Scaling-up AI Systems:
Insights from Computational Complexity

Bart Selman & Carla Gomes
Cornell University

Ryan Williams
Stanford → MIT
Computational Complexity Hierarchy

**EXP-complete:**
- games like Go, …

**PSPACE-complete:**
- QBF, *planning*, chess (bounded), …

**#P-complete/hard:**
- #SAT, sampling, *probabilistic inference*, …

**NP-complete:**
- SAT, deep learning, *propositional reasoning*, scheduling …

**P-complete:**
- circuit-value, …

**In P:**
- DB, sorting, shortest path, …

---

What are the implications for human understanding of machine intelligence?
Hypothesis: Even though machines are moving to higher levels of the computational complexity hierarchy, it may not necessarily be the case that humans won’t be able to understand their behaviors/decisions.

Why? In earlier work, we showed how automated reasoning on very large reasoning problems (millions of variables) can often be understood in terms of the behavior of a small set (a few dozen) of key variables (“backdoor variables”). The machine can provide the backdoor variables (i.e., explains itself).

<table>
<thead>
<tr>
<th>$B(n)$</th>
<th>deterministic</th>
<th>randomized</th>
<th>heuristic</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n/k$</td>
<td>small $exp(n)$</td>
<td>smaller $exp(n)$</td>
<td>tiny $exp(n)$</td>
</tr>
<tr>
<td>$O(\log n)$</td>
<td>$(\frac{n}{\log n})^{O(\log n)}$</td>
<td>$(\frac{n}{\log n})^{O(\log n)}$</td>
<td>$poly(n)$</td>
</tr>
<tr>
<td>$O(1)$</td>
<td>$poly(n)$</td>
<td>$poly(n)$</td>
<td>$poly(n)$</td>
</tr>
</tbody>
</table>
So, at least in the context of automated reasoning, there is hope for human understanding of complex machine reasoning.

More formally: When do we have short witnesses for complex computational tasks? (NP-complete / #P / PSPACE; typical case)

Concrete challenge: Recent 8 GB machine proof of the Erdős discrepancy conjecture (conj. stated in 1932; solved 2014). Does a human accessible version exist? (likely… we conjecture)

Focus: Human understanding of super-intelligent machines, cont.

For example, 1160 +/- 1 elements, all sums of sub-sequences stay between -2 and +2. No 1161-sequence exists!