Toward Modeling and Teaching Legal Case-based Adaptation with Expert Examples

Kevin Ashley[1] Collin Lynch[2], Niels Pinkwart[3], Vincent Aleven[4]

[1] Intelligent Systems Program (ISP), Learning Research and Development Center & School of Law, University of Pittsburgh, Pittsburgh, Pennsylvania, USA, ashley@pitt.edu
[2] ISP, University of Pittsburgh, Pittsburgh, Pennsylvania, USA, collinl@cs.pitt.edu
[3]CSI, Clausthal U. of Technology, Clausthal, Germany, niels.pinkwart@tu-clausthal.de
[4]Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, aleven@cs.cmu.edu

Abstract. Studying examples of expert case-based adaptation could advance computational modeling but only if the examples can be succinctly represented and reliably interpreted. Supreme Court justices pose hypothetical cases, often adapting precedents, to evaluate if a proposed rule for deciding a problem needs to be adapted. This paper describes a diagrammatic representation of adaptive reasoning with hypothetical cases based on a process model. Since the diagrams are interpretations of argument texts, there is no one "correct" diagram, and reliability could be a challenge. An experiment assessed the reliability of expert grading of diagrams prepared by students reconstructing examples of hypothetical reasoning. Preliminary results indicate significant areas of agreement, including with respect to the ways tests are modified in response to hypotheticals, but slight agreement as to the role and import of hypotheticals. These results suggest that the diagrammatic representation will support studying and modeling the examples of case-based adaptation, but that the diagramming support needs to make certain features more explicit.

Keywords: Case-based adaptation, Hypothetical reasoning, Legal reasoning

1 Introduction

A repository of transcripts of human experts solving problems through case-based adaptation can be a valuable resource for CBR research as a modeling and teaching tool. Given the continuing dearth of empirical data about how humans modify cases to solve problems [24], this resource could support developing computational models of and teaching case-based adaptation [9, p. 7].

The oral arguments of the U.S. Supreme Court (SCOTUS) are one such repository. The transcripts record the use by human expert decision-makers (i.e., the Justices) of case-based reasoning to explore a space of possible solutions as they respond to the recommendations urged by advocates. Each transcript is an extended argument about how to decide the case in the form of a "multilogue" between one advocate at a time and the nine Justices. The arguments are inherently case-based and include proposing

tests (i.e., rules) for deciding the case, drawing analogies to past cases (i.e., precedents), justifying the analogies in terms of principles and policies underlying the legal domain, challenging the proposed tests by posing hypothetical cases and responding to the hypotheticals, for instance, by modifying the proposed test [3].

Designing the hypotheticals and modifying the tests are a kind of case-based adaptation that we hope to study empirically and to model computationally. In order to flesh out a model for generating hypotheticals and adapting tests, we need more, and more detailed, SCOTUS examples and a way to represent them. For our LARGO program, we developed a diagrammatic representation capturing arguments involving hypothetical reasoning in a succinct way that is partially interpretable by the program.

Given our modeling and pedagogical goals, it is important that the argument diagrams are interpretable in a reliable way. Humans must be able to understand and evaluate argument diagrams reliably in order to model the examples, especially if the diagrams will someday be an input/output medium for a program that instantiates the computational model. Inter-rater reliability in interpreting the argument diagrams is, therefore, a precondition for making further progress in modeling and teaching.

Since the diagrams are interpretations of argument texts, however, there will not be a single "correct" diagram. These are complex real examples of case-based adaptation, expressed in text of which the diagrams are interpretations. The texts may be incomplete, and even if not, multiple reasonable interpretations of the texts are normal. Legal problems are ill-defined; there is no one right answer but often competing reasonable arguments employing different interpretations of open-textured terms. That is the reason hypothetical reasoning is important as a technique for dealing with open textured legal terms. This implies that reasonable people may differ as to the description of the role and import of a hypothetical or the level of abstraction with which to formulate the proposed tests.

Thus, it is an empirical question whether the diagrams can be interpreted reliably. An experiment assessed the reliability of expert grading of diagrams prepared by students as they reconstructed examples of hypothetical reasoning in SCOTUS oral arguments. This paper presents preliminary results with respect to inter-rater reliability. In Section 2, we present an example of hypothetical reasoning that highlights the case-based adaptation and a process model of hypothetical argument that provides a high-level account of it. Section 3 relates the current work to previous work on case-based adaptation and reasoning with examples and hypotheticals. Section 4 illustrates the diagrammatic representation of the same example using our LARGO program, an intelligent tutoring system (ITS) designed to teach law students the process of hypothetical argument. The experiment to assess the reliability of interpreting LARGO diagrams of hypothetical reasoning is described in Section 5, where the results are presented and discussed. Conclusions follow in Section 6.

2 Reasoning with hypothetical cases and adaptation

The resolution of a case before a court may be subject to conflicting legal principles. The resolution comprises: (1) a result (e.g., the winner is the party that brings suit, the plaintiff, or the opponent against whom suit is brought, the defendant); (2) a rule that

generates that result when applied to the case facts; and (3) a justification of the result and the rule as consistent with precedents and principles/policies.

Hypothetical reasoning involves generating and testing a rule for deciding the dispute. The proposed test is a hypothesis about how to decide the case in the form of rule the advocate proposes and defends as consistent with past cases and underlying principles/policies. A hypothetical is an imagined case that involves such a hypothesis (i.e., a proposed test) and is designed to explore its meaning or challenge it.

The process of hypothetical reasoning incorporates case-based adaptation, both in the design of an appropriate hypothetical and in the modification of the test. While the hypotheticals are figments of the Justices' imagination, often, they are adaptations of the facts of the current case or past cases. The hypothetical often is designed so that the proposed test applies but reaches a result that contradicts one or more of the underlying principles/policies. That is, the test is too broad. In other situations it is constructed so that the test does not apply but should do so according to one or more of the principles/policies (i.e., the test is too narrow.) In response to the hypothetical, an advocate may adapt the test, narrowing or broadening it as appropriate.

2.1 Example of Reasoning with Hypotheticals

We have assembled examples of hypothetical reasoning from a variety of SCOTUS oral arguments, including cases involving freedom of religion, the search warrant requirement, copyright infringement, and, as illustrated here, personal jurisdiction. A standard topic addressed in first year legal process courses, personal jurisdiction refers to the power of a court under the U.S. Constitution to compel a party from outside the state in which the court is located to appear and defend a lawsuit. The underlying legal principles/polices include the due process concern of ensuring fairness to the defendant in requiring him to appear in court within the state versus the state's interest in adjudicating issues and disputes affecting its residents.

The example is based on the petitioner's argument in *Asahi Metal Industry Co. v. Superior Court of California*, 480 U.S. 102 (1987), a case that involved an issue of personal jurisdiction. Specifically, the question is whether Asahi, a Japanese company, may be called into a California court to answer in a civil suit for injuries caused by a blowout of an allegedly faulty motorcycle tire of which Asahi manufactured one component, the tube's valve assembly. Over a fourteen year period, Asahi sold at least 100,000 tire valve assemblies to Cheng Shin, the Taiwanese tube manufacturer. Evidence suggested that Asahi was aware that at least some of its tire valve assemblies would end up in the United States. Furthermore, about twenty percent of the Cheng Shin tires exported to the United States were sold in California.

SCOTUS oral arguments occur after the parties have submitted briefs but before the Justices decide a case or draft an opinion; each side's advocate has one half hour to press his case before the Justices. Since the Court may not be bound to follow the rule in a precedent, or may reinterpret a rule, much of the "action" involves debating about how best to formulate/interpret a rule for deciding the case. For instance, Asahi's advocate, Mr. Staring (Mr. S) argued (in ll. 36-43) that even if it were foreseeable that Asahi's products would end up in California (CA), that was not enough to subject Asahi to the jurisdiction of a CA court. In Fig. 1, box [a] shows an

interpretation of Mr. S's point as a proposed test for deciding the case in favor of his client. It is an "interpretation", because, as frequently occurs in oral argument, the advocate's and Justices' positions need to be inferred from what they say; in the oral medium under extreme time pressure, their comments are often very brief.

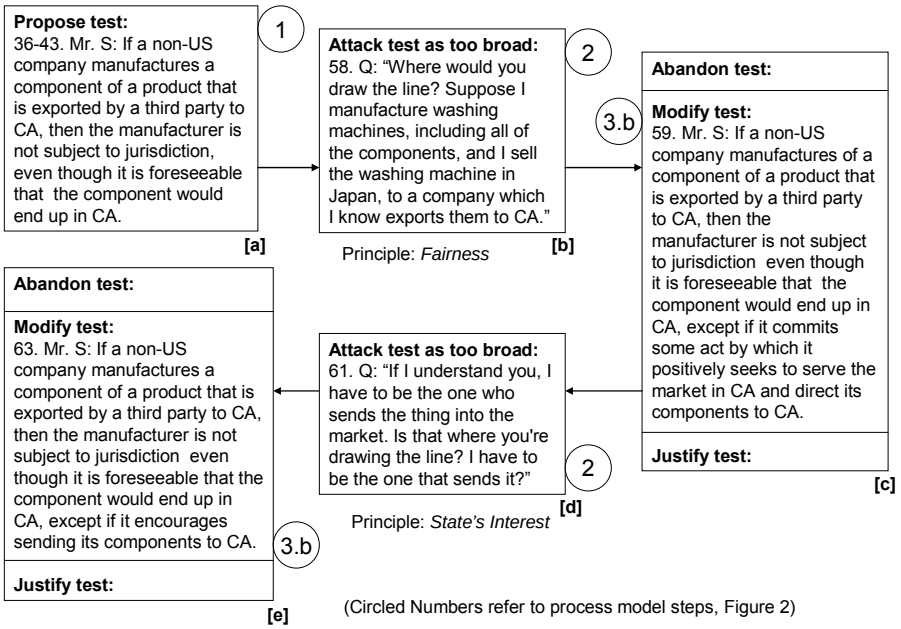

Fig. 1. Posing Hypotheticals to Attack Proposed Test as Too Broad and Modifying Tests

This proposed test leads the Justice (in box [b], l. 58 – the line numbers are included to show the proximity of the moves in the transcript) to respond with a hypothetical that suggests the advocate's test is too broad in defining what would *not* be sufficient for jurisdiction and that a line needs to be drawn somewhere. (The circled numbers in Fig. 1 refer to the steps in the process model of Fig. 2. Posing a hypothetical to challenge the test as too broad is Step 2.) Surely, it would be sufficient if the manufacturer knew that his purchasing exporter was going to send them to California. The change in facts may seem innocuous, but it has a significant effect. In posing the hypothetical, the Justice implies that it clearly would satisfy due process fairness to subject a manufacturer to jurisdiction in a state where he *knew* his product would be shipped. In response, Mr. S. makes his proposed test more specific (in box [c]) by introducing an exception requiring some kind of positive act by the manufacturer to serve the California market. (The exception to the test's definition of what is *not* subject to jurisdiction acts as a limitation or narrowing of this test.)

The advocate's narrowing of the test does not sufficiently address one Justice's concerns about where to draw a line; he worries that the modified test, with its exception requiring some positive act, is still too broad in defining what is not subject to California's jurisdiction. In box [d], the Justice poses a hypothetical implying that

subjecting a manufacturer to a state's jurisdiction only if he sends the product to the state, limits too severely the state's interest in enabling its citizens to redress injuries through its courts. In response, box [e], Mr. S. again narrows his test by expanding the exception somewhat. He concedes that if the manufacturer encourages the sending of the product to California, then it should be subject to California's jurisdiction. At the same time, Mr. S. would maintain, Asahi did nothing to encourage the sending of the tube assemblies to California, distinguishing the case at hand from the hypothetical.

2.2 Process Model of Hypothetical Argument

The Process Model of Hypothetical Argument presented in Fig. 2, provides a partial account of the hypothetical reasoning examples we have encountered in SCOTUS oral arguments [3], including the *Asahi* example in Fig. 1. An advocate proposes a test (i.e., a general rule) for deciding the case at hand (step 1 in Fig. 2 illustrated in box [a], Fig. 1). The Justices challenge the proposed test, posing a hypothetical case in order to determine how the proposed test would handle it. Their goal may be to critique the proposed test as too broad (step 2 in Fig. 2, illustrated in boxes [b] and [d], Fig. 1) or too narrow (steps 2′ and 3′ below the ellipsis, Fig. 2). Alternatively, the Justices may pose a hypothetical case, not in order to critique the test, but simply to explore if the test applies to the hypothetical case and with what result. In modeling this more exploratory use of hypothetical cases, one can relax certain criteria in step 2 or 2′, but we do not pursue that here.

In responding at step 3 (or 3′), Fig. 2, the advocate may: (a) stick with his test, justifying it as correctly deciding the case at hand despite the hypothetical, (b) modify the test so that it still assigns the advocate's preferred result in the case but also accommodates the hypothetical, or (c) give up the test and propose another. In the *Asahi* example, Fig. 1, the advocate modifies the test in boxes [c] and [e].

The Process Model incorporates some traditional case-based moves. When responding that the test is not too broad (step 3), justifying the test (3.a) involves analogizing the hypothetical and case; modifying the test (3.b) involves distinguishing the hypothetical from the case. Responding that the test is not too narrow (step 3′) involves just the reverse: distinguishing in 3′.a and analogizing in 3′.b. The analogizing involves pointing out relevant shared facts that are reasons for deciding the case and hypothetical the same way. Distinguishing involves pointing out relevant unshared facts that are reasons for deciding the real and hypothetical cases differently.

Facts are relevant, and thus suitable for analogizing and distinguishing, if and to the extent that they matter given the principles/policies of the law. When case facts connect to the law's and regulations' underlying principles/policies, they justify deciding the case consistently with those principles/policies. These principles/policies embody the goals that laws and legal regulations are designed to achieve, for example, to avoid intentionally-inflicted personal injuries, encourage economic competition, discourage frivolous lawsuits, or protect citizens from arbitrary government power. This last is the goal of the law of personal jurisdiction: the due process concern with fairness protects out-of-state citizens from having to defend themselves in court in states to which they have no substantial connections.

→ **1. Propose test: For proponent, propose test for deciding the current fact situation (cfs):** Construct a proposed test that leads to a favorable decision in the cfs and is consistent with applicable underlying legal principles/policies and important past cases, and give reasons.

← **2. Pose hypothetical: For interlocutor, pose hypothetical example to probe if proposed test is too *broad*:** Construct a hypothetical example that:
- (a) emphasizes some normatively relevant aspect of the cfs and
- (b) to which the proposed test applies and assigns the same result as to the cfs, but
- (c) where, given legal principles/policies, that result is normatively wrong in the hypothetical.

→ **3. Respond: For proponent, respond to interlocutor's hypothetical showing test too broad:**
- (3.a) Justify the proposed test: Analogize the hypothetical example and the cfs and argue that they both should have the result assigned by the proposed test. *Or*
- (3.b) Modify the proposed test: Distinguish the hypothetical example from the cfs, argue that they should have different results and that the proposed test yields the right result in the cfs, and add a condition or limit a concept definition so that the narrowed test still applies to the cfs but does not apply to, or leads to a different result for, the hypothetical example. *Or*
- (3.c) Abandon the proposed test and return to (1) (i.e., construct a different proposed test that leads to a favorable decision in the cfs and is consistent with applicable underlying legal principles/policies, important past cases, and hypotheticals…)

…

← **2′. Pose hypothetical: For interlocutor, pose hypothetical example to probe if proposed test is too *narrow*:** Construct a hypothetical example that:
- (a) emphasizes some normatively relevant aspect of the cfs, and
- (b) that normatively should have the same result as the cfs, but
- (c) to which the test does not apply or assigns a different result.

→ **3′. Respond: For proponent, respond to hypothetical example showing test too narrow:**
- (3′.a) Justify the proposed test: Distinguish the hypothetical and the cfs, arguing that they should not have the same result or that they should have the same result but for different reasons. *Or*
- (3′.b) Modify the proposed test: Analogize the hypothetical example to the cfs, conceding that the result should be the same in each and arguing that the proposed test yields the right result in the cfs, and eliminate a condition or expand a concept definition so that the test applies to both the cfs and the hypothetical example and leads to the same result in each. *Or*
- (3′.c) Abandon the proposed test and return to (1) (i.e., construct a different proposed test that leads to a favorable decision in the cfs and is consistent with applicable underlying legal principles/policies, important past cases, and hypotheticals…)

Fig. 2. Process Model of Hypothetical Argument

2.3 Case-Based Adaptation in the Process Model

The Process Model also incorporates more complex moves involving case-based adaptation. In step 2 (or 2′) the hypothetical case is designed to demonstrate that the test is too broad (or too narrow). Frequently, the seed for the hypothetical lies in the

facts of the case at hand (i.e., the cfs) or of a relevant precedent. The Justices appear to focus on some legally relevant aspect and adapt the seed so that the proposed test applies to the hypothetical and assigns it the same result as the advocate proposes for the case at hand, but where that result would be wrong in light of the underlying legal principles/policies. Similar adaptations occur in step 2′, but the hypothetical is designed so that normatively, it should have the same result as the cfs but does not because the proposed test does not apply or assigns a different result.

Case-based adaptation also occurs in step 3.b (or 3′.b), where the advocate responds to the hypothetical by modifying the proposed test. Having distinguished the hypothetical case from the cfs and argued that they should have different results, the advocate adapts the test by adding a condition or limiting a concept definition so that the narrowed test still applies to the cfs but no longer applies to the hypothetical or leads to a different result. Similar adaptations occur in step 3′.b. Having analogized the hypothetical case and cfs, conceding that the result should be the same in each, the advocate broadens the test, eliminating a condition or expanding a concept definition so that the revised test applies. Although the thing that is modified is the test, the adaptation is still clearly case-based. In each step, the modifications are informed and guided by the distinctions or analogies between the case and hypothetical. Since a Justice designed the hypothetical, these analogies and distinctions indicate his concerns; the modifications to the test are designed to allay those concerns.

One sees both kinds of adaptation in the *Asahi* example of Fig. 1. The Justice's first hypothetical, box [b], changes: (1) the manufacturer of a component part into a manufacturer of the whole product; (2) the assumption that it is foreseeable the product would end up in CA into the company's knowing that it will be exported to CA. The first change simplifies the analysis for purposes of argument; any complexities due to the fact that Asahi is the manufacturer of only a component part are temporarily set aside. Arguably, there might be some reason for treating component parts manufacturers more leniently. The second change makes a clearer case for finding that it is fair to subject the manufacturer to personal jurisdiction in CA; it is not just foreseeable that his product will end up there, he *knows* it will.

The advocate's two adaptations are also interesting. Through successive broadening of the exception, each narrows the scope of who is not covered by personal jurisdiction. The first exception, Fig. 1, box [c] covers only those who somehow "commit some act [that] positively seeks to serve the market in CA and direct its components to CA." The second, box [e] is broader; one need only encourage the sending of the component to CA. The impetus for the second adaptation is responding to the Justice's "sending" hypothetical, used both to clarify Mr. S's somewhat obtuse "positive act" requirement and to establish a boundary on extending personal jurisdiction. It is as if the Justice said, "You don't really mean to suggest that personal jurisdiction only applies to one who actually sends the product into CA? CA's interest in protecting its citizens extends farther than that, doesn't it?"

As the example suggests, the details of the adaptations are quite subtle and involve the integration of extensive background knowledge, much of which remains implicit. In fact, this may be the reason Justices employ hypothetical reasoning; it is a remarkably succinct (some might say laconic) way of plumbing the complex implications of a proposed rule. The Process Model skims the surface of these subtleties; extending the model depends on studying more examples in greater detail.

3 Related Work

Reasoning with hypothetical cases is a staple of SCOTUS arguments and common law decision making [4; 8; 19], American legal education [5; 22; 23 pp. 66, 68, 75], civil law (i.e., continental European) legal reasoning [14, pp. 528f], ethical reasoning [7] and mathematical discovery [10]. The process model of hypothetical argument of Fig. 2 adapts patterns of hypothetical reasoning observed in legal opinions to a dialogue between an advocate and a judge [4, p. 100]. It adapts three common modes of responding to hypotheticals in order to resolve the dissonance created when a proposed test reaches an arguably undesirable result in a hypothetical [5, pp. 120f]. It focuses on accommodating the conflicting underlying principles at stake [7, pp. 221-8]. The model is similar to Lakatos' mathematical reasoning method of proof and refutations [10, p. 50]. SCOTUS oral arguments are working examples of reasoners' employing hypothetical counterexamples as in the artificial Socratic tutorial dialogue Lakatos reconstructed from centuries-long communications of mathematicians.

As noted, the Process Model provides a high-level account for two types of case adaptation, designing the hypothetical and modifying the proposed test; the above sources, however, do not explain how these subtle adaptations are performed. CBR research on adaptation provides some help. Adaptations like the above can be categorized in terms of the adaptation methods and strategies in [9, p. 395]. Clearly, substitution is involved, but it is a kind of model- and explanation-based substitution based on the knowledge that "knowing" is not only more specific than "foreseeable" but also more strongly supports a legal inference of personal responsibility for the consequences. The other adaptations are based on the knowledge that "sending" is a kind of "positive act" and that "encouraging the sending" is broader than actually sending. Based on other examples, we have suggested the ontological requirements for modeling this kind of model-based substitution with domain facts and factors, legal concepts, principles and policies, and various orderings capturing the kind of legal knowledge illustrated above [2].

A major open question, however, involves the mechanisms to control inferences and moves associated with hypothetical reasoning. The example in Fig. 1 suggests a rhetorical strategy; it shows the beginnings of a slippery slope as the Justice maneuvers Mr. S into needing to explain why supplying component parts to products one knows at least some of which will enter CA is not the very kind of "encouraging the sending" that, according to Mr. S's last test, would subject Asahi to jurisdiction in CA. Some AI research in computationally modeling Lakatos' methods of proof and refutations [6; 15; 12] and reasoning with examples and hypotheticals [1; 20; 21] provides insights into the control problem.

Applying these insights intelligently, however, requires studying many real-world examples. Conducting that kind of empirical study requires a means for adequately representing the examples, namely LARGO diagrams to which we now turn.

4 Representing Hypothetical Reasoning Diagrammatically

In order to extend the Process Model to provide a more detailed account of case-based

adaptation, to implement the Model computationally, and to teach students this process of hypothetical reasoning, a succinct representation of the examples is useful. This is especially true since the oral argument examples are described in text, are often distributed across multiple argument "moves" (i.e., turns taken by advocates and Justices), and involve background novel that is only implicit in the transcripts.

We have developed a diagrammatic representation of argument moves involving hypothetical cases, based on our Process Model of Hypothetical Argument [3]. Using the LARGO (Legal ARgument Graph Observer) intelligent tutoring system, students can represent in diagrammatic form portions of SCOTUS oral arguments that relate to hypothetical reasoning [16, 18]. LARGO is intended to help law students learn the process of arguing with hypotheticals by diagrammatically reconstructing examples of SCOTUS oral arguments according to the Process Model. Fig. 3 shows a student's LARGO diagram representing the same portion of the *Asahi* oral argument discussed in Fig. 1. A scrollable pane (not shown) contains the argument transcript. A student prepared the diagram by selecting and connecting the elements and relations and linking the latter to corresponding passages with a text highlighting feature. There are elements for representing the facts of the case for decision, proposed tests, hypotheticals, and five kinds of relations among them: modifying a test, distinguishing or analogizing a hypothetical, a hypothetical's leading to a test or modification, and a generic relation. The test element is structured to encourage students to prepare a logical formulation with slots for "if", "then", "and", "unless", and "even though".

A somewhat simplified version of the Process Model, together with educationally targeted feedback messages, has been implemented, but a full implementation of the model that would allow the program to make arguments has not been completed. Although LARGO cannot make or respond to hypothetical arguments, it does give advice to students, based on the Process Model, about their argument diagrams. Whenever a student selects the Advice button (not shown), the program provides three new hints on improving the diagram or reflecting on its significance. The advice concerns where to look in the transcript for passages that should be represented in the diagram, how to repair or augment portions of the diagram that appear not to conform to the Process Model, and what patterns of diagram elements appear to be worth reflecting about in terms of the model. In LARGO, a "graph grammar" of rules enforces the expectations embodied in the Process Model. The grammar parses the diagram represented in graph notation [17] in order to flag parts of the diagram where the elements and relations miss relevant parts of the text, do not conform to the Process Model, or are complete enough to warrant reflection.

The graph grammar rules employ classification concepts including a number that focus on CBR functions, for example, distinguishing (or analogizing) without providing reasons, using a general relation between a hypothetical and the cfs rather than analogizing or distinguishing, a hypothetical in isolation (offering an opportunity to enquire if it should connect to a test) and a hypothetical connected to multiple tests (offering an opportunity to discuss if the hypothetical played a role in the modification of one test to another). LARGO's version of the Process Model does not (yet) explicitly cover the ideas of broadening / narrowing tests and the ways in which hypotheticals are crafted to solicit these test revisions.

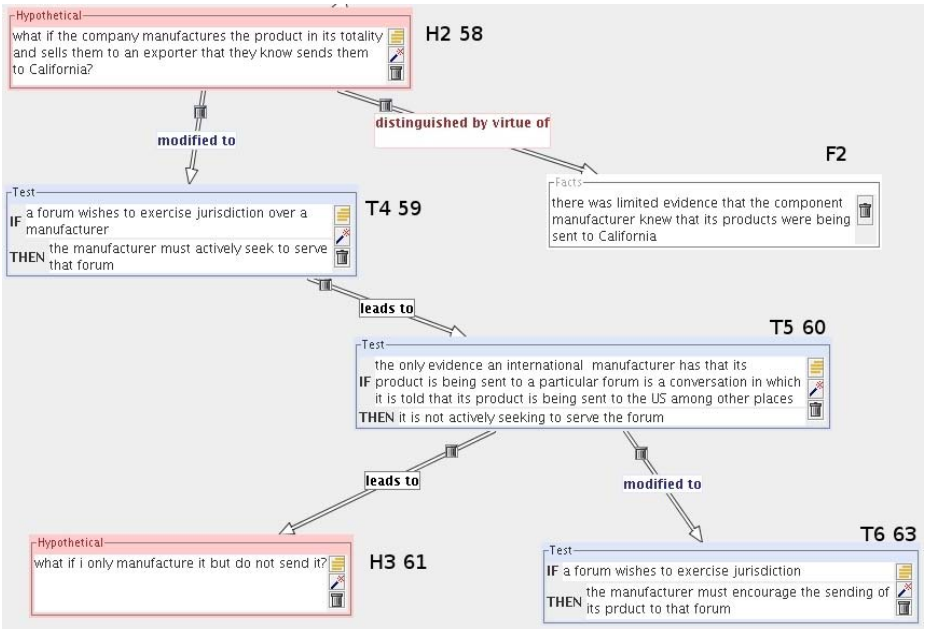

Fig. 3. Sample LARGO Diagram of *Asahi* Oral Argument

LARGO's advice is couched as a recommendation rather than as a declaration that something is incorrect. The program does not have a "definitive" argument representation; an instructor's marked-up transcript only indicates where process-model-related components are located in the text. Sometimes, multiple ways of representing argument moves are reasonable, for instance, where different diagrammers interpret the tests at different levels of abstraction. In addition, a Justice may move on to another topic before the advocate can finish; the diagram will be incomplete according to the model but it accurately reconstructs the argument.

The student's diagram in Fig. 3 satisfies some conventions in the Process Model but violates others. Some relations are mislabeled or omitted. For instance, hypothetical H2 should not be labeled as being modified into test T4. A hypothetical leads to a proposed test's modification. The student shows that H2 is distinguished from the case facts, but leaves the "distinguished by virtue of" relation unfilled. Where a student has analogized or distinguished a hypothetical as in Fig. 3, LARGO encourages him to explain why this matters (e.g., "Usually, attorneys should give a reason why the distinction matters from a legal viewpoint. For instance, does it matter in terms of the principles and policies underlying the issue? Please enter this in the highlighted distinction relation."). This student has not done so.

5 Experiment to Assess Reliability of Interpreting Diagrams

Given our intention to employ LARGO in computational modeling and teaching, the question is whether humans can interpret the argument diagrams reliably. We have

been comparing argument diagrams created by first and third year law students. The first year students used LARGO as part of a study to determine if the system helped students learn skills of hypothetical reasoning better than a more traditional approach involving reading and note-taking but not diagramming [18]. The third year students used the system in the same ways and context as the first years. In [3] we presented evidence that features of LARGO argument diagrams are correlated with two independent measures related to argumentation ability: standardized test scores that assess ability to evaluate reasoning and arguments and students' number of years in law school. LARGO diagram features, including advice-related classification concepts mentioned above, are also correlated with post-test performance [13].

5.1 Experimental Procedure

This experiment involved grading argument diagrams prepared by first and third year law students at the University of Pittsburgh. First year students are typically recent college graduates. Since the third year is the last year of a law school education, it is fair to assume that third year students are more expert in their understanding of legal argument than first years. We used the full set of first diagrams produced by all students who completed the study. This comprised 33 diagrams, prepared in fall 2007 by first year students in their first semester legal process course as part of their regular coursework, and 23 diagrams prepared by volunteer third year students in the middle of their final year. Unlike first-years, the third year students were selected from the top half of their class in terms of law school GPAs and prepared their diagrams for pay outside of class work. The third-years, however, performed the same tasks as the first year students: a pre-test and instruction with LARGO, sessions diagramming three SCOTUS cases, and a post-test, all spread over four two-hour sessions.

Two senior law school professors graded the diagrams following a double-blind procedure. The graders were not aware of whether any diagram was prepared by a first-year or third-year student. The procedure was as follows:

1. Both graders trained on LARGO using the same cases as the students. The graders each produced their own diagrams for the three cases. When grading student diagrams, each grader had his own diagram available.
2. The graders first graded a sample of 6 diagrams drawn from a different study using a draft set of criteria. They graded the diagrams independently and then met to discuss the results and refine the criteria. This ensured that they agreed on and understood the criteria and led to some minor revisions of the criteria.
3. Each grader received the diagrams-to-grade in anonymized form; each diagram had a randomly assigned ID that did not identify the diagrams' author or group. Each grader's diagrams-to-grade were shuffled to ensure that they did not grade them in the same order. Each grader also had the oral argument transcript for which the diagrams were constructed. Annotations on each diagram indicated whether or not an element was linked to the transcript text, and if so to what segment (see Fig. 3, top right lined icon of test and hypothetical elements).
4. Each grader partitioned the diagrams into three bins: poor, medium and good. He then divided each bin into better and worse. This binning resulted in a six-point grading of diagrams based on an initial gestalt inspection. The binning

was designed to avoid the reassessment phenomenon in which graders routinely alter their criteria as they grade a set of materials.

5. Each grader reshuffled the diagrams and (a) assigned detailed grades according to three categories of General Grading Criteria (i.e., coverage, correctness, and comprehension), Table 1; (b) graded each Test and Hypothetical element in the diagram independently according to criteria specific to each type of element, Table 2; and assigned an overall grade to each diagram on a 12 point scale reflecting their by then more complete judgment of the diagram's quality. (One grader assigned overall grades on a 6 point scale. In all of the analyses below, overall grades have been rescaled for comparison.) As Tables 1 and 2 indicate, many of the grading criteria pertain directly to how well student diagrams reflect features of the Process Model of adaptation with hypothetical cases.

Table 1. General Grading Criteria and Inter-rater Agreement.

Category	Criterion	κ
Coverage	How well does the diagram cover …	
	1. … all of the essential tests in the argument?	0.05***
	2. … all of the essential hypotheticals in the argument?	0.75***
	3. … all of the essential relationships in the argument?	0.62***
	4. How well are the diagram components related to the appropriate facts of the case?	0.56***
	5. … the argument components as a whole?	0.71***
Correct-ness	How well does the diagram…	
	1. … reflect the ways in which the hypotheticals challenge the tests?	0.64***
	2. … reflect the ways in which tests are modified in response to hypotheticals?	0.69***
	3. … reflect analogizing and distinguishing of hypotheticals with respect to other hypotheticals and essential case facts?	0.35**
	4. … capture the role of policies and principles in the argument (e.g., in analogizing and distinguishing)?	0.28**
	5. Overall, how correctly does the diagram represent the argument?	0.7***
Compre-hension	How well does the student understand…	
	1. … this particular argument both in factual and procedural terms?	0.59***
	2. … the role of proposed tests in legal argument?	0.71***
	3. … the role of hypothetical cases in argument?	0.09***
	4. … the process of analogizing and distinguishing hypothetical cases?	0.3*
	5. … the general process of arguing with tests and hypotheticals?	0.07***
	6. … the role of policies and principles in arguments of this type?	0.3**

5.2 Preliminary Results and Discussion

Inter-rater reliability is often measured in terms of the kappa coefficient, which ranges between -1 and 1. How high a kappa value must be to indicate agreement is subject to debate and varies according to the domain, task, and purpose of the grading. Given the lack of domain-specific guidance, we adopted the standards in [11] for strength of agreement for the kappa coefficient: ≤ 0=poor, .01–.20=slight, .21–.40=fair, .41–.60=moderate, .61–.80=substantial, and .81–1=almost perfect.

In analyzing the grades, a comparison of the gestalt rankings using Spearman's Rho, shown in Table 3, line (1), reveals a strong correlation between the graders' scores. As shown in line (2), these rankings were also highly correlated with the

graders' final grades, an indication that the more detailed grading process tended to confirm initial assessments rather than alter them. Finally, there was strong inter-grader agreement on the final grades as shown in line (3). For the overall grades we aligned the grades, converting one grader's overall grade to a 12 point scale and correcting the sets to compensate for a difference in mean grades. We then computed agreement using Cohen's weighted kappa with squared weights. Under the standard in [11], the kappa value in Table 3, line (3) indicates "substantial agreement."

Table 2. Test/Hypothetical Grading Criteria and Inter-rater Agreement.

Category	Criterion	κ
Test Element	1. Is the test summary test like (formulated as a logical rule with applicable conditions and a relevant legal conclusion for deciding an issue or the case)?	0.48***
	2. Is the test linked to an appropriate segment of the argument?	0.02***
	3. Is this test correctly related to the relevant preceding tests?	0.58***
	4. Is this test correctly related to the relevant hypotheticals?	0.62***
	5. How well does the diagram capture the role this test plays in the argument?	0.51***
Hypothet-ical Element	1. How well does the summary reflect the hypothetical posed in the text?	0.15*
	2. Is this hypothetical correctly related to the relevant test nodes?	0.04***
	3. How well does the diagram capture the role of this hypothetical in the argument with respect to challenging the tests? For instance, does it capture the judge's implication with the hypothetical (i.e., probing the test as too broad, too narrow, or exploring what the test means)?	0.03***
	4. How well does the diagram capture the analogizing and distinguishing of this hypothetical with respect to other hypotheticals and essential case facts?	0.01

Table 3. Grader Agreement.

Measure	*Agreement*
(1) Inter-grader ranking agreement	$\rho = 0.71$, $p < .001$
(2) Intra-grader rank-score agreement	$\kappa = 0.73$ for grader 1, $p < .001$ $\kappa = 0.84$ for grader 2, $p < .001$
(3) Inter-grader score agreement	$\kappa = 0.74$, $p < .001$

The levels of agreement with respect to the General Grading Criteria most relevant to hypothetical reasoning and case-based adaptation vary. There is substantial agreement on coverage of essential hypotheticals (Coverage 2), correctness showing ways hypotheticals challenge tests and ways in which tests are modified in response to hypotheticals (Correctness 1, 3), comprehension of the role of proposed tests (Comprehension 2), and whether the test is correctly related to relevant hypotheticals (Table 2, Test Element 4). There is moderate agreement with respect to whether the diagram captures the role of a test (Table 2, Test Element 5).

There is only fair agreement, however, concerning the correctness and comprehension of how the diagram reflects analogizing and distinguishing hypotheticals or the role of policies and principles (Correctness 3, 4; Comprehension 4, 6). Agreement is slight re: comprehension of the argument role of hypothetical cases and of the general process of arguing with tests and hypotheticals (Comprehension 3, 5), and how well the summary reflects the test and the hypothetical is related to relevant tests, how well the diagram captures the role of the

hypothetical, and how well the hypothetical is analogized and distinguished (Table 2, Hypothetical Element 1 – 4).

Generally, these inter-grader agreement results suggest that the diagrams can be interpreted reliably for purposes of instruction and modeling of some aspects of the Process Model, but that the role and import of hypotheticals is problematic. Of course, these are preliminary results, dealing with diagrams of only the first of three cases. The other two sets of diagrams have been graded, but the data are still being entered and analyzed. Since *Asahi* was the first case graded, the graders may have been uncertain about the criteria associated with the role and import of hypotheticals; the graders may have converged later as they gained practice grading.

In addition, as noted, representing the role and import of hypotheticals is subtle. The diagrams were constructed with our first version of LARGO. We were aware that our tool was unrefined for representing the role of principles/policies in informing analogizing and distinguishing and for representing details about how hypotheticals challenge tests as too broad or narrow. We are exploring the use of pull-down menus with which students annotate the kinds of links between hypotheticals and tests shown in Fig. 3 with information about the role and import of the hypothetical.

6 Conclusions

The Supreme Court oral arguments are a repository of examples of hypothetical reasoning and case-based adaptation. A hypothetical case is designed to help evaluate if a test or rule proposed for deciding a problem is consistent with underlying principles/policies and often leads to adaptation of the test to improve consistency. Studying these examples could advance computational modeling of case-based adaptation, especially inference control, strategic reasoning, and creative design in support of case-based adaptation, and aid in teaching the process. A key requirement for progress in modeling and teaching, however, is a means for succinctly representing these examples in a way that humans can interpret reliably.

This paper has described a diagrammatic representation of hypothetical reasoning based on a process model that explains important features of the oral argument examples. An experiment was undertaken to assess the reliability of expert grading of diagrams prepared by students as they reconstructed examples of hypothetical reasoning in the oral arguments. Preliminary results indicate some significant areas of agreement, including with respect to the correctness of ways tests are modified in response to hypotheticals. With respect to other features associated with case-based adaptation such as the role and import of hypotheticals, agreement was slight. These results suggest that the diagrammatic representation will support studying and modeling the examples of case-based adaptation, but that the diagramming support needs to make certain features more explicit.

The researchers plan to reevaluate the results once grading data for two additional cases are analyzed, and to improve the ways in which the diagrams reflect the role and import of the hypotheticals in arguments. They also plan to computationally model realistic legal arguments involving adaptation with hypotheticals.

Acknowledgments. NSF Grant IIS-0412830, Hypothesis Formation and Testing in an Interpretive Domain, supported this work.

References

1. Ashley, K. (1990) *Modeling Legal Argument: Reasoning with Cases and Hypotheticals*. The MIT Press.
2. Ashley, K. (2009) What a Legal CBR Ontology Should Provide. *Proceedings of the 22d Int'l FLAIRS Conf. Case-Based Reasoning Track*. Sanibel Island, FL, May.
3. Ashley, K., Lynch, C., Pinkwart, N., Aleven, V. (2008) A Process Model of Legal Argument with Hypotheticals. In *Legal Knowledge and Info. Sys.*, Proc. Jurix 2008. 1-10.
4. Eisenberg, M. (1988) *The Nature of the Common Law*, Harvard U. Press.
5. Gewirtz, P. The Jurisprudence of Hypotheticals, *J. of Legal Education*, 32: 120-124 (1982).
6. Hayes-Roth, R. (1983) "Using proofs and refutations to learn from experience", in *Machine Learning: An A. I. Approach*, eds., R. Michalski, et al., 221-240, Tioga, Palo Alto.
7. Hurley, S. (1990) Coherence, Hypothetical Cases, and Precedent, *Oxford J. Legal Studies*, 10: 221-251.
8. Johnson, T. (2004) *Oral Arguments and Decision Making on the U. S. Supreme Court*, SUNY.
9. Kolodner, J. (1993) *Case-Based Reasoning*, Morgan Kaugman: San Mateo, CA.
10. Lakatos, I. (1976) *Proofs and Refutations*, London: Cambridge University Press.
11. Landis, J. and Koch, G. (1977) The measurement of observer agreement for categorical data. *Biometrics*. 33:159-174.
12. Lenat, D. B., and Brown, J. S. (1984). "Why AM and EURISKO appear to work." *Artificial Intelligence* 23(3):269--294.
13. Lynch, C., Pinkwart, N., Ashley, K. and Aleven, V. (2008) What do argument diagrams tell us about students' aptitude or experience? Workshop on ITSs for Ill-structured Domains, ITS-2008, Montreal.
14. MacCormick D. and Summers, R. (ed.) (1997) *Interpreting Precedents* Ashgate/Dartmouth.
15. Pease, A., Colton, S., Smaill, A. and Lee, J. (2002) Lakatos and Machine Creativity In Proceedings of the ECAI Creative Systems Workshop.
16. Pinkwart, N., Aleven, V., Ashley, K. and Lynch, C. (2007) Evaluating legal argument instruction with graphical representations using LARGO. In Proc. AIED2007. July.
17. Pinkwart, N., Ashley, A., Aleven, V. and Lynch, C. (2008) Graph Grammars: an ITS Technology for Diagram Representations. In Proc. 21st Int'l FLAIRS Conf., ITS Track, Coral Gables. May.
18. Pinkwart, N., Lynch, C., Ashley, K. and Aleven, V. (2008) Reevaluating LARGO in the Classroom: Are Diagrams Better than Text for Teaching Argumentation Skills? In Proc. ITS-08, 90-100.
19. Prettyman, Jr., E. (1984) The Supreme Court's Use of Hypothetical Questions at Oral Argument, *Catholic University Law Review*, 33: 555-591.
20. Rissland, E. (1981) "Constrained Example Generation". COINS TR 81-24. U. Mass.
21. Rissland, E. (1989) "Dimension-based Analysis of Hypotheticals from Supreme Court Oral Argument" In *Proc. 2d Int'l Conf. on Artificial Intelligence and Law*. 111-120 ACM Press.
22. Stuckey, R. et al. (2007) *Best Practices for Legal Education*, 214-215, Clin. Leg. Ed. Assc.
23. Sullivan, W., Colby, A., Wegner, J., Bond, L. and Shulman, L. (2007) *Educating Lawyers*, 62, 66, 68, 75 The Carnegie Foundation for the Advancement of Teaching.
24. Visser, W. (1995) Reuse of Knowledge: Empirical Studies. In (Veloso and Aamodt, eds.) *Case-Based Reasoning Research and Development* (ICCBR-95) LNAI 1010. pp. 335-346.