Exploring AGI Scenarios

Shahar Avin
sa478@cam.ac.uk
AGI strategy

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

1. *The Technological Singularity* by Murray Shanahan
2. *Superintelligence: Paths, Dangers, Strategies* by Nick Bostrom
3. *The Age of Em: Work, Love, and Life when Robots Rule the Earth* by Robin Hanson
4. *Life 3.0: Being Human in the Age of Artificial Intelligence* by Max Tegmark
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


http://www.decisionproblem.com/paperclips/
SPACECRAFT SCIENTIST/ENGINEER

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 on 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

- Specification (Define purpose of the system)
- Robustness (Design system to withstand perturbations)
- Assurance (Monitor and control system activity)

**Design**
- Bugs & inconsistencies
- Ambiguities
- Side-effects
- High-level specification languages
- Preference learning
- Design protocols

**Prevention and Risk**
- Risk sensitivity
- Uncertainty estimates
- Safety margins
- Safe exploration
- Cautious generalisation
- Verification
- Adversaries

**Emergent**
- Wireheading
- Delusions
- Malware and sub-agents
- Detecting emergent behaviour

**Recovery and Stability**
- Instability
- Error-correction
- Fail-safe mechanisms
- Distributional shift
- Graceful degradation

**Monitoring**
- Interpretability
- Behavioural screening
- Activity traces
- Estimates of causal influence
- Machine theory of mind
- Tripwires & honeypots

**Enforcement**
- Interruptibility
- Boxing
- Authorisation system
- Encryption
- 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

- Accenture
- 😊 Affectiva
- Amazon
- Apple
- Baidu

- Cogitai
- DeepMind
- Telekom
- ElementAI
- Facebook

- Google
- IBM
- Intel
- McKinsey & Company
- Microsoft
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

2018

- May: AI R&D Strategy
- June: National Strategy for AI
- Fall 2018: Germany's AI Strategy
Landscape factors

Society and culture

The New York Times

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

The economy

Automated for the people
Automation risk by job type, %

- Food preparation
- Construction
- Cleaning
- Driving
- Agricultural labour
- Garment manufacturing
- Personal service
- Sales
- Customer service
- Business administration
- Information technology
- Science & engineering
- Healthcare
- Hospitality & retail management
- Upper management & politics
- Teaching
- Source: OECD

Economist.com

Productivity growth and hourly compensation growth, 1948–2017

Alphabet Inc Class C
NASDAQ: GOOG

1,040.05 USD +4.44 (0.43%) ↑
Jan 2, 12:51 PM EST - Disclaimer

1 day 5 days 1 month 6 months YTD 1 year 5 years Max

1,037.08 USD Dec 28, 2018
Landscape factors

The environment

AI for Earth

AI for Earth is a Microsoft program aimed at empowering people and organizations to solve global environmental challenges by increasing access to AI tools and educational opportunities, while accelerating innovation.
Landscape factors

Security

---

Graph showing the Pareto Front and possible solutions in a 2-dimensional objective space. The graph indicates the infeasible region and the dominated solutions.
How do we explore and communicate these futures?

Scenario role-play