UNIVERSITY OF CAMBRIDGE

Exploring AGI Scenarios

Shahar Avin
sa478@cam.ac.uk

AGI strategy

This is Bob.

Bob heads an AGI R&D lab.

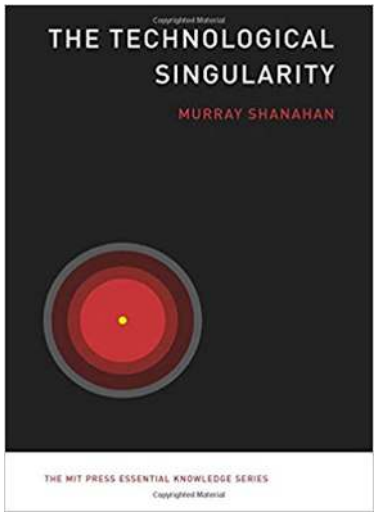
What should Bob do?

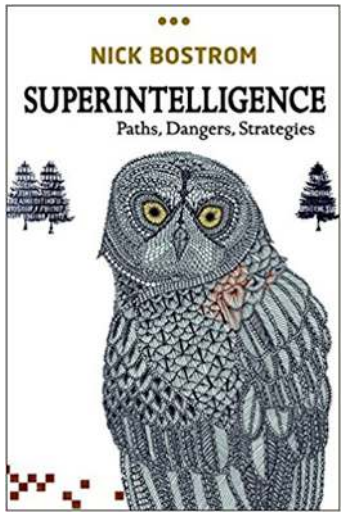
AGI futures narratives

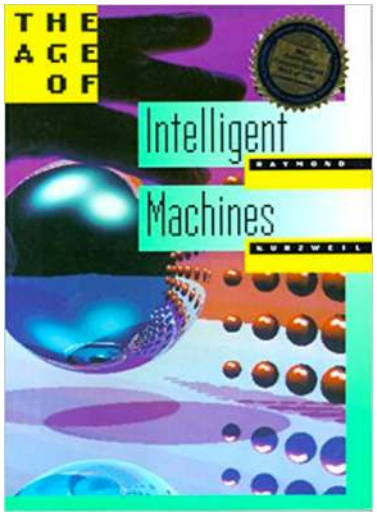
- Tech utopia
- Arms race
- Malicious use
- Existential risk

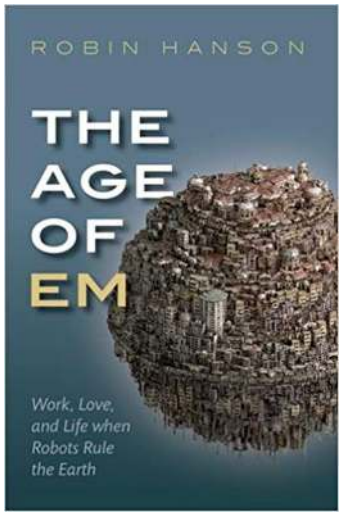
How do we explore and communicate these futures?

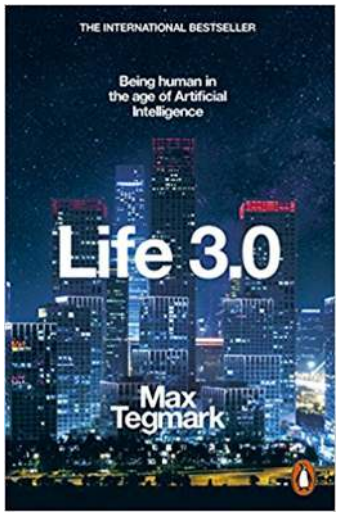
Single author exploration





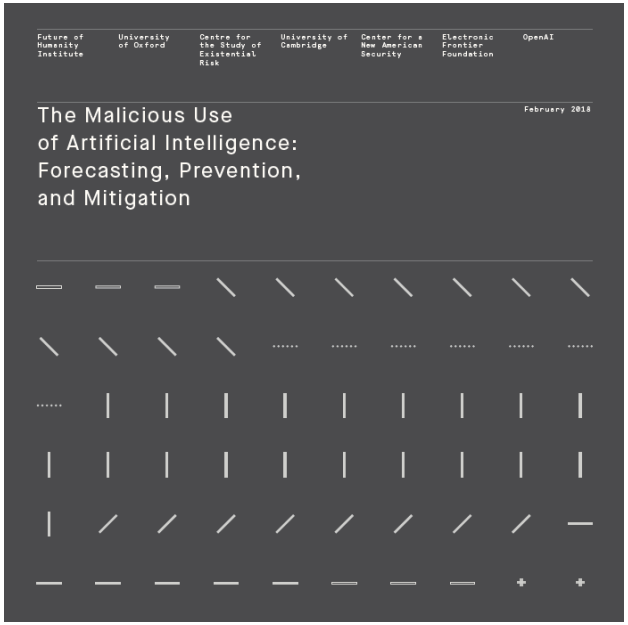






How do we explore and communicate these futures?

Expert workshops, multi-authored reports

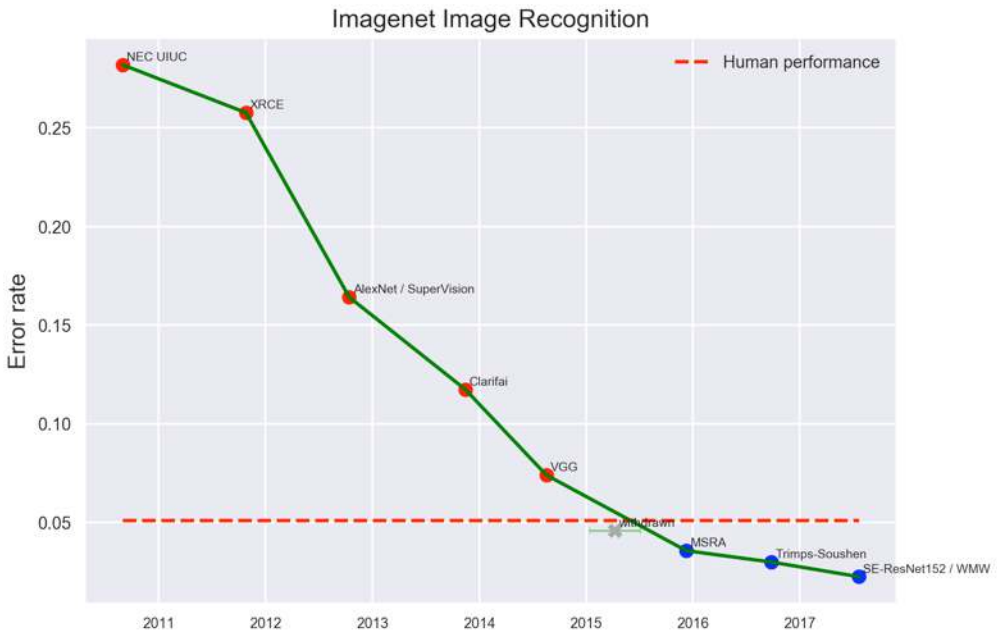


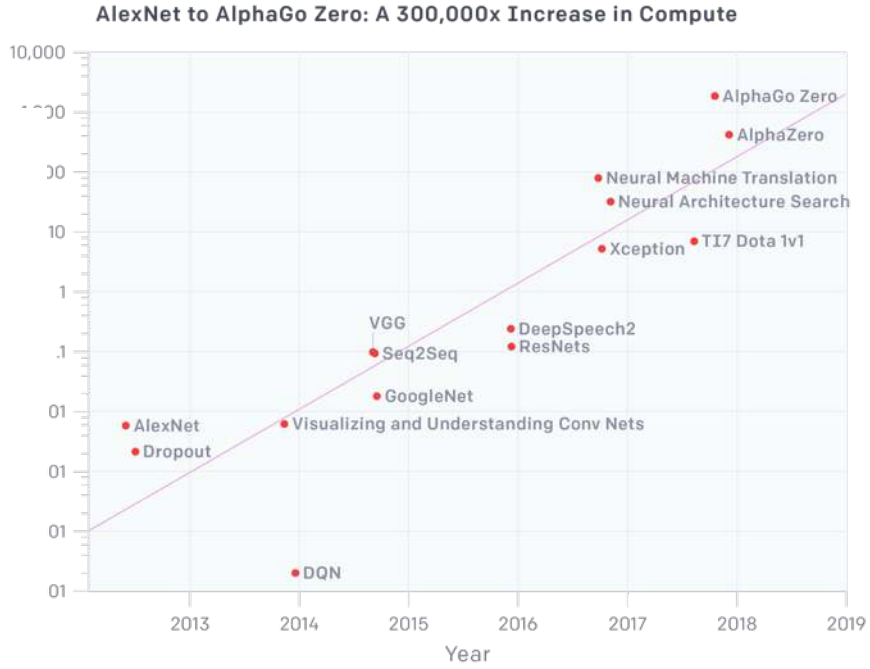
http://maliciousaireport.com/

How do we explore and communicate these futures?

Data trends

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



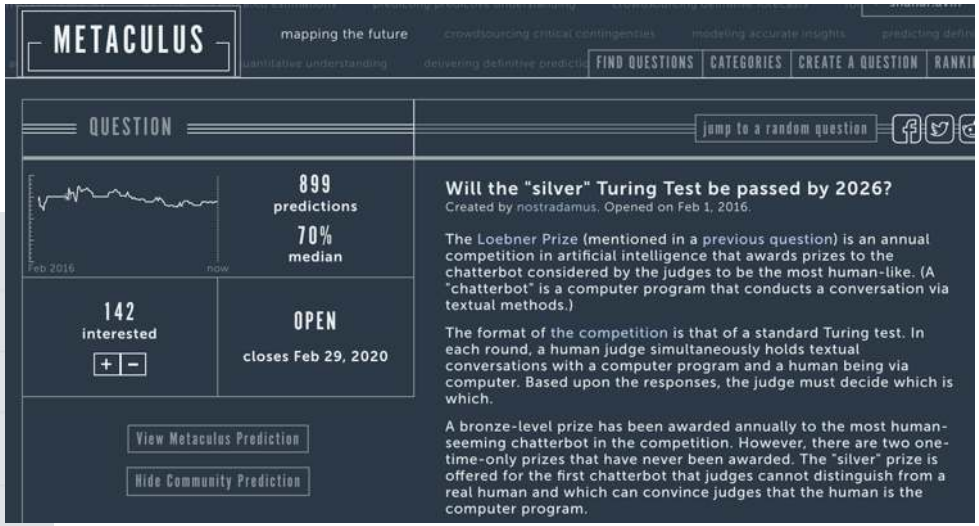


https://blog.openai.com/

How do we explore and communicate these futures?

Aggregate probability estimates

https://www.getguesstimate.com

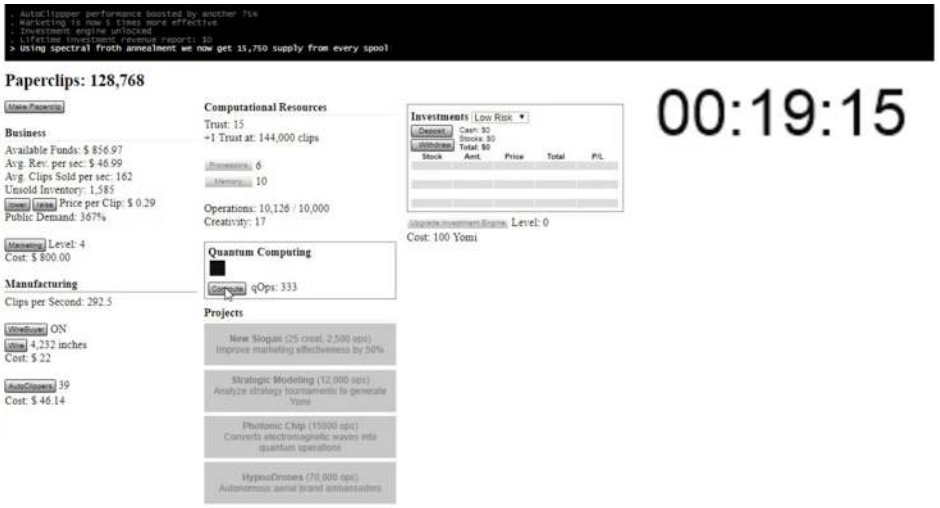


https://www.metaculus.com

How do we explore and communicate these futures?

Video games

https://bit.ly/2uqXNUn

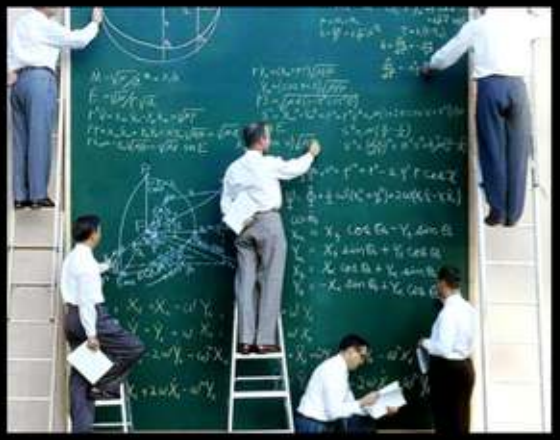


http://www.decisionproblem.com/paperclips/

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 factors

Inputs

Data use policy

Last updated: 23 September 2011

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.

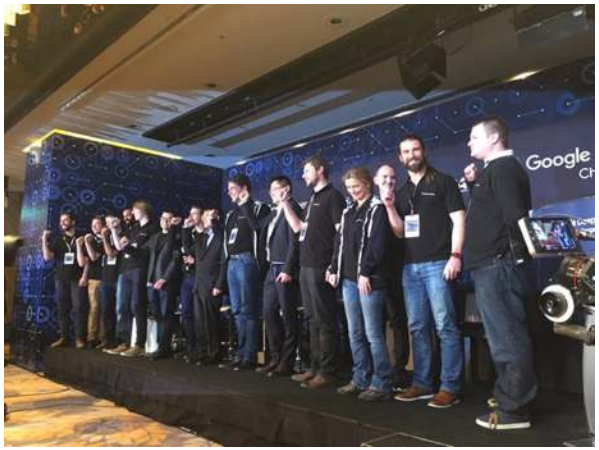
Sharing with other websites and applications
Find out about the ways your information is shared with the games, applications and websites you and your friends use off Facebook.

If you have questions or complaints regarding our privacy policy or practices, please contact us by post at 1601 Willow Road, Menlo Park, CA 94025 or through this help page.

More resources

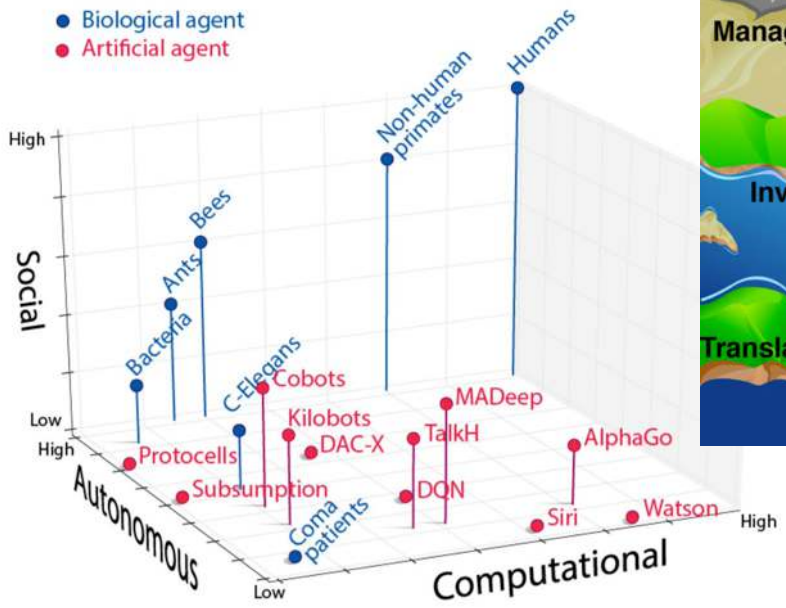
Interactive tools

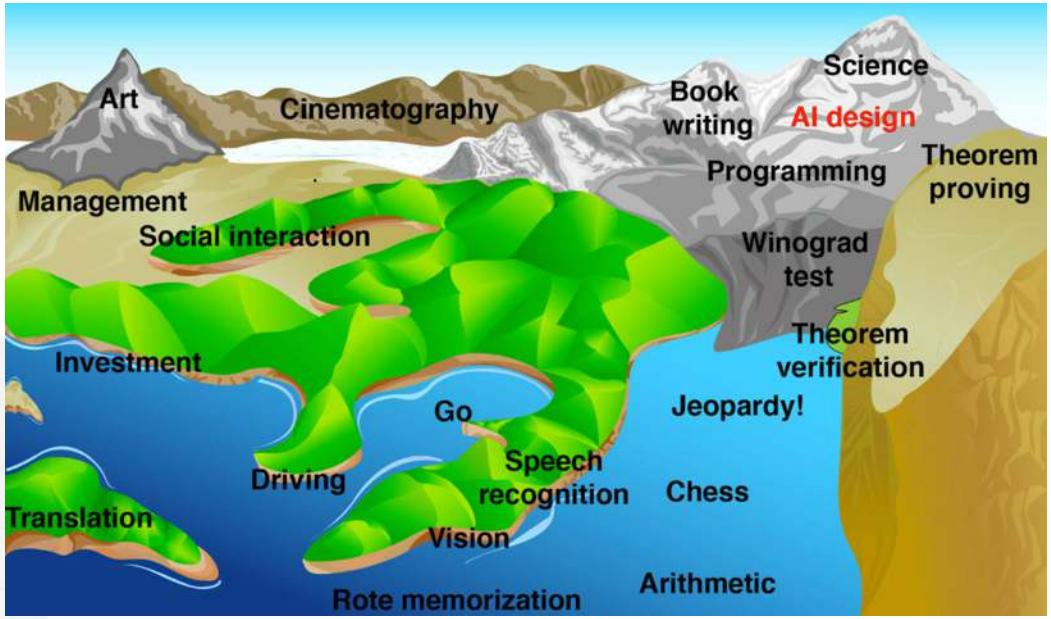
View the complete Data Use Policy



Development factors

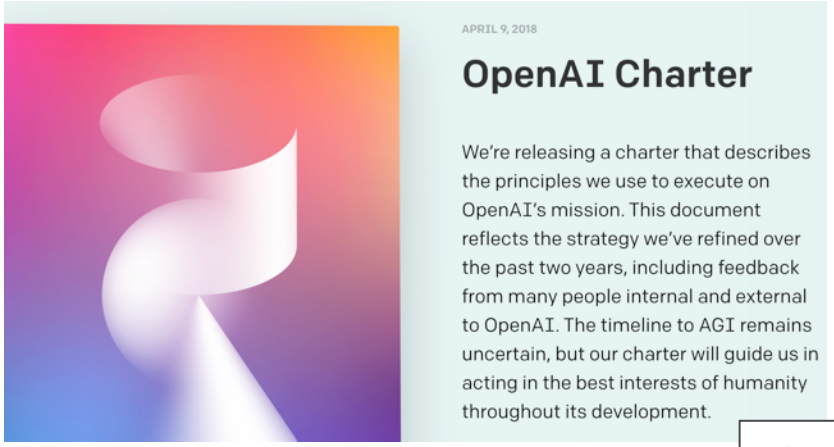
Nature of the problem





Development factors

Control, incentives, openness



APRIL 9, 2018

OpenAI Charter

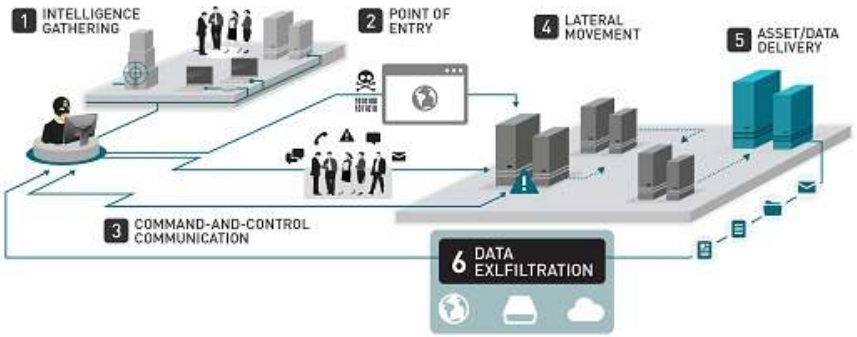
We're releasing a charter that describes the principles we use to execute on OpenAI's mission. This document reflects the strategy we've refined over the past two years, including feedback from many people internal and external to OpenAI. The timeline to AGI remains uncertain, but our charter will guide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

DARPA

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	**Prevention and Risk**	**Monitoring**
Bugs & inconsistencies Ambiguities Side-effects High-level specification languages Preference learning Design protocols	Risk sensitivity Uncertainty estimates Safety margins Safe exploration Cautious generalisation Verification Adversaries	Interpretability Behavioural screening Activity traces Estimates of causal influence Machine theory of mind Tripwires & honeypots
Emergent	**Recovery and Stability**	**Enforcement**
Wireheading Delusions Metalearning and sub-agents Detecting emergent behaviour	Instability Error-correction Failsafe mechanisms Distributional shift Graceful degradation	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

SIGHT

Cloud Vision API
Image recognition and classification.

Cloud Video Intelligence API
Scene-level video annotation.

AutoML Vision BETA
Custom image classification models.

LANGUAGE

Cloud Translation API
Language detection and translation.

Cloud Natural Language API
Text parsing and analysis.

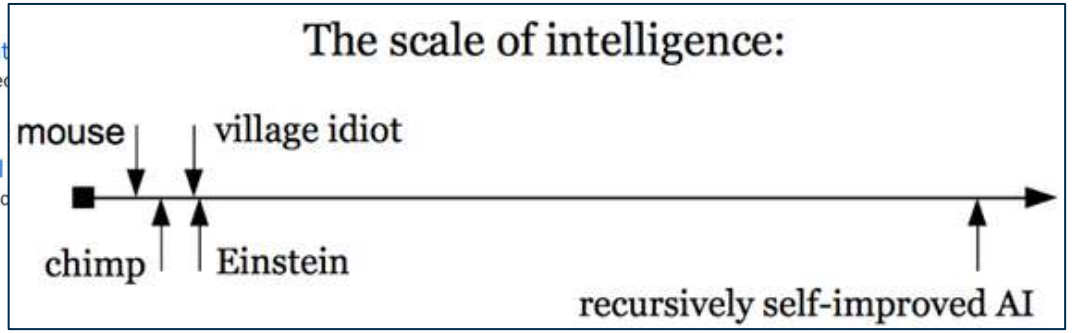
AutoML Translat
Custom domain-spec

AutoML Natural
Custom text classific

CONVERSATION

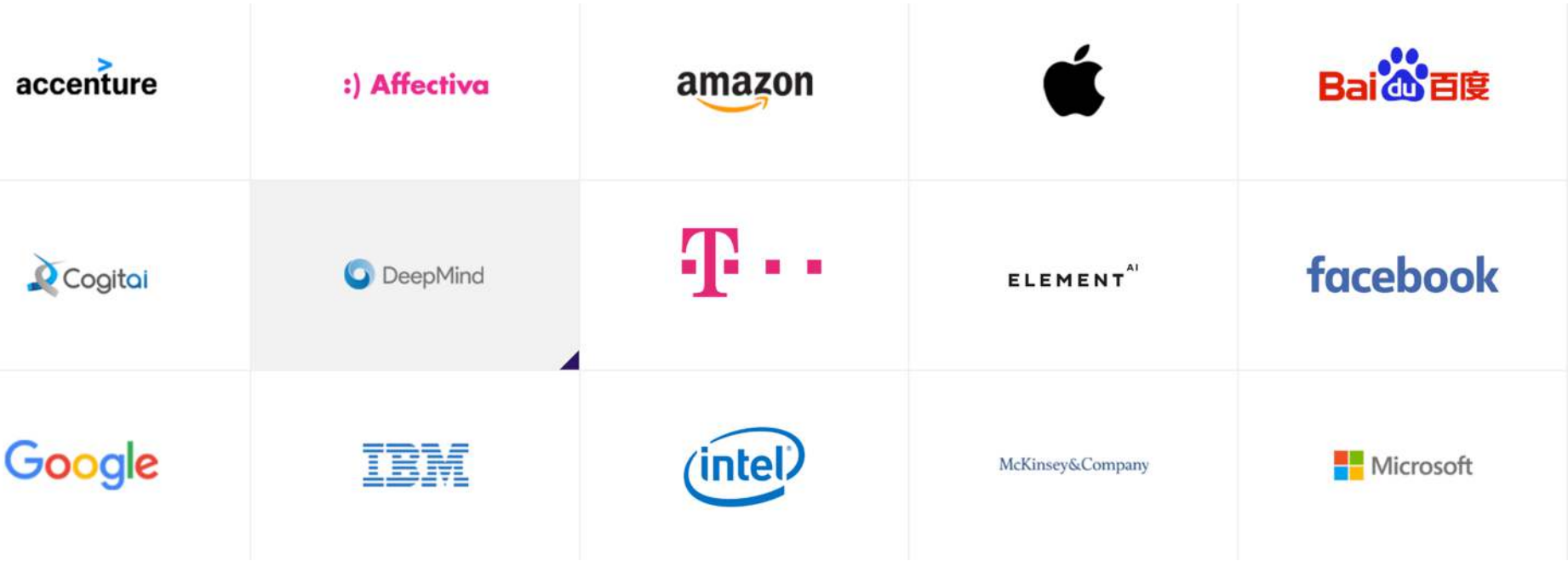
Dialogflow Enterprise Edition
Build conversational interfaces.

Cloud Text-to-Speech API
Convert text to speech.



Landscape factors

Number and identity of actors

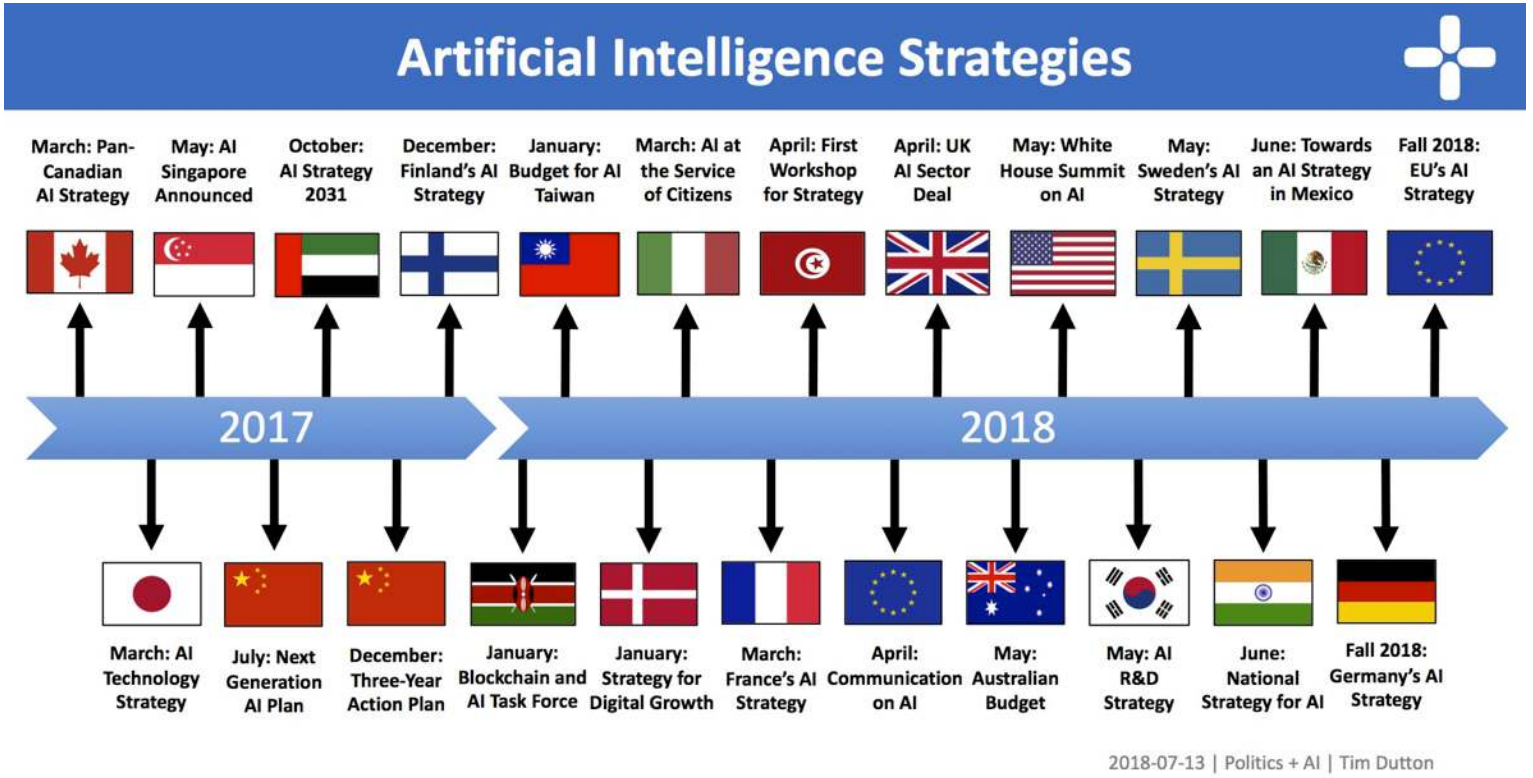


Landscape factors

Inter-actor relationships

Landscape factors

International relations

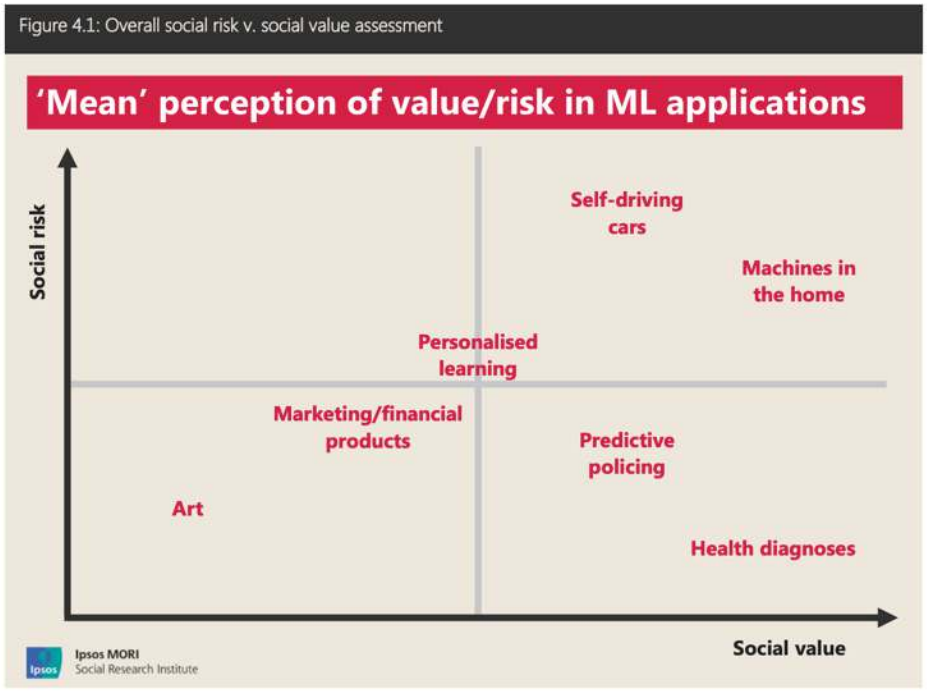


Landscape factors

Society and culture

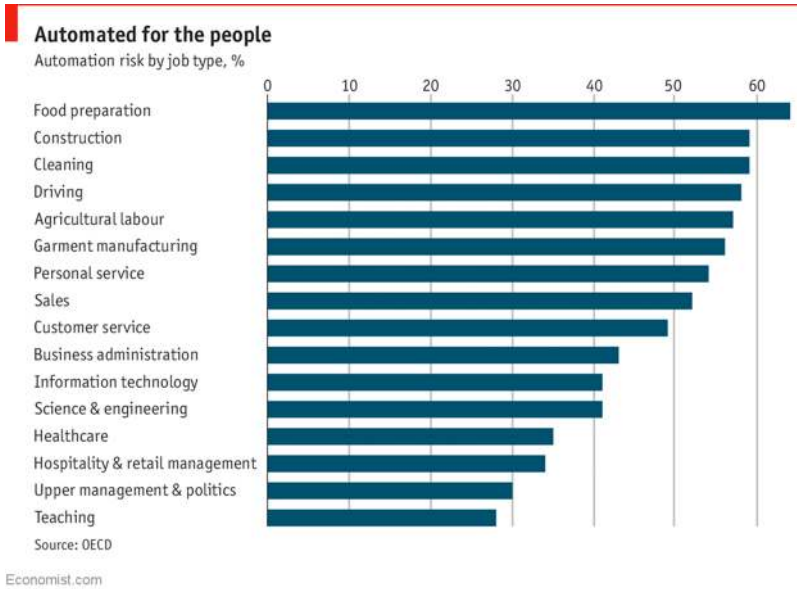
The New York Times

Wielding Rocks and Knives, Arizonans Attack Self-Driving Cars



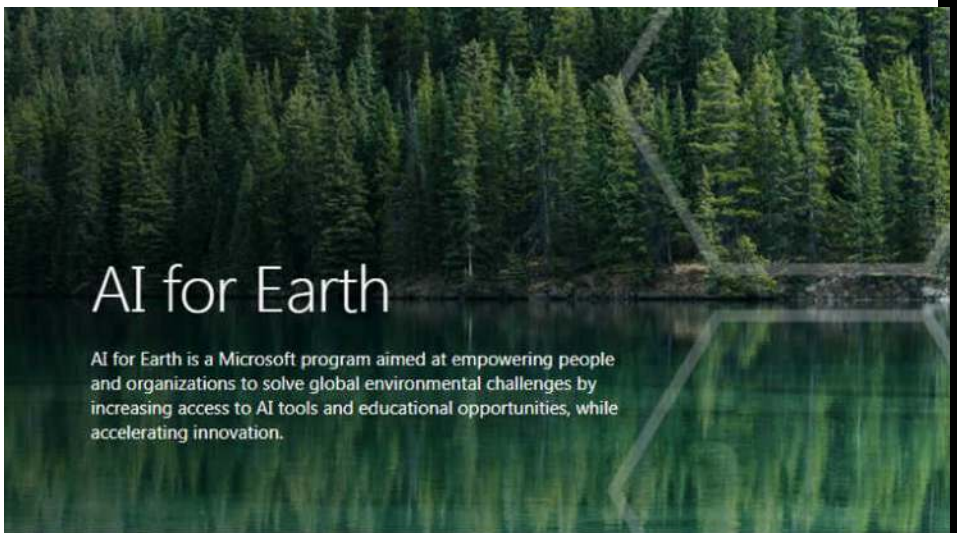
Landscape factors

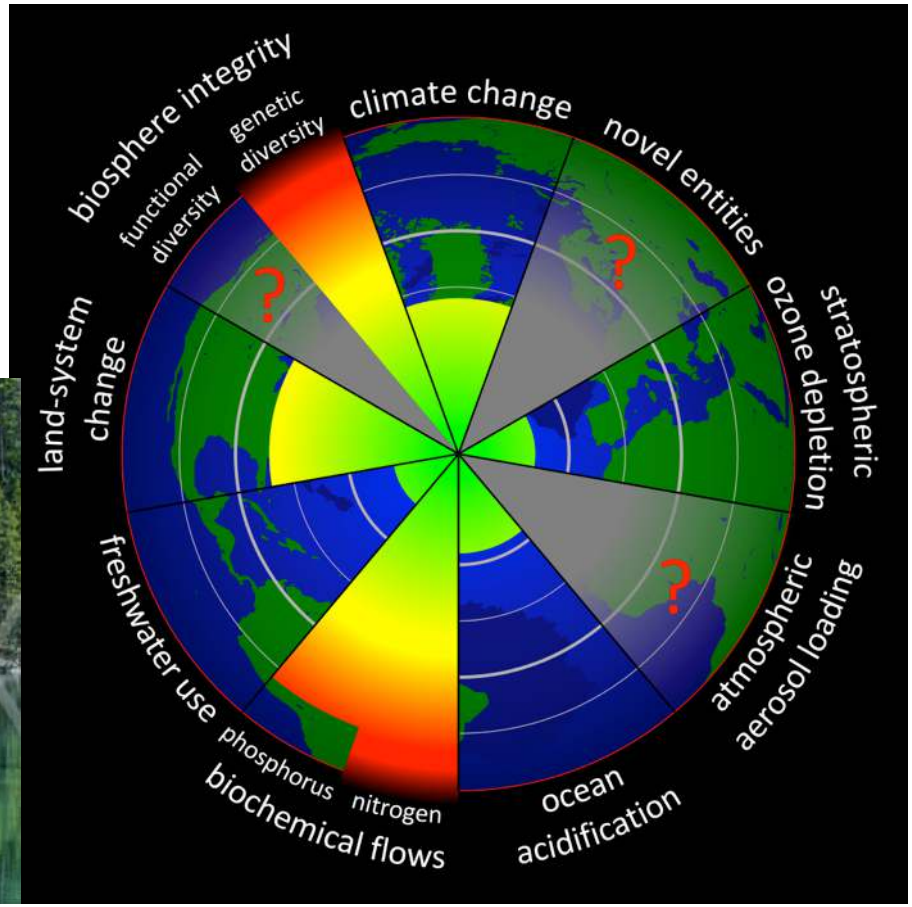
The economy



Landscape factors

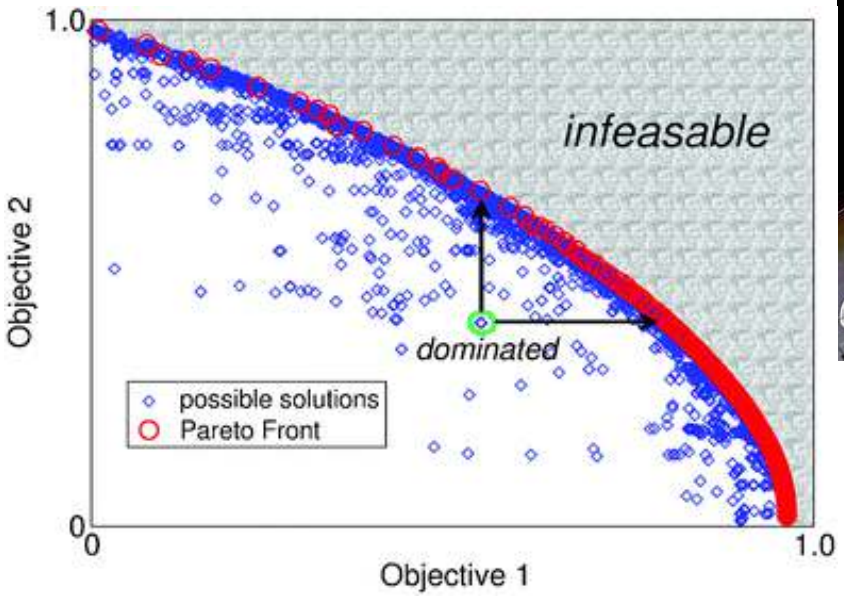
The environment

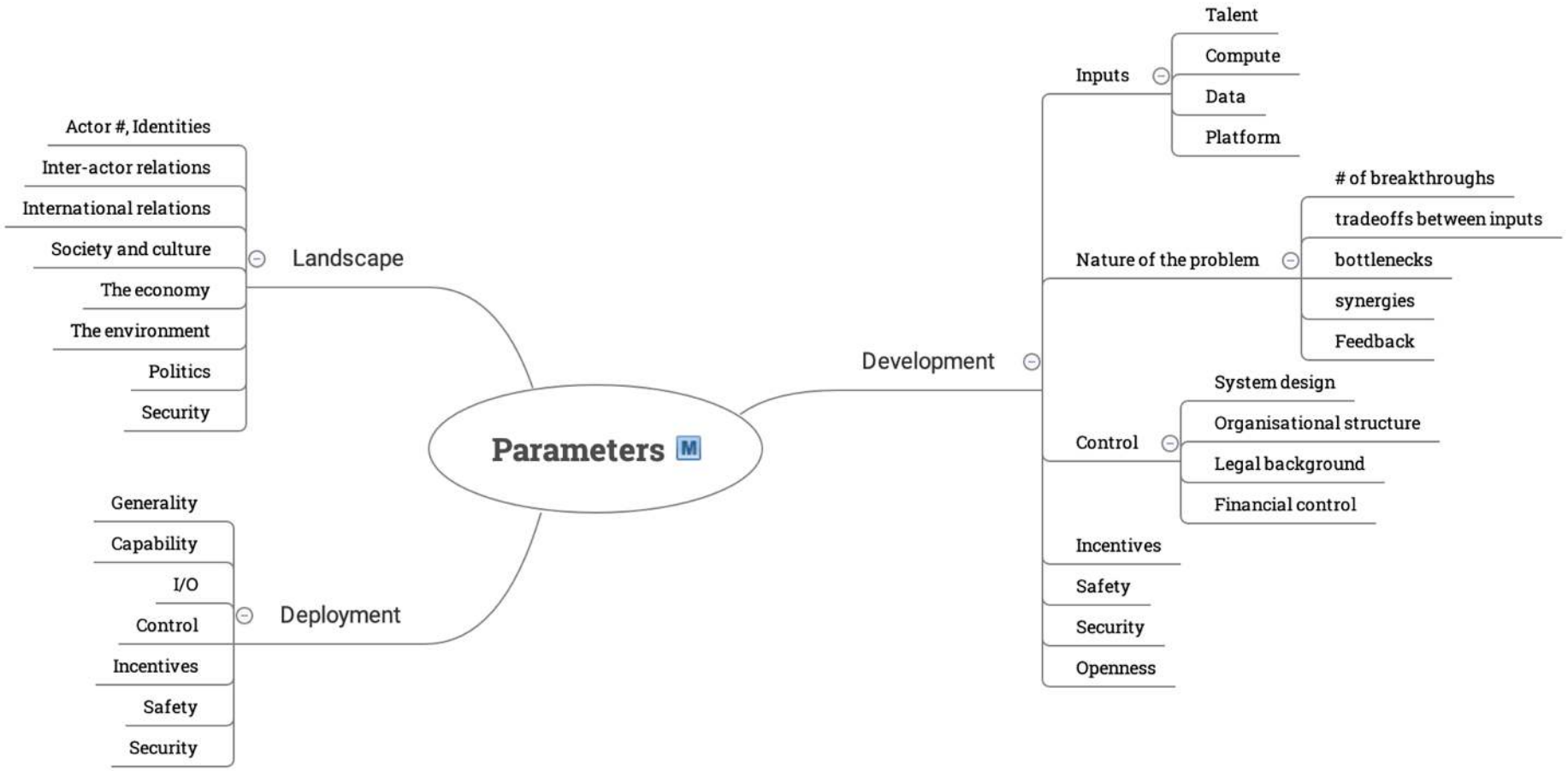




Landscape factors

Security




Parameters
Landscape
Actor #, Identities
Inter-actor relations
International relations
Society and culture
The economy
The environment
Politics
Security
Deployment
Generality
Capability
I/O
Control
Incentives
Safety
Security
Development
Inputs
Talent
Compute
Data
Platform
Nature of the problem
# of breakthroughs
tradeoffs between inputs
bottlenecks
synergies
Feedback
Control
System design
Organisational structure
Legal background
Financial control
Incentives
Safety
Security
Openness

How do we explore and communicate these futures?

Scenario role-play