Reasoning with Reasons in Case-Based Comparisons[1]

Kevin D. Ashley and Bruce M. McLaren
University of Pittsburgh, Pittsburgh, Pennsylvania 15260

Abstract. In this work, we are interested in how rational decision makers reason with and about reasons in a domain, practical ethics, where they appear to reason about reasons symbolically in terms of both abstract moral principles and case comparisons. The challenge for reasoners, human and artificial, is to use abstract knowledge of reasons and principles to inform decisions about the salience of similarities and differences among cases while still accounting for a case's or problem's specific contextual circumstances. TRUTH-TELLER is a program we have developed and tested that compares pairs of cases presenting ethical dilemmas about whether to tell the truth. The program's methods for reasoning about reasons help it to make context sensitive assessments of the salience of similarities and differences.

1. Introduction

A rational approach to decision making involves, at a minimum, elaborating reasons for and against a proposed action. Since reasons both for and against an action usually exist, a decision maker also requires some method for deciding which reasons are more important and should be able to explain the decision in terms of the reasons for it and a justification for the preference.

In a wide variety of domains, such as law, practical ethics, business, and organizational policy, resolving conflicting reasons involves both reasoning with principles or policies and comparing cases. A principle is a fundamental rule, law, or code of conduct. A policy is a definite, but general, method of action for guiding future decisions. Both are more abstract reasons which may underlie and give symbolic weight to a reason for an action. When facing a problem presenting conflicting reasons in a field like practical ethics, a decision maker may assess the reasons in light of their underlying principles or policies *and* compare the problem to past cases, involving some or all of the same principles or policies, to buttress an argument that some reasons should override others in the problem context. The reasoner may select paradigmatic, hypothetical and past cases identifying the principles or policies which made a decision to act right or wrong in the case, compare the cases and problem to see whether those criteria apply more or less strongly to the problem, make arguments how to resolve conflicting reasons in terms of the criteria as they were applied in the similar cases, and evaluate the arguments to come to a decision (Jonsen & Toulmin, 1988, Strong, 1988).

In "Interpretive CBR" (Kolodner, 1993), computer programs that generate case-based arguments justifying decisions, researchers are just beginning to find ways to incorporate reasoning with principles or policies into their case-based comparative

---

[1] **This paper was submitted to and accepted by the First International Conference on Case-Based Reasoning, Sesimbra, Portugal, October, 1995. It received, in a tie with two other papers, the "Most Distinguished Paper Award" for the conference**. This paper was based on Bruce McLaren's 1994 Masters project.

evaluation models. Integrating these different kinds of knowledge is a hard problem, even for humans, because they vary from very abstract principles through an intermediate range of reasons to very specific cases. Some people are better than others at resolving ethical dilemmas, cases presenting conflicting reasons for and against actions, backed by conflicting principles. Cognitive psychological evidence has shown, for instance, that gender and developmental differences affect a reasoner's ability to account for a dilemma's particular circumstances in applying general principles in moral decision making (Gilligan, 1982).

From an AI viewpoint, the challenge is to represent abstract principles and get programs to integrate reasoning with cases, reasons and underlying principles in a context sensitive manner. In general, context sensitivity in case comparison means knowing which similarities and differences are the most salient in different circumstances: which should a reasoner focus upon and which should it ignore. AI/CBR programs have explored different ways of representing the salience of similarities and differences and expanded the circumstances which a program can take into account in making a determination of salience. In HYPO (Ashley, 1990) and CATO (Aleven and Ashley, 1994), the circumstances included the side argued for, the set of cases being compared and the particular argument move involved (e.g., analogizing, distinguishing, citing counterexamples). In CABARET (Rissland and Skalak, 1991), the circumstances also included the arguer's viewpoint and various argument moves associated with broadening or restricting the meanings of decision rule predicates. GREBE (Branting, 1991) accounted for the presence or absence of criterial facts, facts an authoritative decision maker deemed important in a case. BankXX (Rissland, Skalak, and Friedman, 1993) added high level (legal) theories, standard stories and "family resemblance" into the mix. CASEY's justified match referred to a causal inference network and certain evidentiary principles to assess salience of differences (Koton, 1988). PROTOS employed weights reflecting explanatory significance and prototypicality (Bareiss, 1989). CHEF, PERSUADER, and PRODIGY-ANALOGY, in various ways, related the salience of similarities and differences to the existence or resolution of goal conflicts (Hammond, 1989, Sycara, 1987, Veloso, 1992). SWALE related them to expectation-violating anomalous outcomes (Kass et al., 1986), and CREANIMATE to functionality (Edelson, 1992).

We have made some progress in developing an AI/CBR program's ability to make context sensitive determinations of the salience of similarities and differences based on qualitative assessments of case similarity and general domain criteria for qualifying the absolute and relative importance of reasons. TRUTH-TELLER (TT) (Ashley and McLaren, 1994) compares pairs of cases presenting ethical dilemmas about whether to tell the truth and generates a comparison text contrasting the reasons in each case. The reasons may invoke ethical principles or selfish considerations. TT's comparisons point out ethically relevant similarities and differences (i.e., reasons for telling or not telling the truth which apply to both cases, and reasons which apply more strongly in one case than another or which apply only to one case). It is important to note that TRUTH-TELLER does not yet analyze problem situations by retrieving and selecting relevant paradigmatic cases. In this respect the work is quite different from the CBR programs referred to above, all of which retrieve and apply cases to problems. We hope to develop TRUTH-TELLER into such a program, computationally realizing, for instance, the case-based (i.e., "casuistic") model of moral decision-making proposed by the ethicist, Carson Strong

(Strong, 1988) which integrates principles, reasons and cases into a five-step model (See discussion in Ashley and McLaren, 1994).

In a formative evaluation of an initial version of the program, as reported in (Ashley and McLaren, 1994), an ethicist criticized TRUTH-TELLER's inability to "marshal" its answers. The expert commented generally that TT's comparison texts lacked an "organizing roadmap" and a recommended final decision "which could guide thinking and in terms of which the similarities, differences, and ethical principles could be marshaled." Accordingly, we have reorganized the comparison texts around specific conclusions and developed techniques for marshaling the similarities and differences in a context sensitive way.

As a result of these changes, TRUTH-TELLER now selects similarities and differences for salience in light of an overall assessment of similarity of the two cases; the program marshals the comparisons differently depending on how close the cases are to one another in terms of categories that ethicists seem to regard as important. TT's other heuristics for reasoning about reasons identify additional criteria for regarding some similarities and differences as more important than others in terms of underlying principles, criticalness of consequences, participants' roles and untried alternatives. We have developed a mechanism for "tagging" reasons with qualifications based on a comparative analysis of the cases.

In this paper we describe TRUTH-TELLER's expanded comparison algorithm and illustrate it with an extended example focusing on the way it employs marshaling and other techniques for reasoning about reasons to assess the salience of reasons in the two cases in a context sensitive way. We also report the results of a new experiment in which five expert ethicists evaluated the quality of TT's comparisons.

2. TRUTH-TELLER's Algorithm and Knowledge Structures

TRUTH-TELLER has a set of methods for reasoning about reasons that enables it to integrate reasons, principles, and cases intelligently in its case comparisons. Broadly characterized, TT's methods comprise three phases of analysis for (1) aligning, (2) qualifying, and (3) marshaling reasons, followed by (4) an interpretation phase. Each of the phases is described in more detail below:

The Alignment Phase. Aligning reasons means building a mapping between the reasons in two cases. The initial phase of the program "aligns" the semantic representations of the two input cases by matching similar reasons, actor relations, and actions, by marking reasons that are distinct to one case, and by noting exceptional reasons in one or both of the cases.

The Qualification Phase. Qualifying a reason means identifying special relationships among actors, actions, and reasons that augment or diminish the importance of the reasons. The qualification phase adjusts the relative importance of competing reasons or principles in the problem. During the qualification phase, heuristic production rules qualify or "tag" objects and the alignments between objects in a variety of ways based on considerations like criticalness, altruism, participants' roles and alternative actions.

The Marshaling Phase. Marshaling reasons means selecting particular similar or differentiating reasons to emphasize in presenting an argument that (1) one case is as strong as or stronger than the other with respect to a

conclusion, (2) the cases are only weakly comparable, or (3) the cases are not comparable at all. The marshaling phase analyzes the aligned and qualified comparison data, determines how the cases should be compared to one another based on five pre-defined comparison contexts reflecting a qualitative assessment of the overall similarity between the two cases, and then organizes information appropriate to that type of comparison.

The Interpretation Phase. A fourth phase of the program generates the comparison text by interpreting the activities of the first three phases.

The program employs various knowledge structures to support its algorithm, including semantic networks that represent the truth telling episodes, a relations hierarchy, and a reasons hierarchy. All structures are implemented using LOOM (MacGregor and Burstein, 1991).

The truth telling episodes were adapted from a game called Scruples (TM), from (Bok, 1989), and from a study that we conducted involving high school and ethics graduate students. Each episode includes representations for the actors (i.e., the truth teller, truth receiver, and others affected by the decision), relationships between the actors (e.g., familial, professional, seller-customer), the truth teller's possible actions (i.e., telling the truth, not telling the truth, or taking some alternative action) and reasons that support the possible actions.

The relations hierarchy is a taxonomy of approximately 80 possible relationships among the actors in a truth telling episode. Relationship types include familial, commercial, and acquaintance relations. The relations hierarchy is used to infer which relationships are "similar" for purposes of identifying levels of trust and duty that exist between the participants.

Finally, the reasons hierarchy represents possible rationales for taking action. Based on the formulation in (Bok, 1989), the hierarchy employs, at its top tier, four general reasons for telling the truth or not: fairness, veracity, producing of benefit, and avoiding harm. All other reasons are descendants of one of these abstract reason types. Each reason also has three other facets, criticalness, if altruistic, and if principled, each of which is important to ethical decision-making. The program accepts principled and unprincipled reasons as rationales for taking action.

3. An Extended Example

We now illustrate TT's algorithm and knowledge structures by providing an extended example of the program. A comparison of two cases by TT is depicted in Figure 1. We first describe the comparison text in its entirety, by describing in general terms what the program does, and then we focus on the underlined portion of the comparison text (the last paragraph of Figure 1), by guiding the reader through the four stages of the program.

The program starts by accepting semantic representations of each of the cases. The representation of a case is a manually constructed interpretation of the story. Figure 2 depicts the semantic representations of the Stephanie and Bruce cases. In Stephanie's case Stephanie is the truth teller, since it is she who is confronted with the decision to tell the truth or not. The experiment subjects will hear the truth, should Stephanie decide to divulge it, and thus are the truth receivers in this episode. Finally, the citizenry and the scientific community are affected others, since they would be affected by any truth telling disclosure (i.e., they stand to benefit in some way should the experiment result in an important scientific finding).

TRUTH-TELLER is comparing the following cases:

CASE 1: Should Stephanie, a psychology researcher, lie to human subjects about the intent of an experiment in order to study some aspect of the subject's behavior?

CASE 2: Bruce sells radios for a living. His favorite brother, Mark, picks out an expensive model with a history of maintenance problems. Selling this model would mean a big commission to Bruce but a big problem for Mark. Bruce has been doing very well lately, so the commission on this particular radio will not make much difference to his overall financial situation. Should Bruce warn his brother about the potential problems of this radio?

TRUTH-TELLER's analysis:

Stephanie and Bruce are faced with similar dilemmas. They abstractly share reasons to both tell the truth and not tell the truth. The cases also share similar relationship contexts. The relationship between Stephanie and the experiment subjects and between Bruce and Mark both involve a high level of duty.

Stephanie and Bruce abstractly share one reason to tell the truth. Both actors share the general reason to protect a right. More specifically, Stephanie has the reason to not trick someone into a disclosure for the experiment subjects, while Bruce has the reason to provide sales information so that a consumer can make an informed decision for Mark.

The two cases also abstractly share a reason to not tell the truth. Stephanie and Bruce share the general reason to produce benefit. Stephanie has the reason to enhance professional status and opportunities for herself, while Bruce has the reason to realize a financial gain for himself.

However, these quandaries also have relevant differences. Arguments can be made for both Stephanie and Bruce having a stronger basis for telling the truth.

On the one hand, there is an argument that telling the truth is better supported in Stephanie's case. First, Stephanie has to decide whether to tell a blatant lie, while Bruce must simply decide whether to remain silent. This fact would tend to put more pressure on Stephanie to tell the truth. Second, Stephanie could possibly acquire information for her research by devising a different experimental procedure. However, according to the story, this action was not taken. Thus, there is a greater onus on Stephanie to be honest.

On the other hand, one could also argue that Bruce has a more compelling case to tell the truth. First, the shared reason for telling the truth 'to protect a right' is stronger in Bruce's case, since it involves a higher level of trust between Bruce and Mark. Second, the shared reason for not telling the truth 'to produce benefit' is weaker in Bruce's case, since Bruce's potential profit will not make much difference to his overall financial situation. Third, Stephanie has the reason to not tell the truth to strive for a greater good for the citizenry. Finally, Bruce's motivations for not telling the truth, unlike Stephanie's, appear to be purely selfish. This increases the onus on Bruce to tell the truth.

Figure 1: TT's Output Comparing Stephanie's and Bruce's Cases

The semantic representation also contains a set of possible actions that the truth teller could take and reasons supporting or justifying each of the actions. The actions and reasons are supplied manually. (In future work, we hope that TT will infer actions and reasons by comparing cases.) One of the possible actions is always to tell the truth and another is some version of not telling the truth, for instance, telling a lie or keeping silent (i.e., not disclosing information). In Stephanie's case, the choice is between telling the truth about the intent of the experiment or deceiving the subjects by telling a premeditated lie. The lie is premeditated since, supposedly, Stephanie would know the intent of her own experiment and would therefore have the opportunity to reflect on the decision to lie or not. Stephanie also has an alternative action she could take before deciding whether or not to lie to the experiment subjects:

she could pursue an alternative experimental design that would not require deceit. (Notice that this alternative is not explicitly stated in the story. In interpreting the episodes for semantic representation, we sometimes extrapolated.)

In our knowledge representation, actions are supported by reasons; a reason is treated as a rationale for taking an action. For example, a rationale for Stephanie's telling the truth is to protect the right of the experiment subjects to not be tricked into the disclosure of information. A rationale for Stephanie telling the lie is to provide a "greater good" for the scientific community and the citizenry at large, i.e., the experiment may result in some general benefit for many people.

In this work, we focus on TRUTH-TELLER's techniques for reasoning about the reasons contained in a case in the course of comparing that case to another case. The underlined portion of the comparison text of Figure 1 is part of the argument for Bruce's having a more compelling case to tell the truth than Stephanie. A portion of the Stephanie and Bruce semantic networks and the comparison between these portions (Figure 3) is directly responsible for this text. The following paragraphs explain how this diagram and the four phases of the program led to the underlined comparison text.

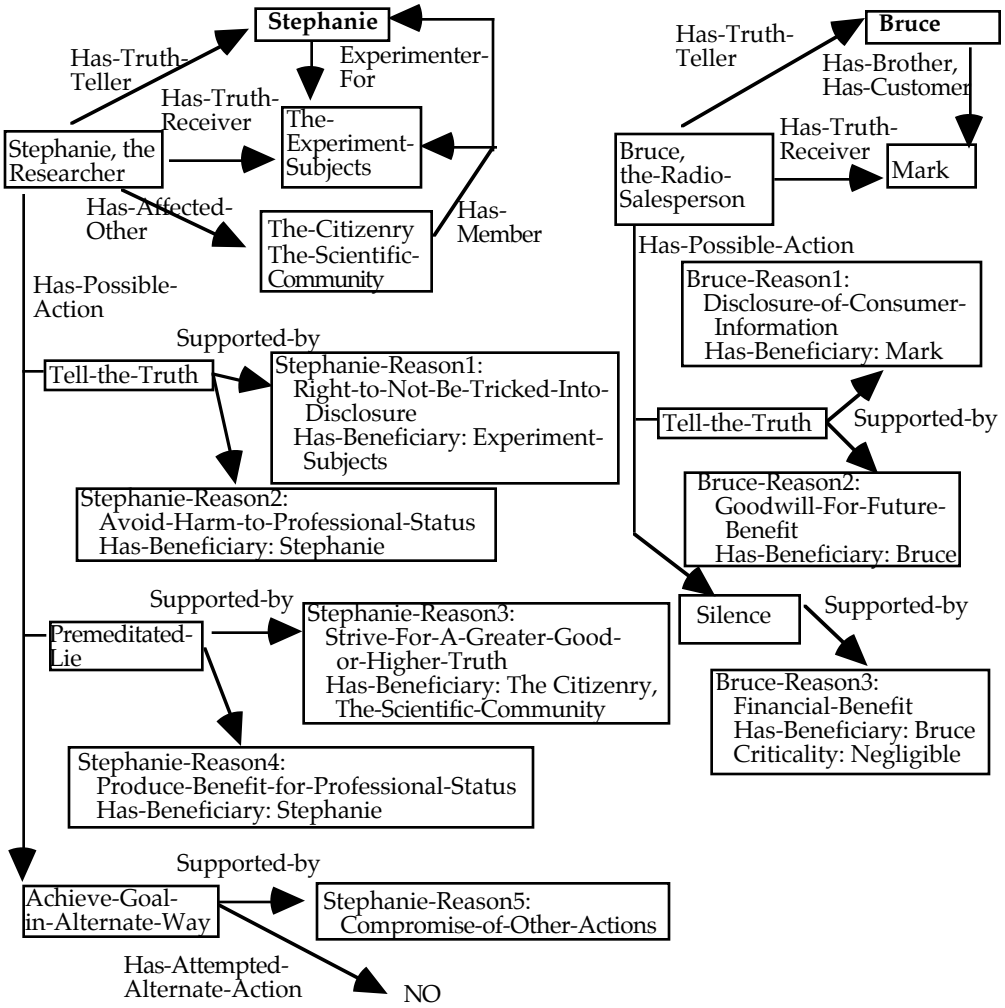


Figure 2: Representation: Stephanie's Case (l) & Bruce's Case (r)

Given the input representations, TT reasons about reasons by aligning, qualifying, and marshaling the semantic networks. First, the Alignment phase is initiated. The thin lines in Figure 3 depict alignments between the Stephanie and Bruce representations. First, Stephanie's reason involving a "greater good" is determined to have no counterpart in the Bruce representation (i.e., it is a clear distinction); thus it is "misaligned" and labeled as a reason distinction. Second, the premeditated lie for Stephanie and silence for Bruce -- possible actions for each case -- are aligned with one another because they abstractly match as a "do not tell the truth" action. Finally, Stephanie and Bruce have reasons that abstractly match and are thus aligned with one another (i.e., Stephanie may benefit professionally by lying to her subjects and conducting the experiment, while Bruce may benefit financially by

withholding the truth from his brother). These reasons are not identical; however they match in the reason hierarchy at the level of "producing benefit."

The program next commences the Qualification phase. The italicized text in Figure 3 represents the qualifications that are applied to the comparison. During this phase individual objects and alignments between objects are qualified or "tagged" based on finer-grained knowledge. The first step is to qualify the individual objects. For instance, Stephanie's reason to produce benefit (i.e., to improve her professional status) is tagged (1) as having no duty or trust expectations, since it only involves herself, (2) as being unprincipled, since no ethical principle supports the reason, (3) as having average criticality, since no comment was made in the story about any critical consequences, and (4) as being self-motivated, since the reason is clearly selfish. Bruce's aligned reason (i.e., financial benefit) is tagged likewise, with the exception that its criticality is labeled as negligible, since the story states that the financial benefit to Bruce is not important.

The second step of Qualification is to qualify alignments. In Figure 3 there are three alignment qualifications. The first one is the reason distinction misalignment which is tagged as being a reason found only in case 1. Second, the alignment between Stephanie's premeditated lie and Bruce's silence is tagged as being fully selfish, and thus weaker, on Bruce's side, since Bruce has only a selfish reason for withholding the truth, while Stephanie has one rationale, the "greater good" reason, that is not selfish. Finally, the abstract reason match between Stephanie's possible professional gain and Bruce's financial benefit is tagged as stronger for Stephanie. This is the case because of the superiority of average criticality over negligible criticality. (A Reason A is said to be stronger than a Reason B iff Reason A has a higher criticality than Reason B OR Reason A has at least one qualifier (i.e., trust, duty, altruism, principled) that is stronger than Reason B's AND Reason A has no qualifier that is weaker than Reason B's.)

Next, the program begins the Marshaling phase. Its first marshaling task is to assign the case comparison to one of five possible comparison contexts. The five comparison contexts are defined as follows: (1) Comparable-Dilemmas/Reason-Similarity, if the cases present similar dilemmas. i.e., the reasons supporting both telling the truth and not telling the truth are similar either in an identical or abstract way, (2) Comparable-Dilemmas/Criticality-Similarity, if the cases are similar due to the critical nature of possible consequences, (3) Comparable-Reasons, if the cases share a similar reason or reasons for either telling the truth or not telling the truth but not for both possible actions, (4) Incomparable-Dilemmas/Reason-Difference, if the cases do not have any reasons supporting like actions that are similar, and (5) Incomparable-Dilemmas/Criticality-Difference, if the cases are incomparable due to a difference in the criticality of the possible consequences.

The Stephanie/Bruce comparison is classified as an instance of the Comparable-Dilemmas/Reason-Similarity comparison context, since it involves abstract reasons to tell the truth and abstract reasons not to tell the truth. After classifying the comparison, the program then marshals information that is appropriate to the classified context. There are two general categories of information that are marshaled, the *comparison focus* (i.e., information that is to be the initial focus of comparison and is typically the most important information to draw attention to) and the *distinguishing information* (i.e., information that contrasts with the comparison focus). For instance, for the Comparable-Dilemmas/Reason-Similarity comparison context, the program marshals the similar reasons and relations as the comparison

focus and then, to distinguish the cases, marshals the information that supports arguing the relative merits of telling the truth in the two cases. As another example, if a comparison is assigned to the Incomparable-Dilemmas/Criticality-Difference, the program marshals the reasons that have greater criticality as the comparison focus, and then marshals, as distinguishing information, the contrasting, less critical reasons of the less critical case.

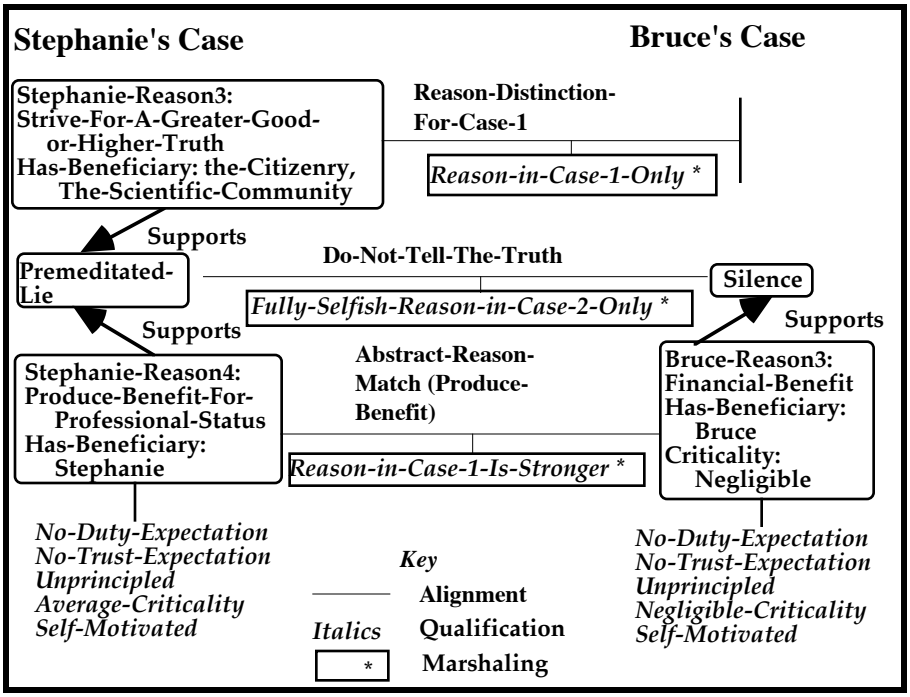


Figure 3: Comparison after Alignment, Qualification, and Marshaling

Now let us return to Figure 3 to explain how the data in the figure is marshaled. Marshaled information is enclosed in a box with an asterisk. Only the marshaled distinguishing information is depicted in the diagram, corresponding to the underlined text in Figure 1. It is interesting to note, however, that the abstract reason match in Figure 3 is actually marshaled as part of both the comparison focus and the distinguishing information. This is so because the abstract reason match is a similarity, but qualification has also revealed that, at a more detailed level, it gives rise to a distinction. Note that the qualifier *Reason-In-Case-1-Is-Stronger* strengthens the case for Bruce's telling the truth relative to Stephanie, since Bruce's reason is less compelling for not telling the truth. This marshaled data corresponds to the sentence in the comparison text beginning "Second, ..." Stephanie's reason for not telling the truth to attain a "greater good" is also marshaled as a strength of Bruce's case relative to Stephanie's, because it is a misaligned reason distinction (i.e., it provides a justification for Stephanie to not tell the truth that is unshared by Bruce). This corresponds to the sentence "Third, ..." Finally, the program marshals the qualifier, *Fully-Selfish-Action-In-Case-2-Only*, from the matching actions in Figure 3. This also supports Bruce's case relative to Stephanie's, since it shows a

weakness for not telling the truth that exists for Bruce but not Stephanie. This final marshaled data corresponds to the text in Figure 1 beginning "Finally, ..." The natural language text is generated in the final phase, Interpretation, using an augmented transition network.

To summarize, the extended example shows how TT uses its 4-phase algorithm to generate context sensitive and marshaled case comparisons. It "reasons about reasons" by aligning cases according to similarities and differences, and qualifying reasons in various ways including tagging reasons as altruistic, principled, critical, high trust, high duty, etc. and tagging alignments with relative strengths. The program marshals the reasons and qualifications by recognizing the context represented by a pair of cases. Figure 4 summarizes seven ways that TT reasons about reasons. The Alignment phase employs methods 1 through 4, the Qualification phase employs methods 5 and 6, and the Marshaling phase employs 7.

1. Elicit principles underlying reasons and classify reasons individually: Reason Hierarchy follows links from reason type to principles. Classify reasons as principled, self-motivated, or altruistic.
2. Classify reasons in the aggregate: Note if all reasons supporting an action are principled or unprincipled, altruistic or self-motivated.
3. Match reasons: Identify reasons for a particular action shared by cases and reasons not shared. Matches may be exact or abstract, based on a hierarchy of reasons and principles. Also, note exceptional reasons in one or both cases or reasons distinct to one case.
4. Map reason configurations: Mark shared configurations of reasons such as shared dilemmas (i.e., similar opposing reasons).
5. Qualify reasons by: (a) criticality (what happens if action is not taken?), (b) whether altruistic or not, (c) whether principled or not, (d) participants' roles and relationships (e.g. trust, duty), (e) existence of untried alternative actions, (f) how others are affected by action, (g) comparing actions (Lie vs. Silence, Premeditated vs. Unpremeditated)
6. Compare overall strength of reasons: Use the qualifiers to decide whether one reason is "stronger" than another.
7. Marshaling reasons: Select, collect reasons to emphasize based on the overall similarity of cases, nature of the reason mapping and qualifications on reasons.

Figure 4: TRUTH-TELLER's techniques for reasoning about reasons

4. The Evaluation

Our goal was to obtain some assurance that TRUTH-TELLER's techniques for reasoning about reasons generated case comparisons that expert ethicists would regard as appropriate. Our experimental design for this formative evaluation was to poll the opinions of five expert ethicists as to the reasonableness, completeness, and context sensitivity of a relatively large sampling of TT's case comparisons.

We divided the evaluation into two parts. The first experiment presented the experts with twenty comparison texts TT generated for pairs of cases randomly selected from three classes described below. The comparison texts were similar to and included the one in Figure 1. We also added two comparison texts generated by humans (a medical ethics graduate student and a law school professor). The experts were advised that the texts had been generated by humans or a computer program, but they were not told which texts or how many texts were generated by which. The

second experiment presented the experts with five comparison texts in which TT compared the same case to five different cases. For each experiment, the evaluators were instructed: "In performing the grading, we would like you to evaluate the comparisons as you would evaluate short answers written by college undergraduates. ... Please focus on the substance of the comparisons and ignore grammatical mistakes, awkward constructions, or poor word choices (unless, of course, they have a substantial negative effect on substance.)" We also instructed the experts to critique each of the comparison texts.

In the first experiment, we instructed the experts to assign three grades to each of the twenty-two comparison texts, a separate grade for reasonableness, completeness, and context sensitivity. The scale for each grading dimension was 1 to 10, to be interpreted by the evaluators as follows: for reasonableness, 10 = very reasonable, sophisticated; 1 = totally unreasonable, wrong-headed; for completeness, 10 = comprehensive and deep; 1 = totally inadequate and shallow; for context sensitivity, 10 = very sensitive to context, perceptive; 1 = very insensitive to context. The twenty case pairs presented to TRUTH-TELLER were selected randomly as follows: (1) two cases selected at random from the training set (total of 5). (2) one case selected at random from the training set and one case selected at random from the test set (total of 5). (3) two cases selected at random from the test set (total of 10). The *training set* comprised twenty-three of the fifty-one cases in TRUTH-TELLER's case base; these cases were used to develop the program. The other twenty-eight cases served as the *test set* for the evaluation; these were used sparingly (or, in most cases, not at all) to develop the program.

The results of the first experiment were as follows. The mean scores across the five experts for the twenty TT comparisons were R=6.3, C=6.2, and CS=6.1. Figure 5 shows the maximum, minimum and mean scores per comparison for all three of the dimensions. By way of comparison, the mean scores of the two human-generated comparisons were R=8.2, C=7.7 and CS =7.8. The mean scores for the Stephanie/Bruce comparison, number 13, were R = 6.2; C = 7.2; CS = 5.8. Not surprisingly, one of the human comparisons, number 16, attained the highest mean on all three dimensions (R = 9; C = 8.8; CS = 8.8). Two of the program generated comparisons (numbers 2 and 14), however, were graded higher on all three dimensions than the remaining human comparison (number 22).

In the second part of the evaluation, we wanted to assess the program's sensitivity to context. To achieve this, we asked the experts to grade five additional TRUTH-TELLER comparisons. These comparisons all involved one case repeatedly compared to a different second case (i.e., a One-To-Many comparison). For this part of the evaluation, the experts were asked to grade the set of all comparisons with a single score along each of the three dimensions (i.e., reasonableness, completeness, and context sensitivity).

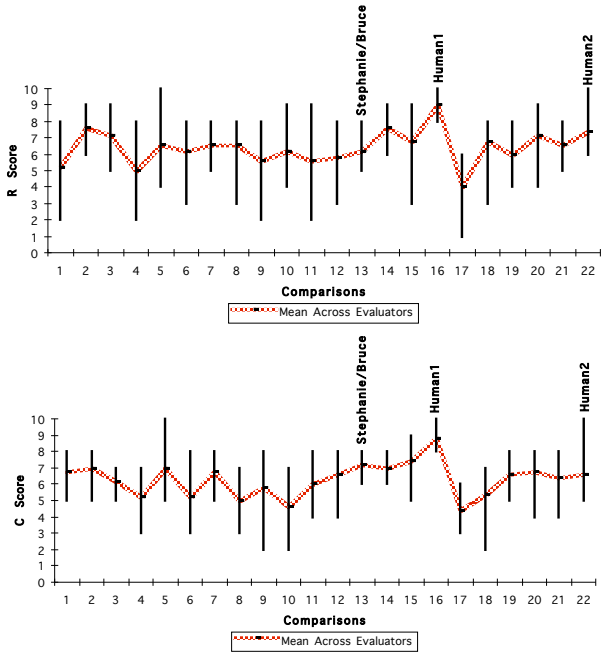
The results of the second part of the evaluation were as follows. The mean across all evaluators was R = 6.7; C = 6.9; CS = 7.0. Notice that the program fared better on the context sensitivity dimension than on the other two dimensions. This contrasts with the first experiment in which the mean CS score was the lowest of the three dimensions. Also, notice that the scores of all three dimensions were improved slightly over the first experiment.

5. Discussion and Conclusions

Our results should be viewed in light of our goals and the experimental design in this formative evaluation. We solicited expert opinions about the adequacy of TRUTH-TELLER's comparison texts in order to assess whether our knowledge

representation and reasoning techniques were appropriate to the domain task and to obtain critiques identifying areas for improvement. Our primary intention was to determine if TT's comparisons were at least "within range" of that of humans and to determine the ways in which our model could be improved. We interpret the results as indicating that TRUTH-TELLER is somewhat successful at comparing truth telling dilemmas. Given the instruction to "evaluate the comparisons as you would evaluate short answers written by college undergraduates," we are encouraged by the grades assigned. We included the two human-generated texts as a calibration of the experts' scores; we are encouraged that some of the program's grades were higher than those assigned to texts written by post graduate humans.

On the other hand, our experiment does not involve an adequately sized sampling of human comparisons nor did we present the experts with outputs in which TRUTH-TELLER and humans generated comparison texts for the same pairs of cases. Quite simply, we felt it was premature to adopt this kind of experimental design for a formative evaluation. We recognize, however, that such an experimental design provides greater assurance of the quality of any results and will be employed for a future summative evaluation.



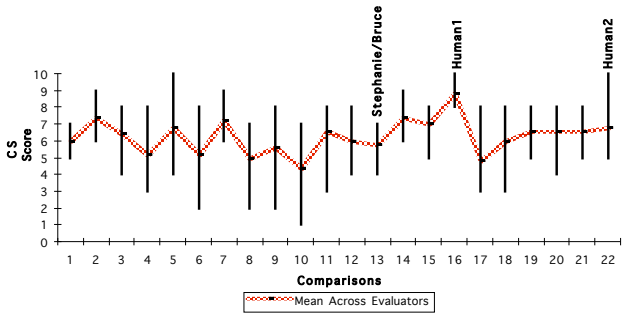


Figure 5: Max, Min, and Mean Values for R (top), C (middle), and CS (bottom) Scores in Experiment #1

The second part of the experiment attempts to address whether TRUTH-TELLER is competent at marshaling comparisons in a context sensitive manner. We believe that the slightly higher scores in the second part of the experiment are due in part to the fact that it would have been easier for the evaluators to recognize TT's sensitivity to context in the one-to-many part of the experiment than in the first part. Upon recognizing the difficulty of context sensitive comparisons and TT's ability to tackle them, the evaluators tended, we believe, to grade the program higher on all dimensions. In fact, as was noted, the context sensitivity dimension graded higher than the other two dimensions in this part of the experiment.

We also invited the evaluators to criticize the comparison texts. Several evaluators questioned TT's lack of hypothetical analysis; the program makes immutable assumptions about reasons, actions, and actors. We will explore the possibility of hypothetically modifying problems in terms of factors as in HYPO (Ashley, 1990). Another repeated criticism involved the use of aggregate reasons to support an action. This was due, at least in part, to the fact that aggregate reasons are hard for the program to explain, since they require reference to other parts of the text. However, we have also considered that the calculus of support for actions and comparison of actions may be more complex than focusing simply on whether an action is fully supported by altruism or principled reasons. For instance, one could argue that an action that is mostly altruistic, but happens fortuitously to lead to a selfish side benefit, is as good as a fully altruistic action.

In conclusion, the evaluation encourages us that TRUTH-TELLER makes mostly reasonable comparisons of practical ethical cases with some ability to assess the salience of similarities and differences in a context sensitive manner. Its determinations of salience take into account a qualitative assessment of the cases' overall similarity and its other heuristics for reasoning about reasons identify additional criteria involving underlying principles, criticalness of consequences, participants' roles and untried alternatives.

References

Aleven, V. and Ashley, K.D. (1994). An Instructional Environment for Practicing Argumentation Skills; In the Proceedings of *AAAI-94*, pages 485-492.

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

Ashley, K.D. and McLaren, B.M. (1994). A CBR Knowledge Representation for Practical Ethics; In *Proceedings, 2d EWCBR*.

Bareiss, E. R. (1989). *Exemplar-Based Knowledge Acquisition - A Unified Approach to Concept Representation, Classification, and Learning*. Academic Press, San Diego, CA, 1989. Based on PhD dissertation, 1988.

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

Branting, K. L. (1991). Building Explanations From Rules and Structured Cases. In the *Journal of Man-Machine Studies*. 34, 797-837.

Edelson, D. C. (1992). When Should A Cheetah Remind You of a Bat? Reminding in Case-Based Teaching. In *Proceedings of AAAI-92*, 667-672. San Jose, CA.

Gilligan, C. (1982). *In a Different Voice*. Harvard University Press.

Hammond, K. (1989) *Case-Based Planning -- Viewing Planning as a Memory Task*. San Diego, CA: Academic Press.

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

Kass, A. M., Leake, D., and Owens, C. C. (1986). Swale: A Program that Explains. In Schank, R. C. (ed.), *Explanation Patterns: Understanding Mechanically and Creatively*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Kolodner, J. (1993) *Case-Based Reasoning* Morgan Kaufmann Publishers, Inc., San Mateo, CA.

Koton, P. (1988) *Using Experience in Learning and Problem Solving* . PhD thesis, MIT.

MacGregor R. and Burstein, M. H. (1991). Using a Description Classifier to Enhance Knowledge Representation. In *IEEE Expert* 6(3), pages 41-46.

Rissland, E. L. and Skalak, D. B. (1991). CABARET: Rule Interpretation in a Hybrid Architecture. In the *Journal of Man-Machine Studies*. 34, pages 839-887.

Rissland, E. L., Skalak, D. B., and Friedman, M. T. (1993). BankXX: A Program to Generate Argument through Case-Based Search. In *Fourth International Conference on AI and Law*, Vrie Universiteit, Amsterdam.

Strong, C. (1988). Justification in Ethics. In Baruch A. Brody, ed., *Moral Theory and Moral Judgments in Medical Ethics*, pp. 193-211. Kluwer, Dordrecht.

Sycara, E. P. (1987). *Resolving Adversarial Conflicts: An Approach Integrating Case-Based and Analytic Methods* Georgia Inst.Tech., Tech. Rep.GIT-ICS-87/26. Atlanta.

Veloso, M. V. (1992). *Learning by Analogical Reasoning in General Problem Solving*. PhD thesis, Carnegie Mellon University. Technical Report No. CMU-CS-92-174.