Exploring AGI Scenarios

Shahar Avin
sa478@cam.ac.uk
This is Bob.

Bob heads an AGI R&D lab.

What should Bob do?
AGI futures narratives

- Tech utopia
- Arms race
- Malicious use
- Existential risk
How do we explore and communicate these futures?

Single author exploration
How do we explore and communicate these futures?

Expert workshops, multi-authored reports

http://maliciousaireport.com/
How do we explore and communicate these futures?

Data trends

https://www.eff.org/ai/metrics

https://blog.openai.com/
How do we explore and communicate these futures?

Aggregate probability estimates

https://www.getguesstimate.com

https://www.metaculus.com
How do we explore and communicate these futures?

Video games

https://bit.ly/2uqXNUUn

http://www.decisionproblem.com/paperclips/
What my friends think I do

What my parents think I do

What society thinks I do

What my boss thinks I do

What I think I do

What I really do
What should we be looking at?

Development

Deployment

Landscape
Development factors

Inputs

Data use policy

Information we receive and how it is used
Learn about the types of information we receive, and how that information is used.

Sharing and finding you on Facebook
Get to know the privacy settings that help you control your information on facebook.com.

Sharing with other websites and applications
Find out about the ways your information is shared with the games, applications and websites you and your friends use off Facebook.

More resources
Interactive tools
View the complete Data Use Policy

Last updated: 23 September 2011
Development factors

Nature of the problem
Development factors

Control, incentives, openness

OpenAI Charter

We're releasing a charter that describes the principles we use to execute on OpenAI's mission. This document reflects the strategy we've refined over the past two years, including feedback from many people internal and external to OpenAI. The timeline to AGI remains uncertain, but our charter will guide us in acting in the best interests of humanity throughout its development.

Solve intelligence. Use it to make the world a better place.
Development factors

Safety and Security

<table>
<thead>
<tr>
<th>Specification</th>
<th>Robustness</th>
<th>Assurance</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Core purpose of the system)</td>
<td>(Design system to withstand perturbations)</td>
<td>(Monitor and control system activity)</td>
</tr>
<tr>
<td>Design</td>
<td>Prevention and Risk</td>
<td>Monitoring</td>
</tr>
<tr>
<td>Bugs &amp; inconsistencies</td>
<td>Risk sensitivity</td>
<td>Interpretability</td>
</tr>
<tr>
<td>Ambiguities</td>
<td>Uncertainty estimates</td>
<td>Behavioural screening</td>
</tr>
<tr>
<td>Side-effects</td>
<td>Safety margins</td>
<td>Activity traces</td>
</tr>
<tr>
<td>High-level specification languages</td>
<td>Safe exploration</td>
<td>Estimates of causal influence</td>
</tr>
<tr>
<td>Preference learning</td>
<td>Cautious generalisation</td>
<td>Machine theory of mind</td>
</tr>
<tr>
<td>Design protocols</td>
<td>Verification</td>
<td>Tripwires &amp; honeypots</td>
</tr>
</tbody>
</table>

| Emergent | Recovery and Stability | Enforcement |
| Wireheading | Instability | Interruptibility |
| Delusions | Error-correction |Backing |
| Meta-learning and sub-agents | Fail-safe mechanisms | Authorisation system |
| Detecting emergent behaviour | Distributed shift | Encryption |
| Graceful degradation | | Human override |

Theory (Modelling and understanding AI systems)
Deployment factors

All of the above (I/O, Control, Safety & Security)!

Plus: generality, capability, domains of application
## Landscape factors

### Number and identity of actors

<table>
<thead>
<tr>
<th>Accenture</th>
<th>:D Affectiva</th>
<th>Amazon</th>
<th>Apple</th>
<th>Baidu 百度</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cogitai</td>
<td></td>
<td></td>
<td></td>
<td>Element AI</td>
</tr>
<tr>
<td>Google</td>
<td>IBM</td>
<td>Intel</td>
<td>McKinsey &amp; Company</td>
<td>Microsoft</td>
</tr>
</tbody>
</table>

[Image of logos for each company mentioned]
Landscape factors

Inter-actor relationships
Landscape factors

International relations

Artificial Intelligence Strategies

- **March**: Pan-Canadian AI Strategy
- **May**: AI Singapore Announced
- **October**: AI Strategy 2031
- **December**: Finland’s AI Strategy
- **January**: Budget for AI Taiwan
- **March**: AI at the Service of Citizens
- **April**: First Workshop for Strategy
- **April**: UK AI Sector Deal
- **May**: White House Summit on AI
- **May**: Sweden’s AI Strategy
- **June**: Towards an AI Strategy in Mexico
- **Fall 2018**: EU’s AI Strategy

**2017**
- **March**: AI Technology Strategy
- **July**: Next Generation AI Plan
- **December**: Three-Year Action Plan
- **January**: Blockchain and AI Task Force
- **January**: Strategy for Digital Growth
- **March**: France’s AI Strategy
- **April**: Communication on AI
- **May**: Australian Budget
- **May**: AI R&D Strategy
- **June**: National Strategy for AI
- **Fall 2018**: Germany’s AI Strategy

**2018**
Landscape factors

Society and culture

The New York Times

Wielding Rocks and Knives, Arizonans Attack Self-Driving Cars
Landscape factors

The economy
Landscape factors

The environment
Landscape factors

Security
How do we explore and communicate these futures?

Scenario role-play