The Significance of Evaluation in AI and Law

A Case Study Re-examining ICAIL Proceedings

Jack G. Conrad
Thomson Reuters Global Resources
Catalyst Lab
Baar, Switzerland CH-6340
jack.g.conrad@thomsonreuters.com

John Zeleznikow*
School of Management and Information Systems
Victoria University
Melbourne, Australia 3086
john.zeleznikow@vu.edu.au

ABSTRACT

This paper examines the presence of performance evaluation in works published at ICAIL conferences since 2000. As such, it is a self-reflexive, meta-level study that investigates the proportion of works that include some form of performance assessment in their contribution. It also reports on the categories of evaluation present as well as their degree. In addition, the paper compares current trends in performance measurement with those of earlier ICAILs, as reported in the Hall and Zeleznikow work on the same topic (ICAIL 2001). The paper also develops an argument for why evaluation in formal Artificial Intelligence and Law reports such as ICAIL proceedings is imperative. It underscores the importance of answering the question: how good is the system?, how reliable is the approach?, or, more succinctly, does it work? The paper argues that the presence of a performance-based ethic within a scientific research community is a sign of maturity and essential scientific rigor. Finally the work references an evaluation checklist and presents a set of recommended best practices for the inclusion of evaluation methods going forward.

Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: Performance evaluation—*efficiency and effectiveness*; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Evaluation, Performance, Measurement, Validation

Keywords

artificial intelligence and law, legal information systems, evaluation, performance assessment, verification [1]

*Co-author of the first self-reflexive work on evaluation at ICAIL (ICAIL 2001).

[1]Although performance evaluation is arguably a more narrow category than evaluation in general, the two are used interchangeably here.

ICAIL '13, June 10-14 2013, Rome, Italy
ACM 978-1-4503-2080-1/13/06.

1. INTRODUCTION

1.1 Motivations

This work revisits a study performed just over a decade ago which investigated the presence of evaluation in published IAAIL papers.[2] That study, conducted by Hall and Zeleznikow [9], examined ICAIL[3] works from 1995, 1997 and 1999 for evidence of performance evaluation, and compared these percentages with those from the very first ICAIL in 1987. In an analogous manner, this work studies published papers from the last six ICAILs (2001-2011) for the presence or absence of evaluation. We define evaluation as a systematic determination of a subject's merit, worth and significance, using criteria governed by a set of standards. It is used to ascertain the degree of achievement or value in regard to the objectives and results of the execution of the system or approach presented. As with Hall and Zeleznikow, this research distinguishes theoretical works from non-theoretical works (assuming that theoretical works typically contain no formal evaluation), and further segments non-theoretical works into evaluated vs. non-evaluated types. The motivation for this study is to determine whether, as the community has evolved over time, it as become more mature in its use of empirical methods for performance evaluation and other forms of self-assessment. The hypothesis of this work is that if a researcher does not answer the fundamental question surrounding his or her efforts – how good is the system? or how reliable is the technique?, or, more succinctly, does it work? and if so, how well? – then how can that researcher expect the broader audience to be convinced of the benefits and utility the published report delivers? [1]

1.2 Previous Work

Hall and Zeleznikow (2002) [10] note that evaluation of human endeavour has a long proud history, from the Stone Age when stone chippers left a track record of gradually improving quality of materials and design through to the last two decades of the 20th century when the Software Engineering community invested considerable effort in the movement towards 'quality' software and established methodologies and standards for software development and evaluation. Karlsson et al. [15] argue that evaluating software is fundamental for the development of useful software systems. Chelimsky [3], in the introduction to her book on evaluation for the 21st century, suggests that systems in any discipline should

[2]IAAIL – Int'l Association for Artificial Intelligence and Law
[3]ICAIL – Int'l Conference on Artificial Intelligence and Law

be evaluated for three main reasons: to demonstrate accountability, gain knowledge and enhance development.

Jadhav and Sonar [13] claim that evaluating and selecting software packages that meet an organisation's requirements is a difficult software engineering process. Selection of a wrong software package can turn out to be costly and adversely affect business processes. They found that there is a lack of a common list of generic software evaluation criteria and its meaning, and there is need to develop a framework consisting of software selection methodology, evaluation technique, evaluation criteria and systems to assist decision-makers in software selection.

Cohen and Howe [4] argue evaluation should be a mechanism of progress both within and across AI research projects. Evaluation can tell us how and why our methods and programs work and so tell us how our research should proceed. For the Artificial Intelligence community (and we believe this is especially true for the AI and Law community) evaluation expedites the understanding of available methods and so their integration into further research. They presented a five-stage model of AI research and developed guidelines for evaluation that are appropriate at each of these five stages.

Reich [21] notes that the evaluation of intelligent systems raises difficult theoretical and pragmatic issues. Evaluations of systems have often been performed in an ad-hoc manner without regard to theoretical concepts associated with the nature of measurement and the identification of appropriate evaluative criteria. To remedy this, he suggests the application of measurement theory concepts to the task.

According to Reich, the determination of appropriate criteria cannot be done without a conceptualisation of the nature of knowledge. He claims knowledge can be defined in two ways: structurally and functionally. In the structural definition knowledge is a static entity that includes facts, rules and models that represent real world phenomena. This definition of knowledge enables the direct measurement of knowledge.

At the first International Conference on Artificial Intelligence and Law in Boston, in 1987, Professor Richard Susskind suggested that if legal knowledge based systems were to move out of the research laboratory and into the marketplace, their evaluation was essential [27]. Legal knowledge based system specific issues concerned with knowledge acquisition – different jurisdictions, judicial discretion and the potential for significant social impact – are identified below. A detailed discussion on these issues is available in Hall (2005) [6], Hall et al. (2003) [8] and Hall and Zeleznikow (2002) [10].

In her work on planning in Artificial Intelligence and Law, resulting in the development of the CHIRON system, Sanders [22] stressed the importance of the evaluation of legal software. She argued that if we are to make progress we must ask what our goals are, and whether they have been achieved. She continues that if we want others to use our systems, then we must persuade them that the output of these systems will be useful and reliable.

Stranieri et al. [26] used a hybrid of machine learning and rules to provide advice about the distribution of marital property under Australian Family Law. Toulmin Argument Structures were used to justify the decision suggested by the resulting Split-Up system. As Stranieri and Zeleznikow [25] noted, if the results of machine learning are to indicate something more than mere data dredging, then developers must provide some justification for their evidence. They discuss evaluation as a means of justifying the results obtained through statistical endeavours. They claim that outside the legal domain, the evaluation of models derived with the use of a data-mining technique is often restricted to the measurement of the generalised performance; the extent to which the model is accurate on examples not in the data set.

Hall and Zeleznikow [9] considered the need for knowledge-based systems in general and legal knowledge-based systems in particular. The proceedings of the 1987, 1995, 1997 and 1999 International Conferences on Artificial Intelligence and Law were analysed to determine the rate of reporting evaluation in non-theoretical papers. Since the field is now more than 25 years old, we believe it is appropriate to examine whether the community has matured to the extent that evaluation of systems or techniques is a common feature of applied research articles.

2. WHY SHOULD KNOWLEDGE-BASED SYSTEMS BE EVALUATED?

Hendriks and Vriens [11] consider the organisational value of knowledge-based systems. They argue that the emphasis in such systems tends to be on the underlying technology rather than the knowledge perspective. They contend that simply equating managing the development and use of Knowledge Based Systems with Knowledge Management is to be avoided. They argue for a three step process:

1. Diagnosis – how can an organisation assess the functionality of Knowledge Based Systems as Knowledge Management measures: where is the gain for the organisation?
2. Development – how should Knowledge Based Systems be designed and implemented given the context of intended usage in the organisation?
3. Assessing and evaluating changes – what are the consequences to the organisation if it decides to deploy the Knowledge Based Systems?

In Cohen and Howe's seminal 1988 AI evaluation paper, the authors assert that evaluation is an essential component of any credible research community that wishes to discover why and how its approaches and systems work. In addition, it permits the direct performance-based comparison of systems with themselves by establishing baselines [4]. Some individuals within the AI and Law community take performance evaluation seriously because they may be developing a commercial system that needs to be the best of its breed, not to mention to avoid litigation based on its results. For this reason, it is not uncommon to find three or four distinct tests performed on the system and documented before certification and release [5, 17].

But what about the case of theoretical works within the community? which is surely a question that will arise. Even though there may not be a resulting artifact to test and compare with other approaches or systems, still some authors have taken great strides to demonstrate the applicability and utility of their methods. Upon presenting new models or techniques that address certain patterns of evidence, Prakken et al., for example, customarily present one or more extended examples to illustrate how their approach works and address the challenges that typically confront them. [19, 20, 2, 14].

3. EVALUATION OF LEGAL KNOWLEDGE-BASED SYSTEMS

Hall et al. [8] claim that established software evaluation methodologies that are not specifically tailored to the legal domain may be unsuitable for use by those who do not have a sufficient software engineering background. The development and evaluation of legal knowledge-based systems is subject to additional challenges beyond those apparent with knowledge-based systems designed to operate in other less open domains. Legal knowledge-based system specific issues concerned with knowledge acquisition, different jurisdictions, judicial discretion and the potential for significant social impact are discussed below.

Domain knowledge acquisition issues in law include the possible lack of a strong consensus on the theories of jurisprudence used to develop the model of legal reasoning, the dynamic nature of the knowledge and its open texture requiring interpretation by experts who often have limited availability.

Legal principles vary between different jurisdictions with some giving more significance to legislation (statutes) and others to precedents (cases). In certain jurisdictions law postulates a fiction of certainty (Koers et al. 1990) [16]. Here a judge does not have the option to decide that there is insufficient knowledge to reach a decision, or to attach a degree of probability to the correctness of a decision under the law. A decision must be made. Judicial discretion can also compound the problem. When there is no concept of 'one correct answer,' two assessors given the same details may arrive at a different decision. Software, in contrast, will arrive deterministically at a reliable and repeatable outcome given the same inputs. This presents a major difficulty when evaluating the validity of legal knowledge-based system.

Legal knowledge-based systems have a potential for a significant social impact both upon individuals and beyond. There is an ethical onus on the knowledge-based system designer, developer and evaluator to be accountable and exercise social responsibility.

How should evaluators of a legal knowledge-based systems frame (plan) their evaluation? This is non-trivial. The VALENS tool, developed in the POWER-program, has been used to verify legal knowledge (Spreeuwenberg et al. 2001) [23]

Many existing evaluation frameworks and methodologies lack specific legal domain content. Typically they are designed for use by evaluators or developers with software engineering expertise and contain terms, references and methods likely to be unfamiliar to the non-computer expert who finds herself charged with evaluating a legal knowledge-based system. These existing resources are not all available in one well-publicised, easily accessed location and they may not cover the complete range of activities required. It is difficult for an evaluator to select bits and pieces from different locations, as these materials have partial overlapping content and differing organisation, with no 'common interface.' A common framework to order and organise such knowledge would be of value to evaluators of such systems. The PhD thesis of Hall [6] proposes a process for evaluating legal knowledge-based systems based upon the context criteria contingency framework guidelines (described in Section 4.2).

3.1 Methodology

In the subsections below, we provide background material on how our evaluation rating system for ICAIL proceedings evolved from the binary classification approach undertaken by Hall and Zeleznikow [9] to the 5-category approach used in the current work. In addition, we describe the seven main categories of evaluation that were used by Hall and Zeleznikow and which we have elected to follow for the purposes of consistency and comparison. It is significant to note that in order to avoid inter-assessor agreement issues, Hall worked with us to ensure that our assignments were consistent with the ones that were used in the earlier study. Furthermore, we had initially envisioned reporting on only full-length ICAIL papers; however, upon observing that there was nearly as much evaluation reported on in the short papers and two-page abstracts, we decided to included these in the study as well (See Table 3 in the Appendix).

3.1.1 Evaluation Ratings (Grades) for ICAIL Papers

That there are distinct challenges and difficulties in evaluating the reliability and correctness of systems and techniques in the legal domain is clear. Our objective in this work, like Hall and Zeleznikow before us [9], is to nonetheless examine the extent to which submissions to past ICAIL conferences have made conscious efforts to evaluate the performance of their experiments and subsequent results in some appropriate form, whether that form be in terms of accuracy, coverage, purity, effectiveness, improved efficiency, or other metrics, such as elapsed time to complete a task. If a researcher in the legal domain wishes to demonstrate credibility to the broader AI and Law scientific community, it is essential that the researcher at least attempt to answer the question – how well does this system or approach work? Evidence of this pursuit along with answers to questions like the above were thus the central motivation behind this investigation. At the highest level, then, we distinguish between those papers that contain some suitable form of evaluation from those that do not, while acknowledging that theoretical papers would not typically contain such a degree of scientific evaluation.

To be able to assess the level of evaluation occurring in ICAIL papers, we thus need to develop a classification system. Hall and Zeleznikow (2001) [9] relied on three high-level categories:

1. Papers dealing with theoretical developments where a discussion of evaluation would not necessarily be expected;

2. Papers describing a system, algorithm or other approach where evaluation was addressed;

3. Papers describing a system, algorithm or other approach with no mention of evaluation.

In order to probe this topic in more detail, we established our categorisation at a slightly finer level than that of Hall and Zeleznikow. Furthermore, unlike Hall and Zeleznikow, we contend that theoretical works can within limits have their utility or applicability if not strict performance assessed. In certain contexts, this can be achieved to some degree through simulated applications or extended illustrations, as exemplified in the works of Prakken mentioned earlier.

For papers describing a system, algorithm or other technique, Hall and Zeleznikow [9] take a binary approach: there

'is' or 'is not' mention or application of evaluation. By contrast, while we have a category in the paper that represents no form of evaluation, we also have four categories for those papers that do mention evaluation – from the discussion or design level to the comprehensively applied. These levels are also associated with corresponding "grades," from A (for thorough) to F (for no presence). These categories are described below.

0. Absent (Grade: F). No mention of evaluation in any form in the submitted work.

1. Discussion (Grade: D). Paper discusses how the proposed system or approach could be evaluated.

2. Basic (Grade: C). A very preliminary and simplistic evaluation is performed on either a portion of the system or portion of the relevant data. May consist of anecdotal assessment evidence and presentation.

3. Moderate (Grade: B). A good faith evaluation effort is performed on the proposed system or approach. As such, it represents just one form of evaluation exercise.

4. Mature/Comprehensive (Grade: A). A credible degree of evaluation is performed on the system or approach, including multiple assessments (across components, relevant content, modular vs. end-to-end, system vs. baseline, system vs. human, and in terms of some combination of the above).

3.1.2 *Forms of Evaluation Identified*

In addition to examining the level of evaluation presented in ICAIL papers, it is also instructive to examine what *forms* of evaluation are undertaken. There are a wide range of such forms considered here. These assessments could be fully automated or performed totally by humans, not to mention a host of other types, including those that measure changes in human performance when given access to the tools or results generated by the given experimental environment. Comparisons can also be made with the performance of humans or the performance of other systems. The range of evaluation forms, largely established by Hall and Zeleznikow in their earlier work, and followed by the current authors for continuity and comparability, are presented below.

- Statistical – for instance, the kind of assessments that could be performed on clustering solutions, which might include measures of purity and entropy.

- Comparison with Other Systems – where the other systems may have the reported top performance to date or simply be used as a baseline measure.

- Comparison based on Human Performance – for some systems or approaches intended to be used as a tool for human practitioners, the performance of the humans may be measured with and without the use of the experimental tool.

- Comparison with Expert Judgment – this would include systems or approaches that use expert assessments for their relevance judgments or other gold standards; thus the reporting of recall and precision would rely on such expert assessments.

- "Foreshadowing" or Discussion of how the approach might be evaluated – in other words, no true evaluation is performed, but there is a discussion on how it could be done.

- Impact on Current Operating Environment – this turned out to be a broad category which overlapped with one or more of the above categories. Some environment will be impacted in practically every case. This was a finding of Hall in her PhD work and why she, and us still more, have so few works here assigned [6].

- Other – those systems with distinct forms of evaluation not covered in the categories above,

3.2 Current Results

In order to permit a useful and instructive degree of continuity and a means of comparing the findings of Hall and Zeleznikow (1987, 1995-1999) with our own (2001-2011), we have extended the Hall and Zeleznikow classifications for the early set of ICAIL works to our later period of study. These latest comparative findings are presented in Figure 1 in terms of Theoretical vs. Non-Theoretical works, with the later category further divided into Evaluated vs. Non-Evaluated for the years 1987-2011.[4] To put this later comparison under further scrutiny, we show the specific set of Evaluated vs. Non-Evaluated papers in Figure 2. And lastly, to underscore the observation that theoretical, which include papers on logical formalisms, can also be investigated for limited degrees of assessment, we apply our lower-level evaluation categories (0-1-2) to these works in Figure 3.

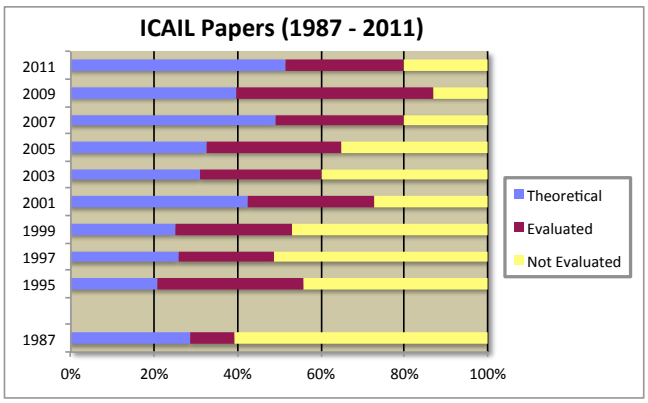

Figure 1: Proportion of Theoretical vs. Evaluated and Non-Evaluated Works

Figure 1 appears to show that the percentage of ICAIL papers in the theoretical category has been steadily increasing since the early days of the conference. Upon closer inspection, however, one can see that it is actually just two conferences, 2007 and 2011, that affirm this perspective. If one were to discount the extent of these two contributions, the trend would not appear to be as dramatic, with the theoretical papers ranging from about 30% to 40% of the overall pool. This distribution would leave between 60% and 70%

[4]Hall and Zeleznikow [9] did not examine the proceedings for 1989 and 1991. Their main focus was on the years 1995, 1997 and 1999, while at the same time comparing characteristics of these years with those for the first ICAIL in 1987.

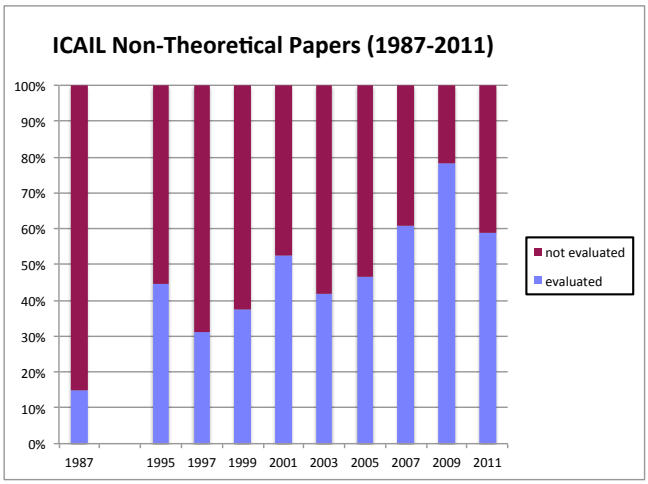

Figure 2: Proportion of Evaluated vs. Non-Evaluated Works

of the works falling into the non-theoretical category. Yet even when considering the distribution in its entirety, the lower bound for non-theoretical works is about 50%, while for most years it is appreciably greater, and often surpassing 70% in the early years of the conference.[5]

Since our interest is aimed at those papers that are either evaluated or not from the main set of papers where evaluation can generally be performed, in Figure 2, we focus only on the non-theoretical papers. Here the trend looks a bit more encouraging, since, if anything, there is a slight upward trend in evaluation, especially over the last three conferences. During the principal years of the Hall and Zeleznikow study (1995-97-99), evaluation for these works fell into the 30% to 45% range. By contrast, 2003 and 2005 aside, evaluation remained in the 60% to 80% range for the last three conferences. Interim years aside, were this trend maintained, clearly this would be a positive development. Perhaps one could argue that it took some time for the Hall and Zeleznikow message to take hold. Of course, one needs to be able to examine the size of the data sets that produce these percentages, and these are presented and addressed in the next section, 3.3 "Comparison with Earlier Results" and Table 1.

Earlier in this work, we asserted that unlike Hall and Zeleznikow, we believe that it is possible to discuss performance assessment, even for theoretical works, albeit, possibly in not quite the same quantitative terms. To this end, our current study focuses on the ICAILs since the seminal Hall and Zeleznikow research (2001) [9] where we additionally scrutinize the published theoretical works for evidence of assessment. Although a majority may still receive a 0 rating (no assessment), there were nonetheless a non-trivial number of works that at least discussed how assessment of the logical models or other types might be conducted (representing a 1 rating, i.e., assessment contemplated). Still other papers made a good faith effort at demonstrating the utility and coverage of their works, either through formal proofs or extended applications/examples or similar types of direct realization (for a rating of 2, initial assessment). The resulting distribution is shown in Figure 3. The positive finding here is that, again, from 2001 to 2011, the trend was for some form of assessment to grow, from the 16% to 20% range in 2001-03 to the 30% to 40% range in 2007-2011, with 2005 being an outlier, with a full 60% of the theoretical works presenting some anticipated or basic assessment. (Note that the comprehensive set of our scores for the current study can be found in Tables 3 and 4 in the Appendix.)

The next topic we examined in this work, closely relying on the foundation established by Hall and Zeleznikow, was the categorical breakdown of the forms of evaluation conducted. The distribution for our recent set of the ICAILs from 2001 to 2011 can be found in the pie chart shown in Figure 4. Here we see that nearly half of the works evaluated did so using some gold data or other forms of judgments provided by domain experts in order to facilitate the assessment. The measurement of system recall and precision, using relevance judgments provided by some form of "experts" is among the most frequent form in this category. The next largest category, at just under a quarter of the set, was what Hall and Zeleznikow termed "Foreshadowed" and what we have described above as at least "Anticipated or Discussed." Close behind this category was "Statistical" which represents some other metric not reliant upon expert judgments, for instance, system measures from a neural network. Coming after this at about a tenth is the category that Hall and Zeleznikow termed "Computer Generated" and which we determined to cover comparisons between the current system, algorithm or approach and a similarly established baseline. Finally we have "Human Performance," which as a category signifies a comparison between human performance for a specific task, for instance, with and without the tool or computer model described in the research. The negligible presence of the category entitled "Environmental Impact" is discussed in the Appendix.

3.3 Comparisons with Earlier Results

In contrast to the evaluation types just described for ICAILs 2001-2011, Hall and Zeleznikow found the following types in their study of ICAILs 1995-99. The percentage of works where evaluation was "Foreshadowed" was roughly the same (just under one-quarter) while the presence of both the "Expert Opinion" categories increased from just under a quarter (then) to just under half of the evaluated works (now).

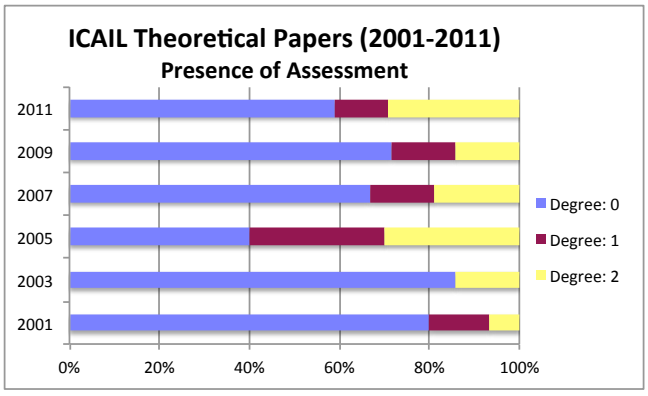

Figure 3: Presence of Assessment in Theoretical Works (2001-2011)

[5]Table 1 bears this out more comprehensively, where its final row ("Total") indicates that 65% of ICAIL papers in this period were non-theoretical.

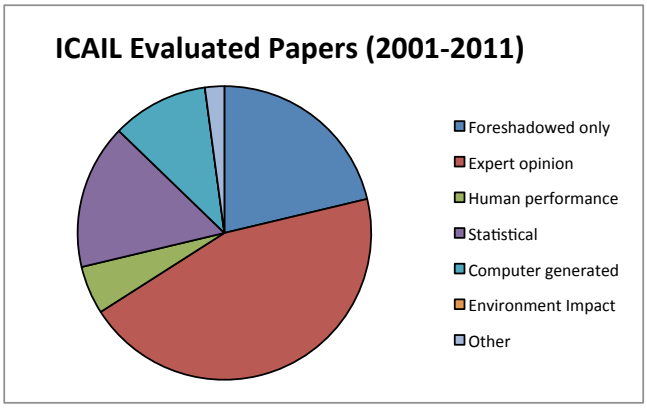

Figure 4: ICAIL – Types of Evaluated Works (2001-2011)

"Statistical" similarly increased from about one-twelfth to one-eighth. What made up the difference was the "Human Performance" category (decreasing from one-quarter to one-sixteenth) and the "Computer Generated" from one-quarter to one-eighth. There are also a few examples of "Environmental Impact" in this set, something that was subsumed by the other categories in our study. The changes in size of these categories across the two studies may be accounted for by experimental trends (e.g., the growing importance of expert-generated gold data) as well as rater subjectivity.

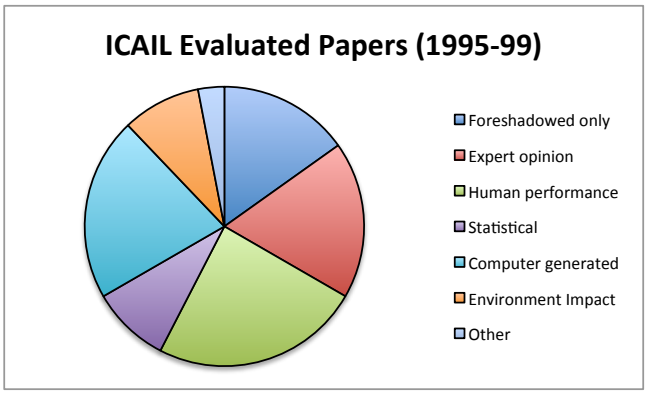

Figure 5: ICAIL – Types of Evaluated Works (1987, 1995-99)

The last major table we present in this section corresponds to the data shown in Figure 1. Table 1 presents the actual counts of all of the evaluation-types observed either in the initial Hall and Zeleznikow study or the current study. It permits quantitative comparisons across the two experiments (1987,1995-99 vs. 2001-11) as well as a focus on the size of the underlying data/result sets. In general, the number of works published for recent ICAILs (excluding 1987) ranged from the mid-to-low 30s to the mid-40s. Keeping our eyes on numerical evaluation patterns, one can observe that the raw number of evaluated works has been increasing (2011 excepted), with a noticeable uptick into double digits following the Hall and Zeleznikow paper focusing on ICAILs prior to 2001. Did their work have an appreciable impact on the community and its evaluation practices? It is hard to say. One thing one can say, however, is that it did not hurt, and may have helped encourage the community to be more mindful of the importance of demonstrating how one's approach can improve a given baseline.

4. DISCUSSION

4.1 Current Findings

In the results presented above, we have observed some encouraging trends by way of modest increases in the presence of evaluation, almost exclusively in the non-theoretical works. At the same time, there may be cause for concern insofar as a scientific research community that champions Artificial Intelligence for the benefit of the legal domain may still have as many as a fifth of its non-theoretical works presenting no performance evaluation at all. Furthermore, if one considers the last ICAIL, where approximately half of the submitted works addressed theoretical subjects and of these 60% made no mention of evaluation, this means that 40% of the conference's published papers may still make no mention of assessment or answer the fundamental questions involving whether the presented work evaluates its performance. The conclusion one is left to draw here is that despite meta-level studies that have been conducted, progress in this area can still be made, in terms of influencing both the authors and the reviewers of AI and Law research.

In the words of the former Chief Research Scientist at Thomson Reuters, evaluation is what we are all about. It is what separates us from other technologists. It is what adds the value to our research. We compare what we design with existing baselines to demonstrate that our approach is better, about the same, or worse, but the point is that we investigate the topic from a measurable, highly quantitative and comparative perspective [12].

4.2 Challenges in Developing and Evaluating AI & Law Systems/Techniques

Stranieri and Zeleznikow [24] examined evaluation strategies to determine the effectiveness of legal knowledge based systems. They claimed that such strategies enable strengths and limitations of systems to be accurately articulated. This facilitates efforts in the research community to develop systems and also promotes the adoption of research prototypes in the commercial world. Hall and Zeleznikow (2001) [9] continued this work by analysing the proceedings of four

Year	Theoretical	%	Evaluated	%	Not Evaluated	%	Total
1987	8	29%	3	11%	17	61%	28
1995	7	21%	12	35%	15	44%	34
1997	10	26%	9	23%	20	51%	39
1999	8	25%	9	28%	15	47%	32
2001	14	42%	10	30%	9	27%	33
2003	14	31%	13	29%	18	40%	45
2005	14	33%	13	33%	15	14%	42
2007	22	49%	14	31%	9	9%	45
2009	15	39%	18	47%	5	13%	38
2011	18	51%	10	29%	7	20%	35
Total	130	35%	111	30%	130	35%	371

Table 1: ICAIL – Table of Theoretical vs. Evaluated and Non-Evaluated Works (1987-2011)

ICAIL conferences. In that paper, their stated goal was to determine the rate of reported evaluations in non-theoretical papers.

Unlike the current research, Hall had a goal of developing an evaluation methodology specifically tailored to legal knowledge-based systems. In her 2003 ICAIL paper (Hall et al., 2003) [8] and her PhD thesis (Hall 2005) [6], Hall introduced the Context, Contingency, Criteria evaluation framework. The framework extends the ISO 14598 evaluation process and ISO 9126 quality model. It takes a contingency approach to evaluation emphasising the importance of context. The framework includes Contingency, Context, Process and Quality Models and supporting materials such as report templates and a strategy for the Formative Rapid Evaluation of Research Knowledge-bases. Validation exercises include its practical application in the evaluations of several legal knowledge-based systems, and desk checking against its requirement specification, codes of good practice and criteria suggested by three methodology evaluation frameworks.

In an Evaluation Context Checklist, Hall (2005) [6] considered:

(a) Background to the evaluation – the conditions that prompted this evaluation, the evaluation significance; the priority of this evaluation; the consequences of failure of the evaluation; is the evaluation mandatory.

(b) Evaluation values / ethos subscribed to – formative or summative or both; objective / subjective or both; quantitative / qualitative or both.

(c) Evaluation process issues – Is it intended to generalize the results of this evaluation to other similar situations; have the content areas of the final evaluation report been agreed upon; Have success criteria for the evaluation itself been defined.

(d) Objectives of the evaluation – What?, Why?, When?, Who? By?, Who for?, Who else is affected?, Where?, How?

(e) Evaluation constraints – Permitted evaluator autonomy; Data constraints; Resourcing constraints.

While we value Hall's work, our goal is far more limited: we limit ourselves to examining whether the AI and Law community has followed Cohen and Howe's [4] exhortation to evaluate the systems it has developed. No attempt is made to develop new evaluation formalisms.

Because we have followed this approach, we can examine the evaluation of the six ICAIL conferences from 2001 to 2011 in far greater detail than the four conferences (1987, 1995, 1997, 1999) that Hall and Zeleznikow [9] explored. Because we have considered six consecutive conferences over a ten-year time-frame, we can also argue that we have developed a longitudinal study.

From Figures 1 and 2 and Table 1 (using the 4 conferences examined by Hall and Zeleznikow and our six, and using Hall and Zeleznikow's classification) we see that there has been a steady increase in theoretical papers. Those papers describing system or algorithmic development have been significantly more heavily evaluated. But because there are now fewer application papers, the absolute number of evaluated papers in recent ICAIL conferences is not significantly higher than in earlier ICAIL conferences.

If the Artificial Intelligence and Law community wishes to remain (or more accurately become) more relevant to legal practitioners, then it needs to develop systems that provide significant new knowledge and support. And such systems need to be evaluated: not just in a rudimentary way, but using several distinct techniques. Such techniques could include statistics, comparison with other systems, comparison based on human performance, comparison with expert judgement, and the impact on the current operating environment.

Using ICAIL conferences as an indicator, this is not currently the case. Unless we act on the advice given by Cohen and Howe [4] a quarter a century ago, we run the risk of our work becoming obsolete.

4.3 Methodology Maturity among AI & Law Practitioners

The Software Engineering community accepts evaluation as an important research activity. It may not be the case that software is being regularly evaluated, but new methods for the evaluation of software *are* constantly being developed. Jadhav and Sonar [13] provide a recent and detailed review of techniques for evaluating software.

It is also worth mentioning explicitly here the Carnegie Mellon Capability Maturity Model [18], as advocated in the Hall thesis [6]. The Capability Maturity Model (CMM) defines a process maturity framework for software development organizations. Some organisations where intelligent legal systems are developed are now in the process of evolving from their research base starting point (corresponding to the CMM Initial level) to the Repeatable level. One of the major work process improvements that will assist this transition is evaluation of the systems they produce. The goal of this process improvement is the production of better intelligent legal systems.

Unfortunately the AI and Law community has not followed this trend. While our analysis of recent ICAIL proceedings shows that the level of evaluation discussed in papers has increased little, what is more dismaying is the lack of papers developing new techniques for evaluating legal knowledge based systems. Since Hall and Zeleznikow (2003) [8], no ICAIL conference article has focused upon the evaluation of legal knowledge based systems.

5. STRATEGY

The most beneficial take-away from our current work is a set of recommendations for how to improve the extent of self-assessment within the community. Such recommendations could take the form of a set of best practices that the community would be encouraged to follow. Examples of such best practices would include:

1. Non-theoretical works presenting a system, algorithm or other approach should conduct and report on performance evaluation wherein the work is compared to known baselines, using, whenever possible, publicly available data sets;

2. Non-theoretical works should also explore how variations to known parameters affect system or algorithm performance;

3. When such empirical tests are not possible, then the authors should sketch out procedures that would permit such self-assessment in the future;

4. Theoretical works have opportunities to demonstrate their strength and utility relative to earlier approaches, for instance, by presenting an extended example where the problem is addressed both by the authors' model as well as by competing approaches, and the pros and cons of each are spelled out.

If such basic procedures as these were adhered to as a matter of common practice, the degree of empiricism and performance quality monitoring would already be demonstrably improved.

6. CONCLUSIONS

This work has conducted anew an instructive and self-reflexive study of the IAAIL community in terms of the percentage of published ICAIL works containing some degree of evaluation. Allowing for a sizable variety of types of evaluation, what the current investigation has found is that despite efforts to educate the community about the benefits of self-assessment of one's work, and the disadvantages from its absence, the proportion of non-theoretical ICAIL works containing some form of evaluation has witnessed but modest change in the last ten years. Whereas in the case of some research forums, evaluation is a basic requirement for publication eligibility, in the case of ICAIL, it has remained permissible to submit a work that does not address the fundamental research question – does it work?, and if so, how well? The subtext of this examination is to encourage members of the community to include evaluation within their own works and to advocate that the ICAIL review committees acknowledge those works that do.

7. FUTURE WORK

The authors would like to be able to corroborate the rate of change in AI and Law community research and reporting behaviors sooner than a full ten years from now. There are a number of different directions such research could take. One involves tracking the prevalence of empirical evaluation in non-theoretical works. Another addresses how often theoretical works make a good faith effort to demonstrate the utility of their approach, and still another less formal statistic includes where in the proceedings the successfully evaluated works reside.

8. ACKNOWLEDGMENTS

The authors wish to extend a special word of thanks to Maria Jean Hall, author of the first study of evaluation at ICAIL, who generously provided data files from the original study and corresponding insights. Her recollections and explanations were invaluable. This work was partially funded by Thomson Reuters Global Resources.

9. REFERENCES

[1] Trevor J. M. Bench-Capon, Michal Araszkiewicz, Kevin D. Ashley, Katie Atkinson, Floris Bex, Filipe Borges, Daniele Bourcier, Paul Bourgine, Jack G. Conrad, Enrico Francesconi, Thomas F. Gordon, Guido Governatori, Jochen L. Leidner, David D. Lewis, Ronald Prescott Loui, L. Thorne McCarty, Henry Prakken, Frank Schilder, Erich Schweighofer, Paul Thompson, Alex Tyrrell, Bart Verheij, Douglas N. Walton, and Adam Zachary Wyner. A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law. *Artificial Intelligence and Law*, 20(3):215–319, 2012.

[2] Floris J. Bex, Henry Prakken, and Bart Verheij:. Formalising argumentative story-based analysis of evidence. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL 2007) (Stanford, CA)*, pages 1–10. IAAIL, ACM Press, 2007.

[3] Eleanor Chelimsky. The coming transformations in evaluation. *Evaluation for the 21st Century: A Handbook*, pages 1–26, 1997.

[4] Paul R. Cohen and Adele E. Howe. How evaluation guides AI research. *AI Magazine*, 9(4):35–43, Winter 1988.

[5] Jack G. Conrad, Khalid Al-Kofahi, Ying Zhao, and George Karypis. Effective document clustering for large heterogeneous law firm collections. In *Proceedings of the 10th International Conference on Artificial Intelligence and Law (ICAIL 2005) (Bologna, Italy)*, pages 177–187. IAAIL, ACM Press, June 2005.

[6] Maria Jean J. Hall. *An Adaptive, Subsumptive Evaluation Framework for Intelligent Legal Systems*. PhD thesis, La Trobe University, Faculty of Science Technology and Engineering, Bundoora, VIC, 3086, Australia, Jan. 2005.

[7] Maria Jean J. Hall. Personal communication, 2013.

[8] Maria Jean J. Hall, Richard Hall, and John Zeleznikow. A process for evaluating legal knowledge-based systems based upon the context criteria contingency-guidelines framework. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law (ICAIL 2003) (Edinburgh, Scotland)*, pages 274–283. IAAIL, ACM Press, June 2003.

[9] Maria Jean J. Hall and John Zeleznikow. Acknowledging insufficiency in evaluation of legal knowledge-based systems: Strategies towards a broad-based evaluation model. In *Proceedings of the 8th International Conference on Artificial Intelligence and Law (ICAL 2001) (St. Louis, Missouri)*, pages 146–156. IAAIL, ACM Press, May 2001.

[10] Maria Jean J. Hall and John Zeleznikow. Current inadequacies in the evaluation of legal knowledge-based systems. the context, criteria, contingency evaluation framework for legal knowledge-based systems. In *Proceedings of Fifth Business Information Systems Conference*, pages 219–227, Poznan, Poland, 2002. ACM.

[11] Paul H.J. Hendriks and Dirk J. Vriens. Knowledge-based systems and knowledge management: friends or foes? *Information & Management*, 35(2):113–125, 1999.

[12] Peter E. Jackson. Personal communication, 2009.

[13] Anil S. Jadhav and Rajendra M. Sonar. Evaluating and selecting software packages: A review. *Information and Software Technology*, 51(3):555–563, 2009.

[14] Sindhu Joseph and Henry Prakken. Coherence-driven argumentation to norm consensus. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law (ICAIL 2009) (Barcelona, Spain)*, pages 58–67. IAAIL, ACM Press, 2009.

[15] Joachim Karlsson, Claes Wohlin, and Bjorn Regnell. An evaluation of methods for prioritizing software requirements. *Information and Software Technology*, 39(14):939–947, 1998.

[16] Albert W. Koers. Criteria for the classification of legal knowledge systems advisory systems on legal questions. *Legal Knowledge Based Systems*, pages 23–35, 1990.

[17] Qiang Lu, Jack G. Conrad, Khalid Al-Kofahi, and William Keenan. Legal document clustering with built-in topic segmentation. In *Proceedings of the 20th International Conference on Information and Knowledge Management (CIKM11)*, pages 383–392. ACM Press, 2011.

[18] Mark C. Paulk, Charles V. Weber, Bill Curtis, and Mary Beth Chrissis. *The Capability Maturity Model: Guidelines for Improving the Software Process*. Addison-Wesley, Reading, MA, 1995. SEI series in software engineering.

[19] Henry Prakken. Modelling reasoning about evidence in legal procedure. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law (ICAIL 2001) (St. Louis, Missouri)*, pages 119–128. IAAIL, ACM

Press, 2001.

[20] Henry Prakken. A study of accrual of arguments, with applications to evidential reasoning. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law (ICAIL 2005) (Bologna, Italy)*, pages 85–94. IAAIL, ACM Press, 2005.

[21] Yoram Reich. Measuring the value of knowledge. *International Journal of Human-Computer Studies*, 42(1):3–30, 1995.

[22] Katheryn E. Sanders. *Chiron: Planning in an Open-textured Domain*. PhD thesis, Brown University, Providence, RI, 1995.

[23] Tom Van Engers Silvie Spreeuwenberg and Rik Gerrits. The role of verification in improving the quality of legal decision-making. In *Fourteenth Annual International Conference on Legal Knowledge and Information Systems. Frontiers in Artificial Intelligence and Applications (JURIX 2001)*, pages 1–15, Amsterdam, Netherlands, 2001. IOS Press.

[24] Andrew Stranieri and John Zeleznikow. Evaluating legal expert systems. In *Proceedings of the 7th International Conference on Artificial Intelligence and Law (ICAIL 1999) (Oslo, Norway)*, pages 18–24. IAAIL, ACM Press, 1999.

[25] Andrew Stranieri and John Zeleznikow. *Knowledge Discovery from Legal Databases*, volume 69. Springer Law and Philosophy Library, Dordrecht, The Netherlands, 2005.

[26] Andrew Stranieri, John Zeleznikow, Mark Gawler, and Bryn Lewis. A hybrid rule–neural approach to the automation of legal reasoning in the discretionary domain of family law in Australia. *Artificial Intelligence and Law*, 7(2-3):153–183, 1999.

[27] Richard E. Susskind. Expert systems in law. In *Proceedings of the 1st International Conference on Artificial Intelligence and Law (ICAIL 1987) (Boston, MA)*, pages 1–8. IAAIL, ACM Press, May 1987.

10. APPENDIX

In subsequent research on the subject of formal evaluation culminating in her dissertation, Hall developed a taxonomy of evaluation forms, expressed in her "Context Criteria Contingency-guidelines Framework" [8], which at the highest level relied on a four quadrant matrix representing:

- Validation and verification;
- Technical infrastructure;
- User credibility; and
- Impact on the operating environment.

Hall came to the conclusion that the top-level category of "Impact on the operating environment" represents a broad category from which several sub-categories stem [7]. In other words, numerous AI & Law applications ultimately have some form of impact on their operating environment. Such "impact" is thus a rather high-level and abstract concept. For that reason, we have tended to classify more recent evaluation efforts under some of the finer-grained, more tractable and quantitative forms of performance assessment described in Section 4. We consequently have not assigned any of the works examined in the current work (2001-2011) to this category in Table 4 below.

Year	Paper-type	Proposal	Theoretical				Non-Theoretical					
		Total	Total	0	1	2	Total	0	1	2	3	4
2001	**Full**		10	8	1	1	10	2	1	3		4
	Short						3	2		1		
	Abstract		4	4			6	3	1	2		
2003	**Full**	1	9	8		1	15	5	3	1		6
	Short						4		1	2	1	
	Abstract		4	4			12	8	1	2	1	
2005	**Full**		7		3	4	13	2	4		1	6
	Short	2	1	1			5	3	1	1		
	Abstract	1	3	3			10	5		2	3	
2007	**Full**		8	2	2	4	11	4		2	2	3
	Short	1	8	8			8	3		2	1	2
	Abstract		5	4	1		4	1	1	2		
2009	**Full**	1	7	4	1	2	14		1	3	6	4
	Short	–	–					–				
	Abstract		7	6	1		9	3	1	2	3	
2011	**Full**		9	5	2	2	8			2	2	4
	Short	1	8	5		3	9	6	1	1		1
	Abstract	–	–				–					
Totals	**(Cum. 237)**	7	89				141					

Table 2: Evaluation Distribution in ICAIL Papers (2001-11): Theoretical vs. Non-Theoretical. [Evaluation Depth Scale: 0=None; 1=Discussion; 2=Initial/Basic; 3=Moderate; 4=Comprehensive/Multiple Forms]

Year	Paper -type	Statis- tical	System Compar- ison	Compare w/ Human Performance	Compare w/ Expert Judgment	Discus- sion	Impact on Operating Environmt	Other	Total
2001	**Full**	1	1	1	5				8
	Short					1			1
	Abstract	1			1	1			3
2003	**Full**		3		4	3			10
	Short	1		1	1	1			4
	Abstract				3	1			4
2005	**Full**	3			4	4			11
	Short			1		1			2
	Abstract				3	2			2
2007	**Full**	1	3		2	2			8
	Short			2	3				5
	Abstract				2	1			3
2009	**Full**	3		1	9	1			14
	Short								–
	Abstract	2	2		1	1			6
2011	**Full**	2			4			2	6
	Short	1	1			1			3
	Abstract								–
Totals		15	10	6	42	20	0	2	103

Table 3: Evaluation-type Distribution in ICAIL Papers (2001-11) [A Subset of Table 3]