The Role of Evaluation in AI and Law

An Examination of its Diverse Forms in the AI and Law Journal

Jack G. Conrad*
Thomson Reuters
Research & Development
Saint Paul, Minnesota 55123 USA
jack.g.conrad@thomsonreuters.com

John Zeleznikow†
Victoria University
College of Business
Melbourne, Australia 3086
john.zeleznikow@vu.edu.au

ABSTRACT

This paper explores the presence and forms of evaluation in articles published in the Artificial Intelligence and Law Journal for the ten-year period from 2005 through 2014. It also compares its findings to previous work conducted on evaluation reported on in the Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL). It thus represents a meta-level study of some of the most significant works produced by the AI and Law community, in this case nearly 140 research articles published in the Journal. The paper also highlights works harnessing performance evaluation as one of their chief scientific tools, and the means by which they use it. This work also extends the argument for why evaluation in formal Artificial Intelligence and Law reports such as the Journal is essential. As in the case of its predecessors, it pursues answers to the questions: how good is the system, algorithm or proposal?, how reliable is the approach or technique?, and, ultimately, does the method work? The paper scrutinizes the role of performance evaluation in scientific research reports, underscoring the argument that a performance-based ‘ethic’ signifies a level of maturity and scientific rigor within a community. In addition, the work examines recent publications that address the same critical issue within the broader field of Artificial Intelligence.

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

*Co-author of the second self-reflexive work on evaluation at ICAIL (ICAIL 2013).
†Co-author of the first two self-reflexive works on evaluation at ICAIL (ICAIL 2001 & 2013).

General Terms

Evaluation, Performance, Measurement, Validation

Keywords

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

1 Introduction

1.1 Motivations

To be accepted as a mature, significant computing discipline, it is vital that research in the area of Artificial Intelligence and Law meets rigorous evaluation standards. While such standards should be met when undertaking theoretical, jurisprudential or argumentation research, it is even more important to conduct comprehensive evaluations when conducting pragmatic research such as machine learning or information retrieval.

Hall and Zeleznikow [10] and Conrad and Zeleznikow [7] conducted an in-depth investigation of the degree of evaluation conducted in papers presented in the International Conference on Artificial Intelligence and Law (from 1987 to 2011). The IAAIL community also publishes the Journal of Artificial Intelligence and Law.

We believe it would be valuable to extend the work on evaluation of articles at ICAIL conferences by examining the articles appearing in the Journal of Artificial Intelligence and Law in the ten-year period from 2005 to 2014.

1.2 Previous Work

1.2.1 Evaluation in Computer Science and Artificial Intelligence

One recent example of significant work on evaluation in Computer Science and Software Engineering is research reported by Vermeeren, et al. [25]. They collected ninety-six user experience evaluation methods and analysed them, among other criteria, based on the product development phase and the studied period of experience. Their analysis reveals development needs for user experience methods, such as early-stage methods, methods for social and collaborative user experience evaluation, establishing practicability and scientific quality.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyright for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

ICAIL ’15, June -12 2015, San Diego, CA, USA
Copyright 2015 ACM 000-0-0000-0000-0...$15.00.

http://www.springer.com/computer/ai/journal/10506
Jadhav and Sonar [13] claim that the improper selection of a software package may result in wrong strategic decisions being made. Selecting a software package that meets the requirements needs a full examination of many conflicting factors and it is a difficult task. This has led researchers to investigate better ways of evaluating and selecting software packages. In their article they reviewed the research performed in the field of evaluating and selecting software packages. Jadhav and Sonar [14] claim that evaluating and selecting software packages that meet an organisation’s requirements is a difficult software engineering process. 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 [5] 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. They claim that for the Artificial Intelligence 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. These involve finding criteria for evaluating research problems, methods, implementations, experiments’ design, and evaluation of the experiments.

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.

McCarthy [17] argues that “Artificial Intelligence is the science and engineering of making intelligent machines.” Hernández-Orallo [11] claims that as a consequence, Artificial Intelligence evaluation should focus on evaluating the intelligence of the artifacts it builds. He states that “intelligence tests” (of whatever kind) are not the everyday evaluation approach for Artificial Intelligence. He argues that the explanation for this is that most Artificial Intelligence research is better identified by Minsky’s more pragmatic definition: “[Artificial Intelligence is] the science of making machines capable of performing tasks that would require intelligence if done by humans.” He claims that as a result, Artificial Intelligence evaluation focusses on checking whether machines do these tasks correctly.

Hernández-Orallo [11] further argues that to his knowledge, there is no comprehensive analysis about how evaluation is performed in AI and how it can be improved and adapted to the challenges of the future. He argues that the way Artificial Intelligence evaluation is commonly performed is through task-oriented evaluation, mostly with a black-box approach. He noted that there is still a huge margin of improvement in the way Artificial Intelligence systems are evaluated. He concludes that Artificial Intelligence requires an accurate, effective, non-anthropocentric, meaningful and computational way of evaluating its progress, by evaluating its artifacts.

1.2.2 Early evaluation projects in the Artificial Intelligence and Law community

In 1987 at the first ICAIL conference in Boston, Dr. (now Professor) Richard Susskind, then a systems developer and currently an author and consultant suggested that if legal Knowledge Based Systems were to move out of the research laboratory and into the marketplace, their evaluation was essential [21].

Aleven and Ashley [1] discussed CATO, an intelligent learning environment designed to help beginning law students learn basic skills of making arguments with cases. CATO was evaluated in the context of a legal writing course at the University of Pittsburgh. Aleven and Ashley found that instruction with CATO leads to statistically significant improvement in students’ basic argumentation skills, comparable to that achieved by a skilled instructor teaching small groups of students. However on a more complex and advanced legal writing assignment, students taught by the legal writing instructor had higher grades.

Stranieri and Zeleznikow [23] evaluated the Split Up system using strategies based on a framework put forward by Reich [21]. Hall and Zeleznikow [10] covered three related areas corresponding to three separate levels of granularity in the evaluation of software: “conventional” software systems, knowledge-based systems (KBS), and legal knowledge-based systems. The paper provided a thorough review of the literature and sketches a picture of the existence (or lack thereof) of evaluation in software, Knowledge Based Systems and Legal Knowledge Based System papers. Secondly, it performed an analysis of the presence of evaluation in works published by ICAIL and compares the then three most recent conferences (1995-97-99) with those of the first ICAIL (1987) to see how the presence and use of evaluation has changed.

And thirdly, the paper presents a number of then existing evaluation-related models, such as fundamental models, traditional verification and validation-related models; the O’Keefe and O’Leary hierarchy of expert system quality and the Capability Maturity Model from Carnegie Mellon University.

Hall, et al. [9] conducted a sustained effort to develop a process for evaluating legal knowledge-based systems based upon the Context Criteria Contingency-guidelines Framework. The framework emphasizes the importance of the evaluation context and goals and integrates these with system properties and contingency guidelines to suggest appropriate evaluation criteria. The evaluation process supports the selection of criteria by manual, semi-automated or automated methods and the design of an architecture to support this choice of appropriate criteria, is presented.

Conrad and Zeleznikow [7] defined evaluation as a systematic determination of a subject’s merit, worth and significance, using criteria governed by a set of standards. They claim that 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 [10], their research distinguishes theoretical works from non-theoretical works (assuming that theo-
The motivation for this study was to determine whether, as the community has evolved over time, it has become more mature in its use of empirical methods for performance evaluation and other forms of self-assessment. The hypothesis of this work was that if a researcher does not answer the fundamental question surrounding his or her efforts — how good is the system, algorithm or approach? 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? They reported that more than 60% of ICAIL papers in 1987 did not have any evaluation, but this number had decreased to 20% for papers appearing in ICAIL 2011.

The paper by Conrad and Zeleznikow [7] provoked much discussion at both the 2013 ICAIL conference and amongst the broader IAAIL Community. For example, Vlek, et al. [26] in examining the use of Bayesian Networks for analysing legal evidence conducted what they claimed is an evaluation of their work citing Conrad and Zeleznikow’s (2013) warning that without an evaluation, no researcher can expect the broader audience to be convinced of the benefits and utility of their work. They claim that the results of their case study shows that their design method is indeed capable of representing narratives in a Bayesian network.

Thus the question worth further investigation is – is there a general trend towards including the issue of evaluation in reported research in the Artificial Intelligence and Law community? Because journals, unlike conferences, have no time deadlines, nor page limits, we would expect journal articles to have a greater focus on evaluation than is the case for conferences.

Even papers focused upon argumentation should consider including an evaluation of the research. For example, Caninada and Amgoud [4] defined principles, called rationality postulates, that can be used to judge the quality of a rule-based argumentation system. They defined two important rationality postulates that should be satisfied: the consistency and the closure of the results returned by that system. They provided a relatively easy way in which these rationality postulates can be warranted for a particular rule-based argumentation system. They defined two important rationality postulates that should be satisfied: the consistency and the closure of the results returned by that system. They provided a relatively easy way in which these rationality postulates can be warranted for a particular rule-based argumentation system developed within a European project on argumentation.

In this paper we will address the question of whether there is a general trend towards including the issue of evaluation in the IAAIL Community by examining papers appearing in the Journal of Artificial Intelligence and Law from the period 2005 to 2014. Like the works of [11] and [7], we examine the patterns in theoretical and non-theoretical works, however, henceforth referring to the latter as empirical works.

1.2.3.1 Application: Automatic Case Classification and Prediction

One article that appeared in the *Journal of Artificial Intelligence and Law* and reported on an extensive set of evaluations was Ashley and Brüninghaus [2]. The paper describes a computer program called SMILE + IBP (Smart Index Learner Plus Issue-Based Prediction) that bridges case-based reasoning and extracting information from texts. The program extracts information from textual descriptions of the facts of decided cases and applies that information to predict the outcomes of new cases. The program attempts to automatically classify textual descriptions of the facts of legal problems in terms of Factors, a set of classification concepts that capture stereotypical fact patterns that effect the strength of a legal claim. Using these classifications, the program can evaluate and explain predictions about a problem’s outcome given a database of previously classified cases. The application domain is trade secret misappropriation.

SMILE + IBP has been subjected to extensive evaluation. The authors undertook two sets of experiments, one to evaluate IBP’s predictions as compared to other approaches to legal prediction, and the other to evaluate how well SMILE learned to classify Factors in case texts.

They present an experiment comparing the IBP program’s predictions with those of alternative prediction algorithms and present hypotheses about the relative utility of the three text representations for accurately classifying new textual cases, namely, that more knowledge-intensive representations outperform bag-of-words and that propositional patterns outperform roles-replaced.

It also describes four experiments carried out to evaluate those hypotheses and reports results in support of their claim that SMILE + IBP analyzes cases input as texts:

1. Which representation enabled SMILE to do the best job of assigning Factors to cases as compared to the human assignments?
2. Which representation enabled SMILE + IBP to do the best job of predicting outcomes of cases as compared to the actual outcomes?
3. How well did SMILE + IBP predict outcomes as compared to a baseline of informed guessing?
4. How well did SMILE + IBP predict outcomes as compared to a baseline method of predicting outcomes directly from texts without using Factors?

They described their experimental design (cross-validation-based), the competing algorithms they would test at each step of their study, and the metrics with which they would compare results. In addition, they performed significance tests on their results to see which performance differences were meaningful. Lastly, they interpreted and discussed their results relative to their stated hypotheses in a thorough, empirical manner.

1.2.3.2 Application: Automatic Deception Detection in Cases

Another recent article in the *Journal of Artificial Intelligence and Law*, by Formaciari and Poesio [8], reported on an elaborate series of experiments and analyses. The paper focuses on methods for evaluating the reliability of statements made by witnesses and defendants in court settings. The authors report on results obtained by using stylometric techniques to identify deceptive statements in a corpus of hearings accumulated by Italian courts. Defendants in these cases were convicted of making false statements, thus some of their statements were decidedly untrue. In addition to
harnessing methods from past studies, now applied to “high stakes” data, the authors tested new techniques. Through their comparative analysis, the authors examined the contribution of a number of separate variables as well as the degree of homogeneity of the dataset.

The initial objective of the researchers was two-fold. First, to collect a set of criminal proceedings void of the shortcomings of earlier sets with which to develop computational models for deception detection, and second to compare results from this data set with those from earlier studies, with respect to accuracy and verbal cues employed. In deploying machine learning techniques, the authors used a number of dimensions associated with the LIWC Italian dictionary\(^3\) as well as their own features tied to word n-grams and POS. The techniques examined were put through a rigorous series of evaluations. The two feature selection strategies the authors invoked included Best Frequencies (assessing separate n-gram frequency lists) as well as the Information Gain metric. Their tests relied upon measures of precision, recall, and F-score and used three baselines (random, heuristic and majority).

Some of the additional variables they investigated included gender (e.g., male speakers only), language level (e.g., native language speakers only) and age (e.g., speakers over 30 years of age). The authors found that all of their models managed to surpass 70% accuracy, but they generally required an amount of training data in the range of 25 hearings minimum in order to surpass the best performing (majority) baseline. Results were also correlated with the percentage of false utterances a given speaker makes and the length of those utterances. The impact of noise (uncertain utterances) was also examined. When uncertain utterances were removed from the modeling, the prediction accuracy gap between the model and the baseline grew by up to 9%. In addition, the authors explored the impact of specific language used when comparing their results with those of past researchers, in terms of both tokens and n-grams.

Fornaciari and Poesio were able to confirm the utility of the features contributing most to detecting false utterances in their environment relative to seminal past studies\(^1\) and identify which combinations of features contributed best to the their models. In summary, they conducted a rigorous, exhaustive, solidly assessed and well-documented scientific study.

2 **Why Should Legal Applications Be Evaluated?**

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\(^2\). 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 \(^6\)\(^16\). 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.\(^19\)\(^20\)\(^3\)\(^15\).

3 **Evaluation of Legal Applications**

In earlier ICAIL works,\(^10\) and\(^7\) discussed evaluation methodologies suitable for the legal domain and the strategies with which AI and Law researchers might frame their evaluation. In this current report, we extend those studies of ICAIL Proceedings to the *Journal of AI and Law*. We first address methodologies to be used to examine the presence and degree of evaluation in these works, and then analyze the data collected and scored.

3.1 **Methodology**

In the subsections below, we provide background descriptions on how our evaluation rating system for the *Journal of AI and Law* evolved from the binary classification approach undertaken by\(^10\) to the 5- and 3- category system used for empirical and theoretical papers, respectively, in the current work. In addition, we describe six categories of evaluation ‘types.’ These differ to some extent from the set used by\(^10\) and\(^7\). Whereas\(^7\) elected to use the same set as that used by\(^10\) for purposes of consistency and comparison, we rely on a modified set as it better conforms to what was encountered in the data, and each category is distinct (as now defined, there is no possibility of overlap or dual assignments).

3.1.1 **Evaluation Ratings (Grades) for JAIL Papers**

Despite the challenges that exist in assessing the effectiveness and efficiencies of systems, algorithms and approaches to problem solutions in the legal domain, we believe there are clear benefits to investigating how well we do as a research community in the underlying science of AI and Law. To that end, our objective in this report is to explore how the papers published by the *Journal of AI and Law* have made credible efforts to evaluate the performance of their work in an appropriate manner. If researchers in AI and Law wish to demonstrate the value and utility of their work to the broader scientific community, it is imperative to be able to answer questions like how well does the approach work? Finding evidence of this pursuit along with answers to questions like those above were among the chief goals of this study. We thus distinguish in this work those papers that perform some appropriate form of evaluation from those that do not, while recognizing that theoretical or proposition papers generally would not contain that degree of scientific evaluation.

We rely on the same level of granularity for empirical works as\(^4\) five categories ranging from no evaluation to mature, multi-faceted evaluation. These levels are also associated with corresponding “grades,” from A (for thorough) to F (for no presence). These categories are described below.

---

\(^3\)*Lexical Inquiry and Word Count Dictionary*
Algorithmic – assessment made in terms of performance of a system such as a multi-agent simulation system

Operational-Usability – assessment of systems operational characteristics or usability aspects

Other – those systems with distinct forms of evaluation not covered in the categories above (e.g., task-based, conversion-based, etc.)

These categories are similar but different from those used employed by [10] and [7], insofar as some of their categories could be assigned to the same research work (e.g., expert opinion and computer generated). As a result, these category definitions were refined to better suit the distinct forms of evaluation we observed in the Journal articles. In the figures presented and discussed below, we show how the Journal of AI and Law articles from the last ten years are distributed across these categories. A master table containing the complete set of these assignments can be found in the Appendix.

3.2 Current Findings

Figure 1 presents a breakdown of articles published in the AI and Law Journal based on their status as theoretical or empirical works. Empirical works are represented in the green and violet bars on the right-hand side while theoretical works are represented in the blue and red bars on the left-hand side. Empirical works that contain no form of evaluation are represented in the violet bars on the far right (E0). Theoretical works that similarly contain no examples or illustrations are represented in the red bars towards the left (T0).

---

3.1.2 Forms of Evaluation Identified

In addition to examining the level of evaluation or illustration shown in the JAIL papers, it is also informative to consider the types of evaluation pursued. There is a broad range of evaluations undertaken by researchers. Because of the distinct nature of evaluation displayed in the Journal in the last decade, we have decided to compile our own set of categories, rather than follow those sets used by Conrad and Zeleznikow during the presentation of their work at ICAIL 2013. The levels can be described as follows.

0. [T0] Absent. No mention of assessment in any form. No substantive examples or illustrations of how the given approach or model would be applied.

1. [T1] Initial Assessment or Illustration. Paper broaches the subject of how the proposed approach or model can be assessed or illustrated.

2. [T2] More complete Demonstration. Paper makes a clear effort to demonstrate the utility or coverage of the approach or model, often by one or more extended examples.

---

Gold Data – evaluation performed with respect to domain expert judgments (e.g., accuracy, precision, recall, F-score, etc.)

Statistical – evaluation performed with respect to some comparison function (e.g., for unsupervised learning: cluster internal-similarity, cosine similarity, etc.)

Manual Assessment – performance is measured by humans via inspection, assessment, review of output

---

In [10] and [7], empirical works were classified as "non-theoretical."
Extraction, both of which could be argued to lie on boundary of theory and applications. So from a research quality perspective, overall this is a positive pattern to observe. In addition, when one examines the theoretical works as represented in Figure 1, one again sees a solid presence of works that have been assessed or illustrated in a significant manner (fraction in blue), whereas the proportion of works that do not have such a presence varies substantially from year to year (fraction in red)\textsuperscript{5}. This variance may again depend on the topics addressed in the issues for a given year. We note that of the two years containing the largest percentage of theoretical works lacking assessment or extended examples (2008 and 2012), both had two of four issues which were published as special issues. The issues in question covered topics including Institutions and Legal Theory, Norms and Laws, in addition to Modeling Legal Cases. At least in the case of the first two, topics tending to be more abstract may play a role in the reduced presence of some form of assessment or extended illustration. Yet in general, one can see that the green and blue portions of the bars dominate, indicating that the role of evaluation and assessment, certainly among empirical works but also among theoretical works, is a sustained practice among the published reports. It also indicates that the reviewers who are working on behalf of the journal take their role in the submission process seriously by ensuring that such validations are present in accepted works.

Since our primary interest is aimed at those works that are either subjected to credible evaluation or not from among the main set of articles where evaluation can generally be performed, in Figure 2, we focus on empirical papers only. Starting at the top of the pie chart and moving clock-wise, one sees the five different levels of evaluation laid out for empirical works, starting with no evaluation and moving around finally to mature and varied evaluation. It is encouraging to see that fully two-thirds of the works shown in the figure represent evaluated works (re: green, violet, light blue sections). And of these, the largest category representing one-third of the figure is the “mature and varied” category. For those works subjected to no evaluation or “discussed” evaluation, more than three times as many works present some form of “discussion” of evaluation than works with no form of evaluation. That only 5% of the articles on empirical subjects contain no evaluation is a positive finding. Clearly, 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 in Table 1.

Earlier in this work, we contended that it is possible to discuss performance assessment in some form even for theoretical works that may not directly involve computing systems. We recognize, however, that for such works, the form that appropriate assessment may take would be different than that performed for empirical works and the computing systems, algorithms and approaches that they address. To this end, our current study, like \[7\] before us, also examines theoretical works published in the journal during the years in question. Given the methodology described above, we assign one of three categories of assessment to the theoretical works: no form of assessment, a preliminary level, and finally a more dedicated or labored level. The resulting distribution is shown in Figure 3. From the figure, one can see that over half the theoretical works examined contain some form of extended example or other illustration of the utility, coverage or robustness of what is proposed. It is also encouraging to see that another 10% of the articles fall into the second category presenting lesser forms of such illustrations or discussions. Finally, just over one-third of the theoretical articles present no form of such examples, illustrations or discussions. The size of this category suggests that there remains room for improvement on the theoretical side in illustrating their contributed value.

The next topic that we examined in this study was the categories of evaluation conducted, in particular, among the empirical works. The distribution among these works is shown in Figure 4. Here we see that nearly half of the works evaluated did so using gold data or other forms of judgments provided by domain experts in order to facilitate the assessment. Classification measurements or measures of accuracy, precision, recall, and similar metrics relying on class assignments, relevance judgments, or similar provided by some form of “experts” is the most frequent form in this category. The next largest category in this study is that of manual assessments, often performed by grad students or research assistants. These are conducted when the trials are small scale and require a basic proof of concept assessment.
Comparisons with Earlier Results

Current Findings

Discussion

4 Discussion

4.1 Current Findings

In the results presented above, for both the ICAIL conference proceedings and more particularly for the Journal of Artificial Intelligence and Law we have observed some encouraging trends by way of modest increases in the presence of evaluation, almost exclusively in the empirical works. At the same time, there may remain a little 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 empirical conference 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 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 [22] 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. [10] continued this work by analysing the proceedings of four ICAIL conferences. In that paper, their stated goal was to determine the rate of reported evaluations in non-theoretical papers.

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[5] exhortation to evaluate the systems it has developed. No attempt is made to develop new evaluation formalisms.

From Table 1 (using the four conferences examined by [10] 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.

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. Empirical 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. Empirical 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 a self-reflexive study of the IAAI community in terms of the percentage of published Journal of AI and Law papers containing some degree of evaluation.

It has also compared current findings with earlier works that examined the presence of evaluation in ICAIL conference proceedings. Allowing for a sizable variety of types of evaluation, what the current investigation has found is that possibly thanks to earlier efforts to educate the community about the benefits of self-assessment of one’s work, and the deficiencies arising from its absence, the proportion of empirical AI and Law Journal articles containing some form of evaluation has been kept a high level over the last ten years. Meanwhile the proportion of theoretical Journal of AI and Law articles containing extended examples and other illustrations of utility still represent the majority of the works shared, though there is clearly still room for improvement.

Over time and with dedicated reviewers who share these understandings, one can remain cautiously optimistic that these trends will continue.

7 Future Work

The authors would like to be able to corroborate the rate of change in AI & Law community research and reporting behaviors more frequently than every ten years. There are a number of different directions such research could take. One involves tracking the prevalence of scientific evaluation in empirical 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 in what areas in the AI and Law proceedings or journal the successfully evaluated works are found.

8 Acknowledgments

The authors wish to thank Catherine Apuy for her tireless help in creating a database of records and annotations corresponding to each of the articles published in the Journal.

9 References


**10 Appendix**

Additional Notes:

The following special Journal contributions were not considered in the following table:

a — In 2010 there was one “critical review” in Issue 3;

b — In 2011 there was one corpus profiled in Issue 4;

c — In 2012, the special collection of reviews celebrating the “ICAIL at 25” appeared in issue 3.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>E0</td>
<td>E1</td>
<td>E2</td>
<td>E3</td>
</tr>
<tr>
<td>2005</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3-4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2006</td>
<td>14</td>
<td>1-2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2007</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>2008</td>
<td>16</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2009</td>
<td>17</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2010</td>
<td>18</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>2011</td>
<td>19</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2012</td>
<td>20</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2013</td>
<td>21</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>2014</td>
<td>22</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>17</td>
<td>6</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 2: Evaluation Distribution in AI & Law Journal Papers (2005-14): Empirical vs. Theoretical. Eval. Depth Scale: 0=None; 1=Discussion; 2=Initial/Bas; 3=Moderate; 4=Comprehensive/Multiple Forms [* = Special Issue]