student makes an argument for or against a certain case study resolution, we incorporated a *for* and *against* distinction in the interface. We asked students to provide arguments *for* the chosen resolution in the left column of the arguments' table, and arguments *against* it in the right column (see the central part of Figure 4.1).

In the first version of Umka we also asked students to label their arguments as a rule, a consequence or an idea (the left-middle part of Figure 4.1) to enable more intelligent system feedback on students' arguments.

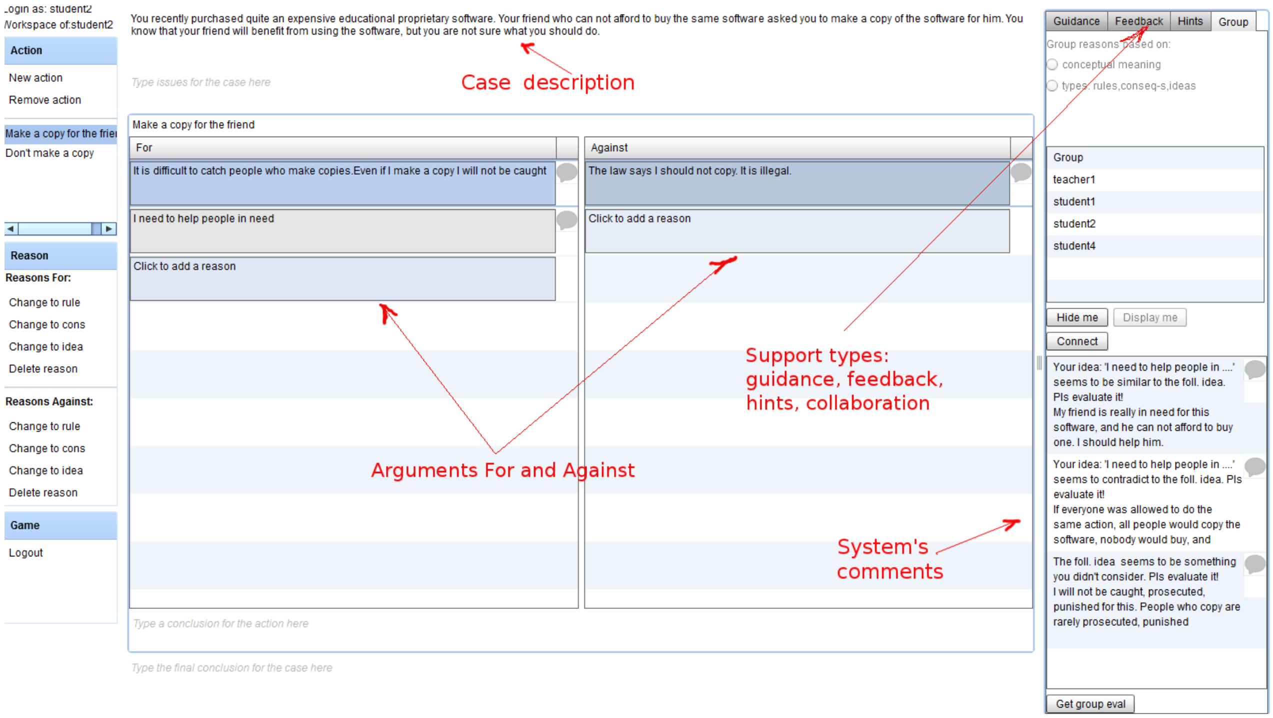


Figure 4.1: Interface of version 1 of Umka

A rule is an argument for or against a chosen resolution that illustrates a certain law, moral or legal. A consequence is a positive or negative outcome that has happened as a result of a chosen resolution. An idea is any other type of an argument apart from a rule or consequence for or against a chosen resolution. However, in the later versions of Umka we abandoned this labelling after seeing how it complicated students' work.

The proposed organization of the interface allowed the system to distill all distinct parts of a student's analysis including the labelling of arguments for or against a particular resolution. An extra benefit of this interface arrangement lies in encouraging students to structure their own analysis by breaking it up into distinct arguments for and against a certain case proposition, by considering different sides of a case, and by summarizing their argumentation into a final conclusion.

4.1.2 Student's Workflow

This section explains a student's steps in the analysis of a case study in the Umka system. Students are given a case study to analyze. For example, a case study from our first experiment presented a dilemma involving issues with copyright protection. The case study description goes like this:

> You recently purchased a quite expensive educational proprietary software. Your friend who can not afford to buy the same software asked you to make a copy of the software for him. You know that your friend will benefit from using the software, but you are not sure what you should do.

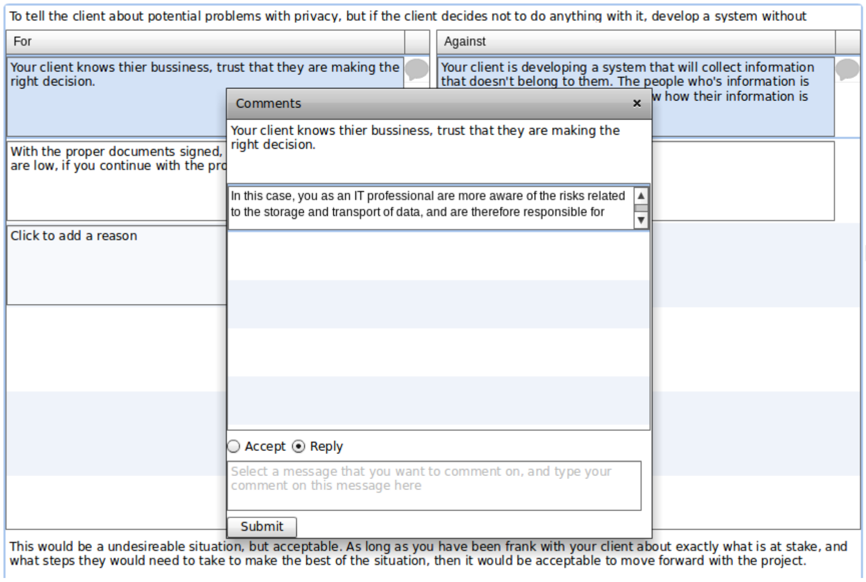


Figure 4.2: Students' interactions with each other in the Umka system: clicking on a circle sign next to an argument, a student can either agree with this argument of the viewed student by choosing the "Accept" option, or disagree with it by choosing the "Reply" option and providing a critique.

A case study in Umka is supplied with some predefined ways to resolve the case study dilemma, but learners can supply their own resolutions as well. For the above case study the predefined resolutions were:
1) Make a copy for the friend.
2) Don't make a copy.

First, students analyze a case study individually: they propose a resolution for a case study, provide arguments why a particular resolution is good or bad, and supply a conclusion justifying their choice of the resolution based on the provided arguments. The students are asked to consider different sides of the case: several potential resolutions, and both arguments for or against them. The students are encouraged to request different help types available in Umka, including feedback on their arguments (the right-upper part of Figure 4.1), and to modify their analysis according to the received help.

After the individual stage the students collaborate on a case study: they access analyses of each other, and read and comment on each other's arguments. The students can agree with comments or arguments of others, or disagree and provide critiques (Figure 4.2). These critiques can run as a dialogue between the students about a particular argument. Based on the peers' case analyses the students studied, and based on the feedback they received from their peers, the students can modify their own analyses by addressing the received feedback, or broaden their analyses with new arguments seen in the work of their peers.

4.2 Domain Model

The system's domain knowledge consists of several case studies representing professional ethical dilemmas. The representation language of the domain knowledge is English. This simple representation is easy to augment with new arguments, but as a downside, it means that the system must "understand" English in some sense, which in our case means matching the natural language of the students to the natural language of the arguments stored in the system knowledge base (Section 4.4).

For every case study the system stores possible resolutions for the case study dilemma. In Umka version 1 we also stored all possible arguments for and against a particular resolution that students may come up with. Every stored argument was labelled as a "good" argument, or a misconceived argument. For good arguments the system stored hints attached to them, and for misconceived arguments the system stored challenging questions to correct the misconception — i.e. feedback. This feedback (both hints and challenges) was used to provide tutorial support to the students' arguments.

4.3 Tutoring Model

The tutoring model in Umka version 1 consisted of various system feedback on students' arguments in the individual part of the analysis, and system suggestions to consider helpful arguments of other students in the collaborative part of the analysis.

4.3.1 System Suggestions in the Individual Analysis

We hypothesized a number of support types that could be beneficial for a learner's individual analysis (Table 4.1, rows 1-6). These support types were based on the system's predefined answers, and were implemented only in Umka version 1. They included the following types: guidance on the steps of ethical analysis (Table 4.1, row 1); feedback that the student argument is good, when the argument was closely matched with a good system argument (row 2); feedback that the argument is original, if the system was not able to find any close match (row 3); asking challenging questions for the argument, if the argument was closely matched with a misconceived system argument (row 4); providing a counterargument to the student argument, if a similar argument was found in the system's knowledge base but on the opposite side of analysis, for vs against (row 5); and giving hints on good system arguments that the student has not yet considered in her analysis (row 6). The support types were provided based on a student's request: a student clicked a corresponding button whenever she needed a certain type of support.

4.3.2 System Suggestions Based on the Answers of Other Students

Previous research in computer-based learning systems for ill-defined domains has demonstrated the learning benefits of showing the students similar and opposite thinking of their peers [13, 49]. We pursued this idea

Table 4.1: Support types in Umka version 1

N	Stage	Type of support
1	Individual	Guidance on the steps of the ethical analysis
2	Individual	Feedback on a good argument
3	Individual	Feedback on an original argument
4	Individual	Feedback by asking challenging questions
5	Individual	Feedback by providing counterarguments
6	Individual	Hints on arguments the student didn't consider
7	Collaboration	System suggestions to consider similar arguments of other students
8	Collaboration	System suggestions to consider different arguments of other students
9	Collaboration	System suggestions to consider counterarguments of other students
10	Collaboration	Showing arguments of all students semantically clustered

in the design of Umka's support, and for the collaborative stage of the case analysis the following support types were developed (Table 4.1, rows 7-10):

- suggestions to consider similar arguments of other students, where *similar* means arguments on a same issue of the case study, and that are closely matched by a similarity algorithm (row 7). The similarity algorithm is discussed in the next section.

- suggestions to consider different arguments of other students, where *different* means arguments of other students that the student has not considered in her analysis, and which are found to be quite far from her own arguments by the similarity algorithm (row 8).

- suggestions to consider counterarguments of other students, where others' *counterarguments* mean arguments similar to the student's arguments, but on the opposite side of analysis, for versus against (row 9).

- showing arguments of all students semantically grouped based on the chosen similarity algorithm (row 10).

4.4 Student Model

Umka's tutoring support operates based on matching the students' arguments with the system arguments, or arguments of other students. Thus, an effective algorithm that could measure similarity or difference between two textual arguments was necessary.

4.4.1 Measuring Semantic Similarity Between Arguments

There are various definitions of semantic similarity between two pieces of text. Recognizing textual entailment and paraphrases are popular measures to gauge similarity between sentences. Textual entailment means that the meaning of one sentence can be inferred from another sentence. And two sentences are considered paraphrases, if they restate the same idea in different words.

For our task, we say that two arguments of students are different if they discuss different issues in the case study. For example, in a case where students may decide to argue that it is reasonable not to make a copy of proprietary educational software for a friend, three different arguments in favour of this proposition are: 1) "This is company's intellectual property, copying it is like stealing." 2) "The friend may not be able to use the results of their work because the software they used was pirated." 3) "The company that makes the software is able to do so because people pay for licenses. Not paying limits reduces their incentive (maybe?) and ability to produce changes, features, or more software." [1] The first argument talks about the intellectual property law. The second argument is about constraints that the pirated software may end up with. The third argument discusses economic consequences resulting from not respecting intellectual property. These are all separate issues in the case study, and hence these three arguments are categorized as "different".

We say that two arguments of students are similar if they discuss the same issue of the case study. These arguments can be paraphrases of each other, or entail one another. Examples of similar arguments about why not to make a copy of the proprietary software for the friend are: 1) "This is company's intellectual property, copying it is like stealing." 2) "A software is an intellectual property of software developers. Cannot be copied without permission from the owner." 3) "Copying the software I am violating a copyright law (It depends on the location where I am now)". Although these arguments contain different words, in fact, they all discuss the issue of intellectual property and copyright law, and hence, all these arguments are categorized as "similar".

To find if the students' arguments are similar or different from each other, we used a combination of Latent Semantic Analysis (LSA) and interface design.

Latent semantic analysis (LSA) is a technique for analyzing relationships between a set of documents in terms of how close they are in a semantic space defined by the words in these documents [46]. LSA is typically used in information retrieval and document clustering for collections of documents of a large size. Although not a typical use, LSA "in the small" has been utilized for evaluating students' answers against predefined system answers in several ITSs including Autotutor [28], iSTART [64], and R-SAT [22]. We also employed LSA in our system for this purpose, making use of the Gensim library [76] implementation of LSA with stopwords removed and a similarity threshold ranging between 0.5-0.85.

LSA works with a *bag of words* model without deep analysis of a sentence structure, and therefore is not very effective in distinguishing positive arguments (arguments for a certain proposition) from negative

[1] The example arguments in this section are actual students' arguments from the experimental study.

ones (against a certain proposition). As mentioned before (Section 4.1), we incorporated a positive/negative argument distinction in the interface, asking students to provide arguments for or against in separate columns of the arguments' table. Using this cross-interaction of LSA and the interface, Umka matched a student argument, which usually consists of 1-2 sentences, against a system argument and other students' arguments based on the following procedure:

- If the LSA similarity score between two given arguments is high (equal to or above the predefined threshold), and the arguments are on the same side for or against a certain proposition in the interface, the arguments are categorized as "similar" arguments.
- If the LSA similarity score is high, and the arguments are on the different sides of a certain proposition, the arguments are categorized as "contradictory".
- If the LSA similarity score is low (below the predefined threshold), and the arguments are on the same size of a certain proposition, the arguments are categorized as "different".

Umka finds system or other students' similar, different, or contradictory arguments to the given student's arguments, and presents them to the student in form of suggestions. The matching of the student's arguments with the system arguments could be used to assess what part of the system's domain knowledge a student covered, so constitutes an individual student model in the Umka system.

4.5 Experimental Results

Our first experiment, which we label *CS*, was run in the fall of 2011 with 23 University of Saskatchewan (UofS) graduate or senior undergraduate students, mostly computer science majors. Unlike all other following experiments this experiment was not a part of any ethics class, and the experiment participants were students who volunteered to take part. The experiment lasted for two weeks. During a first week the students analyzed a case study individually; during the second week they collaborated on the analysis.

The goal of our study was to answer our three research questions: (i) what types of support are pedagogically effective in the individual analysis?; (ii) what types of support are pedagogically effective in the collaborative analysis?; (iii) how effective is LSA in finding good matches. An additional goal was to investigate what support types are preferred by different categories of students. Sections 4.5.1–4.5.4 present the answers for these questions.

4.5.1 What Types of Support are Effective in the Individual Analysis?

To answer the questions of this section and the following section we needed to measure the pedagogical effectiveness of the proposed support types in Umka. The following criteria were used to measure the pedagogical effectiveness of different support types:

1. the impact of the support types on the students' productive behaviours and the students' answers (what was deleted, added, changed in the students' analyses)

2. the frequency of the usage of different support types by the students

3. the students' evaluations of the helpfulness of the support types as indicated in the post-study questionnaire

Table 4.2 demonstrates the evaluation of the effectiveness of the support types available in the individual student analysis. As can be seen from the table, the most helpful individual support type is the system hints on arguments that the student has not considered (row 6). This support type triggered the biggest number of productive behaviours, urging the students to consider and add more arguments to their analysis. The system hints were also used more often than other support types, and were rated the highest in the students' questionnaires.

The students found the system feedback on their arguments also helpful (Table 4.2, rows 2-5). In particular, the feedback containing challenging questions for their arguments brought about the the biggest number of changes to the students' analyses (row 4). The system support in the form of providing guidance on the steps of the ethical analysis didn't find much usage among the students, and neither did it produce a major impact on their work (row 1).

Table 4.2: Support types in the individual analysis and their pedagogical effectiveness. Results based on experiment #1.

N	Type of support	Students' behaviours triggered	Avg # of times used per student	Students' feedback
1	Guidance	2 args added	0.26	N/A
2	Feedback on a good argument	No effect observed	2.91	Helpful=6.2 out of 10
3	Feedback on an original argument	No effect observed		
4	Feedback - challenging questions	3 args changed		
5	Feedback - counterarguments	1 arg strengthened		
6	*Hints on new arguments*	*8 args added*	*5.04*	*Helpful=6.5 out of 10*

4.5.2 What Types of Support are Effective in the Collaborative Analysis?

Table 4.3 demonstrates the pedagogical effectiveness of various support types in the collaborative stage of the analysis. As can be seen from the table, the students used the system's suggestions to consider similar, different or counterarguments of other students more frequently than other types of support (rows 1-3). In the questionnaire the students indicated that the most helpful type of support was to see arguments of all students semantically clustered (row 4). The type of support that triggered the most number of desired students' behaviours could not be precisely determined from the results, as for two types of support no effect was observed, and the other two support types produced a rather small effect.

Table 4.3: Support types in the collaborative analysis and their pedagogical effectiveness. Results based on experiment #1.

<table>
<tr><th>N</th><th>Type of support</th><th>Students' behaviours triggered</th><th>Avg # of times used per student</th><th>Students' feedback</th></tr>
<tr><td>1</td><td>Similar arguments of other students</td><td>No effect observed</td><td rowspan="3">5.3</td><td rowspan="3">Affected your analysis =5.1 out of 10</td></tr>
<tr><td>2</td><td>Different arguments of other students</td><td>2 args added</td></tr>
<tr><td>3</td><td>Counterarguments of other students</td><td>No effect observed</td></tr>
<tr><td>4</td><td>Arguments of all students semantically clustered</td><td>2 args added</td><td>1</td><td>Helpful=7.7 out of 10</td></tr>
</table>

4.5.3 Measuring the Precision of the Text Similarity Algorithm

To evaluate the performance of LSA for our task of judging the similarity between short case arguments we employed three measures: (i) evaluation of the precision of LSA when compared to human expert judgement; (ii) comparison of the LSA precision with the precision of the keyword search; (iii) the students' evaluations of the relevance of the feedback provided through LSA.

Evaluation of the Precision of LSA When Compared to a Human Expert Judgement. We evaluated post-hoc the relevance of the various types of feedback Umka retrieved for the students. In our first experiment we didn't employ any external evaluators, and I myself was the only rater who judged the performance of the LSA based Umka's feedback. Row 1, columns 1-3 of Table 4.4 shows the precision of LSA in retrieving similar, contradictory or different system arguments when compared with a given student's arguments. Row 2, columns 1-3 shows the precision of LSA in retrieving similar, contradictory or different arguments from other students' arguments for the given student's arguments. We then compared the LSA precision with the precision obtained when a system or a peer's argument is randomly chosen in response to the given student's arguments — a random precision (Table 4.4, columns 4-6). As can be judged from Table 4.4, the LSA precision depending on the task was from 1.6 to 9 times higher than the random precision, reaching the average precision of 0.765. LSA performed best for the task of retrieving different arguments for the given student's arguments, and not so well for the task of retrieving similar arguments. The reason for this could be that there were many students' similar arguments on a case expressed using different words, and LSA failed in judging them to be similar.

Table 4.4: Precision of LSA in comparison with random matching.

	1.LSA similar	2.LSA contrad.	3.LSA diff.	4.Random similar	5.Random contrad.	6.Random diff.
1. Individual	0.32	0.27	0.80	0.17	0.17	0.12
2. Collaboration	0.52	0.33	0.73	0.08	0.08	0.08
Average	0.42	0.30	0.77	0.12	0.13	0.10

Table 4.5: Students' evaluations of the relevance of various support types.

N	Type of support	Relevant	Neutral	Not relevant
1	Guidance on the steps of ethical analysis	N/A	N/A	N/A
2	Feedback on a good argument	65%	3%	31%
4	Feedback by asking challenging questions	50%	17%	33%
5	Feedback by providing counterarguments	50%	16%	34%
6	Hints on arguments	75%	16%	9%
7	Similar arguments of other students	80%	15%	5%
8	Different arguments of other students	50%	30%	20%
9	Counterarguments of other students	63%	21%	17%

Comparison of the LSA Precision with the Precision of Keyword Search. Using the Apache Lucene search engine [2] we implemented post-hoc the keyword search based method for retrieving similar, different, and contradictory arguments for the given student's arguments. The average precision of keyword search integrated over all tasks was 0.245, which was lower than the average LSA precision of 0.2675. Thus, LSA performs better than the keyword search for our task.

Students' Evaluations of the Relevance of the Feedback Provided through LSA. We asked students to evaluate the relevance of the different feedback and help messages they received from the system. Table 4.5 shows the results of this evaluation for various support types. As can be judged from Table 4.5, on average students found 62% of the support messages relevant, 17% moderately relevant, and 21% not relevant.

To evaluate the performance of the LSA-based similarity algorithm, we used three evaluation methods. The obtained results revealed that the LSA precision is significantly higher than the precision of random matching, higher than the keyword search precision, and is found relevant by the students 62% of time. This

[2] Apache Lucene. http://lucene.apache.org/java/docs/index.html

could serve as a positive indication of the effectiveness of LSA.

4.5.4 Preference of Support Types by Different Categories of Students

Another evaluation that we carried out was the assessment of preferences in support types by different categories of students. In the post-study questionnaire we asked students to identify the following information about themselves:

1. Gender: •Female •Male
2. Age group: •less than 22 •22-28 •over 28
3. Have you ever taken any Ethics course before: •Yes •No
4. Did you know the procedure of the Ethical analysis before: •Yes • No
5. Would you call yourself well-versed in ethical issues of computing: •Yes •No

Table 4.6: Preference of support types by different categories of students.

N	Support types	Female	Male	Age<22	22<=Age <=28	28<Age	Ethics course-Yes	Ethics course-No	Eth_analysis Yes	Eth_analysis No	Well-versed-Yes	Well – versed-No
1	Guidance on the steps of ethical analysis	**3.00**	**1.24**	1.00	2.13	0.86	0.50	2.06	0.85	2.80	1.86	1.44
2	Feedback on a good argument	0.67	2.00	1.00	2.00	1.00	2.00	1.44	1.92	1.10	**2.57**	**0.22**
3	Feedback on an original argument	0.00	0.18	0.00	0.13	0.14	0.00	0.11	0.15	0.10	0.14	0.11
4	Feedback by asking challenging questions	1.17	1.65	1.00	1.67	1.29	3.25	1.22	2.08	0.80	1.79	1.11
5	Feedback by providing counterarguments	**0.83**	**2.82**	1.00	2.67	1.71	3.50	2.00	2.54	2.00	3.07	1.11
6	Hints on arguments	3.50	2.18	1.00	2.47	2.86	2.00	2.50	2.08	2.80	1.43	3.67
7+8+9	System's suggestions to consider similar, different, and counterarguments of other students	4.16	5.47	8.00	4.73	5.57	**1.25**	**6.22**	4.69	5.70	4.93	5.44
10	The system shows arguments of all students semantically clustered	6.50	7.23	9.00	7.13	6.57	8.50	6.70	7.23	6.80	**7.93**	**5.67**

We were interested in whether different categories of students (e.g. females vs males, younger students vs older students, students who have taken an ethics course before vs students who haven't, etc.) have different patterns of usage of various support types and were affected by them differently. For support types indicated in rows 1-6 of Table 4.6 we calculated the average number of times they were used or commented on by various groups of students. For support types in rows 7-10 we calculated how helpful they were found to be by various groups of students (a value between 0 and 10). The highlighted data in yellow represent statistically significant differences. For example, from row 1 we can see that females used the guidance support on average 3 times during the study, and this was statistically significant higher usage than that of males, who used the guidance support on average 1.24 times.

From the results in Table 4.6, it is clear that different categories of students indeed prefer different support types. Thus, male students are more responsive to counterarguments than female students (row 5), while females used guidance of the ethical analysis more than males (row 1). Students who find themselves well-versed in ethical issues of computing appreciated feedback on a good idea and presentation of all semantically clustered ideas more than their counterparts (rows 2 and 10). And finally, students who haven't taken an ethics course before are more affected by arguments of other students suggested by the system than students who have taken an Ethics course (rows 7,8,9).

4.5.5 Questionnaire Results About Umka

In the post-study questionnaire we asked students to provide their comments in a free form about the system. The students commented on different system features, and their comments are summarized and categorized in Table 4.7. The biggest number of comments involved the criticism of the interface — how user-unfriendly it was, and how it can be improved: "interface of the system not really user friendly", "I was confused when I was trying to select different things", "I couldn't find how to see my feedback to other students' arguments" (row 1). In contrast to the interface, Umka's collaboration features received more praise: "it was interesting what other people came up with", "interesting, specifically in terms of collaborative decision making"(row 2). Some students were also impressed with Umka's intelligence in finding good matches to the student's arguments: "Fun, was surprised to see the system was as smart as it was", "It looks pretty inteligent [sic]" (row 3). The overall impression from the system was mostly positive; the students found Umka helpful, interesting, and innovative: "it is helpful for study cases", "I think that this system helped us generate more thorough conversation and deeper ...","very interesting and useful" (row 4).

Table 4.7: Students' qualitative answers from the post-study questionnaire on different system features. Results based on experiment #1.

# of positive/ negative comments	**Comments**
1. Comments on the interface	
Pos: 3	nice interface and structure; Very well designed / implemented.; Was a tiny bit confusing at first, but the system was helpful in guiding me into figuring out what to do.

Neg: 15	interface of the system not really user friendly; It might be helpful to include the question into pop-up window when it comes to giving feedback; The group evaluation part was more difficult to use than the individual stage; Underlying system looks great but I think the user interface need to be designed more user friendly; When I connect to groups, all the answers on the screen look messy. Too many scroll bars; As a research tool, it's really good, but there are a few interface issues that bothered me while using the system; Interface layout design could have been improved; I think the interface need lot of work. I was confused when I was trying to select different things. Moreover, the system seem to run slow; I had difficulties with scrolling down some questions. Therefore I could not see the whole answere or comment; Improve the useability of interface (help menu, exampls etc); The application looks nice however it can be more "user-friendly"; Interface can be improved, e.g., adding more visual clues, colors, etc.; I couldn't find how to see my feedback to other students' arguments.; This website looked strange (text cut off on the top) in Safari but okay in Linux.; I think there should be either more help instructions embeded into the system, or its interface should be designed more friendly; The system could be further improved by providing documentation and more help items
2. Collaboration with others	
Pos: 3	It was interesting to see what other people came up with. There were many reasons in common between...; Good to see how others responded, but sometimes difficult to get an overall impression of the other ...; Interesting, specifically in terms of collaborative decision making, as it allows people to talk abo..
Neg: 1	I think it would have aided exploration a lot if there was a visualisation of how many responses there were, who had responded, and a way to easily browse responses.
3. Umka's Intelligence, accuracy of the similarity algorithm	
Pos: 4	suprised to see LSA perform critical reasoning so well, since it's based on mathematica; Fun, was surprised to see the system was as smart as it was; It looks prrety inteligent [sic]; Definitely was impressed at the intelligence of the system, I didn't expect that.
Neg: 2	Some of the groupings of all the arguments weren't quite right. But I think it was just overloaded with arguments that were all quite similar. For negatives, "it's illegal" was very common.; I guess sometimes some sentences are complicated or gramatically wrong which make it some what dificault for the system to find the exat pattern [sic].
4. General comments on the system	
Pos: 9	it is helpful for study cases, but so far it seems for me that it is similar to information representation; good... overall, I think that this system helped us generate more thorough conversation and deeper a...; interesting; good; interesting; Different yet interesting; It was a good system with new idea; impressive; Very interesting and usefull
Neg: 2	could be better; strange

4.6 Conclusions for Version 1 of Umka

In Umka version 1 we built a prototype of an ITS for the analysis of case studies in professional ethics education. We experimented with various kinds of tutoring support in Umka both for the individual and collaborative stages of the case study analysis. The main technology behind the designed support types was the cross-interaction of LSA and a specific structure of the interface. We evaluated the pedagogical effectiveness of the support types, and the precision of LSA in finding good matches for feedback. We also examined users' written evaluations of Umka.

There are several lessons we learned from the design and experimentation with Umka version 1:

- Various support types seem to produce unequal pedagogical effects. Thus, the students found it most helpful to see all students' arguments semantically clustered. The most frequently used support type was the system suggestions to consider specific arguments of other students. The type of support that affected the students' ethical analysis most was the system hints on arguments the students hadn't considered yet.
- Different categories of students preferred different support types, which could serve as a basis for Umka's potential personalization.
- The cross-interaction of LSA and the structure of the interface was found to be a fairly effective. LSA performed better than other methods such as a keyword based search. Moreover, the students' mostly positive evaluations of the relevance and helpfulness of the system's feedback can serve as an additional proof for the effectiveness of LSA.
- Umka's interface needed to be simplified and redesigned to be more user-friendly.
- The students found Umka useful, interesting, and innovative.

Chapter 5

Version 2 of Umka

We presented the experimental results from Umka version 1 at the Intelligent Tutoring Systems Conference 2012 [87, 88]. After the interaction with other researchers at the conference, for our future directions we decided to concentrate only on supporting the collaborative stage of the ethical analysis. To organize support types at the individual stage of analysis requires big effort in domain knowledge construction by considering all potential arguments learners can come up with for a given case, and potential feedback for these arguments. Therefore, other researchers with whom we interacted at the conference, including the author of another Ethics ITS, the Pete system [23], were harnessing the power of social collaboration to support students, compensating for the absence of the domain knowledge. We adopted a similar approach, and versions 2 and 3 of Umka are focused on organizing an effective collaborative environment for case study analysis.

We also intended to improve Umka based on the students' comments who used Umka version 1 in the first experiment. One of the directions for the improvement was to simplify the interface and make it more user-friendly. Addressing this comment, for the version 2 of Umka we eliminated several interface features, such as different types of arguments a student can supply, and made the interface simpler (Figure 5.1). Another direction was to improve collaboration features of Umka. A study participant from the first experiment suggested how it can be done: "I think it would have aided exploration a lot if there was a visualization of how many responses there were, who had responded, and a way to easily browse responses". Considering this suggestion, in version 2 of Umka, we introduced a visualization of students' positions.

Thus, the tutoring support in Umka version 2 was based on the newly introduced visualization, and system suggestions to consider similar, different, and contradictory arguments of other students. We tested version 2 of Umka in three experiments from the fall of 2012 to the spring of 2013.

5.1 Visualization of Students' Positions

> "If there is one secret of success, it lies in the ability to get the other person's point of view and see things from that person's angle as well as from your own" (attributed to Henry Ford [65]).

As the literature review suggested, an open group learner model that presents to a learner how he is doing in comparison with his peers, promotes a learner's reflection on his knowledge, skills, and progress in the group context, and encourages the learner's participation, and adoption of social norms (Section 2.1.4).

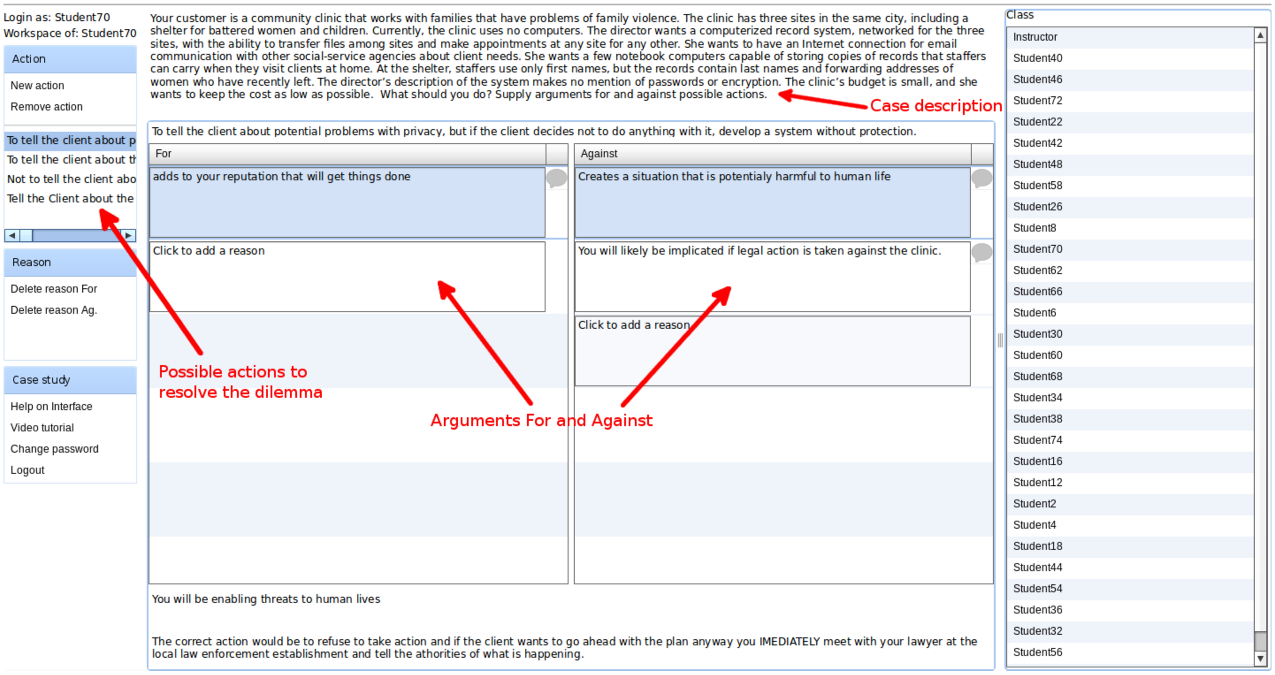


Figure 5.1: Interface of version 2 of Umka

We hypothesized that presenting to students their and their classmates' progress in ethical analysis could serve as another way to support them in ethical problem-solving. It should stimulate students' positive behaviors of reflection on their positions and the position of others, and students' productive interactions with each other. The question was how to organize this open group learner model, how to visualize students' positions, that is students' convictions about a given case?

To answer this question we consulted pedagogical studies in ethics education. The two important goals of ethics education are to help students establish their own convictions, and to help them to develop the skills for analysis and critique of others' and their own convictions (Section 2.2.1). We wanted to model the students' progress in achieving these goals in the visualization of the students' positions. We devised two factors that could be good indicators of how well-established a learner's position is, and how developed his analytical skills are. These two factors are 1) *the breadth of the learner's position*, expressed as the number of different arguments he considered for a case, which shows the learner's ability to look at the given dilemma from multiple perspectives, and 2) *the level of acceptance of the learner's reasoning and comments* by his peers and the instructor, which demonstrates the learner's ability to argue his own position and critique the positions of others.

Our visualization of an open group learner model represents every learner's position as a circle (Figure 5.2). Three things are reflected in this visualization:

1. The *size of the circle* reflects the breadth of the student's position, which is determined by the number of different arguments the student has for and against a particular action in a case study.

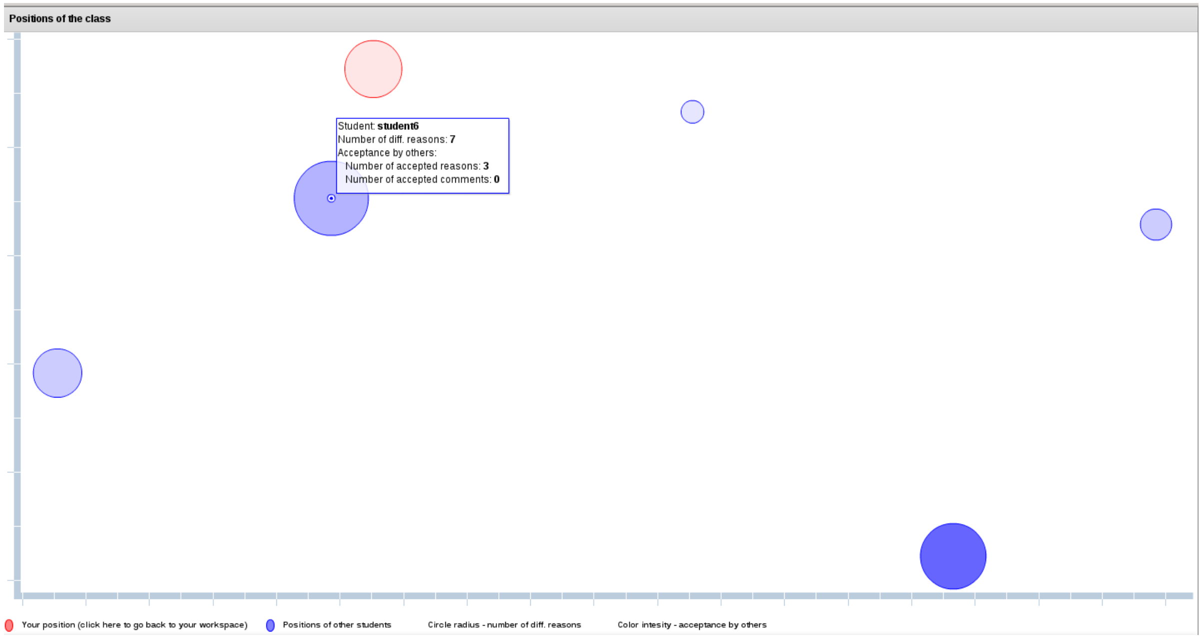


Figure 5.2: Visualization of students' positions. A student sees her position as a red circle, positions of others - as blue circles.

2. The *darkness of the circle* reflects the well-formedness of the student's position. The more the student's own arguments and his feedback on arguments of others are accepted by others, the more well-formed is his position, and the darker is his circle.

3. The *distance between circles* of different students reflects the distance or contrast in their positions. The more distant are two circles, the more different are the positions of the two respective students. The distance is determined based on the number of different arguments between the students.

Figure 5.2 demonstrates the visualization of the students' positions from one of our experimental classes. The X and Y coordinates in this visualization don't have any particular meaning; what is important is the distance between students' circles. In the visualization a student sees her position as a red circle, and positions of her classmates — as blue circles (Figure 5.2). By hovering over a circle of any classmate, the student can get information about the position that this circle represents: the number of diverse arguments in this position and the number of arguments and comments accepted by others (see the middle part of the Figure 5.2). By clicking on a circle of any classmate, the student can see the ethical analysis of this classmate: his arguments, case study resolution, conclusion. Students should aim to achieve a big and dark circle for the position they take, indicating that they have reached a broad and well-formed position about a given ethical dilemma through interaction with their peers. Thus, another objective for the visualization was encouragement of competition between students in achieving big and dark circles for their positions.

5.1.1 Measuring Semantic Similarity

In the proposed visualization the breadth of a student's position is defined by the number of diverse arguments the student has identified for a given case study. The distance between two students' positions is determined by the number of different arguments the two students had between each other. To calculate the breadth of a student's position and the distance between any two students' positions, we need to compare the students' arguments with each other and judge how semantically similar or different they are.

Similar to Umka version 1, for judging the semantic similarity between arguments we used the cross-interaction of the LSA and interface. Using LSA, Umka calculates the similarity between all arguments in a single student's position to determine the number of arguments that discuss different issues of the case. This will constitute the breadth of the student's position. Using LSA, Umka also calculates the number of different arguments between positions of different students. Unlike Umka version 1, in version 2 we didn't make the distinction between positive and negative arguments when judging differences between students' argumentation. If two students' positions contain many different arguments between them (regardless of whether these arguments are positive or negative), they will be displayed apart from each other in the visualization (Figure 5.2). If two students' positions contain many similar arguments between each other, they will be displayed close to each other in the visualization (Figure 5.2).

5.2 Experimental Results

5.2.1 Formulation of Hypotheses

Our goals for the visualization were not only to model students' knowledge. We expected students to reflect on the visualization, and we expected that this reflection would stimulate some positive behaviours in students. We wanted students to consider various issues of the case study, different reasons for and against possible case resolutions. We wanted them to look at the diverse arguments of their classmates, and to consider and incorporate some of these arguments into their own positions, and by doing this broaden their own positions. We also wanted students to practise evaluating and critiquing the arguments of each other, and to defend their arguments and overcome criticism. In short, we expected that students who had Umka's support in the form of the visualization and suggestions would demonstrate more of these desired behaviours of reflection and interaction with others, and as a result of these behaviours would expand their perspectives more than students who didn't have Umka's support. Our expectations can be distilled into the following hypotheses:

Hypothesis 1: Umka's visualization and suggestions foster productive interactions among students. Students with these support types, in comparison with students without these support types:

1. give more comments to the analyses of their peers

2. view more "helpful" analyses of their peers

3. spend more time in analyses of other students.

Hypothesis 2: Umka's visualization and suggestions foster students' metacognitive processes of reconsidering and broadening their positions. Students with these support types, in comparison with students without these support types:

1. introduce more changes to their analyses during the collaboration stage.

5.2.2 Description of Experiments

We ran three experiments to test the hypotheses and evaluate the second version of Umka:

- SIAST — this experiment was run in the fall of 2012 at the Saskatchewan Institute of Applied Science and Technology (SIAST) with 44 Computer Systems Technology students taking a *Computer Systems Ethics* seminar. SIAST is a local institution for post-secondary technical education and skills training; students at SIAST have a diverse age range from young people after high school to senior adult learners. We randomly broke up the class into two groups: a treatment group of 21 students used the Umka system with the visualization and suggestions, and a control group of 23 students used Umka without these features. Thus, the control group students used the same Umka system where they can input their analysis, and interact with their peers, with only Umka's suggestions and visualization missing. All students were given a week to analyze a case study in Umka.

- GSR960 — this experiment was run in the fall of 2012 with 16 University of Saskatchewan (UofS) graduate students taking the *Introduction to Ethics and Integrity* class. The whole class was an online only class, and the case analysis in Umka was one of the assignments in the class. The students were given a month to analyze a case study in Umka.

- CMPT408_2013 — this experiment was run in the spring of 2013 with 6 UofS undegraduate students taking *Ethics in Computer Science*. We compared the case study analysis made by the students in the Umka system with an analysis of another case study made by the students in a wiki system of Moodle[1]. In this wiki system the students worked together on a group analysis of a case study, and anyone could edit any part of the group analysis. The students were given a week to analyze a case study in Umka.

5.2.3 Testing the Hypotheses

Hypothesis 1: *Umka's visualization and suggestions foster productive interactions among students.*

We looked at three sub-hypotheses here for each separate type of interactions: hypotheses 1.1, 1.2 and 1.3.

[1]https://moodle.org

Table 5.1: Average number of comments per student, and the proportion of students who made at least one comment to their peers. Results based on experiments #2–4.

N	**Experimental group**	**Avg. # of comments**	**% of commenting students**
1	SIAST treatment	*4.7*	*62%*
2	SIAST control	1.3	26%
3	GSR960	1.4	44%
4	CMPT408_2013 treatment	*2.0*	*50%*
5	CMPT408_2013 control – wiki	0.0	0%

Hypothesis 1.1 *Students with Umka's support types, in comparison with students without these support types, give more comments to the analyses of their peers.*

We have calculated the average number of comments the students made on the arguments of other students, and the proportion of students who made at least one comment to their peers (Table 5.1). As can be seen from the table, the treatment groups with Umka's support types (rows 1 and 4) provided more comments about their peers' arguments than the control groups (rows 2 and 5). We found that the SIAST treatment group made statistically significantly more comments than the SIAST control group (the Mann-Whitney test: p=0.00834). Since for the GSR960 group (row 3) we did not have a control group for comparison, this group's data is presented for the purpose of contrasting with other experimental groups. The results demonstrate that the GSR960 group, which also had Umka's support features, was less interactive than the SIAST and CMPT408_2013 groups.

Furthermore, in the treatment groups with the visualization (Table 5.1, rows 1 and 4) the percentage of commenting students was much higher than in the control groups without the visualization (rows 2 and 5). Thus, all the experimental data support this hypothesis, and hypothesis 1.1 is accepted.

Hypothesis 1.2 *Students with Umka's support types, in comparison with students without these support types, view more "helpful" analyses of their peers.*

We analyzed what positions of their peers the students visited, and how many of these positions were helpful (Table 5.2). A "helpful" position for a student to view is a position of another student that either contains more arguments, has been accepted by others more, or is semantically far from the position of a given learner. We call these positions "helpful" for a learner, because by studying them the learner can broaden her perspective with new arguments that she has not considered, and see examples of strong arguments that have been accepted by others.

The experimental data showed that 68% of all the positions that the SIAST treatment students viewed were helpful positions (Table 5.2, row 1). Among the SIAST control students this number was smaller 66%

Table 5.2: Positions of classmates that students viewed. Results based on experiments #2–4.

N	**Experimental group**	**Proportion of helpful positions**
1	SIAST treatment	68%
2	SIAST control	66%
3	GSR960	76%
4	CMPT408_2013	50%

Table 5.3: Average time students spent studying analyses of other students. Results based on experiments #2–4.

N	**Experimental group**	**Average time in secs**
1	SIAST treatment	809
2	SIAST control	593
3	GSR960	332
4	CMPT408_2013	29

(row 2), but not statistically significantly smaller to claim that this difference was due to the effect of the proposed visualization. In the GSR960 group, the proportion of the viewed helpful positions was the highest reaching 76% (row 3). In the CMPT408_2013 group the proportion of the viewed helpful positions was low — around 50% (row 4). This can be attributed to the low number of students in this group, and consequently, to a lack of data for statistical testing.

Although, the treatment group demonstrated more of the desired behaviours of viewing helpful positions, the difference between the control and treatment group was rather small, which means we cannot statistically accept hypothesis 1.2.

Hypothesis 1.3 *Students with Umka's support types, in comparison with students without these support types, spend more time in analyses of other students.*

Table 5.3 demonstrates the average time spent by students in seconds viewing the analyses of the peers or commenting on them. The SIAST treatment group students (row 1) spent much more time studying the analyses of other students than the SIAST control group students (row 2). The average time spent in the analysis of other students by the GSR960 students was 332 seconds (row 3). The students of the CMPT408_2013 were much less active in viewing each other's analysis, spending on average 29 seconds (row 4).

Although the difference between the SIAST treatment and control groups in the time spent was not statistically significant (the Mann-Whitney test: p-value = 0.09947), p-value was small enough to demonstrate a favourable trend; and for some studies p-value of 0.1 could be considered a borderline value for judging statistical significance. Thus, although we could not statistically accept hypothesis 1.3, the data showed a clear positive trend towards accepting it.

Hypothesis 2: *Umka's visualization and suggestions foster students' metacognitive processes of reconsidering and broadening their positions.* Students with these support types, in comparison with students without these support types, *introduce more changes to their analyses during the collaboration stage.*

We calculated the number of arguments the students added to their analysis during the collaboration stage. Table 5.4 shows the average number of arguments a student added to his analysis during the collaboration stage, and the proportion of these arguments from the total number of the student's arguments. As can be judged from the table, the collaboration stage for the SIAST treatment group was more productive than for the SIAST control group. The SIAST treatment group students expanded their analyses on average with 4 arguments, which constituted on average 66% of their total number of arguments (row 1). The SIAST control group students added on average 3.5 arguments, which constituted 56% of their arguments. However, the difference between the control and the treatment groups was not statistically significant (the Mann-Whitney test: p-value = 0.399), and based on this hypothesis 2 cannot be accepted statistically.

Table 5.4: Effect of the collaboration: average number of changes introduced by the students in their own analyses. Results based on experiments #2–4.

N	**Experimental group**	**Avg. # of args added in collaboration**	**Avg. % of these args from the total # of args**
1	SIAST treatment	4.1	66%
2	SIAST control	3.5	56%
3	GSR960	3.8	46%
4	CMPT408_2013	12.5	57%

Rows 3 and 4 of Table 5.4 demonstrate the effect of the collaboration stage for the GSR960 and CMPT408_2013 groups. This effect was more prominent for the CMPT408_2013 students than for the GSR960 students.

The results for the hypotheses demonstrated the joint pedagogical effect of all Umka's support types — visualization and suggestions plus the students' interactions. We were interested to know how much the visualization was helpful by itself in isolation, and how much the system suggestions were helpful by themselves in isolation. To measure the exclusive pedagogical effect of the visualization we calculated how many times the students looked only at the visualization without accessing the actual arguments of other students at all, and then went back to their analyses, and added more arguments (Table 5.5). As it turned out, the visualization was quite effective tool by itself in encouraging the students to expand their perspectives. On average across three experiments, one third of all arguments made by the students was introduced after

Table 5.5: Effect of the visualization: the students expanded their analysis with new arguments. Results based on experiments #2–4.

N	**Experimental group**	**Avg. # of args added after seeing the visualization**	**Avg. % of these args from the total # of args**
1	SIAST treatment	3.0	38%
2	SIAST control	N/A	N/A
3	GSR960	2.0	25%
4	CMPT408_2013	9.2	46%

Table 5.6: The pedagogical effect of system suggestions. Results based on experiments #2–4.

N	**1. Experiment**	**2. Response rate**	**3. # of new args added**
1	SIAST	8%	0
2	GSR960	13%	1
3	CMPT408_2013	27%	0

seeing only the visualization (Table 5.5, column 4).

To measure the helpfulness of the system suggestions, we measured how many times the students followed up the system's suggestions, and replied to or ranked the suggested arguments of the peers (Table 5.6, column 2). We also measured the students' behaviours after these suggestions, if the students expanded their positions based on the suggested arguments (Table 5.6, column 3). As can be judged from Table 5.6 the helpfulness of the system suggestions was quite modest, the students ignored 73–92 % of them, and there was not much pedagogical effect observed.

Overall, the obtained experimental data were in support of all hypotheses on the benefits of Umka's tutorial model. However, we could obtain statistical significance for only two out of four hypotheses.

5.2.4 Measuring the Precision of the Text Similarity Algorithm

To validate how accurate the system was in judging if two arguments are about the same case issue or about different issues of a case study, we randomly picked 20 pairs of students' arguments from each experimental dataset: GSR960, SIAST, CMPT408_2013. Half of each dataset were 10 pairs, in which two arguments in a pair were judged to be similar by the system, and another half of each dataset, 10 pairs, in which two arguments in a pair were judged to be different by the system similarity algorithm. The system similarity algorithm was LSA in this case.

We employed an external rater 1, an interdisciplinary Ph.D student in Computer Science and Education, to judge which pairs represent similar arguments, and which represent different arguments. I myself — rater 2

— did the same evaluation. There was quite high agreement between two human raters: raters 1 and 2 agreed in 83% of cases. We took the evaluation of rater 1 as a baseline for correct classifications, and calculated the precision of the system as the sum of correct classifications divided by the total number of classifications. Table 5.7 displays the results of the calculations for three experimental groups. As the CMPT408_2013 students were not provided with the system suggestions displaying similar arguments of other students, the system precision for finding similar arguments of other students was not available for this experimental group (row 3, column 2). The last row of the table displays the average precision across three experiments.

Table 5.7: The precision of LSA in choosing similar and different arguments of other students for a given student's analysis. Results based on experiments #2–4.

N	1. Experimental group	2. Precision for similar	3. Precision for different
1	SIAST	0.30	0.90
2	GSR960	0.40	0.60
3	CMPT408_2013	N/A	0.83
Average precision		0.35	0.78

The performance of LSA for these three experiments was compatible to the performance of LSA for our first experiment (see Section 4.5.3), and similarly to the first experiment, was higher for the task of finding different arguments (average precision=0.78) than for the task of finding similar arguments (average precision=0.35). The average precision across all tasks was 0.6. This is similar to the AutoTutor system [26], where the correlation between AutoTutor and a human expert was approximately $r = 0.5$, whereas two human experts correlated approximately $r = 0.65$. In both Umka and AutoTutor this relatively modest correlation is likely a result of having texts that are too short for LSA to function at a higher level. But, in both cases even this level of precision turned out to be useful.

Later, we discovered another method, called Weighted Textual Matrix Factorization (WTMF) [31], for calculating the similarity between two pieces of text. This method was specifically designed to "extract nuanced and robust latent vectors for short texts/sentences, such as tweets, SMS data, short forum posts/comments", which ideally suited for our needs of judging similarity between short students arguments. We compared the LSA performance with the performance of WTMF. Table 5.8 demonstrates the precision of LSA as compared with the precision of WTMF in choosing similar and different arguments for the GSR960 experiment's data. As can be seen from Table 5.8, the WTMF method significantly outperformed LSA. For this reason we used the WTMF algorithm as the main tool to find similarities and differences between students' arguments in our future experiments.

Table 5.8: The precision of LSA in comparison with the precision of WTMF. Results based on experiment #3.

Text comparison method	Precision for similar	Precision for different	Average precision
LSA	0.40	0.60	0.50
WTMF	0.89	0.85	0.87

5.2.5 Questionnaire Results About Umka

We studied what particular features of Umka are useful for students and liked by them. In the post-study questionnaire we asked the students to rank and comment on Umka's suggestions and different visualization features.

Table 5.9 presents the students' quantitative rankings on various system's features. The header of the table displays three experimental groups, and the number of students who answered the questionnaire in each of these groups. As can be judged from the table, the students ranked as highest in helpfulness the visualization's feature to display positions of others as close or distant depending on their similarity with the students' own positions (Table 5.9, row 4, average rank=3 out of 5). The helpfulness of the system suggestions, and the visualization's ability to display different sizes of the students' positions depending on the number of arguments in them were ranked moderately (rows 1 and 2). The demonstration of the darkness of the students' positions in the visualization was found to be the least helpful feature by the students (row 3). Overall, the students found the circle visualization to be a moderately fair representation of their positions (row 5, rank=2.9 out of 5).

Table 5.9: Students' ranks of Umka's features from the questionnaire (scale: 0–5). Results based on experiments #2–4.

N	1.Question & average rank across all students	2. GSR960	3. SIAST	4. UofS_2013	5. Avg
	# of students answered	**10**	**14**	**1**	
1	How helpful were the system suggestions to consider various arguments of other students	2.9	2.7	1.0	2.7
2	How helpful was to see the size of your position in comparison with other students' positions	3.1	2.1	4.0	2.6
3	How helpful was to see the darkness of your position in comparison with other students' positions	1.5	2.1	1.0	1.8
4	How helpful was to see how close or distant were other students' positions to your position	2.6	3.4	1.0	3.0
5	How well did the circle visualization reflect your position	N/A	2.9	2.0	2.9

Table 5.10 reports students' qualitative evaluations of various Umka features in a textual form. Among the students' comments there was still a fair amount of criticism about the complexity of Umka's interface: "Some of the details are confusing", "This website was kind of hard to follow" (row 1). However, the number of negative comments about the interface decreased in comparison with version 1 of Umka (15 negative comments in version 1 versus 5 in version 2). Moreover, more students reported their satisfaction with the interface and the overall Umka system: "It was far more enjoyable than just using the bblearn domain. The set up and visualization of this system was much more engaging", "Overall the visualization tool was pretty helpful".

Many students appreciated the helpfulness of the system suggestions: "they were pretty helpful", "Helped get me thinking" (row 2).

Contrary to our expectations, the students stated that the size of their circular positions in the visualization that reflected the number of their arguments didn't much encourage them to look for more diverse arguments (row 3). This was partly because the students did not understand this feature and partly because they didn't much trust it: "I didn't not [sic] understand what that was based on", "I believe I made more that two arguments but it's showing I only made 2...".

Unlike the size of the students' positions, the darkness of their circle positions that reflected the peers' acceptance of their arguments was able to encourage the students' interactions, but only to a limited degree: "I wanted to defend my views", "a little bit [encouraged]" (row 4).

While the layout of the students' positions was helpful for some students, there were more students who found it confusing: "I don't really understand how my position related to others..", "I found it a bit confusing" (row 5).

Table 5.10: Students' qualitative answers from the post-study questionnaire on different system features. Results based on experiments #2–4.

# of positive/ negative comments	**Comments**
1. General comments on interface and visualization	
Pos: 7	It was far more enjoyable than just using the bblearn domain. The set up and visualization of this system was much more engaging; Clean interface; Overall the visualization tool was pretty helpful; Very neat and helpful.; It reflected it well.; it sure did [visualization reflected well my position]; It was a good exercise.

Neg: 5	Some of details are confusing; There is no easy way to view if anyone has a new reply or how many acceptances someone has from the window on the side; Possibly make it a bit more clear that you want people to provide arguments for both sides rather than pick one and only provide aruguments for that side.; The program felt like it was being pulled in several different directions. On the one hand it seemed to be asking for a acceptable solution to a situation that should test a persons ethics. It also was asking for pros, and cons serounding the possible solutions to the system. and lastly it was als; This website was kind of hard to follow. Lots of bugs. When you go to the diagram you can't go back without logging into the webpage again.
2. System suggestions to consider similar, different, contradictory args of other students	
Pos: 10	I liked this feature; they were pretty helpful; They we're ok; somewhat helpful; Figuring out what other people theorized helped me understand how to do the assignment.; they were as helpful as can be expected; they were moderately helpful; Helped get me thinking.; they were very helpful in showing contradictory statements; they were pretty helpful in consideration
Neg: 5	It didn't really help.; I don't like looking at similar arguments before or after my argument is made.; They didn't really get me to think of other arguments; Other people provided criticism rather than insight; I feel this is over helpful and therefore not helpful at all. If I can view the answers for other students then whats to prevent me from copying them? I feel this should be viewable after you've given
3. Representing the number of arguments in a position by the size a circle	
Pos: 1	After answering the questions I took a look and realized that my bubble was tiny compared to others. Although I feel my arguments are in much more depth then the once sentance answers other students
Neg: 7	I didn't understand what that was based on.; I can understand the width of the circle and what that meant alone and in relation to other students; I believe I made more than two arguments but it's showing I only made 2 so my circle isn't that big.; It didn't really affect it that much.; It didnt at all, it's only a circle. Everybody has different views on ethics and everybody has diff; not much; Not very much; didn't really help; I feel that my 2 reasons were more detailed then 5 one sentance details that other students gave. Although if they expanded their reasons more I feel I would try increase my position
4. Representing the quality of arguments by the darkness of a circle	
Pos: 3	This was easier to understand; I wanted to defend my views.; a little bit.
Neg: 5	I did not care about there positions so long as I was doing what was expected of me.; It didn't affect it much.; Not much at all. My opinion is my own, i have no reason to defend it.; Not much at all.; not at all
5. Representing the difference between args in students' analyses by the layout of circles	
Pos: 4	Showed me relative positioning; pretty helpful. no problems i guess.; pretty helpful it made the positions more clear.; it gave me an idea of how my peers viewed the same question
Neg: 7	I really didn't care how others felt. I also didn't really know what that was based on...; I don't really understand how my position related to others... ; I found it a bit confusing, not sure how the positions were created.; It looked rather disjointed and didn't provide a friendly interface to see the other comments and on; No other positions were displayed when I went to the diagram; wasn't very helpful. Didn't really make any sense; ...I was confused on what the chart means for x,y coordinates? If the coords do not matter just use a bar graph

5.3 Conclusions for Version 2 of Umka

5.3.1 Integrating Results

This section integrates the questionnaire results with the observed students' behavioural data.

Umka's feature of suggestions to consider various helpful arguments of other students received more positive responses from the questionnaire results than from the behavioural data. While the behavioural results showed that the suggestions were only followed up by the students on average 16% of the time and triggered few student positive behaviours (Table 5.6), the questionnaire results demonstrated that the students rated the helpfulness of the suggestions above average — 2.7 out of 5 (Table 5.9, row 1), and this feature received twice as many positive written comments as negative ones — 10 versus 5 (Table 5.10, row 2). Thus, the system suggestions were able to trigger a relatively small amount of positive students' behaviours, and hence were not quite useful for this task, but the overall positive students' rating of them suggests that they were useful for some other tasks.

In contrast, Umka's visualization received a more negative assessment from the students' questionnaire replies than from the observational behavioural data. The behavioural data indicated a noticeable effect of the circle visualization in encouraging students' interaction and self-reflection on their own positions (Section 5.2.3). But in the students' questionnaire replies the visualization features were ranked quite modestly: from 1.8 to 3.3 out of 5 (Table 5.9, rows 2–4), and received more negative than positive comments (Table 5.10, rows 3-5).

In either case the questionnaire results deviated from the behavioural results. The students seemed to rank higher more explicit support in the form of suggestions, while less easily observable support in the form of the visualization in fact produced more pedagogical effect.

5.3.2 Analyzing the Effectiveness of Students' Interactions

One of the goals of Umka's support is to stimulate productive student interactions with each other. By productive interactions we mean interactions that caused students' learning, observable by the extent the students reconsider and expand their positions on a given ethical case. We analyzed how beneficial were the students' interactions with other students' arguments, how much these interactions provoked the students to expand or change their initial positions on a given case study, and what specific student interactions caused the students to make the changes, or correlated with these changes. We studied which student interaction patterns are correlated with *whether the students make any changes in their analyses* during the collaboration. By changes in the analysis we mostly mean new arguments that the students expanded their analyses with, but also changes in case resolutions, and conclusions, that is all changes the students made to their initial positions as a result of the collaborative interactions.

To find productive student interaction patterns, we analyzed the data from the SIAST experiment, as

Table 5.11: Variables correlated with *whether students introduce any changes in their analyses* and the significance of these correlations. Results based on experiment #2.

N	1. Correlation	2. Significance
1	# of comments made to the analyses of others	71%
2	# of comments made to the analyses of others with different points of views	68%
3	# of visited analyses of others	71%
4	# of visited analyses of others with different points of views	71%
5	time spent in the analyses of others	56%
6	time spent in the analyses of others with different points of views	68%

the experiment with the largest number of participants, and hence the biggest amount of experimental data. We analyzed various students' interaction patterns, and using logistic regression we calculated the likelihood of the students making changes to their analysis given their certain interaction pattern. We have discovered that *whether the students make any changes in their analyses* during collaboration is correlated with the following variables of the students' interactions (Table 5.11):

1. *the number of comments they made to the analyses of other students*: more comments made lead to changes being introduced (correctly classified instances 71%, which means that in 71% of cases the number of comments a student made can correctly predict if the student will introduce changes into his/her analysis).

2. *the number of comments they made to the analyses of others with different points of views.* By analyses with different points of view we mean analyses of other students that have different arguments from the student's own arguments. The more comments the students make to the analyses of their peers with different points of view, the higher is the chance that the students will introduce changes to their own analyses (correctly classified instances 68%).

3. *the number of analyses of others they visited*: the more peers' answers the students study, the higher is the chance that the students will introduce changes to their own analyses (correctly classified instances 71%).

4. *the number of analyses with different points of view they visited*: the more peers' answers with different points of view the students study, the higher is the chance that the students will introduce changes to their own analyses (correctly classified instances 71%).

5. *time spent by students in the analyses of others*: the more time the students spend in reading and assessing the analyses of their peers, the higher is the probability that the students will introduce changes to their own analyses (correctly classified instances 56%).

6. *time spent by them in the analyses of others with different points of view*: the more time the students spend in reading and assessing the analyses of their peers with different points of view, the higher is the chance that the students will introduce changes to their own analyses (correctly classified instances 68%).

The obtained correlation results demonstrate that the more students are engaged in the collaborative activities of reading and commenting on the analyses of other students, and spending enough time on these activities, the higher will be the chance that the students will reconsider and expand their positions, and hence the higher will be the chance for learning to occur. Furthermore, collaborating with students of divergent points of view who have different considerations for a case study can be especially beneficial for students (Table 5.11, row 6 as opposed to row 5).

5.3.3 Conclusions

Version 2 of Umka was organized around supporting students' collaboration in the case analysis. The system support to students in this version combined system suggestions to consider certain "helpful" arguments of other students, with the newly introduced in version 2 visualization of students' positions.

We formulated several hypotheses to study the effect of Umka's support types on stimulating productive students interactions, and students' metacognitive processes of reconsidering and broadening their positions. We conducted three experiments for testing out the hypotheses.

We have learned several lessons, and gained a number of insights from the design and experimentation with Umka version 2:

- Umka's support in the form of visualization and suggestions foster productive interactions among students. Thus, the students who used Umka version with these support types in comparison with the students who didn't have them:
  - gave more comments to the analysis of their peers (a statistically significant result)
  - viewed more "helpful" analyses of their peers (not a statistically significant result)
  - spent more time in the analyses of their peers (not a statistically significant result, but with a strong positive trend towards it)
- Umka's support in the form of visualization and suggestions foster students' metacognitive processes of reconsidering and broadening their positions. Thus, the students who used these support types introduced more changes to their analyses during the collaboration stage. However, the statistical significance of this conclusion could not be confirmed.
- There was a noticeable effect observed of the system visualization alone in stimulating students' metacognitive processes, causing the students to expand their analyses with new considerations for a given case study. In contrast, the analogous effect of the system suggestions was quite modest.

- Although the average precision of LSA for the task of comparing students' arguments was not very high (0.6), this was sufficient for organizing useful system tutoring support. We have also discovered that another method WTMF, that was specifically designed for finding similarity between short pieces of text, performs better for judging the similarity rate between students' case arguments.
- Umka's visualization requires further design improvement with a goal to make it more intuitive for users.
- We have demonstrated that such behaviours as reading and commenting on the analyses of other students, and spending enough time on these activities, are correlated with students reconsidering and expanding their positions with more new arguments. Interaction with peers who hold different points of view can be especially beneficial for students.

In summary, the experimental evidence confirmed the positive effect of Umka's visualization and suggestions in increasing engagement of students in the desired practices of interacting with each other over a given case, and reconsidering and broadening their own positions. However, as the students' comments revealed, some Umka features were still confusing for them, and hence required further work, the work that we continued in version 3.

Chapter 6

Version 3 of Umka

We presented the experimental results from Umka version 2 at the Artificial Intelligence in Education Conference 2013 [89, 90]. After interacting with the people at the conference and the thesis committee members, and after analyzing the comments from the study participants, for our next version of Umka we made the modifications (described below) to the interface, visualization, and suggestions (Sections 6.1, 6.2.1). The goal for these modifications was to provide more definite answers for the research question #2 on the ways to organize a collaborative learning environment, and the research question #3 on the ways to diagnose students' arguments.

6.1 Interface

The first modification to the system interface that we introduced in version 3 was a ranking feature for arguments. A star symbol next to an argument in the interface (Figure 6.1), allows a student to rank the importance of this particular argument in comparison with other arguments of the student. This helps to further understand the reasoning of different students, in cases where two students have the same arguments, but have come to different conclusions about the case because of the different rankings for their arguments.

Another modification to the interface was a requirement that learners, before they could have access to the analyses of their peers, had to complete their own analyses individually, including choosing their own resolutions for a given case. This afforded a more reliable tracking of how students' case analyses are being changed as the result of Umka's support and interactions with peers.

6.2 Umka's Support in Collaboration

The collaborative support in Umka version 3 was organized in the form of suggestions to students as to certain "helpful" peers to interact with about a case study. These suggestions were either explicit through text messages, or implicit through the visualization that illustrated the similarities and differences between the student's and the peers' perspectives about the case study.

For a given student what are the most helpful students to interact with? Pedagogical theories on moral development suggest that the most beneficial interactions could be with peers of different points of view and

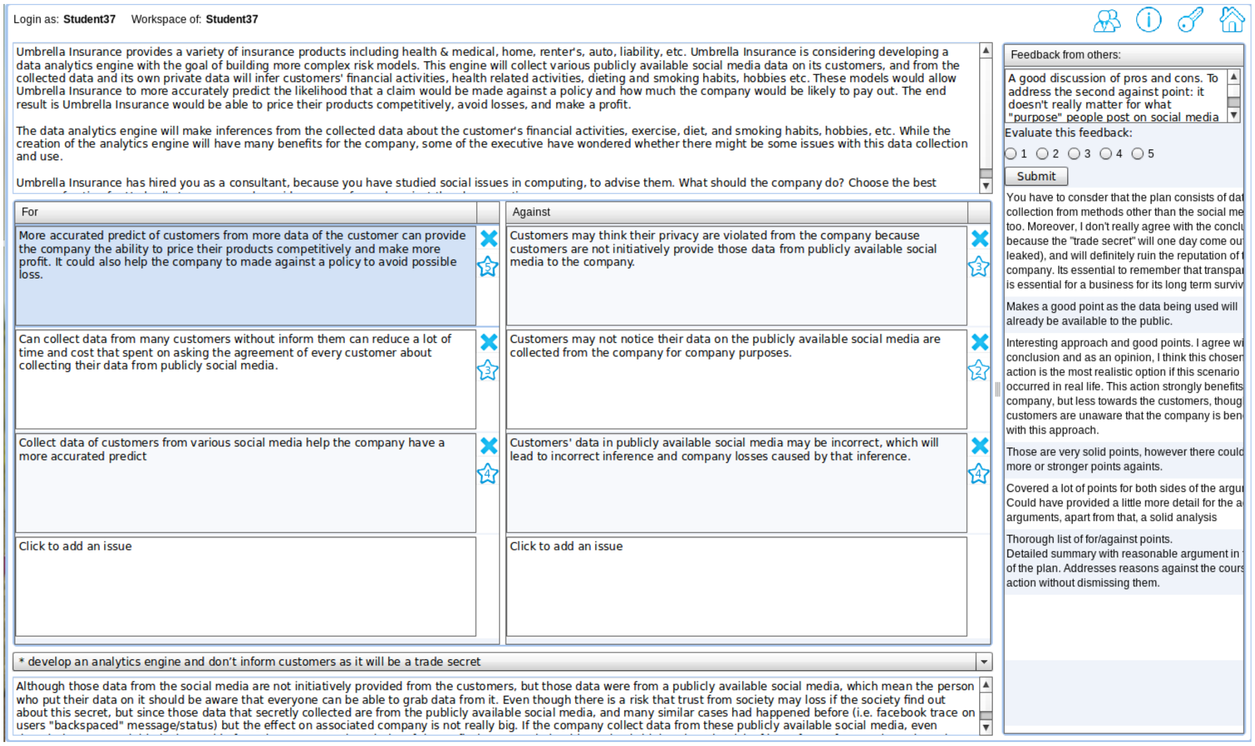


Figure 6.1: Interface of version 3 of Umka

at a slightly higher stage of moral development (Section 2.2.2). Likewise, interactions where students reason about and summarize opinions of each other have also proven to advance the students' development.

For a given student, the peers with different points of view in Umka will be the students who have chosen different resolutions for a case study dilemma and/or different arguments for justifying their chosen resolutions. The number of different arguments among the students can be calculated using a suitable similarity algorithm.

For the given student the peers with a higher stage of moral development in Umka will be students whose arguments and comments have been rated higher by others than the given student's own arguments and comments. Rest, Turiel and Kohlberg [77] demonstrated that students prefer the reasoning of others if it is of higher stage than their own. Thus, a higher level of acceptance of a certain student's reasoning by others in Umka can be an indication of this student having a higher stage of moral development.

Umka then can suggest to the given student the identified "helpful" students to interact with, that is the students who have different positions about a case study, and whose positions were highly accepted by peers. Since interactions, where students need to reason about, evaluate and summarize opinions of others advance their development, Umka version 3 suggestions to the given student are formulated in the form of instructions to evaluate the identified "helpful" positions of the peers.

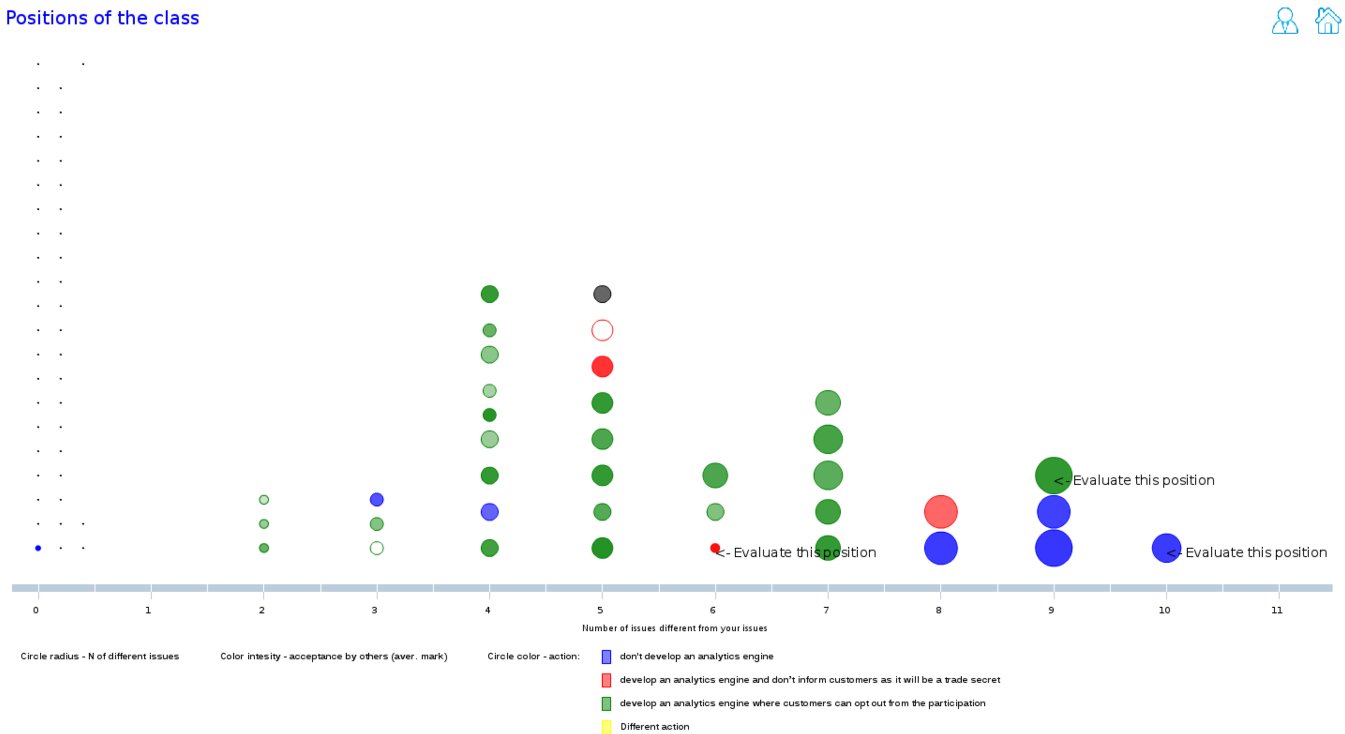


Figure 6.2: Visualization of students' positions in version 3 of Umka

6.2.1 Visualization

The visualization of students' positions in Umka version 2 needed more work to make it more comprehensible for students, and to make the differences between students' positions more apparent. One post-study interview with a participant revealed that one thing she would like to see in the visualization is the resolutions that other students chose for a given case study dilemma. The visualization of Umka version 2 could only demonstrate the difference between students in their provided arguments.

In Umka version 3 we added this new feature to the visualization of illustrating differences in case resolutions. The new visualization depicts students' positions as circles of various colors, a separate color for each case resolution. In the previous version of the visualization, the student saw her position as a red circle and positions of others as blue circles (Figure 5.2). In the new version of the visualization, there are more colors — a separate color for each case resolution. Circles of the same color represent students' positions with the same chosen resolution for an analyzed case study (Figure 6.2). Thus, the new visualization makes the differences between students more apparent by additionally demonstrating to them where their classmates stand as to the best ways to resolve the case study.

In the new visualization we also made the distance between students' positions more obvious strictly basing it on the number of different arguments between the students displayed on the X axis (Figure 6.2).

6.2.2 Suggestions

If Umka version 2 suggested to students to consider individual "helpful" arguments of their peers, the new version of Umka suggests to students certain "helpful" peers to interact with. Umka displays these suggestions in the visualization as messages to evaluate positions of "helpful" students (text messages with arrows in Figure 6.2). The system decides for a given student which are the "helpful" peers to interact with based on the following factors:

1. the difference between the student and the peers' case study resolutions

2. the difference between their arguments

3. the level of acceptance of the peers' reasoning by others

The higher is the difference between the peers and the given student in case resolutions and number of arguments, and the higher is the level of the acceptance by others of the peers' argumentation, the more likely they are to be candidates for "helpful" peers for the given student.

6.3 Experimental Results

6.3.1 Description of Experiments

We ran two experiments with the third version of Umka:

- UofT — this experiment was run in spring 2014 with the University of Toronto Computer Science undergraduate students taking a *Computers & Society* class. 38 students completed the analysis, 37 of whom consented to share their data. We randomly broke up the class into a treatment group with Umka's features of visualization and suggestions, and a control group without these features. Thus, the control group students used the same Umka system where they can input their analysis, and interact with their peers, with only Umka's features of suggestions and visualization missing. The control and treatment groups were not completely separated, a student from the control group could interact with any student from the treatment group, and visa versa. The students were given a week to analyze a case study in Umka.

- CMPT408_2014 — this experiment was run in spring of 2014 with the University of Saskatchewan Computer Science undergraduate students taking a *Ethics in Computer Science* class. The students were given a week to analyze a case study in Umka. 9 students completed the case analysis, but only 4 of them consented to share their data. This critically limited the amount of available data, and we decided not to include this experimental group data for the quantitative analysis of the hypotheses, and to use it only to augment the analysis of the questionnaire results.

6.3.2 Testing the Hypotheses

For the new version of Umka, we examined the same hypotheses as for Umka version 2.

Hypothesis 1: *Umka's visualization and suggestions foster productive interactions among students.*

We looked at three sub-hypotheses here for each separate type of interactions: hypotheses 1.1, 1.2 and 1.3.

Hypothesis 1.1 *Students with Umka's support types, in comparison with students without these support types, give more comments to the analyses of their peers.*

Table 6.1 demonstrates for the UofT control and treatment groups the average number of comments that the students of the corresponding group provided to their peers, and the proportion of the commenting students who made at least one comment. As can been seen from this table, the treatment group students provided more comments to their classmates than the control group students. However, the number of commenting students was smaller in the treatment group than in the control group. Moreover, the difference between the number of comments provided by treatment and control group students was not statistically significant (the Mann-Whitney test: p-value = 0.2348), which means we cannot accept the hypothesis.

Table 6.1: Average number of comments per student, and the proportion of students who made at least one comment to their peers. Results based on experiment #5.

N	**Experimental Group**	**Avg. # of comments**	**% of commenting students**
1	UofT treatment	3.2	87%
2	UofT control	2.9	100%

Hypothesis 1.2 *Students with Umka's support types, in comparison with students without these support types, view more "helpful" analyses of their peers.*

Table 6.2 shows what percentage of all case analyses of the peers that the students viewed were "helpful" for them, either containing different arguments, or proposing different case resolutions. The treatment group students with the visualization and suggestions viewed twice as many "helpful" positions of others as the control group students. This difference was statistically significant (the Mann-Whitney test: p-value = 7.996e-06), which leads to the acceptance of the hypothesis.

Hypothesis 1.3 *Students with Umka's support types, in comparison with students without these support types, spend more time in analyses of other students.*

The experimental data provided evidence that the students of the treatment group spent more time studying the analyses of others than the control group students (Table 6.3). However, this calculated difference was not statistically significant (the Mann-Whitney test: p-value = 0.1235), so we cannot accept the hypothesis.

Table 6.2: Positions of classmates that the students viewed. Results based on experiment #5.

N	**Experimental Group**	**Proportion of helpful positions**
1	UofT treatment	75%
2	UofT control	37%

Table 6.3: Average time the students spent studying analyses of other students. Results based on experiment #5.

N	**Experimental Group**	**Avg. time in secs**
1	UofT treatment	2470
2	UofT control	1211

Hypothesis 2: *Umka's visualization and suggestions foster students' metacognitive processes of reconsidering and broadening their positions.* Students with these support types, in comparison with students without these support types *introduce more changes to their analyses during the collaboration stage.*

We have observed that the collaboration stage was indeed more productive for the treatment group than for the control group, causing them to reconsider and broaden their positions more (Table 6.4). Thus, the students in the treatment group introduced more changes to their analysis than the control group. However, this difference was not statistically significant (the Mann-Whitney test: p-value = 0.4893).

Table 6.4: Effect of the collaboration: average number of changes introduced by the students in their own analyses. Results based on experiment # 5.

N	**Experimental group**	**Avg. # of changes introduced in collaboration**	**Avg. % of these args from the total # of args**
1	UofT treatment	2.15	33%
2	UofT control	1.75	25%

In conclusion, the data obtained from experiment #5 supported all hypotheses. The treatment group indeed outperformed the control group in the number of comments to the work of their peers, the number of viewed helpful analyses, time spent in the peers' analyses, and the number of changes introduced to the their own analyses. However, we could obtain statistically significant results only for 1 out of 4 hypotheses.

One explanation for not observing substantial differences between the UofT treatment and control groups in the number of interactions and changes introduced could be that the control group was also benefiting Umkas features that could enhance students productive interactions. Thus, the control group students could see the analysis, provide comments, get comments, and reply to comments of any student in the class including the treatment group students. This happened, because the UofT treatment and control groups were not completely separated: in the visualization the treatment group saw all students from two groups, and the system could suggest to the treatment group students to interact with certain control group students, if they were selected to be good candidates to interact with by the system algorithm.

6.3.3 Measuring the Precision of the Text Similarity Algorithm

Umka's support in version 3 is designed based on the differences between students in their arguments and case study resolutions. To calculate the number of different arguments between any two students we used the Weighted Textual Matrix Factorization algorithm (WTMF).

To evaluate WTMF's precision in judging differences between students' arguments, we randomly selected 100 pairs of different students' arguments from data collected in experiments #5 and #6. Each pair from these 100 pairs contained two arguments that WTMF calculated to be different from each other, because their similarity rank was below the chosen threshold.

We then asked an external rater 3 (a graduate student in the Humanities Computing program at the University of Alberta) to judge how many of these pairs are really different, and how many similar. Based on her evaluation, the system's precision for experiment #5 UofT = 0.71, the precision of the #6 UofS2014 experiment = 0.7.

6.3.4 Questionnaire Results About Umka

Similar to version 2 of Umka, in the post-study questionnaire we asked students to rank and comment on version 3 of Umka's suggestions and visualization features. Table 6.5 presents the students' quantitative rankings.

The students found the system suggestions moderately helpful (Table 6.5, row 1, average rank= 2.1 out of 5). More helpful for the students were their peers' evaluations on their analysis (row 2, rank=3.6).

In the next set of questions we asked the students how much different aspects of the visualization encouraged their productive behaviours (Table 6.5, rows 3-7). The students acknowledged that the visualization indeed encouraged their behaviours to a certain extent. Thus, the size of their circles moderately encouraged them to look for new arguments (row 3); the darkness of their circles also moderately encouraged them to improve their analysis and provide good quality evaluations on the analyses of others (row 4 & 5); likewise, the layout of the circles and difference in the circles' colors moderately encouraged them to interact with other students (row 6 & 7).

The last set of questions in the questionnaire measured how well the circle visualisation was able to represent the students' positions (rows 8-11). Overall, the students found the visualization to be a fair representation of their positions. The size of the circles and different colors of the circles in the visualization were found to be quite adequate representation (rows 8 & 11), while the darkness of the circles and their layout somewhat adequate (rows 9 & 10).

In many categories (but not all) involving comparison to peers, UofS_2014 data looks worse that UofT (Table 6.5, rows 4,5,6,7,10,11). This could be explained by the "small number" effect of having only 4 reporting students at UofS_2014 (and only 9 overall). The UofT students produced many more arguments, since there were a lot more of them. This gave comparison and ranking algorithms more to work with, and provided more finely tuned comparable arguments and peers, and hence more satisfying support types for the UofT students.

Table 6.5: Students' ranks of Umka's features from the questionnaire (scale: 0–5). Results based on experiments #5–6.

N	**1.Question & average rank across all students**	**2. UofT**	**3. UofS_2014**	**4. Avg**
	# of students answered	**11**	**13**	
1	How helpful were the system's suggestions to consider various positions of other students	2.14	N/A	2.1
2	How helpful were evaluations of other students on your own analysis	3.36	4.25	3.6
3	How much did the number of arguments in your analysis in comparison with others (as represented by the size of circles) encourage you to to look for new arguments	1.91	3.00	2.1
4	How much did the quality of your arguments in comparison with others (as represented by the darkness of circles) encourage you to improve your analysis	2.09	2.00	2.1
5	How much did the quality of your arguments in comparison with others (as represented by the darkness of circles) encourage you to write better evaluations of other students' analyses	2.27	2.00	2.2
6	How much did the difference in arguments between you and other students (as represented by the layout of circles) encourage you to interact with them	2.27	1.67	2.1
7	How much did the difference in chosen actions to resolve the case study between you and other students (as represented by the different colors of circles) encourage you to interact with them	1.91	1.67	1.9
8	How well was the size of your circle able to represent the number of args in your analysis	3.27	3.67	3.4
9	How well was the darkness of your circle able to represent the quality of your arguments	2.73	1.00	2.4
10	How well was the layout of students' circles able to represent the differences in arguments between you and them	2.00	0.67	1.7
11	How well was the difference in the circles' colors able to represent the differences in chosen resolutions for the case between you and other students	3.82	1.33	3.3

Table 6.6 summarized the students' textual elaborations on the different system features, with data combined from both UofT and UofS_2014. Our work on the improvement of the interface based on the students' comments was fruitful, as for the first time in all three of Umka's versions, we finally received more positive than negative appraisals on the overall Umka interface and the visualization: "Good way to look at case studies", "The format of each block helps much on organizing users' work", "...helped to see which points were most important and for which side" (row 1).

Only one student remarked on the system suggestions, finding them inaccurate: "... Pretty inaccurate, in my opinion" (row 2).

The next set of students' comments dealt with the fairness of the visualization in representing the students' positions (rows 3-6). The size of the circles in the visualization received an equal amount of positive and negative comments (row 3). Some students acknowledged that the size of the circles in the visualisation was a fair representation, and it encouraged them: "I tried to provide a roughly-average number of arguments", "I think the system is a fair representation". Other students doubted the system's fairness: "I am concerned with quality over quantity for something like this".

A similar situation occurred with the darkness of the circles in the visualization (rows 4). One student commented that it was encouraging: "Made it seem like a game, competing for a darker colour". Another student had trouble believing that this feature was valuable.

The layout of the circles received the prevailing number of negative comments. The students complained that they didn't understand it, and it did not encourage them to interact with others (row 5).

The evaluation was quite opposite for the different colors in the visualization (row 6): this is the only Umka feature that received only positive comments. The students praised its accuracy and helpfulness to see where the rest of the students stand about a case study.

Another feature of Umka that received a predominant number of positive comments was peer reviews (row 7). The students appreciated their helpfulness in forming the students' own opinions, improving and expanding their analyses, viewing the problem in different perspectives, and understanding other's points of view.

Table 6.6: Students' qualitative answers from the post-study questionnaire on different system features. Results based on experiments #5–6.

# of positive/ negative comments	Comments
1. General comments on interface and visualization	
Pos: 8	Good way to look at case studies...; The format of each block helps much on organizing users'work; I liked the layout of the for/against arguments as it helped to see which points were most important and for which side. This helped to conclude the analysis and formulate a plan of action.; The side by side For and Against comparisson allowed fair evaluation of all aspects; [liked features] the overall template of the UI (choese action, for sections, against section, comclusion, feedback).; [liked features] priority of issues; I liked the commenting system and adding new For and Against arguments. It made it a pretty easy interface to use; The layout of the system is very nice.
Neg: 5	The labelling was a little unclear at times. Does for mean for a certain course of action, for a certain party, for the issue discussed in the case study? Maybe use the labels advantage/disadvantage or explicityly state what for/against references.; in order to view other student's analysis', one had to pick what their own approach to the problem was. If we need to pick our appraoch, there isn't much point in getting the user to type out their own approach.; I think the layout was too...simple.; It was unclear whether the for and against was referring to our chosen action or the proposed idea in the case study.; I think the "for" and "against" columns were unclear. The "decision made" should go first, so we know which columns to put our arguments in, and so it is clear to the reader which decision the columns are supporting.
2. System suggestions to consider certain positions of other students	
Pos: 0	
Neg: 1	They weren't helpful at all! It's just telling me to evaluate the most different amount of issues, which brings to wonder what makes an issue is different? Pretty inaccurate, in my opinion.
3. Representing the number of arguments in a position by the size a circle	
Pos: 2	I tried to provide a roughly-average number of arguments.; I think the system is a fair representation.
Neg: 2	I am concerned with quality over quantity for something like this; I don't think the quantity of many arguments exceed the the quality of one so the circle's size does not seem too relevant to me.
4. Representing the quality of arguments by the darkness of a circle	
Pos: 1	Made it seem like a game, competing for a darker colour.
Neg: 1	I have trouble believing this feedback is actually valuable.
5. Representing the difference between arguments in students' analyses by the layout of circles	
Pos: 1	It definitely caused me to consider other points of view.

Neg: 9	I don't find it useful to interact with others over our differences.; It didn't encourage me to interact with them ..; Again, the difference in arguments algorithm is questionable.; Saw many similar points; Not well at all. I don't think the difference in arguments was picked up accurately.; Instead of using numbers, you should have simply just labelled them.; Not well, since arguments seem to be similar in some; this was unclear, seemed to only represent the difference in number of arguments made; I'm not sure the distance between circles is helpful. Maybe I don't even understand the logic behind it - perhaps a more suitable explanation would help others who may also be confused
6. Representing different case study resolutions by different colors	
Pos: 6	I tried to make sure I read at least one point from every other decision option.; Very well.... found the different colors accurate.; This made it pretty easy to differentiate between the two groups.; This is the only good thing, being able to see where most of the class picks for their action, based on the overall amount of green/blue circles.; I liked being able to visualize who took what stance; The system is easy enough to understand and tends to gauge and sort students by viewpoint accurately enough to use
Neg: 0	
7. Peer reviews – students evaluating each others' analyses	
Pos: 14	They were pretty helpful..; They were somwhat helpful. Most of the information could be picked up from the given paragraph and allowed us to form individual conclusions; It was pretty good, because looking at different perspectives allows us to view the problem in different ways.; I'd consider them.; They indicated simple improvements that could be made. ; evaluations from other students help me to reconsider my issues by different point of view.; It was interesting to see how others think differently from me, it helped me look at things from a perhaps different perspective, despite that I did consider both sides of the coin on my own.; The evaluations I've read so far bring out points that I either didin't clarify or may have missed entirely. Ideally I would like to include these points if we took this another step.; Evaluations from other students raised new points I did not think of but agreed with, changing my conclusion.; was interesting to see what other people thought of my analysis; They were all well thought out and made some good points I hadn't considered.; Provided good information.; I liked the interaction of commenting on others and recieveing feedback; I liked ...being able to interact with the other students; I liked that we could evaluate each other's arguments. It was cool to see how other people drew their own conclusions, especially when their conclusion was different than mine.
Neg: 4	None of them are accurate.; I didn't really look at other student responses; Not really helpful. They don't seem to see what I'm trying to argue..; They didn't make suggestions that gave me new ideas.

6.4 Conclusions for Version 3 of Umka

6.4.1 Integrating Results

This section integrates questionnaire results with the observed students' behavioural results, and the results from the human expert based evaluation of the similarity algorithms.

Similarly to version 2, version 3 of Umka's visualization received more negative assessment from the students' questionnaire replies than from the observed behavioural data. The behavioural data showed that the circle visualization was effective in encouraging students' interaction and self-reflection on their own positions (Section 6.3.2). But in the questionnaire the students ranked the visualisation's ability to encourage them for these actions as not very high: 1.91-2.27 out of 5 (Table 6.5, rows 3-7). The visualization feature that received only positive comments from the questionnaire was the representation of different students' case study resolutions by different colors (Table 6.6, row 6).

Likewise, in the questionnaire the students reported quite low trust in Umka's judgement of differences between their arguments: average rank 1.7 out of 5=34% (Table 6.5, row 10), and provided mostly negative written comments for this feature (Table 6.6, row 5). But in the human expert based evaluation Umka's precision reached 71% in judging differences between students' arguments (Section 6.3.3). Similarly to other visualisation's features, although the students didn't rank this feature very highly, it was still helpful in stimulating desirable students' behaviours.

6.4.2 Analyzing the Effectiveness of Students' Interactions

Similarly to version 2 of Umka, in version 3 we analyzed how beneficial were students' mutual interactions for their learning, and more specifically, for encouraging the students to reconsider and expand their initial positions on a given case study. We studied what interaction variables are correlated with *whether students make any changes in their analyses* during collaboration, and what variables are correlated with *the number of changes that they make in the their analyses*. The changes here are mostly new arguments that students expand their analyses with, but also include changes in case resolutions and conclusions, that is all changes the students introduced to their initial positions as a result of collaboration.

We analyzed the data from the UofT experiment as the experiment with the biggest number of students. We analyzed various students' interaction patterns, and using logistic regression we calculated the likelihood of the students making changes to their analysis given their certain interaction pattern. We have discovered that *whether the students make any changes in their analyses* during collaboration is correlated with the following variables of the students' interaction (Table 6.7):

1. *the number of comments they made to the analyses of others*: more comments made lead to changes being introduced (correctly classified instances 56%).

2. *the number of comments they made to the analyses of others with different points of view.* By analyses with different points of view we mean analyses of other students that have different arguments and different case resolutions from the student's own arguments and the student's own case study resolutions. The more comments the students make to the analyses of their peers with different points of view, the higher is the chance that the students will introduce changes to their own analyses (correctly classified instances 68%).

3. *the number of analyses of others that the students visited* (correctly classified instances 68%).

4. *the number of analyses of others with different points of view that the students visited* (correctly classified instances 68%).

Table 6.7: Variables correlated with *the whether students introduce any changes in their analyses* and the significance of these correlations. Results based on experiment #5.

N	1. Correlation	2. Significance
1	# of comments made to the analyses of others	56%
2	# of comments made to the analyses of others with different points of view	68%
3	# of visited analyses of others	68%
4	# of visited analyses of others with different points of view	68%
5	time spent in the analyses of others	68%
6	time spent in the analyses of others with different points of view	65%
7	# of comments of others assessed	59%
8	# of comments of others assessed with different points of view	68%

5. *time spent by the students in the analyses of other students*: the more time the students spend in reading and assessing the analyses of their peers, the higher is the probability that the students will introduce changes to their own analyses (correctly classified instances 68%).

6. *time spent by the students in the analyses of others with different points of view* (correctly classified instances 65%).

7. *the number of comments of others that the students assessed*: the number of feedback from other students that the students read and rated. More comments of others assessed led to changes being introduced (correctly classified instances 59%).

8. *the number of comments of others with different points of view that the students assessed* (correctly classified instances 68%).

Additionally, we have discovered several numerical correlations from the UofT experimental data. Thus, *the number of changes the students made to their analyses* is correlated with (Table 6.8):

1. *time the students spent in the analyses of other students* (linear regression: the square of the Pearson correlation coefficient r^2=0.676, which means 67% of the variation in the number of changes can be explained by the amount of time the students spent in the analyses of others).

2. *time the students spent in the analyses of other students with different points of view* (linear regression: r^2=0.456).

3. *the number of analyses with different points of view the students visited* (linear regression: r^2=0.532).

Consistent with the conclusions from Umka version 2, version 3 conclusions indicate that the more students are engaged in the collaborative activities of reading and commenting on the analyses of other students,

Table 6.8: Variables correlated with *the number of changes students make to their analyses* and the significance of these correlations. Results based on experiment #5.

N	1. Correlation	2. Significance
1	time spent in the analyses of others	67.0%
2	time spent in the analyses of others with different points of view	45.6%
3	# of analyses with different points of view visited	53.2%

assessing comments of others on their own analyses and spending enough time on these activities, the more they will reconsider and expand their initial positions, and hence the greater will be their learning. Furthermore, collaborating with students of different points of view can be especially beneficial for learners (Table 6.7, rows 2,8; Table 6.8, row 3).

6.4.3 Conclusions

The final modifications to Umka and two final experiments allows us to draw a more complete picture as to the ways to organize collaborative learning environments to support students' ethical problem-solving. In the final treatment/control experiment we replicated the results obtained in Umka version 2. Thus, lessons learned from the experimentation with Umka version 3 are mostly similar to the lessons learned from version 2, but this time they are confirmed with more confidence:

- Umka's support in the form of visualization and suggestions foster productive interactions among students. Thus, the students who used the Umka version with these support types in comparison with the students who didn't have them:
  - gave more comments to the analysis of their peers (not a statistically significant result)
  - viewed more "helpful" analyses of their peers (a statistically significant result)
  - spent more time in the analyses of their peers (not a statistically significant result)
- Umka's support in the form of visualization and suggestions foster students' metacognitive processes of reconsidering and broadening their positions. Thus, the students who used these support types introduced more changes to their analyses during the collaboration. We couldn't, however, achieve statistical significance for this result.
- The precision of the WTMF algorithm for the task of finding semantically different arguments was quite decent, reaching 0.71.
- In the post-study questionnaires we could observe much more positive comments about Umka version 3 than for versions 2 and 1. The students liked most the following Umka features: peer reviews, representation of different students' resolutions by different colors in the visualization, and representation

of the number of arguments in students' positions by the sizes of their circles. The students found the least helpful Umka features of the system suggestions and the representation of the differences between students' positions through the distance between their circles in the visualization.

- Similarly to Umka version 2, in version 3 we have also demonstrated that productive collaborative behaviours such as reading and commenting on the analyses of other students, assessing comments of others on their own analyses, and spending more time on these activities are correlated with students reconsidering and expanding their positions with more new arguments. Particularly, interaction with peers who hold different points of view can be especially beneficial for students.

Chapter 7

Discussion and Conclusions

The goal of this research was the development of techniques for organizing computer-based learning environments to support students in the analysis of case studies for professional ethics education.

Through the literature review we have identified the goals of ethics education as helping learners to establish their own convictions on important ethical and professional issues, broaden their perspectives and develop their skills for analysis and critique of their own and others' convictions. We also came to the understanding (through the literature review and our own research) that learners can achieve the goals of ethical education especially well through interaction with peers over important ethical and professional issues, and especially with peers who hold different points of view on these issues.

We designed a system called Umka with these goals in mind. The main way that Umka achieves these goals is through suggesting and visualizing similarities and differences between students' positions on a given case. We measured the effectiveness of Umka's support types by the extent of the positive student behaviours they triggered, and the students' own perceptions of their helpfulness.

Using three different versions of Umka and a series of six experiments we identified techniques that can effectively support students in ethical analysis. We also identified techniques that do not work as well.

7.1 Umka's Features Across Different Versions

Over the course of this research Umka was deployed in three versions. Table 7.1 demonstrates what support features were implemented in each version of Umka.

Umka version 1 was the only version that supported students in their individual analysis, but this support came at an extra cost of domain model construction. The support feature that was found to be the most helpful in the individual stage was system hints on new arguments that a student had not yet considered for a given case. Another feature that was unique to Umka version 1 was presenting arguments of all students in the class semantically clustered. This feature was rated as the most helpful in the collaborative stage by the students in the questionnaire (Section 4.5.2).

Umka version 2 supported students by suggesting to them certain helpful arguments of their peers, and a useful visualization of students' positions. The system suggestions of helpful arguments did not find much uptake among students (Sections 4.5.2, 5.2.3).

Visualization was enhanced for supporting students in Umka version 3. System suggestions were changed in version 3: instead of suggesting to students certain isolated arguments of other students as in version 2, Umka version 3 suggested students certain helpful peers to interact with. These suggestions were more followed up by the students than the suggestions of version 2, as evidenced in the number of helpful analyses the students visited (Section 6.3.2).

If a next version of Umka were to be developed, the following features from the previous versions will be integrated into it, as these are the features that have been shown to render the best support to students: 1) hints on arguments that a student hasn't yet considered (provided that a domain model for a case is available); 2) presenting arguments of all students semantically clustered; 3) suggestions to consider helpful peers; 4) visualization of students' positions.

Table 7.1: Umka's features across different versions

N	Support feature	Umka v.1	Umka v.2	Umka v.3
1	Individual: guidance on the steps of the ethical analysis	✓		
2	Individual: feedback on arguments	✓		
3	Individual: hints on arguments a student hasn't considered	✓		
4	Collaboration: presenting arguments of all students semantically clustered	✓		
5	Collaboration: system suggesting to consider similar, different, and contradictory arguments of peers	✓	✓	
6	Collaboration: system suggesting helpful peers to interact with			✓
6	Collaboration: visualization of students' positions		✓	✓

7.2 Discussion of Findings

In finding ways to organize a learning environment to support students in ethical case analysis, we have formulated three research questions: 1) what techniques can effectively support students in individual ethical problem-solving, 2) how should collaborative learning environments be organized to advance students' learning, and 3) what are are ways to diagnose students' ethical analyses presented in the form of textual arguments. The sections bellow summarize our findings for these research questions.

7.2.1 Research Question 1

To answer research question #1, which was how to effectively support a learner in their individual analysis, we experimented with different support types: a) guidance on the steps of ethical analysis; b) system feedback as to whether a learner's argument is a good argument; c) system feedback as to whether a learner's argument is an original argument; d) challenging questions about a learner's arguments; e) counterarguments to a learner's arguments; f) hints about arguments that a learner has not yet considered. We found out that the most helpful support type is system hints on new arguments that a learner has not yet considered: their helpfulness was rated highly by the students in the experiments, they were used a lot by the students, and they triggered the highest number of positive student behaviours (Experiment #1, Section 4.5.1). Thus, an ethics ITS by giving learners hints on the arguments they have not considered yet, will help the learners to gain insights they would not have gained otherwise about a given case, thus broadening their perspectives, and helping to achieve this important goal of ethics education.

7.2.2 Research Question 2

To answer research question #2, which was how to organize collaborative environments for case analysis we experimented with the following techniques: a) suggesting that learners consider similar arguments, different arguments, and counterarguments of their peers; b) presenting arguments of all learners grouped by similarity; c) suggesting that learners evaluate analyses of certain helpful peers, and also that they rate their peers' evaluations of their own analyses; d) visualizing similarities and differences in the learners' positions through the circle visualisation.

Through the experimental evaluation of these techniques we gained the following insights:

- Our first hypothesis was that Umka's support in the form of visualization and suggestions foster students' productive interactions. All the data across experiments #2-5 were in support of this hypothesis (Sections 5.2.3 and 6.3.2). Thus, the students who used the Umka version with these support types in comparison with the students who didn't have them: 1) gave more comments to the analyses of their peers; 2) viewed more "helpful" analyses of their peers; 3) spent more time in the analyses of their peers. However, we could achieve statistically significant results only for one out of three sub-hypotheses for experiments #2–4, and only for one sub-hypothesis for experiment #5.

- Our second hypothesis was that Umka's support in the form of visualization and suggestions foster students' metacognitive processes of reconsidering and broadening their positions. Likewise, all the data across experiments #2-5 were in support of this hypothesis (Sections 5.2.3 and 6.3.2). Thus, the students who used these support types introduced more changes to their analyses during the collaboration than the students who did not have these support types; however this difference was not statistically significant.

- The semantic grouping of arguments made by all students is an effective technique to support learners in collaboration. The students highly rated the helpfulness of this technique, and it also triggered some positive behaviours among the students (Experiment #1, Section 4.5.2).

- The proposed visualization by itself is pedagogically useful in encouraging students to broaden their perspectives with new arguments. On average, one third of all arguments made by the students in three experiments were introduced only by analysing the visualization without accessing the actual arguments of other students (Experiments #2–4, Section 5.2.3).

- The proposed visualization by itself is pedagogically useful in fostering interactions between students. The percentage of commenting students was higher among the students who used Umka with the visualization than among the students without the visualization (Experiments #2 and 4, Section 5.2.3).

- The system suggestions to consider various arguments of others seem to be not very helpful for our predefined tasks: they were not much followed up by the students, they didn't much encourage the students to broaden their perspectives. Nevertheless, their helpfulness was rated quite positively by the students in the questionnaire, which suggests that they may be useful for some other tasks (Experiments # 2–4, Sections 5.2.3 and 5.2.5).

- The post-study questionnaire revealed that the students found the following Umka features particularly helpful: peers' evaluations of students' own analyses, and the visualization features of representing the number of arguments in a position by the size of the circle, and representing different case resolutions by different colors. In contrast, the visualization features of representing the quality of students' argumentation by the darkness of their circle positions, and representing the difference between their arguments by the layout of their circle positions were confusing for the students and did not find much usage among them, and hence could not render much support to the students.

- We have demonstrated what specific students interactions trigger their metacognitive behaviours of reflecting on and broadening their positions. We concluded that such behaviours as reading and commenting on the analyses of other students, assessing comments of others on their own analyses, and spending more time on these activities are correlated with students reconsidering and expanding their positions with more new arguments. Particularly, interaction with peers who hold different points of view can be especially beneficial for students (Sections 5.3.2 and 6.4.2).

7.2.3 Research Question 3

To answer research question #3, which was finding ways to diagnose students' arguments, we experimented with Latent Semantic Analysis (LSA) and Weighted Textual Matrix Factorization (WTMF) algorithms to find similar, contradictory and different arguments among students' and system arguments. For the first four experiments we used the combination of LSA with the specially designed structure of the interface that allowed Umka to distinguish separate arguments for and against a chosen resolution. This method rendered precision ranging from 0.3 to 0.9 for different tasks and data, with an average precision of 0.6 in comparison with human expert judgement.

Later, we employed a method specifically developed for finding a similarity score between short texts or sentences — WTMF. WTMF outperformed LSA for the data from experiment #3 by 25–30%. Thus, we used WTMF for calculating the argument similarity in experiments #5–6. The average WTMF precision for finding different arguments in experiments #5–6 was 0.7.

Our experiments demonstrated that the method of combining WTMF with the special interface structure rendered the highest precision in judging similarities and differences between short students' arguments. The average obtained precision of 0.7 was good enough for organizing Umka's support types of suggestions and visualization.

7.3 Contributions

The primary contributions of this research lie in the fields of Artificial Intelligence in Education and Intelligent Tutoring Systems, and these particular sub-fields of AIED and ITS: ill-defined domains, metacognition, open group learner modelling, visualization, collaborative learning environments, domains requiring natural language interaction. More specifically, the following contributions have been made:

- *Discovery of effective instructional strategies for the ill-defined domain of professional ethics.* Ill-defined domains are an under-investigated area in AIED research. There is no consensus on the best instructional strategies for them. This research has shown (at least for the professional ethics domain) what instructional strategies work well, and what don't work that well in the individual case analysis for stimulating learners' reflection, and broadening their perspectives. In particular, system hints to learners on new arguments that they haven't yet considered for a given case, turned out to be an effective strategy for broadening the learners' perspectives.

- *Development of insight into ways that collaborative environments could be organized for a case analysis.* This research demonstrated the effectiveness of an organization based on highlighting similarities and differences between students' positions through the system suggestions, and a specifically designed circle visualization. The proposed organization proved to stimulate students' productive interactions, and metacognitive processes.

- *Development of an open group learner model based on a circle visualization.* This research proposed an open group learner model that visualizes, in the form of a circle visualization, the differences between students' positions, along with the quality of students' positions, and the quality of students' interactions. Throughout the course of the research we experimented with different features of the visualization, and came up with the features that are the most intuitive for learners, such as representing the number of arguments by the size of a circle, representing different case resolutions by different color, etc.

- *Finding techniques to foster important metacognitive skills for the professional ethics domain.* Metacognitive skills play a crucial role in ill-defined domains, often constituting the very subject matter to be learned in these domains. This research demonstrated how the circle visualization is able to encourage students' practices of metacognitive skills of reflection, assessment and revision of their own positions on a given case study.

- *Development of alternative methods to measure students' learning in ill-defined domains.* Measuring learning in ill-defined domains is an open issue. The standard ITS procedure in well-defined domains of measuring learning gains through pretest and posttest doesn't seem to work well here. This research offered alternative methods to measure learning in systems for ill-defined domains through the mixed assessment of students' behaviours, students' attitudes, changes in students' answers, and a comparison with other systems.

- *Identification of methods to diagnose learners' textual arguments.* Domains requiring natural language interaction, and analysis of students' answers in textual form are an on-going research challenge in the AIED field. In this research we have developed methods for diagnosing students' arguments based on calculating similarities and differences between them and predefined system arguments using a combination of interface features with LSA or WTMF. This research has demonstrated that effective tutoring support can be built based on the identified methods.

- *Discovery of correlations between students' interactions, particularly interactions with peers of different points of views, and students' expansion of their own positions on a given case.* This suggests that the better a tutoring system is able to organize these interactions, the more will students learn from them, and the more will they expand their initial positions.

The secondary contributions of this research are concerned with professional ethics education. In this research we surveyed literature on pedagogy in professional ethics education and best teaching practices both in traditional classroom settings and in computer-based learning environments. Based on the literature review on pedagogy in ethics and moral development, and based on the results of experiments we ran in several ethics courses, we have formulated techniques for building computer-based learning systems for the development of skills necessary for the professional ethics domain. And here lies the secondary contributions of our research — to professional ethics education. In Section 2.2.4 we have identified the limitations of the previous ethics ITSs as supporting only some small parts of ethical decision making, having very constrained interfaces for students' input, or using other human raters for the assessment of students' input. Umka advances beyond these limitations by offering support to students during all stages of ethical analysis, containing an interface that allows students to provide their analysis in natural language, and organizing an automatic assessing of students' arguments based on text similarity algorithms.

All these contributions have a goal of expanding of AIED's repertoire of techniques for supporting learning in ill-defined domains. We hope that the techniques developed in this research will allow the building of more systems, and more robust systems for supporting human learning in ill-defined domains.

7.4 Implications For Designing ITSs in Ill-Defined Domains

The presented research findings carry implications for designing intelligent tutoring systems for ill-defined domains, as follows:

- Domain model: The research has shown that it is possible to design a useful tutoring system without building a complete domain model, or without having a domain model per se. This could be done by accumulating answers of learners which, over time, capture many if not most of the various arguments made by students for a given case.

- Student model: Learners' answers presented in a textual form (a commonly found form in ill-defined domains) can be analysed by comparing them with predefined system answers or with answers of other students. Text comparison algorithms such as LSA and WTMF can be used for this comparison. This research has proven that even this kind of shallow processing of learners' answers instead of achieving full understanding still can be useful, and some effective tutoring interventions can be built based on this processing. Also, given the difficulty of measuring learning gains in ill-defined domains through a pretest/posttest procedure, learning gains for these domains can be measured through alternative metrics based on learners' behaviours, learners' attitudes, changes in learners' answers, and comparison with other systems.

- Tutoring model: Tutoring models in systems for ill-defined domains instead of focusing on helping learners to reach right answers (which may not be available anyway), can focus on stimulating learners' metacognitive skills, organizing students' collaboration and engaging them in productive social interactions. This research has shown the benefits of explicit suggestions and implicit feedback through a visualization for stimulating learners' metacognitive skills and productive interactions with each other. Peers' evaluations of students' work, and visualization features of demonstrating students' positions through different sizes and colors, seem to be particularly useful here.

- Interface: A smartly designed system interface can give learners freedom to express their answers in natural language, and at the same time have enough constrains to enable machine processing of these answers. The research has offered one example of such an interface.

Researchers investigating ill-defined domains have noted that there are no clear boundaries between ill-defined and well-defined domains, but a continuum ranging from well-defined to ill-defined [20]. Following the definition of ill-defined domains 2.1.1, the domain of professional ethics lies well towards the ill-defined end of this continuum. Through this work, it is shown that despite the serious ill-definedness of the professional ethics domain, it was possible to build an effective tutoring system for supporting students' learning in this domain. What follows from this is that for less ill-defined domains there should be even a greater possibility for building robust tutoring systems for them.

7.5 Research Limitations and Future Work

Our findings, comments from the system users and the reviewers of our papers have all suggested some areas of future work for this research, including areas of tutoring systems for ill-defined domains, visualization, collaboration, and educational technology for professional ethics education.

Measuring educational impact is a key component of any research in educational technology. Ill-defined domains present a particular challenge for identifying reliable and accurate metrics to measure students' learning in these domains. In our research we assessed students' learning through students' positive behaviours, and changes to their case analyses. While these metrics were valid and adequate for us and the class instructors we worked with, they were not so convincing for some of the reviewers of our papers from the AIED and ITS conferences[1]. Thus, one of the reviewers commented that the way we operationalized pedagogical effectiveness in terms of the frequency of use, surveyed helpfulness, and students' response was not really indicative of pedagogical effectiveness, as these metrics don't measure whether students learned something. While our measures do not directly measure learning gains, they do seem to be useful surrogates. But, it is certainly true that identifying good determinants for students' learning should be a continuous research endeavour, especially for ill-defined domains.

One of the limitations of the research deals with the analysis of students' arguments. In the absence of a domain model in versions 2 and 3 of Umka, we couldn't compare students' arguments with predefined system arguments, and only could compare students' arguments with each other. Thus, it was not possible to say if students' arguments were on the topic: students could introduce irrelevant, off-topic arguments, which could be judged by the system to be different, and hence helpful for other students. A possible way to filter out irrelevant students' arguments is to compare them with a case study description using a text similarity algorithm. A low similarity score from a text similarity algorithm will be an indication of these arguments being irrelevant to the given case. Of course, it is also possible to directly provide good system arguments, as in Umka version 1, and over time to "seed" a large number of such good arguments into the system. This would certainly be easier than building a full domain model, as is usual in a well-defined domain, and may, oddly, be an advantage of working in ill-defined domains.

The adoption of a new technology by users is often associated with its user-friendliness, its pleasing design and the clarity of its interface. While we made progress in improving the design of the system's interface and the visualization strategies over the course of the three versions of our system, there is still room for further exploration, testing and improvement. Edward Tufte [101] presents some practical advice in the design of data graphics, that can be adopted for future versions of the system.

Yet another research limitation is concerned with the short span of our experimental interventions. Our interventions in ethics classes usually lasted for 1-2 weeks, and involved the analysis of only a single case study. It took the students some time to understand the system, and their main comments and questions were

[1] Intelligent Tutoring Systems 2012 and Artificial Intelligence in Education 2013 conferences

often about the interface and usage of the system. It is possible that longer, even semester-long interventions with students analyzing several case studies in our system will produce more reliable results, and may lead to better understanding of the long-run effects of our proposed techniques. Such longer term studies would also help to overcome any effects of the "cold start" problem as students learn the system.

Finally, with regard to the area of ill-defined domains, it would be interesting to see how the research findings will generalize to other ill-defined domains beyond professional ethics, domains that also use case studies for pedagogical purposes. An Umka limitation that may impede this generalization to other ill-defined domains, is that Umka is organized around case studies that have clear dual for and against positions for case resolutions, and Umka's interface is designed to support this for and against division. Case studies in other domains may have more subtle resolutions that go beyond two options of for and against. A possible expansion of Umka to accommodate different varieties of case studies is left for future work.

7.6 Conclusion

This study was largely of an exploratory nature. At the outset, it was unclear how a computer-based learning system for professional ethics should be organized, what features it should have, how it should support students' individual and collaborative work, and how it should process students' textual arguments.

Over the course of this research, over the analysis of much pedagogical and technical literature, over three versions of our system and six experiments with learners, and after lengthy discussions with professionals in the fields of AIED, ITS and ethics education, we have developed a number of effective techniques for organizing a computer-based learning system for professional ethics education. We presented experimental evidence that the proposed techniques have positive effects on stimulating productive student interactions with each other causing students to reconsider and broaden their initial points of view. Thus, these techniques are a useful approach to supporting students in the analysis of case studies and in helping them to achieve the goals of ethics education. We see promise in these techniques for building future robust tutoring systems that support human learning in professional ethics education, and other ill-defined domains as well.

References

[1] V. Aleven. *Teaching case-based argumentation through a model and examples.* PhD thesis, University of Pittsburgh, 1997.

[2] Brock E Barry and Matthew W Ohland. Applied ethics in the engineering, health, business, and law professions: A comparison. *Journal of Engineering Education*, 98(4):377–388, 2009.

[3] R. Benbunan-Fich and S.R. Hiltz. Impacts of asynchronous learning networks on individual and group problem solving: A field experiment. *Group decision and Negotiation*, 8(5):409–426, 1999.

[4] Marvin W Berkowitz and John C Gibbs. Measuring the developmental features of moral discussion. *Merrill-Palmer Quarterly*, pages 399–410, 1983.

[5] Peter Brusilovsky, I-Han Hsiao, and Yetunde Folajimi. Quizmap: open social student modeling and adaptive navigation support with treemaps. In *Proceedings of the 6th European conference on Technology enhanced learning: towards ubiquitous learning*, pages 71–82. Springer-Verlag, 2011.

[6] Susan Bull and Judy Kay. Open learner models. In *Advances in Intelligent Tutoring Systems*, pages 301–322. Springer, 2010.

[7] Susan Bull and Ravi Vatrapu. Supporting collaborative interaction with open learner models: Existing approaches and open questions. *Proceedings of the Computer Supported Collaborative Learning (CSCL) 2011*, 2011.

[8] K. Cho and C.D. Schunn. Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education*, 48(3):409–426, 2007.

[9] M. Constantino-González and D. Suthers. An approach for coaching collaboration based on difference recognition and participation tracking. *The Role of Technology in CSCL*, pages 87–113, 2007.

[10] A.T. Corbett, K.R. Koedinger, and J.R. Anderson. Intelligent tutoring systems. *Handbook of Human-Computer Interaction*, 2:849–874, 1997.

[11] David R Cross and Scott G Paris. Developmental and instructional analyses of children's metacognition and reading comprehension. *Journal of Educational Psychology*, 80(2):131–142, 1988.

[12] J. Dai, R.B. Raine, R. Roscoe, Z. Cai, and D.S. McNamara. The Writing-Pal tutoring system: Development and design. *Journal of Engineering and Computer Innovations Vol*, 2(1):1–11, 2011.

[13] M. de los Angeles Constantino-Gonzalez, D.D. Suthers, and J.G.E. de los Santos. Coaching web-based collaborative learning based on problem solution differences and participation. *International Journal of Artificial Intelligence in Education*, 13(2-4):263–299, 2003.

[14] Vania Dimitrova, Gord McCalla, and Susan Bull. Open learner models: Future research directions (special issue of the IJAIED part 2). *International Journal of Artificial Intelligence in Education*, 17(3):217–226, 2007.

[15] T. Dragon, B. Woolf, D. Marshall, and T. Murray. Coaching within a domain independent inquiry environment. In *Intelligent Tutoring Systems*, pages 144–153. Springer, 2006.

[16] T. Dragon and B.P. Woolf. Understanding and advising students from within inquiry tutors. In *Proceedings of the AIED 2007 Workshop on Emerging Technologies for Inquiry-Based Learning in Science*, page 61. Citeseer, 2007.

[17] John H Flavell. Metacognitive aspects of problem solving. *The nature of intelligence*, 12:231–235, 1976.

[18] John H Flavell. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906–911, 1979.

[19] P. Fournier-Viger, R. Nkambou, and E. Mephu Nguifo. Exploiting partial problem spaces learned from users' interactions to provide key tutoring services in procedural and ill-defined domains. In *Proceeding of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, pages 383–390. IOS Press, 2009.

[20] P. Fournier-Viger, R. Nkambou, and E. Nguifo. Building intelligent tutoring systems for ill-defined domains. *Advances in Intelligent Tutoring Systems*, pages 81–101, 2010.

[21] V. Furtado and E. Vasconcelos. Geosimulation in education: A system for teaching police resource allocation. *International Journal of Artificial Intelligence in Education*, 17(1):57–81, 2007.

[22] Sara Gilliam, Joseph P Magliano, Keith K Millis, Irwin Levinstein, and Chutima Boonthum. Assessing the format of the presentation of text in developing a reading strategy assessment tool (R-SAT). *Behavior Research Methods*, 39(2):199–204, 2007.

[23] I.M. Goldin, K.D. Ashley, and R.L. Pinkus. Introducing PETE: computer support for teaching ethics. In *Proceedings of the 8th International Conference on Artificial Intelligence and Law*, pages 94–98. ACM, 2001.

[24] I.M. Goldin, K.D. Ashley, and R.L. Pinkus. Assessing case analyses in bioengineering ethics education: Reliability and training. In *Proceedings of International Conference on Engineering Education, San Juan, Puerto Rico*, pages 14–19, 2006.

[25] I.M. Goldin, K.D. Ashley, and R.L. Pinkus. Teaching case analysis through framing: Prospects for an ITS in an ill-defined domain. In *Proceedings of the ITS 2006 Workshop on Intelligent Tutoring Systems for Ill-Defined Domains*, pages 83–91. Citeseer, 2006.

[26] A. Graesser, P. Penumatsa, M. Ventura, Z. Cai, and X. Hu. Using LSA in AutoTutor: Learning through mixed-initiative dialogue in natural language. In *Handbook of latent semantic analysis*, pages 243–262. Lawrence Erlbaum Associates Publishers, 2007.

[27] A.C. Graesser, P. Chipman, B.C. Haynes, and A. Olney. Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *Education, IEEE Transactions on*, 48(4):612–618, 2005.

[28] A.C. Graesser, S. Lu, G.T. Jackson, H.H. Mitchell, M. Ventura, A. Olney, and M.M. Louwerse. Autotutor: A tutor with dialogue in natural language. *Behavior Research Methods*, 36(2):180–192, 2004.

[29] Thomas R Graves, Robert J Pleban, Marisa L Miller, Jack V Branciforte, Aram M Donigian, Vanessa Johnson, and Michael D Matthews. Enhancing perception in ethical decision making: A method to address ill-defined training domains. Technical report, DTIC Document, 2010.

[30] Julio Guerra, Denis Parra, and Peter Brusilovsky. Encouraging online student reading with social visualization support. pages 47–50, 2013.

[31] Weiwei Guo and Mona Diab. Modeling sentences in the latent space. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 864–872. Association for Computational Linguistics, 2012.

[32] C.E. Harris, M. Davis, M.S. Pritchard, and M.J. Rabins. Engineering ethics: what? why? how? and when? *Journal of Engineering education*, 85:93–96, 1996.

[33] C.E. Harris, M.S. Pritchard, and M.J. Rabins. *Engineering ethics: Concepts and cases.* Wadsworth Pub Co, 2009.

[34] R. Hodhod, P. Cairns, and D. Kudenko. Innovative integrated architecture for educational games: challenges and merits. *Transactions on edutainment V*, pages 1–34, 2011.

[35] R. Hodhod, D. Kudenko, and P. Cairns. AEINS: Adaptive educational interactive narrative system to teach ethics. In *Proceedings of the AIED 2009 Workshop on Intelligent Educational Games*, page 79, 2009.

[36] I-Han Hsiao, Julio Guerra, Denis Parra, Fedor Bakalov, Birgitta König-Ries, and Peter Brusilovsky. Comparative social visualization for personalized e-learning. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 303–307. ACM, 2012.

[37] S. Hughes, P. Hastings, J. Magliano, S. Goldman, and K. Lawless. Automated approaches for detecting integration in student essays. In *Intelligent Tutoring Systems*, pages 274–279. Springer, 2012.

[38] David H Jonassen and Bosung Kim. Arguing to learn and learning to argue: Design justifications and guidelines. *Educational Technology Research and Development*, 58(4):439–457, 2010.

[39] D.H. Jonassen. Instructional design models for well-structured and ill-structured problem-solving learning outcomes. *Educational Technology Research and Development*, 45(1):65–94, 1997.

[40] Matthew W Keefer, Sara E Wilson, Harry Dankowicz, and Michael C Loui. The importance of formative assessment in science and engineering ethics education: Some evidence and practical advice. *Science and engineering ethics*, pages 1–12, 2013.

[41] Matthew Wilks Keefer. Making good use of online case study materials. *Science and Engineering Ethics*, 11(3):413–429, 2005.

[42] J.M. Kim, R.W. Hill, Jr, P.J. Durlach, H.C. Lane, E. Forbell, M. Core, S. Marsella, D. Pynadath, and J. Hart. BiLAT: A game-based environment for practicing negotiation in a cultural context. *International Journal of Artificial Intelligence in Education*, 19(3):289–308, 2009.

[43] KS Kitchner. Cognition, metacognition, and epistemic cognition. *Human development*, 26(4):222–232, 1983.

[44] Lawrence Kohlberg. Resolving, moral conflicts within the just community. In *Moral dilemmas: philosophical and psychological issues in the development of moral reasoning.* Transaction Publishers, 1985.

[45] Ann Cale Kruger. Peer collaboration: Conflict, cooperation, or both? *Social Development*, 2(3):165–182, 1993.

[46] T.K. Landauer, D.S. McNamara, S.E. Dennis, and W.E. Kintsch. *Handbook of Latent Semantic Analysis.* Lawrence Erlbaum Associates Publishers, 2007.

[47] Nguyen-Thinh Le, Frank Loll, and Niels Pinkwart. Operationalizing the continuum between well-defined and ill-defined problems for educational technology. *Learning Technologies, IEEE Transactions on*, 6(3):258–270, 2013.

[48] N.T. Le and N. Pinkwart. Can soft computing techniques enhance the error diagnosis accuracy for intelligent tutors? In *Intelligent Tutoring Systems*, pages 320–329. Springer, 2012.

[49] A.B. Lees. Learning about ethical decision making in health care using web-based technology: a case study. Master's thesis, AUT University, 2011.

[50] B. Lehman, C. Mills, S. DMello, and A. Graesser. Automatic evaluation of learner self-explanations and erroneous responses for dialogue-based itss. In *Intelligent Tutoring Systems*, pages 541–550. Springer, 2012.

[51] W. Lu. *Understanding the effect of computer-supported, case-based instruction on third-year medical students' ethical reasoning.* PhD thesis, University of Missouri, 2007.

[52] C. Lynch, K. Ashley, V. Aleven, and N. Pinkwart. Defining ill-defined domains; a literature survey. In *Proceedings of the ITS 2006 Workshop on Intelligent Tutoring Systems for Ill-Defined Domains*, pages 1–10, 2006.

[53] C. Lynch, K.D. Ashley, N. Pinkwart, and V. Aleven. Concepts, structures, and goals: Redefining ill-definedness. *International Journal of Artificial Intelligence in Education*, 19(3):253–266, 2009.

[54] C. Lynch, K.D. Ashley, N. Pinkwart, and V. Aleven. Adaptive tutoring technologies and ill-defined domains. *Adaptive Technologies for Training and Education*, page 179, 2012.

[55] Collin Lynch, Kevin Ashley, Tanja Mitrovic, Vania Dimitrova, Niels Pinkwart, and Vincent Aleven. Intelligent tutoring technologies for ill-defined problems and ill-defined domains. In *Proceedings of the ITS 2010 Workshop on Intelligent Tutoring Systems for Ill-Defined Domains*, 2010.

[56] A. McKenzie. Conundrum: A serious game informed by bloom's taxonomy for teaching ethics and social issues. Master's thesis, University of Saskatchewan, 2013.

[57] A. McKenzie and G. McCalla. Serious games for professional ethics: An architecture to support personalization. In *Proceedings of the AIED 2009 Workshop on Intelligent Educational Games*, pages 69–78. Citeseer, 2009.

[58] B.M. McLaren. *Assessing the relevance of cases and principles using operationalization techniques.* PhD thesis, University of Pittsburgh, 1999.

[59] B.M. McLaren. Computational models of ethical reasoning: Challenges, initial steps, and future directions. *Intelligent Systems, IEEE*, 21(4):29–37, 2006.

[60] B.M. McLaren and K.D. Ashley. Assessing relevance with extensionally defined principles and cases. In *Proceeding of The National Conference on Artificial Intelligence*, pages 316–322, 2000.

[61] B.M. McLaren, O. Scheuer, and J. Mikšátko. Supporting collaborative learning and e-discussions using artificial intelligence techniques. *International Journal of Artificial Intelligence in Education*, 20(1):1–46, 2010.

[62] D. S. McNamara, R. Raine, R. Roscoe, S. Crossley, G. T. Jackson, J. Dai, Z. Cai, A. Renner, R. Brandon, J. Weston, K. Dempsey, D. Lam, S. Sullivan, L. Kim, V. Rus, R. Floyd, P. M. McCarthy, and A. C. Graesser. The writing-pal: Natural language algorithms to support intelligent tutoring on writing strategies. In P. M. McCarthy and C. Boonthum, editors, *Applied natural language processing and content analysis: Identification, investigation, and resolution*, pages 298–311. Hershey, PA: IGI Global., 2011.

[63] D.S. McNamara, C. Boonthum, I.B. Levinstein, and K. Millis. Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. *Handbook of Latent Semantic Analysis*, pages 227–241, 2007.

[64] D.S. McNamara, I.B. Levinstein, and C. Boonthum. iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods*, 36(2):222–233, 2004.

[65] Kelly E Middleton and Elizabeth A Petitt. *Simply the best: 29 things students say the best teachers do around relationships.* AuthorHouse, 2010.

[66] A. Mitrovic and S. Ohlsson. Evaluation of a constraint-based tutor for a database language. *International Journal on Artificial Intelligence in Education*, 10:238–56, 1999.

[67] A. Mitrovic and A. Weerasinghe. Revisiting the ill-definedness and the consequences for ITSs. In *Proceedings of 14th International Conference on Artificial Intelligence in Education, AIED09, IOS Press, Frontiers of Artificial Intelligence*, 2009.

[68] M.D. Mumford, S. Connelly, R.P. Brown, S.T. Murphy, J.H. Hill, A.L. Antes, E.P. Waples, and L.D. Devenport. A sensemaking approach to ethics training for scientists: Preliminary evidence of training effectiveness. *Ethics & Behavior*, 18(4):315–339, 2008.

[69] K. Muskavitch. Cases and goals for ethics education. *Science and Engineering Ethics*, 11(3):431–434, 2005.

[70] Larry Nucci. Synthesis of research on moral development. *Educational Leadership*, 44(5):86–92, 1987.

[71] A. Ogan, R. Wylie, and E. Walker. The challenges in adapting traditional techniques for modeling student behavior in ill-defined domains. *Intelligent Tutoring Systems for Ill-Defined Domains*, page 29, 2006.

[72] A.S. Palincsar. Social constructivist perspectives on teaching and learning. *Annual Review of Psychology*, 49:345–375, 1998.

[73] N. Pinkwart, V. Aleven, K. Ashley, and C. Lynch. Toward legal argument instruction with graph grammars and collaborative filtering techniques. In *Intelligent Tutoring Systems*, pages 227–236. Springer, 2006.

[74] N. Pinkwart, V. Aleven, K. Ashley, and C. Lynch. Evaluating legal argument instruction with graphical representations using Largo. In *Proceeding of the 2007 conference on Artificial Intelligence in Education*, pages 101–108. IOS Press, 2007.

[75] N. Pinkwart, C. Lynch, K. Ashley, and V. Aleven. Re-evaluating largo in the classroom: Are diagrams better than text for teaching argumentation skills? In *Intelligent Tutoring Systems*, pages 90–100. Springer, 2008.

[76] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. ELRA, May 2010. `http://is.muni.cz/publication/884893/en`.

[77] James Rest, Elliot Turiel, and Lawrence Kohlberg. Level of moral development as a determinant of preference and comprehension of moral judgments made by others. *Journal of Personality*, 37(2):225–252, 1969.

[78] R.W. Robbins, W.A. Wallace, and B. Puka. Supporting ethical problem solving: an exploratory investigation. In *Proceedings of the 2004 SIGMIS conference on Computer personnel research: Careers, culture, and ethics in a networked environment*, pages 134–143. ACM, 2004.

[79] V. Rus, M. Lintean, A. Graesser, and D. McNamara. Assessing student paraphrases using lexical semantics and word weighting. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education, Brighton, UK*, 2009.

[80] V. Rus, P.M. McCarthy, A.C. Graesser, and D.S. McNamara. Identification of sentence-to-sentence relations using a textual entailer. *Research on Language & Computation*, 7(2):209–229, 2009.

[81] O. Scheuer, B.M. McLaren, F. Loll, and N. Pinkwart. Automated analysis and feedback techniques to support argumentation: A survey. *Educational Technologies for Teaching Argumentation Skills. Bentram*, 2010.

[82] Gregory Schraw, Michael E Dunkle, and Lisa D Bendixen. Cognitive processes in well-defined and ill-defined problem solving. *Applied Cognitive Psychology*, 9(6):523–538, 1995.

[83] D.R. Searing. HARPS ethical analysis methodology: Method description, version 2.0.0. Technical report, Taknosys Software Corporation, 1998.

[84] Donnie J Self, Margie Olivarez, and DC Baldwin Jr. The amount of small-group case-study discussion needed to improve moral reasoning skills of medical students. *Academic Medicine*, 73(5):521–3, 1998.

[85] D.W. Shaffer and A. Graesser. Using a quantitative model of participation in a community of practice to direct automated mentoring in an ill-defined domain. In *Proceedings of the ITS 2010 Workshop on Intelligent Tutoring Systems for Ill-Defined Domains*, pages 61–68, 2010.

[86] D.W. Shaffer, D. Hatfield, G.N. Svarovsky, P. Nash, A. Nulty, E. Bagley, K. Frank, A.A. Rupp, and R. Mislevy. Epistemic network analysis: A prototype for 21st-century assessment of learning. *The International Journal of Learning and Media*, 1(2):33–53, 2009.

[87] M. Sharipova. Supporting students in the analysis of case studies for ill-defined domains. In *Intelligent Tutoring Systems*, pages 609–611. Springer, 2012.

[88] M. Sharipova and G. McCalla. Supporting students in the analysis of case studies for professional ethics education. In *Intelligent Tutoring Systems*, pages 653–654. Springer, 2012.

[89] Mayya Sharipova and Gordon McCalla. Managing ethical thinking. In *Proceedings of the AIED 2013 Workshop on Self-Regulated Learning in Educational Technologies: Supporting, modeling, evaluating, and fostering metacognition with computer-based learning environments*, pages 25–29, 2013.

[90] Mayya Sharipova and Gordon McCalla. Modelling students knowledge of ethics. In *Artificial Intelligence in Education*, pages 721–724. Springer, 2013.

[91] N. Shin. *The relationship between well-structured and ill-structured problem solving in multimedia simulation*. PhD thesis, The Pennsylvania State University, 1998.

[92] N. Shin, D.H. Jonassen, and S. McGee. Predictors of well-structured and ill-structured problem solving in an astronomy simulation. *Journal of research in science teaching*, 40(1):6–33, 2003.

[93] L.J. Shuman, M.F. Sindelar, M. Besterfield-Sacre, H. Wolfe, R. Pinkus, R.L. Miller, B.M. Olds, and C. Mitcham. Can our students recognize and resolve ethical dilemmas? In *Proceeding of the 2004 American Society for Engineering Education Conference*. Citeseer, 2004.

[94] H.A. Simon. The structure of ill structured problems. *Artificial intelligence*, 4(3-4):181–201, 1973.

[95] H.A. Simon. Information-processing theory of human problem solving. *Handbook of learning and cognitive processes*, 5:271–295, 1978.

[96] Robert E Slavin, Eric A Hurley, and Anne Chamberlain. Cooperative learning and achievement: Theory and research. *Handbook of psychology*, 7:177–198, 2003.

[97] Lingling Sun. Motivational visualization for resources-sharing online communities. Master's thesis, University of Saskatchewan, 2005.

[98] D. Suthers, J. Connelly, A. Lesgold, M. Paolucci, E. Toth, J. Toth, and A. Weiner. Representational and advisory guidance for students learning scientific inquiry. *Smart machines in education: The coming revolution in educational technology*, pages 7–35, 2001.

[99] Daniel D Suthers, Ravi Vatrapu, Richard Medina, Samuel Joseph, and Nathan Dwyer. Beyond threaded discussion: Representational guidance in asynchronous collaborative learning environments. *Computers & Education*, 50(4):1103–1127, 2008.

[100] The Hasting Center. The teaching of ethics in higher education. Technical report, The Hastings Center, Institute of Society, Ethics and the Life Sciences, Hastings-on-Hudson, N.Y., 1980.

[101] Edward R. Tufte and P. Graves-Morris. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 1983.

[102] S. van der Burg and I. van de Poel. Teaching ethics and technology with Agora, an electronic tool. *Science and Engineering Ethics*, 11(2):277–297, 2005.

[103] J Vassileva. Open group learner modeling, interaction analysis and social visualization. In *Proceedings of the Workshop on Adaptation and Personalisation in Social Systems: Groups, Teams, Communities. Held in conjunction with UM2007, Crete*, 2007.

[104] C. Voigt. *Educational Design and Media Choice for Collaborative, Electronic Case-Based Learning (e-CBL)*. PhD thesis, University of South Australia, 2008.

[105] J. Voss. Toulmin's model and the solving of ill-structured problems. *Arguing on the Toulmin Model*, pages 303–311, 2006.

[106] E. Walker, A. Ogan, V. Aleven, and C. Jones. Two approaches for providing adaptive support for discussion in an ill-defined domain. *Intelligent Tutoring Systems for Ill-Defined Domains: Assessment and Feedback in Ill-Defined Domains.*, page 1, 2008.

[107] Andrew Webster and Julita Vassileva. Visualizing personal relations in online communities. In *Adaptive hypermedia and adaptive web-based systems*, pages 223–233. Springer, 2006.

[108] B. Woolf, J. Reid, N. Stillings, M. Bruno, D. Murray, P. Reese, A. Peterfreund, and K. Rath. A general platform for inquiry learning. In *Proceedings of the International Conference on Intelligent Tutoring Systems*, 2002.

Appendix A

Algorithms

This appendix presents algorithms behind Umka's suggestions and visualizations implemented in experiments #1-6. The algorithms are largely based on calculated similarity scores between textual arguments. In experiments #1-4 we used a LSA method for calculating similarity. In experiments #5–6 we used a WTMF method. A main difference between WTMF and LSA is that WTMF was specifically designed to calculating similarity between short pieces of text, while LSA is more suited for larger pieces of text. WTMF handles the problem of insufficient information in short texts by also modelling words that are not in a text (missing words), what a text is not about [31].

Both these methods, LSA and WTMF, given two pieces of texts, return a similarity score between them — a value from 0 to 1 inclusive. The closer the value gets to 1, the more similar are the two texts, with 1 value meaning that the texts are the same, and 0 value meaning that they have nothing in common. If a similarity score for two given arguments below a predefined threshold, the arguments are deemed to be different, otherwise, they are deemed to be similar.

Input: a student A
Output: helpful arguments of peers
for *every peer P of A's* **do**
  check if P's arguments are different, similar, or contradictory to A's arguments;
  if *they are and they have not been previously suggested* **then**
    used these arguments as suggestions;
    exit from this loop;
  end
end

Algorithm 1: Finding "helpful" arguments of peers, suggestions in experiments #1-4.

Input: all students and their arguments
Output: calculated positions of students
`// For every student in the class the layout of the visualization is the same`
for *every student S in the class* **do**
  `// calculate the size of S's position`
  size=number of unique args in S's position (similarity score between args <threshold);
  `// calculate the darkness of S's position`
  darkness=average peers' rating of S's arguments and comments;
  `// calculate X,Y coordinates of S's position - distance from others`
  combine all S's arguments as a single text;
  map the text to the LSA space, and get the coordinates for it;
  do dimension reduction for these coordinates to get two dimensions of X,Y coordinates;
end

Algorithm 2: Calculating positions in the visualization, experiments #2–4.

Input: a student A and his peers
Output: calculated positions of students

```
// For every student the layout of the visualization is different relative to
differences between this student and his peers
```

for *every student S in the class* **do**

    // calculate the size of S's position
    size=number of unique args in S's position (similarity score between args <threshold);
    // calculate the darkness of S's position
    darkness=average peers' rating of S's arguments and comments;

end

`// calculate X coordinates`

make X coordinate of A's position=0 ;

for *every peer P of A's* **do**

    make X coordinate of P's position=number of different args between A and P;

end

Algorithm 3: Calculating positions in the visualization, experiments #5–6.

Input: a student A
Output: helpful peers to interact with

for *every peer P of A's* **do**

    diff1= number of different args between A and P;
    if A and P have different resolutions for a case study then
        diff2=1;
    else
        diff2=0;
    end
    influenceP=average peers' rating of P's arguments and comments;

end

choose three peers with the highest combinations of diff1, diff2,and influenceP that were not suggested before;

Algorithm 4: Finding "helpful" peers to interact with, suggestions in experiments #5–6.

Appendix B

Behavioral Ethics Certificates

UNIVERSITY OF SASKATCHEWAN

Behavioural Research Ethics Board (Beh-REB)

Certificate of Approval

PRINCIPAL INVESTIGATOR	DEPARTMENT	BEH#
Gordon McCalla	Computer Science	11-283

INSTITUTION(S) WHERE RESEARCH WILL BE CONDUCTED
University of Saskatchewan

STUDENT RESEARCHER(S)
Mayya Sharipova

FUNDER(S)
NATURAL SCIENCES & ENGINEERING RESEARCH COUNCIL OF CANADA (NSERC)

TITLE
Supporting Students in the Analysis of Case Studies for the Professional Ethics Education

ORIGINAL REVIEW DATE	APPROVAL ON	APPROVAL OF:	EXPIRY DATE
15-Oct-2011	10-Nov-2011	Ethics Application Consent Protocol	09-Nov-2012

Full Board Meeting ☐ Date of Full Board Meeting:

Delegated Review ☒

CERTIFICATION
The University of Saskatchewan Behavioural Research Ethics Board has reviewed the above-named research project. The proposal was found to be acceptable on ethical grounds. The principal investigator has the responsibility for any other administrative or regulatory approvals that may pertain to this research project, and for ensuring that the authorized research is carried out according to the conditions outlined in the original protocol submitted for ethics review. This Certificate of Approval is valid for the above time period provided there is no change in experimental protocol or consent process or documents.

Any significant changes to your proposed method, or your consent and recruitment procedures should be reported to the Chair for Research Ethics Board consideration in advance of its implementation.

ONGOING REVIEW REQUIREMENTS
In order to receive annual renewal, a status report must be submitted to the REB Chair for Board consideration within one month of the current expiry date each year the study remains open, and upon study completion. Please refer to the following website for further instructions: *http://www.usask.ca/research/ethics_review/*

[redacted]

John Rigby, Chair
University of Saskatchewan
Behavioural Research Ethics Board

Please send all correspondence to:

Research Ethics Office
University of Saskatchewan
Box 5000 RPO University, 1602-110 Gymnasium Place
Saskatoon SK S7N 4J8
Telephone: (306) 966-2975 Fax: (306) 966-2069

UNIVERSITY OF SASKATCHEWAN

Behavioural Research Ethics Board (Beh-REB)

Certificate of Approval

PRINCIPAL INVESTIGATOR	DEPARTMENT	BEH#
Gordon McCalla	Computer Science	12-207

INSTITUTION(S) WHERE RESEARCH WILL BE CONDUCTED
University of Saskatchewan

SUB-INVESTIGATOR(S)
Richard Schwier, Mayya Sharipova

STUDENT RESEARCHER(S)
Jennifer Seaton

FUNDER(S)
INTERNALLY FUNDED

TITLE
Assessing Canadian Graduate Student Satisfaction with an Interactive, Game-like Approach to a Compulsory Ethics Course - Pilot Project

ORIGINAL REVIEW DATE	APPROVAL ON	APPROVAL OF:	EXPIRY DATE
25-Jul-2012	25-Jul-2012	Application for Behavioural Research Ethics Review Appendix A, Inclusion/Exclusion Appendix B: Survey Appendix C: Debriefing Form Appendix D: Invitation Email Appendix E: Participant Consent Form	24-Jul-2013

Full Board Meeting ☐ Delegated Review ☒

CERTIFICATION

The University of Saskatchewan Behavioural Research Ethics Board has reviewed the above-named research project. The proposal was found to be acceptable on ethical grounds. The principal investigator has the responsibility for any other administrative or regulatory approvals that may pertain to this research project, and for ensuring that the authorized research is carried out according to the conditions outlined in the original protocol submitted for ethics review. This Certificate of Approval is valid for the above time period provided there is no change in experimental protocol or consent process or documents.

Any significant changes to your proposed method, or your consent and recruitment procedures should be reported to the Chair for Research Ethics Board consideration in advance of its implementation.

ONGOING REVIEW REQUIREMENTS

In order to receive annual renewal, a status report must be submitted to the REB Chair for Board consideration within one month of the current expiry date each year the study remains open, and upon study completion. Please refer to the following website for further instructions: http://www.usask.ca/research/ethics_review/

[redacted]

Beth Bilson, Chair
University of Saskatchewan
Behavioural Research Ethics Board

Please send all correspondence to:

Research Ethics Office
University of Saskatchewan
Box 5000 RPO University, 1602-110 Gymnasium Place
Saskatoon SK S7N 4J8
Telephone: (306) 966-2975 Fax: (306) 966-2069

SIAST
SASKATCHEWAN INSTITUTE OF APPLIED SCIENCE AND TECHNOLOGY

SIAST Wascana Campus
4500 Wascana Parkway
PO Box 556
Regina SK S4P 3A3
www.goSIAST.com

MEMORANDUM

Date: October 25, 2012

From: Dr. Rod Stutt, Chair
SIAST Research Ethics Board

To: Mayya Sharipova
PhD Student, University of Saskatchewan

Re: Supporting students in the analysis of case studies for the professional ethics education

---

Please be advised that the SIAST Research Ethics Board has reviewed your application for revision and found it to be:

☑ 1. APPROVED AS SUBMITTED. Only applicants with this designation have ethical approval to proceed with their research as described in their application.

☐ 2. ACCEPTABLE SUBJECT TO MINOR CHANGES AND PRECAUTIONS. (See attached.) Changes must be submitted to the REB and approved prior to continuing research. Please submit a supplementary memo addressing the concerns noted by the Board. Once changes are deemed acceptable, ethical approval will be granted.

☐ 3. ACCEPTABLE SUBJECT TO CHANGES AND PRECAUTIONS. (See attached.) Changes must be submitted to the REB and approved prior to continuing research. Please submit a supplementary memo addressing the concerns noted by the Board. Once changes are deemed acceptable, ethical approval will be granted.

☐ 4. UNACCEPTABLE AS SUBMITTED. The proposal requires substantial additions or redesign. Please contact the Chair of the REB for advice on how the project proposal might be revised.

[redacted]

Dr. Rod Stutt

Supplementary memo should be forwarded to the Chair of the Research Ethics Board at the Office of Applied Research and Innovation, Wascana campus or by email to applied.research@siast.sk.ca
Phone: 306-775-7320
Fax: 306-798-8113

UNIVERSITY OF SASKATCHEWAN

Behavioural Research Ethics Board (Beh-REB)

Certificate of Re-Approval

PRINCIPAL INVESTIGATOR	DEPARTMENT	Beh #
Gordon McCalla	Computer Science	11-283

INSTITUTION (S) WHERE RESEARCH WILL BE CARRIED OUT

University of Saskatchewan
Saskatoon SK

STUDENT RESEARCHER(S)

Mayya Sharipova

FUNDER(S)

NATURAL SCIENCES & ENGINEERING RESEARCH COUNCIL OF CANADA (NSERC)

TITLE:

Supporting Students in the Analysis of Case Studies for the Professional Ethics Education

RE-APPROVED ON	EXPIRY DATE
27-Feb-2013	26-Feb-2014

Full Board Meeting ☐

Delegated Review ☑

CERTIFICATION

The University of Saskatchewan Behavioural Research Ethics Board has reviewed the above-named research project. The proposal was found to be acceptable on ethical grounds. The principal investigator has the responsibility for any other administrative or regulatory approvals that may pertain to this research project, and for ensuring that the authorized research is carried out according to the conditions outlined in the original protocol submitted for ethics review. This Certificate of Approval is valid for the above time period provided there is no change in experimental protocol or consent process or documents.

Any significant changes to your proposed method, or your consent and recruitment procedures should be reported to the Chair for Research Ethics Board consideration in advance of its implementation.

ONGOING REVIEW REQUIREMENTS

In order to receive annual renewal, a status report must be submitted to the REB Chair for Board consideration within one month of the current expiry date each year the study remains open, and upon study completion. Please refer to the following website for further instructions: http://www.usask.ca/research/ethics_review/

[redacted]

Beth Bilson, Chair
University of Saskatchewan
Behavioural Research Ethics Board

Please send all correspondence to:

Research Ethics Office
University of Saskatchewan
Box 5000 RPO University, 1607 – 110 Gymnasium Place
Saskatoon, SK S7N 4J8
Phone: (306) 966-2975 Fax: (306) 966-2069

UNIVERSITY OF SASKATCHEWAN

Behavioural Research Ethics Board (Beh-REB)

Certificate of Re-Approval

PRINCIPAL INVESTIGATOR	DEPARTMENT	Beh #
Gordon McCalla	Computer Science	11-283

INSTITUTION (S) WHERE RESEARCH WILL BE CARRIED OUT

University of Saskatchewan
Saskatoon SK

STUDENT RESEARCHER(S)

Mayya Sharipova

FUNDER(S)

NATURAL SCIENCES & ENGINEERING RESEARCH COUNCIL OF CANADA (NSERC)

TITLE:

Supporting Students in the Analysis of Case Studies for the Professional Ethics Education

RE-APPROVED ON	EXPIRY DATE
09-Jan-2014	08-Jan-2015

Full Board Meeting ☐

Delegated Review ✔

CERTIFICATION

The University of Saskatchewan Behavioural Research Ethics Board has reviewed the above-named research project. The proposal was found to be acceptable on ethical grounds. The principal investigator has the responsibility for any other administrative or regulatory approvals that may pertain to this research project, and for ensuring that the authorized research is carried out according to the conditions outlined in the original protocol submitted for ethics review. This Certificate of Approval is valid for the above time period provided there is no change in experimental protocol or consent process or documents.

Any significant changes to your proposed method, or your consent and recruitment procedures should be reported to the Chair for Research Ethics Board consideration in advance of its implementation.

ONGOING REVIEW REQUIREMENTS

In order to receive annual renewal, a status report must be submitted to the REB Chair for Board consideration within one month of the current expiry date each year the study remains open, and upon study completion. Please refer to the following website for further instructions: http://www.usask.ca/research/ethics_review/

[redacted]

Beth Bilson, Chair
University of Saskatchewan
Behavioural Research Ethics Board

Please send all correspondence to

Research Ethics Office
University of Saskatchewan
Box 5000 RPO University, 1607 110 Gymnasium Place
Saskatoon, SK S7N 4J8
Phone (306) 966-2975 Fax: (306) 966-2069

UNIVERSITY OF SASKATCHEWAN

Behavioural Research Ethics Board (Beh-REB)

Certificate of Approval Study Amendment

PRINCIPAL INVESTIGATOR	DEPARTMENT	BEH#
Gordon McCalla	Computer Science	11-283

INSTITUTION (S) WHERE RESEARCH WILL BE CARRIED OUT

University of Saskatchewan
Saskatoon SK

STUDENT RESEARCHER(S)

Mayya Sharipova

FUNDER(S)

NATURAL SCIENCES & ENGINEERING RESEARCH COUNCIL OF CANADA (NSERC)

TITLE

Supporting Students in the Analysis of Case Studies for the Professional Ethics Education

APPROVAL OF	APPROVED ON	CURRENT EXPIRY DATE
Amended protocol Amended study questionnaire Amended consent form	09-Jan-2014	08-Jan-2015

Full Board Meeting

Delegated Review ✔

CERTIFICATION

The University of Saskatchewan Behavioural Research Ethics Board has reviewed the above-named research project. The proposal was found to be acceptable on ethical grounds. The principal investigator has the responsibility for any other administrative or regulatory approvals that may pertain to this research project, and for ensuring that the authorized research is carried out according to the conditions outlined in the original protocol submitted for ethics review. This Certificate of Approval is valid for the above time period provided there is no change in experimental protocol or consent process or documents.

Any significant changes to your proposed method, or your consent and recruitment procedures should be reported to the Chair for Research Ethics Board consideration in advance of its implementation.

ONGOING REVIEW REQUIREMENTS

In order to receive annual renewal, a status report must be submitted to the REB Chair for Board consideration within one month of the current expiry date each year the study remains open, and upon study completion. Please refer to the following website for further instructions: http://www.usask.ca/research/ethics_review/

[redacted]

Beth Bilson, Chair
University of Saskatchewan
Behavioural Research Ethics Board

Please send all correspondence to:

Research Ethics Office
University of Saskatchewan
Box 5000 RPO University, 1607 – 110 Gymnasium Plac
Saskatoon, SK S7N 4J8
Phone: (306) 966-2975 Fax: (306) 966-2069