The Role of Evaluation in AI and Law

An Examination of its Different 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, VIC 8001 Australia
john.zeleznikow@vu.edu.au

ABSTRACT

This paper explores the presence and forms of evaluation in articles published in the journal *Artificial Intelligence and Law* for the ten-year period from 2005 through 2014. It represents a meta-level study of some the most significant works produced by the AI and Law community, in this case nearly 140 research articles published in the *AI and Law* journal. It also compares its findings to previous work conducted on evaluation appearing in the *Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL)*. In addition, the paper highlights works harnessing performance evaluation as one of their chief scientific tools and the means by which they use it. It extends the argument for why evaluation is essential in formal Artificial Intelligence and Law reports such as those in the journal. As in the case of two earlier works on the topic, 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 investigates 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

General Terms

Evaluation, Performance, Measurement, Validation

Keywords

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

ICAIL '15, June 08-12, 2015, San Diego, CA, USA
ACM 978-1-4503-3522-5/15/06
http://dx.doi.org/10.1145/2746090.2746116

1. INTRODUCTION

1.1 Motivations

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

Hall and Zeleznikow [6] and Conrad and Zeleznikow [5] performed in-depth investigations on the degree to which evaluation was performed in papers presented in the *Proceedings of International Conference on Artificial Intelligence and Law (ICAIL)* (from 1987 to 2011). The IAAIL association that organizes ICAIL[1] also publishes the journal *Artificial Intelligence and Law.*[2] Given the above foundation, we believe it would be additionally instructive to extend this earlier work on evaluation of articles in ICAIL proceedings by examining the articles appearing in the journal *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

Cohen and Howe [3] argue that 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 present a five-stage model of AI research and develop 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.

The remainder of this section on Previous Work can be found on pages two and three of the full-length version of this paper here: http://tinyurl.com/lyffqvy

Thus the question worth further investigation is – is there a general trend towards including the issue of evaluation

[1]IAAIL is the International Association for Artificial Intelligence and Law
[2]http://www.springer.com/computer/ai/journal/10506

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, Caminada and Amgoud [2] 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 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 *Artificial Intelligence and Law* from the period 2005 to 2014. Like the works of [6] and [5], we examine the patterns in theoretical and non-theoretical works, however, henceforth referring to the latter as empirical works.

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 [3]. 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 [4, 9].

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. [10, 1, 8].

3. EVALUATION OF LEGAL APPLICATIONS

In earlier ICAIL works, [6] and [5] 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 *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 *AI and Law* evolved from the binary classification approach undertaken by [6] 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 [6] and [5]. Whereas [5] elected to use the same set as that used by [6] 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 *AI and Law* have made credible efforts to evaluate the performance of their work in an appropriate manner.[3] 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 suitable form of evaluation from those that do not, while recognizing that theoretical or proposition papers generally would not contain the same types of scientific evaluation.

We rely on the same level of granularity for empirical works as [5] – 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.

0. [E0] Absent (Grade: F). No mention of evaluation in any form in the published work.
1. [E1] Discussion (Grade: D). Paper discusses how the proposed system or approach could be evaluated.
2. [E2] 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. [E3] 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. [E4] 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] In headings and in figures here and to follow, 'JAIL' is used to refer economically to the journal *AI and Law*.

For a means of examining the degree of assessment or practical "illustration" carried out by authors of theoretical papers or research propositions, we use a simplified, three-tier scale that was 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 way of one or more extended examples.

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 [5] and [6] for the ICAIL Proceedings reports. The six *types* of evaluation that we have recorded are based on the most common and reasonable types we have observed in the data we examined. They include the following:

- 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
- 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 employed by [6] and [5], 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 *AI and Law* journal 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 (see http://tinyurl.com/qc278qt).

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.[4] 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 substantive examples or illustrations are represented in the red bars towards the left (T0).

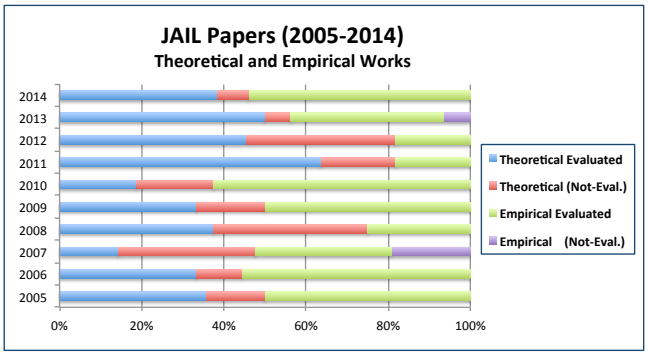

Figure 1: Proportion of Theoretical vs. Empirical Works (Evaluated and Non-Evaluated)

If one examines the empirical works represented on the right-hand side in Figure 1, it is clear that the vast majority of these works do include evaluation as part of their overall research function (fraction in green). It is rare to see a year in the figure where there is a significant presence of non-evaluated works (fraction in violet). In the one year where that appears to be an exception, 2007, it could be pointed out that three of the four issues were special issues, including Legal Knowledge Modeling and Legal Knowledge 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).[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

[4] In [6] and [5], empirical works were classified as "non-theoretical."

[5] We remind the reader that for theoretical works, we view evaluation as a form of assessment accomplished through extended examples and illustrations of the proposal's overall utility, coverage and robustness.

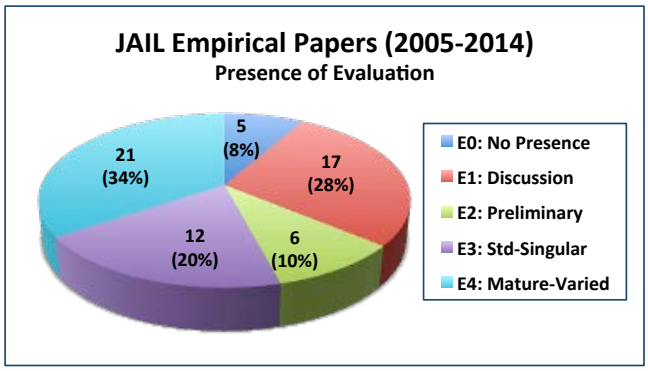

Figure 2: Empirical Works: Degrees of Evaluation (2005-2014)

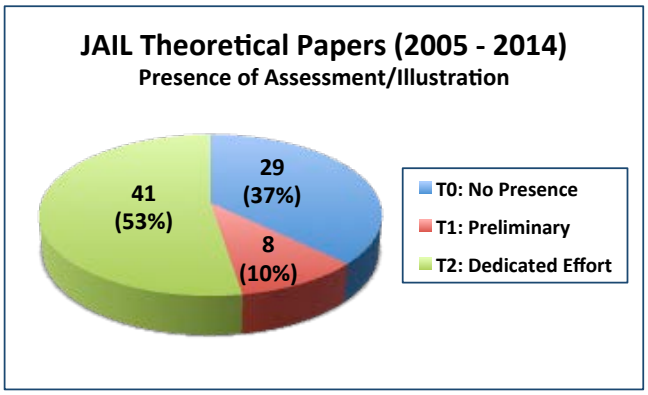

Figure 3: Theoretical Works: Presence of Assessment/Illustration (2005-2014)

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 [5] 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. Together these two forms comprise two-thirds of all of the evaluations performed. The next two categories representing about 10% of the field each include statistical, referring to measures used on areas such as unsupervised learning (e.g., clustering where measures like internal or external similarity may be recorded, or cosine similarity for vector-based systems) and various kinds of algorithmic assessment (e.g., of agent-based systems where the performance of agent simulation algorithms is being assessed). This last category may have resulted from the fact that there has been at least one special issue of the Journal focusing on agent-based systems. There is also the category of operational or usability topics where system complexity or system usability is tracked. Finally, included in the "Other" category would be task-based assessments (e.g., student performance with and without a tool) and system-based assessments judging how well a graph analysis and conversion was performed. Are these categories general enough that they could be applied to any decade of Journal articles? For the larger categories, yes, whereas possibly not for the smaller categories which have been influenced by the sometimes narrower research topics that required them.

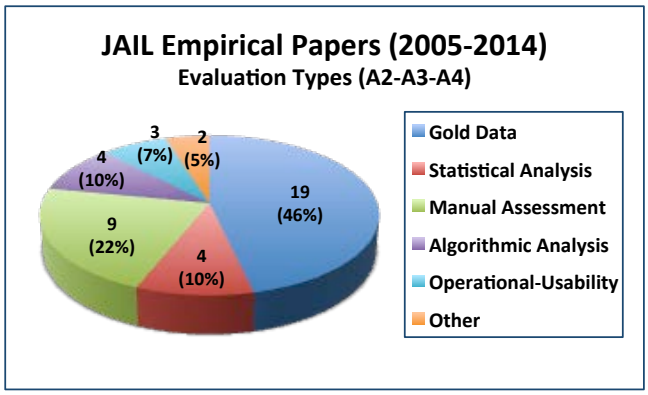

Figure 4: JAIL – Types of Evaluated Works (2005-2014)

3.3 Comparisons with Earlier Results

The table we present in this section is comparable to the yearly breakdowns shown in Figure 1. Table 1 summarizes the findings of [6] and [5]. At the time, it permitted a side-by-side comparison of the researchers' first study (1987, 1995-99) and the second (2001-11) as well as a focus on the size of the underlying data sets. In general, the number of works published at ICAILs (excluding 1987) ranged from the mid-to-low 30s to the mid-40s.[6] If one examines the numerical evaluation patterns, one can observe that the raw number of evaluated works has been increasing (2011 excepted), with a considerable increase into double digits following [6] in 2001. Whether or not this work had an appreciable impact on the community's evaluation practices, at least as evidenced by the ICAIL proceedings, is an open question. However, when we compare that study and its findings with those from this study, we can make a number of instructive observations. First of all, the size of the two data sets is different, 238 entries examined in the ICAIL works versus 139 by the current work. And whereas both studies cover exactly ten years, we recognize that the research reported in this paper is more narrowly focused and current (also covering 2012-2014), whereas just over half of the year's examined by the prior work pre-date 2005, the first year of the current study. While the presence of evaluation has been observed to be increasing in recent years within the non-theoretical body of work in ICAIL, this has never been an issue for the journal, at least in the decade examined in the present work. As observed earlier, if one examines the empirical works shown on the right-hand side of the bars in Figure 1, it is clear that the great majority of these works do include evaluation as part of their overall research function (fraction in green). It is rare to see a year in Figure 1 where there is a significant presence of non-evaluated works (fraction in violet). So whereas it may be viewed as a strength to have a tradition where even theoretical works include extended illustrations and discussions of practical coverage, it is clearly a positive take-away to witness that the vast majority of all empirical works presented in the Journal have been accompanied by a significant degree of evaluation.

[6]Authors of [5] had initially envisioned reporting on only full-length papers; however, upon observing that there was nearly as much evaluation reported on in the short papers and two-page abstracts, they decided to include these other works in their study as well.

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)

4. DISCUSSION

4.1 Interpretation of Findings

In the results presented above, for both the ICAIL conference proceedings and more particularly for the journal *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 some 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 [5],. 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 [7].

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

Stranieri and Zeleznikow [11] 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. [6] 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 more constrained: we limit ourselves to examining whether the AI and Law community has followed Cohen and Howe's [3] 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 [6] and the six by [5], 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 advised 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 AND FUTURE WORK

This work has conducted a self-reflexive study of evaluation in the IAAIL community in terms of the percentage of published papers in the journal *AI and Law* 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, the current investigation has found 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 articles in the journal containing extended examples and other illustrations of utility still represent the majority of the works presented, though there remains room for improvement. Over time and with dedicated reviewers who share these understandings, one can remain cautiously confident that these trends will continue.

In future work, besides additional internal comparisons over time, we would like to begin to compare patterns in the AI and Law community with those in other AI subfields.

7. REFERENCES

[1] F. J. Bex, H. Prakken, and B. 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.

[2] M. Caminada and L. Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 17(5):286–310, 2007.

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

[4] J. G. Conrad, K. Al-Kofahi, Y. Zhao, and G. 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.

[5] J. G. Conrad and J. Zeleznikow. Acknowledging insufficiency in evaluation of legal knowledge-based systems: Strategies towards a broad-based evaluation model. In *Proceedings of the 14th International Conference on Artificial Intelligence and Law (ICAL 2013) (Rome, Italy)*, pages 186–191. IAAIL, ACM Press, June 2013.

[6] M. J. J. Hall and J. 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 (ICAIL 2001) (St. Louis, Missouri)*, pages 146–156. IAAIL, ACM Press, May 2001.

[7] P. E. Jackson. Personal communication, 2009.

[8] S. Joseph and H. 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.

[9] Q. Lu, J. G. Conrad, K. Al-Kofahi, and W. 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.

[10] H. 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.

[11] A. Stranieri and J. 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.