The Comprehensive AI Services framework

Eric Drexler
Future of Humanity Institute
University of Oxford

Beneficial AGI Workshop
3 January 2019
Puerto Rico

Reframing Superintelligence
Comprehensive AI Services
as General Intelligence
K. Eric Drexler
Technical Report #2019-1
Some basic questions

Where do AI systems come from? — Research and development
What do AI systems do today? — Perform tasks (i.e., provide services)
What will AI systems be able to do? — Automate human tasks (and more)
What is research and development? — A bunch of tasks to automate
Where does AI R&D automation lead? — Recursive technology improvement
Where does that lead? — Comprehensive AI services (CAIS)

Automated service development is general intelligence
- The “C” in CAIS does the work of “G” in AGI

Mostly descriptive, partly prescriptive
<table>
<thead>
<tr>
<th>Characteristics</th>
<th>...in AGI-agent models</th>
<th>...in AI-service models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical embodiments:</td>
<td>Distinct AI agents</td>
<td>Networked AI services</td>
</tr>
<tr>
<td>Action evaluation:</td>
<td>Global utility function</td>
<td>Task objective function</td>
</tr>
<tr>
<td>Default goals:</td>
<td>Resources, autonomy</td>
<td>Task completion</td>
</tr>
<tr>
<td>Stronger optimization:</td>
<td>More powerful, general</td>
<td>More effective, focused</td>
</tr>
<tr>
<td>Recursive improvement:</td>
<td>Self-improving agent</td>
<td>Automated AI R&amp;D</td>
</tr>
<tr>
<td>Takeoff threshold:</td>
<td>Agent capabilities</td>
<td>Technological capabilities</td>
</tr>
<tr>
<td>Takeoff dynamics:</td>
<td>Autonomous, abrupt</td>
<td>Managed, potentially swift</td>
</tr>
<tr>
<td>Developer experience:</td>
<td>Unprecedented, novel</td>
<td>Extensive, incremental</td>
</tr>
<tr>
<td>Autonomous AGI:</td>
<td>Uniquely valuable</td>
<td>Needlessly risky</td>
</tr>
</tbody>
</table>
Safety-relevant affordances:

- Implementation relationships (A optimizing B to provide service C)
- Optimization as a constraint (components perform known tasks)
- Separation of learning and performance (stable competencies)
- Competition among systems (enables comparison, choice)
- Adversarial systems (red-team/blue-team, proposers and critics)
- Matching tasks to services in semantically interpretable spaces
- Employing predictive models of human concerns and (dis)approval
Descriptive and prescriptive...

*Description:*

Consider patterns of system development and structure

*Prescription:*

Exploit affordances of system development and structure