THE SIGNIFICANCE OF EVALUATION IN AI & LAW
A Case Study Re-examining ICAIL Proceedings

Jack G. Conrad
Thomson Reuters Global Resources
Catalyst Lab
Baar, Switzerland  6340

John Zeleznikow
Victoria University
School of Management & Info Systems
Melbourne, Australia  3086

14th Int’l Conf. on Artificial Intelligence & Law
Rome, Italy – 10-14 June 2013
OUTLINE

• Background – Original Study of ICAIL Proceedings

• Update – How We’ve Performed Since
First Study of Evaluation in ICAIL Proceedings

- Marie Jean J. Hall and John Zeleznikow
- In Proceedings of the 8th International Conference on Artificial Intelligence and Law (ICAIL 2001) (St. Louis, Missouri), pp. 147-156

“Acknowledging the Insufficiency in the Evaluation of Legal Knowledge-based Systems”

- Verification and validation
- Beyond verification and validation
- Assessment and evaluation

- Conventional Software Systems
- Knowledge-based Systems
- Legal Knowledge-based Systems

- Papers categorized by: “Theoretical”, “Evaluated”, “Not Evaluated”
- Focus on last two
- Also examined type of evaluation used

Evaluation

Software Systems

Survey of ICAIL Proceedings

Evaluation Strategies

- An evaluation methodology
- An Evaluation Context Checklist
- Strategies beyond development of methodology

Blue – general examination
Orange – domain-specific examination
OUTLINE

• Background – Original Study of ICAIL Proceedings

• Update – How We’ve Performed Since
Current Study of Evaluation in ICAIL Proceedings

- A self-reflexive, meta-level study
- Examines the presence of evaluation in works published at ICAIL since 2000 (2001 – 2011)
- Proportion of works that include some form of performance evaluation and their degree
- Compares these recent trends with those identified by Hall and Zeleznikow (ICAIL 2001)
- Develops an argument for why evaluation in formal AI and Law reports is significant
Current Study of Evaluation in ICAIL Proceedings

• **Objective:** Determine how the presence of evaluation at the community level has evolved over time

• **Motivation:** Investigate along one significant dimension if community has matured in use of empirical assessment

• **Proposition:** If fundamental questions unanswered – How good is the system? How reliable is the technique? Does it work? – how can the researcher convince the broader community of the benefits and utility of the work?

• **Definition:** *Evaluation* – systematic determination of subject’s merit, worth, significance using criteria governed by a set of standards
References


Theoretical vs. Evaluated & Non-Evaluated Works

ICAiL Papers (1987 - 2011)

- 2011
- 2009
- 2007
- 2005
- 2003
- 2001
- 1999
- 1997
- 1995
- 1987

0% 20% 40% 60% 80% 100%

- Theoretical
- Evaluated
- Not Evaluated
Evaluation in Non-Theoretical Works

ICAiL Non-Theoretical Papers (1987-2011)

- not evaluated
- evaluated
Presence of Assessment in Theoretical Works

ICAIR Theoretical Papers (2001-2011)

Presence of Assessment

<table>
<thead>
<tr>
<th>Year</th>
<th>Degree 0</th>
<th>Degree 1</th>
<th>Degree 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>2009</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>2007</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>2005</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>2003</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>2001</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Deg 0 = no assessment
Deg 1 = initial assessment
Deg 2 = effort to measure utility or coverage
Type of Evaluated Works (current study)

ICAIL Evaluated Papers (2001-2011)

- Foreshadowed only
- Expert opinion
- Human performance
- Statistical
- Computer generated
- Environment Impact
- Other
### Theoretical vs. Evaluated & Non-Evaluated Works

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

- 0 – **Absent** (F) – no mention of evaluation in any form
- 1 – **Discussion** (D) – discusses how the system or approach could be evaluated
- 2 – **Basic** (C) – preliminary, simply evaluation is performed on a portion of system or data, or evidence of anecdotal assessment
- 3 – **Moderate** (B) – significant evaluation effort is performed on the system or approach
- 4 – **Mature/Comprehensive** (A) – credible degree of evaluation performed, us. multiple assessments
  - E.g., modular vs. end-to-end; vs. baselines; vs. humans
Concluding Remarks

- Current ICAIL evaluation landscape leaves room for improvement
- Short of full-fledged experiments, sketches of how future evaluation should be conducted can be helpful
- Even theoretical works can have extended examples and illustrations of coverage
- To be a mature research community exercising scientific rigor, multi-faceted, in-depth evaluation is required
Current Study of Evaluation in ICAIL Proceedings

- Jack G. Conrad and John Zeleznikow
- The Significance of Evaluation in AI and Law: A Case Study Re-examining ICAIL Proceedings
- In Proceedings of the 14th International Conference on Artificial Intelligence and Law (ICAIL 2013) (Rome, Italy), pp. 186-191

Focus: ICAIL Proceedings 2001 - 2011

Full-length Version & Recent Paper Classifications

http://www.conradweb.org/~jackg/publications.html
Final Perspectives

• “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 comparable perspective.” -- PEJ

• “One might question whether conferences are the correct place to report evaluation. Conferences might be seen as a place to float initial ideas that report new advances: these are reported so that they can be developed and evaluated subsequently. If everything were required to be evaluated where would such progress come from? Also conference papers are very space limited: often there is too little room to report both the system and detailed experiments.” -- Anonymous Reviewer
THE SIGNIFICANCE OF EVALUATION IN AI & LAW
A Case Study Re-examining ICAIL Proceedings

Jack G. Conrad
Thomson Reuters Global Resources
Catalyst Lab
Baar, Switzerland 6340

John Zeleznikow
Victoria University
School of Management & Info Systems
Melbourne, Australia 3086

Questions & Discussion