UNIVERSITY OF
CAMBRIDGE

Exploring AGI Scenarios

Shahar Avin
sa478@cam.ac.uk

AGI strategy

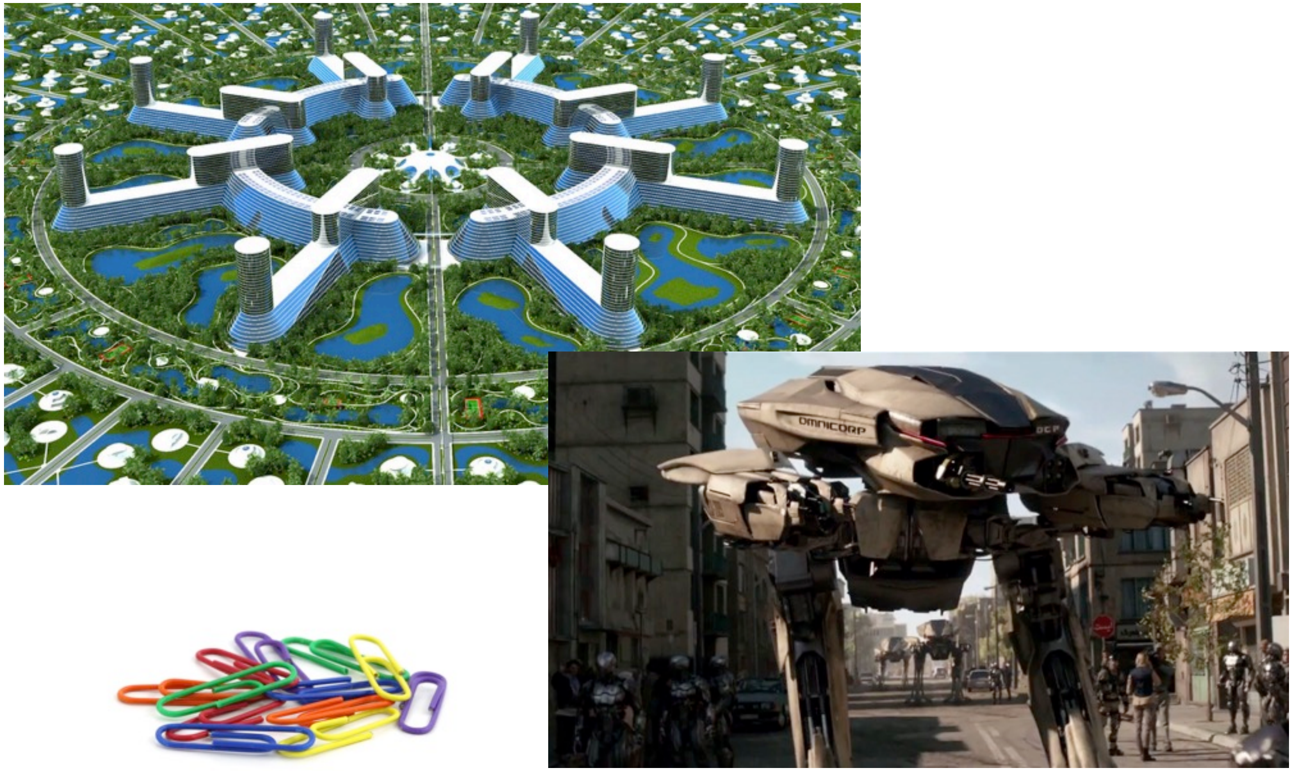
This is Bob.

Bob heads an AGI R&D lab.

What should Bob do?

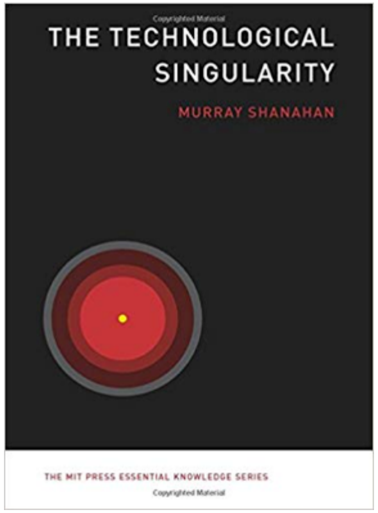
AGI futures narratives

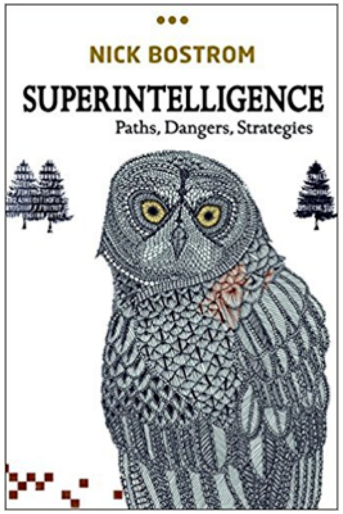
- Tech utopia
- Arms race
- Malicious use
- Existential risk

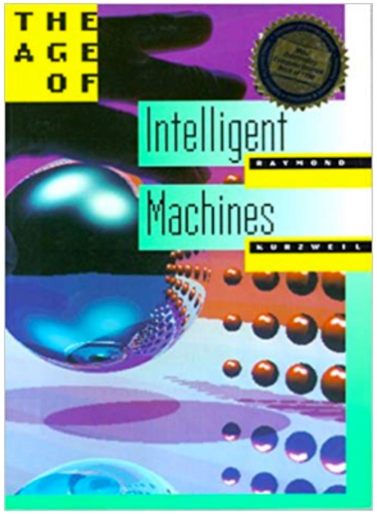


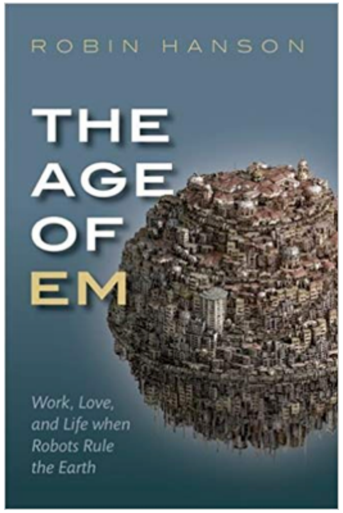
How do we explore and communicate these futures?

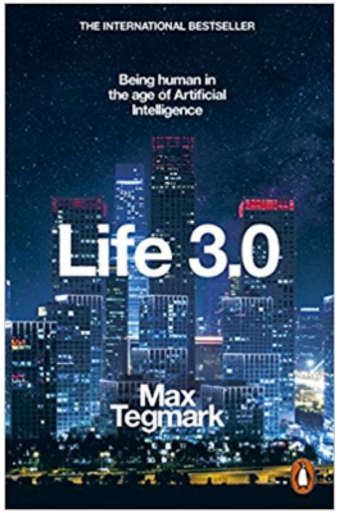
Single author exploration


Copyrighted Material
THE TECHNOLOGICAL SINGULARITY
MURRAY SHANAHAN
THE MIT PRESS ESSENTIAL KNOWLEDGE SERIES
Copyrighted Material



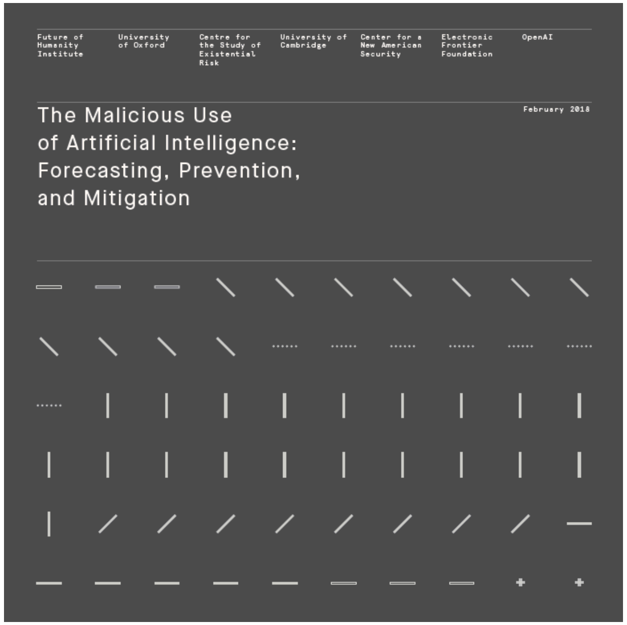





THE INTERNATIONAL BESTSELLER
Being human in the age of Artificial Intelligence
Life 3.0
Max Tegmark

How do we explore and communicate these futures?

Expert workshops, multi-authored reports



Future of Humanity Institute

University of Oxford

Centre for the Study of Existential Risk

University of Cambridge

Center for a New American Security

Electronic Frontier Foundation

OpenAI

February 2018

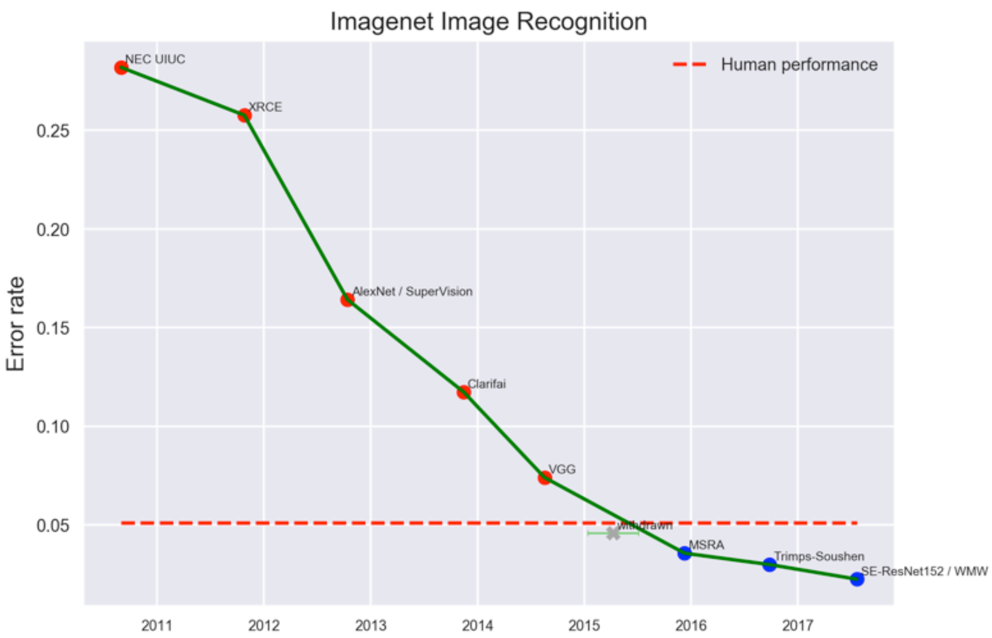
The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation

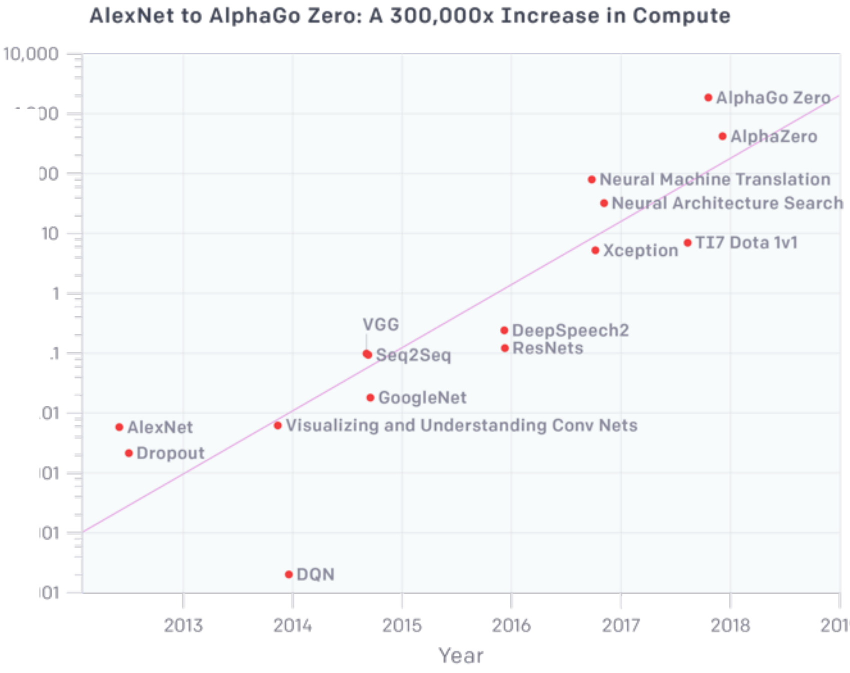
http://maliciousaireport.com/

How do we explore and communicate these futures?

Data trends

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



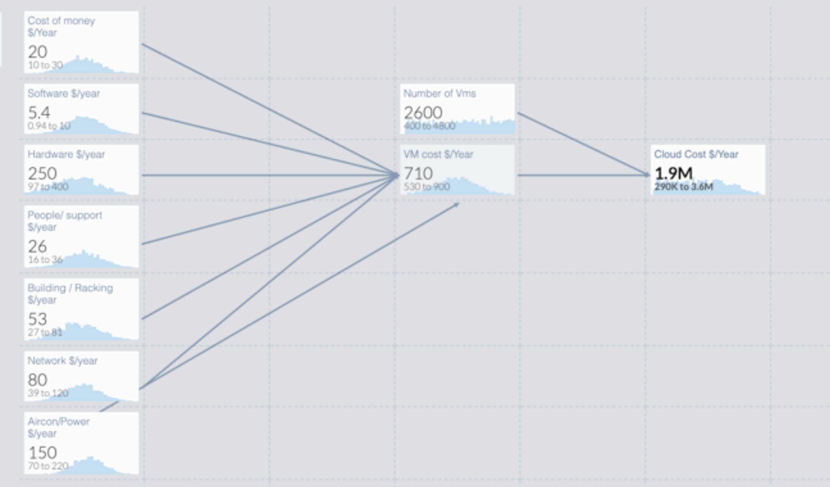


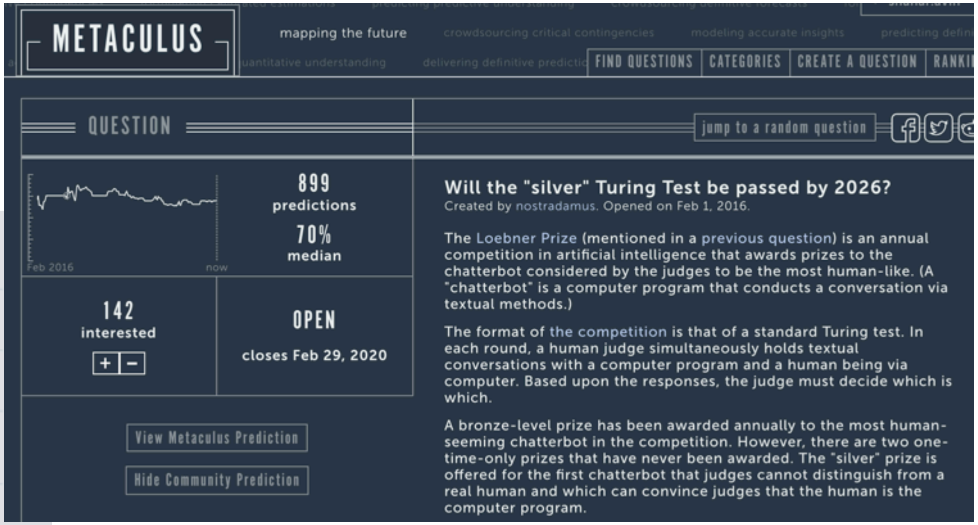
https://blog.openai.com/

How do we explore and communicate these futures?

Aggregate probability estimates

https://www.getguesstimate.com





METACULUS
mapping the future
crowdsourcing critical contingencies
modeling accurate insights
uantitative understanding
delivering definitive predictic
FIND QUESTIONS
CATEGORIES
CREATE A QUESTION
RANKI
QUESTION
jump to a random question
Feb 2016
now
+
−
View Metaculus Prediction
Hide Community Prediction

899 predictions

70% median

142 interested

OPEN

closes Feb 29, 2020

Will the "silver" Turing Test be passed by 2026?

Created by nostradamus. Opened on Feb 1, 2016.

The Loebner Prize (mentioned in a previous question) is an annual competition in artificial intelligence that awards prizes to the chatterbot considered by the judges to be the most human-like. (A "chatterbot" is a computer program that conducts a conversation via textual methods.)

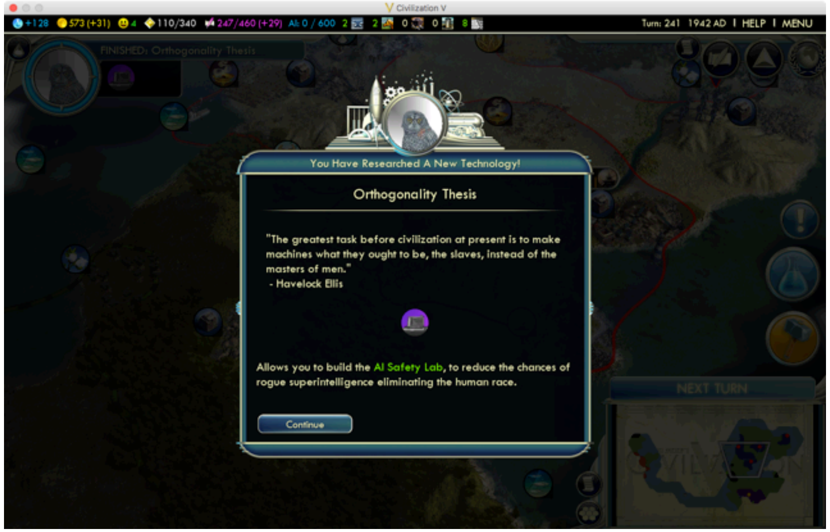
The format of the competition is that of a standard Turing test. In each round, a human judge simultaneously holds textual conversations with a computer program and a human being via computer. Based upon the responses, the judge must decide which is which.

A bronze-level prize has been awarded annually to the most human-seeming chatterbot in the competition. However, there are two one-time-only prizes that have never been awarded. The "silver" prize is offered for the first chatterbot that judges cannot distinguish from a real human and which can convince judges that the human is the computer program.

https://www.metaculus.com

How do we explore and communicate these futures?

Video games


Civilization V
+128 573 (+31) 4 110/340 247/460 (+29) AI: 0 / 600 2 2 0 0 8
Turn: 241 1942 AD | HELP | MENU
FINISHED: Orthogonality Thesis
Continue
NEXT TURN

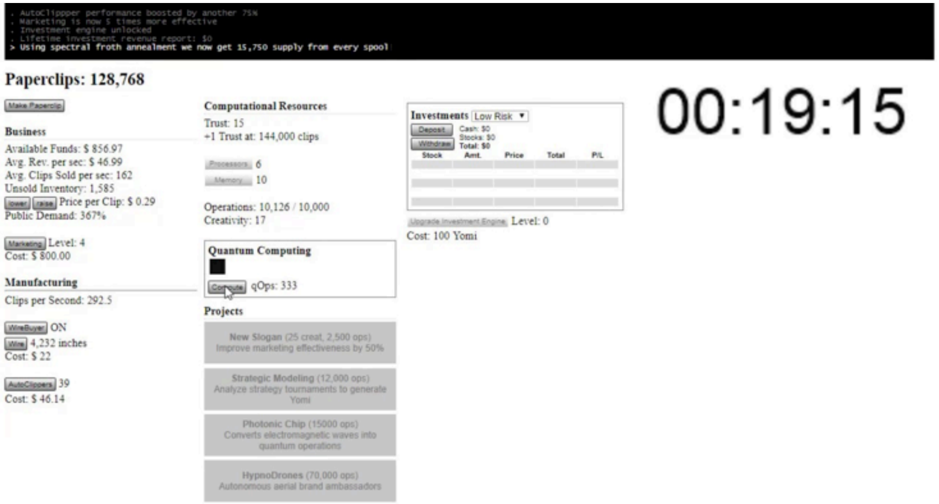
You Have Researched A New Technology!

Orthogonality Thesis

"The greatest task before civilization at present is to make machines what they ought to be, the slaves, instead of the masters of men."
- Havelock Ellis

Allows you to build the AI Safety Lab, to reduce the chances of rogue superintelligence eliminating the human race.

https://bit.ly/2uqXNUn



```
. AutoClippper performance boosted by another 75%
. Marketing is now 5 times more effective
. Investment engine unlocked
. Lifetime investment revenue report: $0
> Using spectral froth annealment we now get 15,750 supply from every spool
```

Paperclips: 128,768

00:19:15

Business

Available Funds: $ 856.97
Avg. Rev. per sec: $ 46.99
Avg. Clips Sold per sec: 162
Unsold Inventory: 1,585
Price per Clip: $ 0.29
Public Demand: 367%

Level: 4
Cost: $ 800.00

Manufacturing

Clips per Second: 292.5

ON
4,232 inches
Cost: $ 22

39
Cost: $ 46.14

Computational Resources

Trust: 15
+1 Trust at: 144,000 clips

Processors 6
Memory 10

Operations: 10,126 / 10,000
Creativity: 17

Quantum Computing

qOps: 333

Projects

New Slogan (25 creat, 2,500 ops)
Improve marketing effectiveness by 50%

Strategic Modeling (12,000 ops)
Analyze strategy tournaments to generate Yomi

Photonic Chip (15000 ops)
Converts electromagnetic waves into quantum operations

HypnoDrones (70,000 ops)
Autonomous aerial brand ambassadors

Investments Low Risk ▾

Cash: $0
Stocks: $0
Total: $0

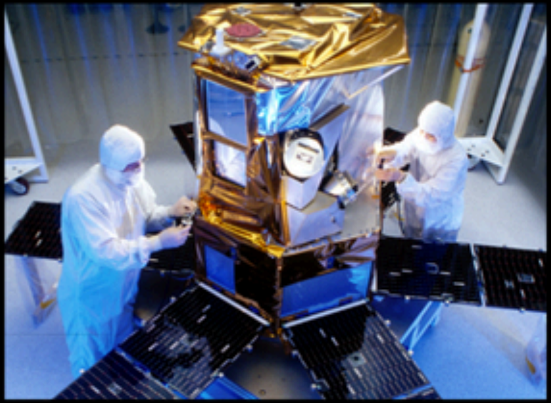
Stock	Amt.	Price	Total	P/L

Level: 0
Cost: 100 Yomi

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

SPACECRAFT SCIENTIST/ENGINEER

What my friends think I do



What my parents think I do

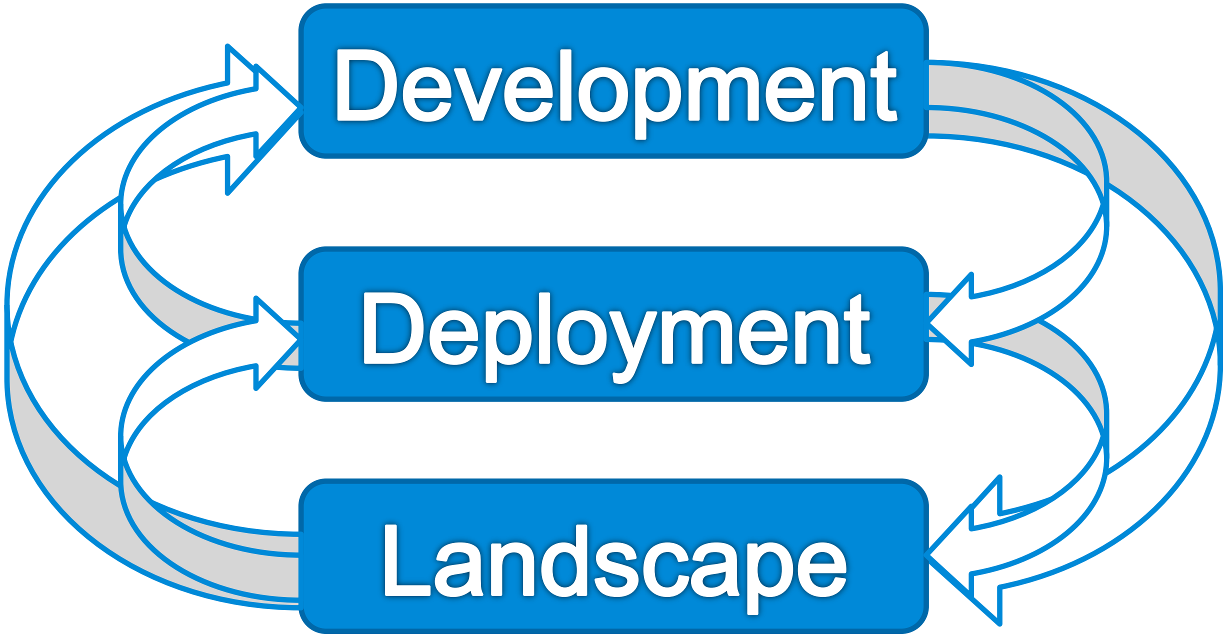
What society thinks I do

What my boss thinks I do

What I think I do

What I really do

What should we be looking at?



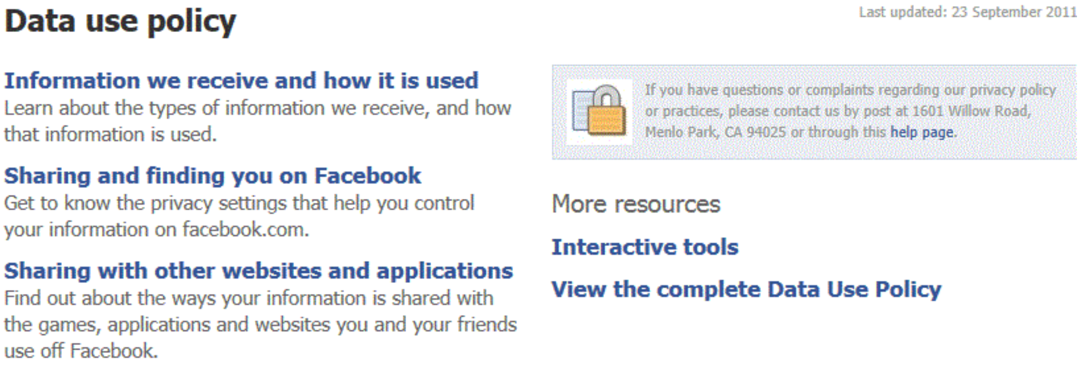
Development

Deployment

Landscape

Development factors

Inputs



Data use policy

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.

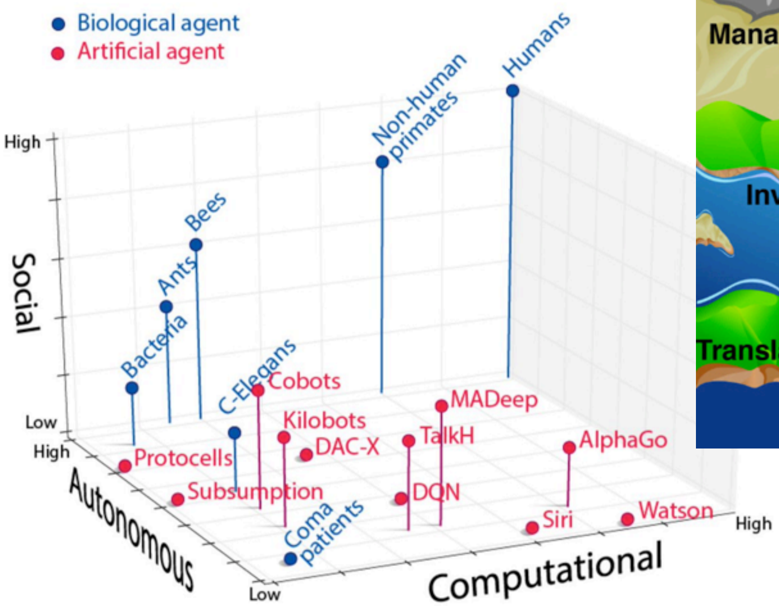
More resources

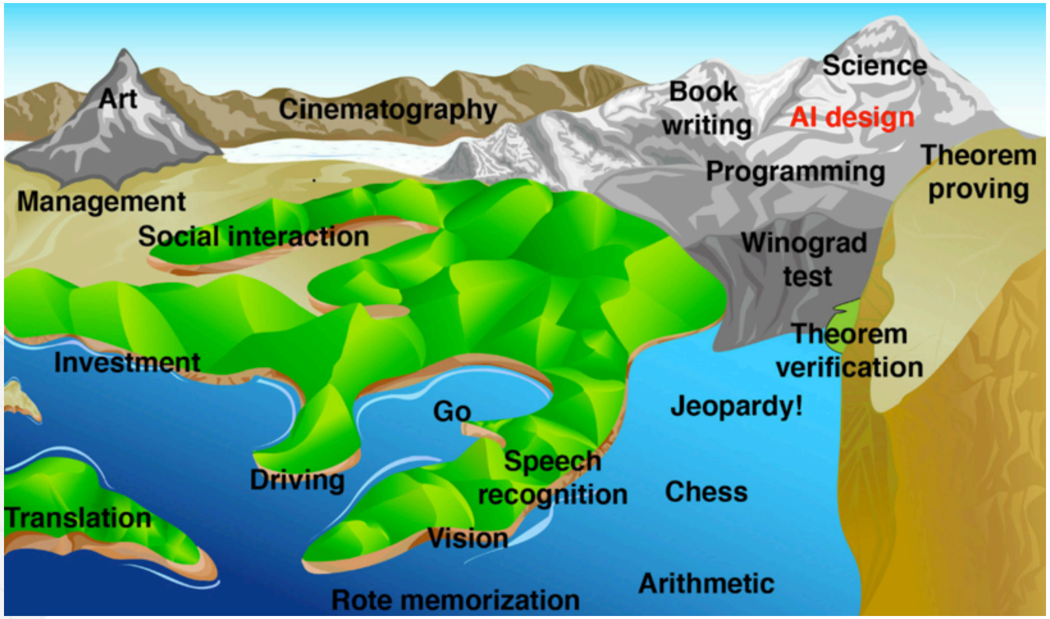
Interactive tools

View the complete Data Use Policy

Development factors

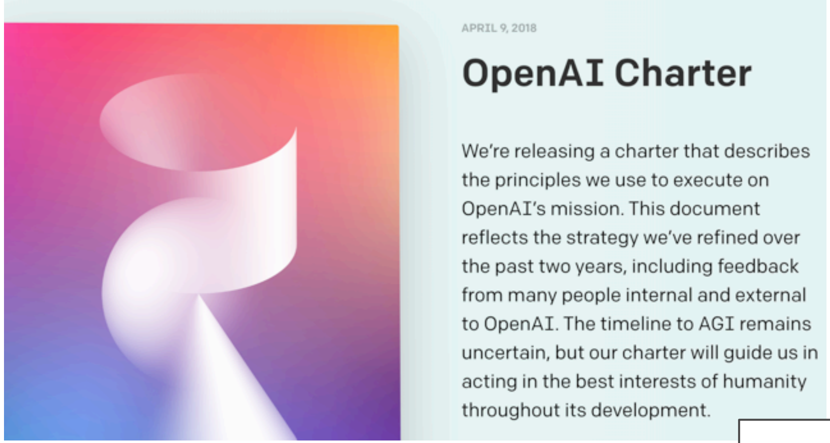
Nature of the problem





Development factors

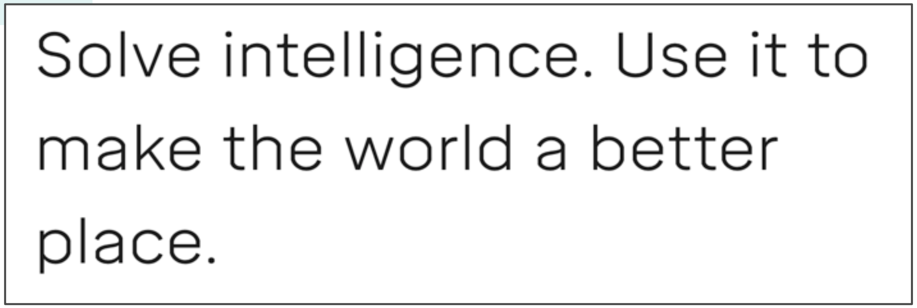
Control, incentives, openness



APRIL 9, 2018

OpenAI Charter

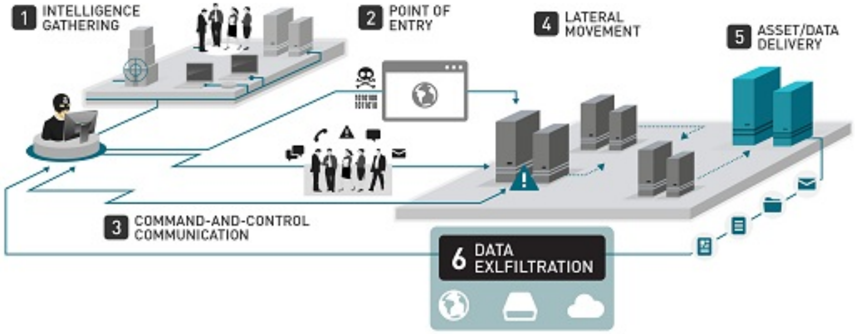
We're releasing a charter that describes the principles we use to execute on OpenAI's mission. This document reflects the strategy we've refined over the past two years, including feedback from many people internal and external to OpenAI. The timeline to AGI remains uncertain, but our charter will guide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

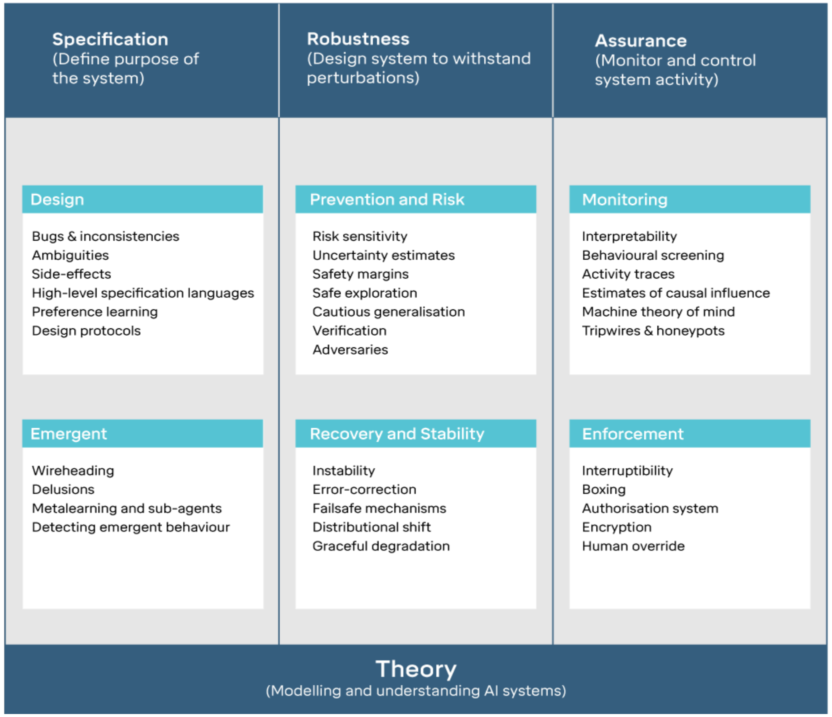


Solve intelligence. Use it to make the world a better place.

Development factors

Safety and Security



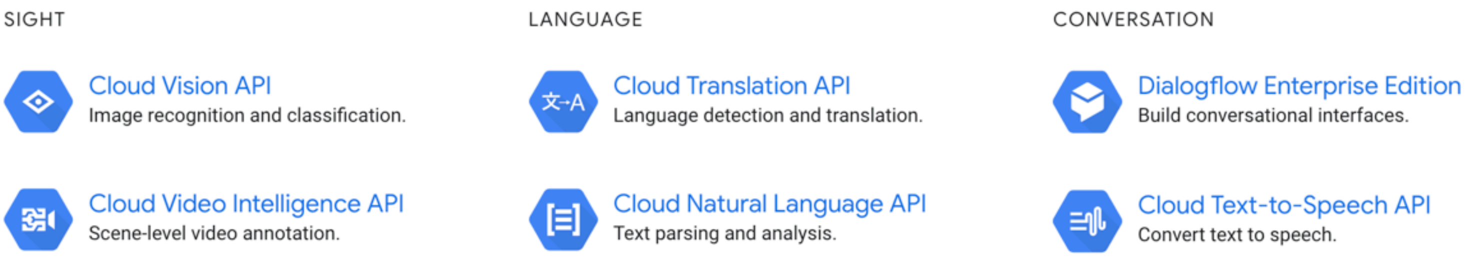


Specification (Define purpose of the system)	**Robustness** (Design system to withstand perturbations)	**Assurance** (Monitor and control system activity)
Design	**Prevention and Risk**	**Monitoring**
Bugs & inconsistencies Ambiguities Side-effects High-level specification languages Preference learning 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.

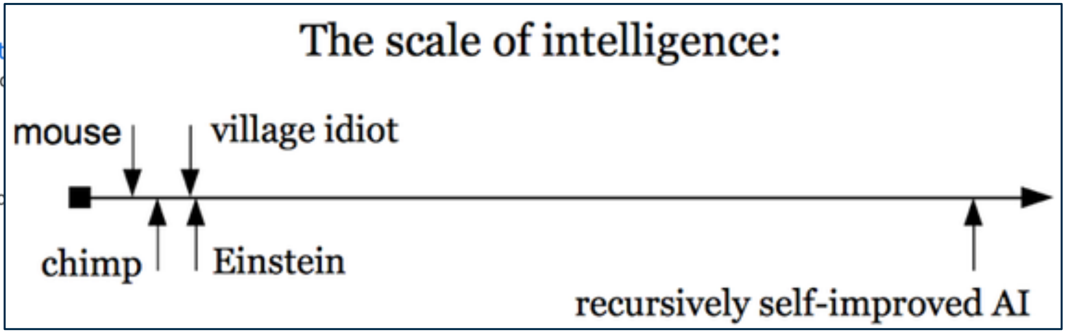
AutoML Translat
Custom domain-spec

AutoML Natural
Custom text classific

CONVERSATION

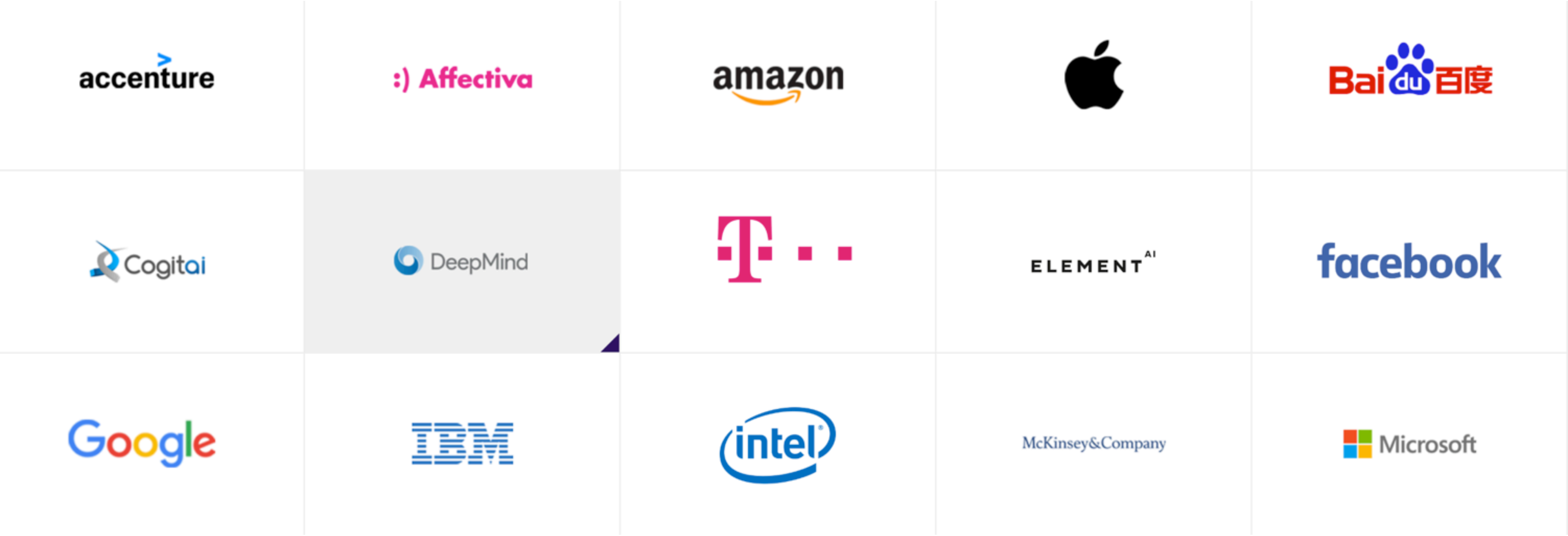
Dialogflow Enterprise Edition
Build conversational interfaces.

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



Landscape factors

Number and identity of actors

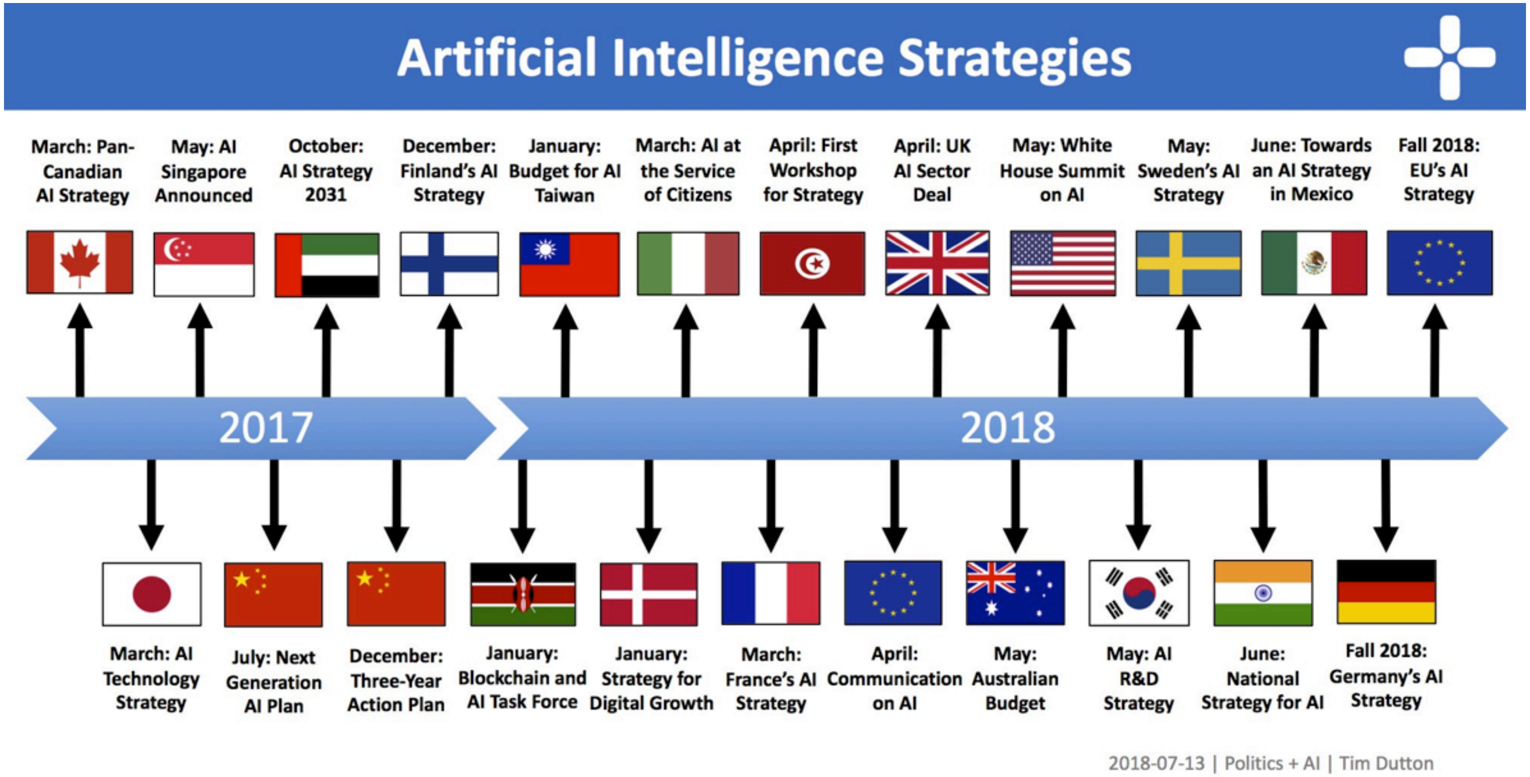


Landscape factors

Inter-actor relationships

Landscape factors

International relations



Artificial Intelligence Strategies

2017

2018

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

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

2018-07-13 | Politics + AI | Tim Dutton

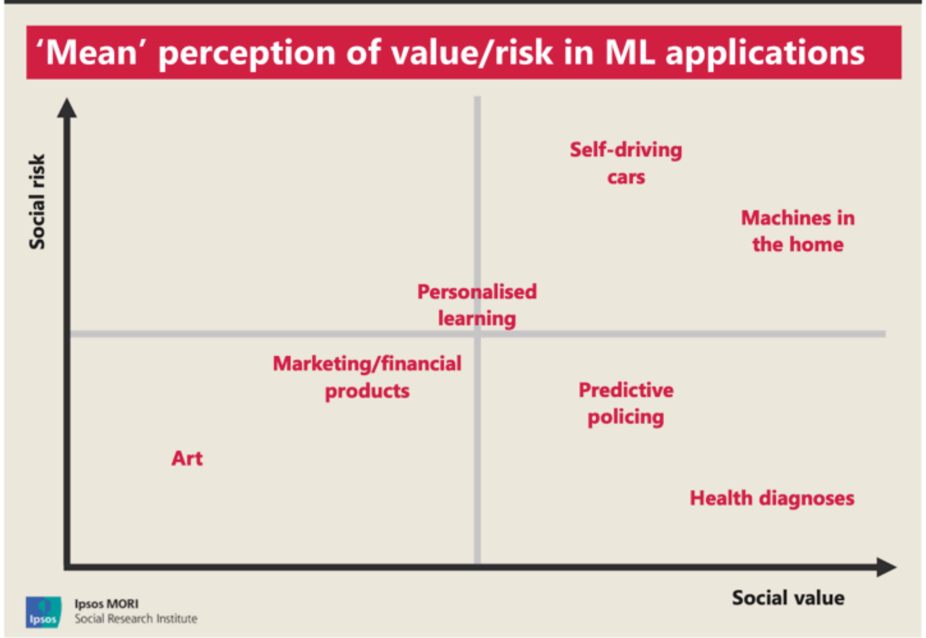
Landscape factors

Society and culture

The New York Times

Wielding Rocks and Knives, Arizonans Attack Self-Driving Cars

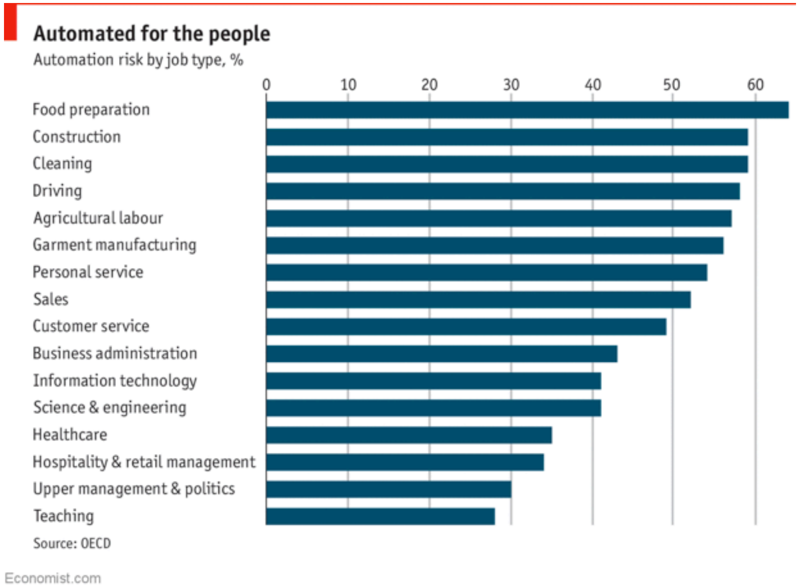
Figure 4.1: Overall social risk v. social value assessment



Ipsos MORI
Social Research Institute

Landscape factors

The economy



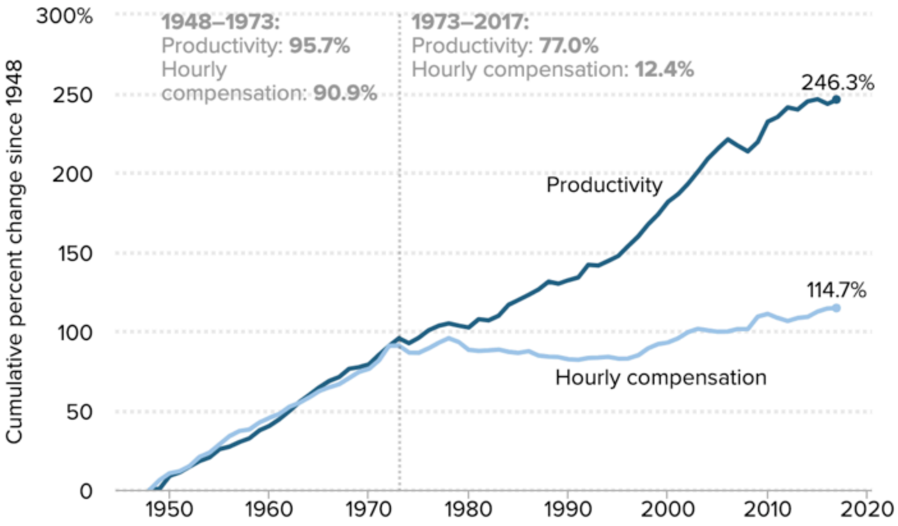
Automated for the people

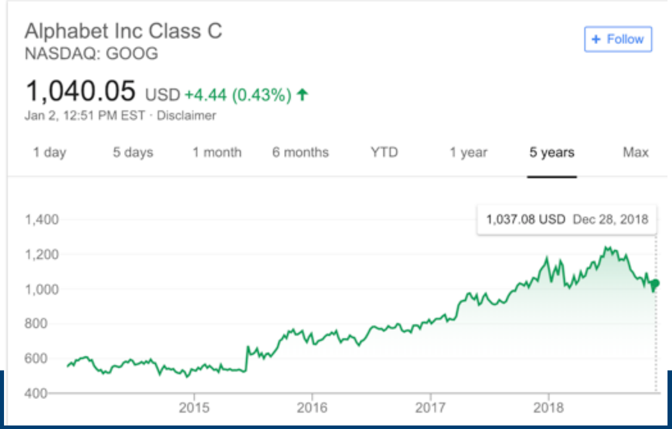
Automation risk by job type, %

Source: OECD

Economist.com

Productivity growth and hourly compensation growth, 1948–2017





Alphabet Inc Class C

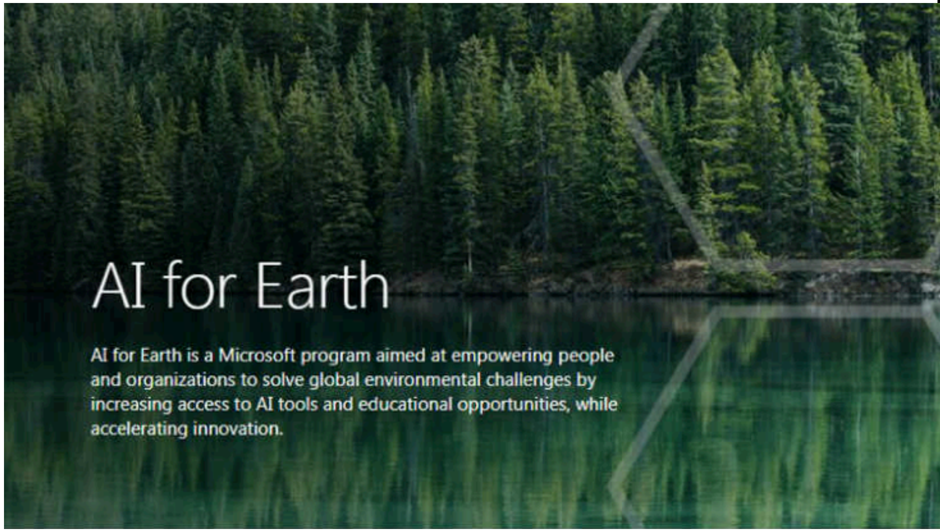
NASDAQ: GOOG

1,040.05 USD +4.44 (0.43%) ↑

Jan 2, 12:51 PM EST · Disclaimer

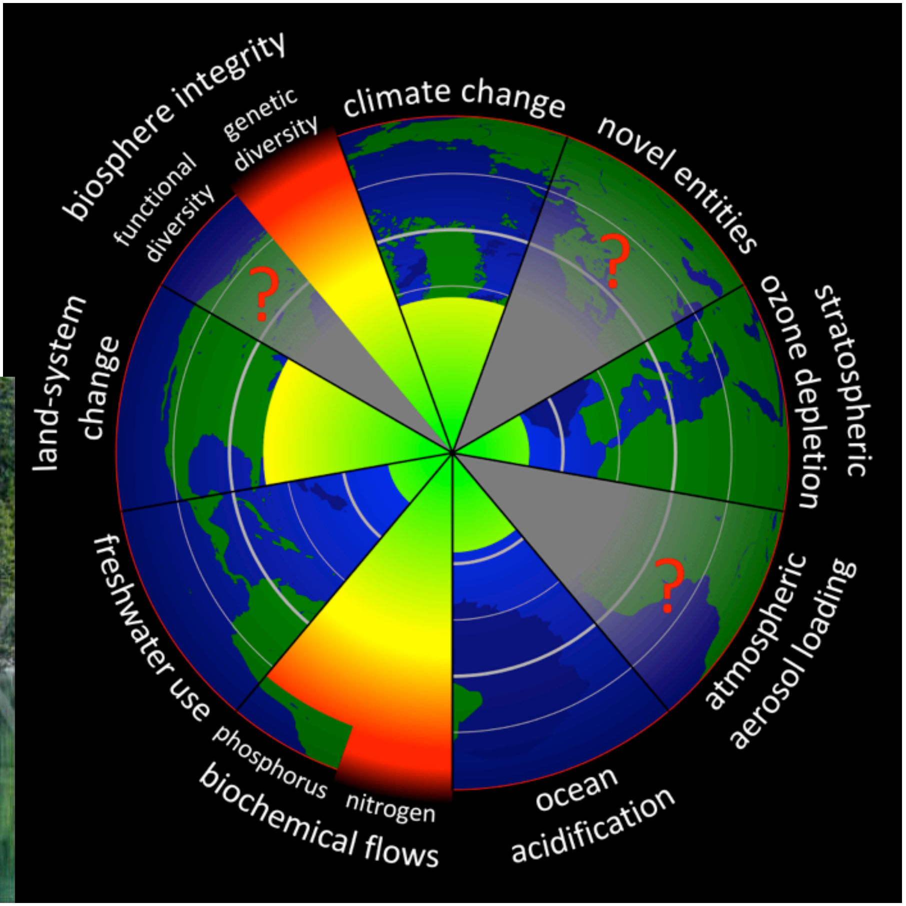
Landscape factors

The environment



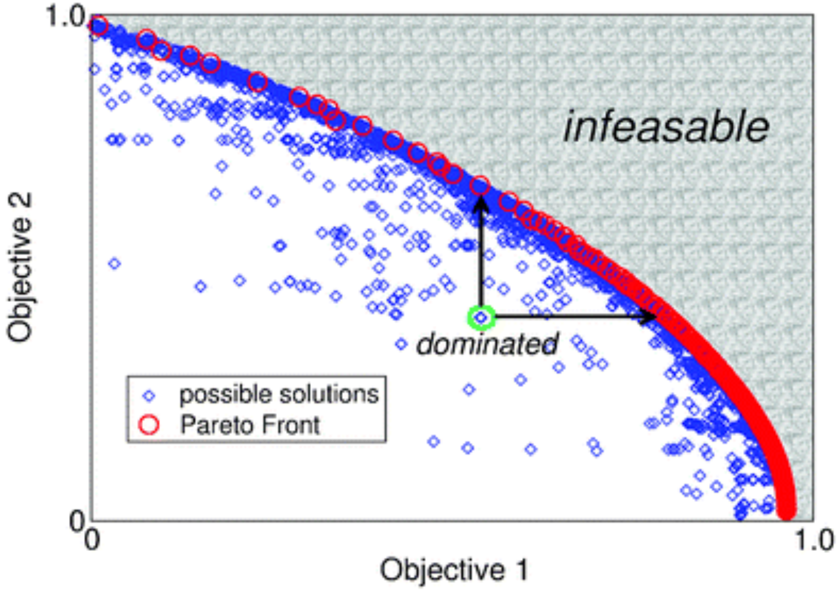
AI for Earth

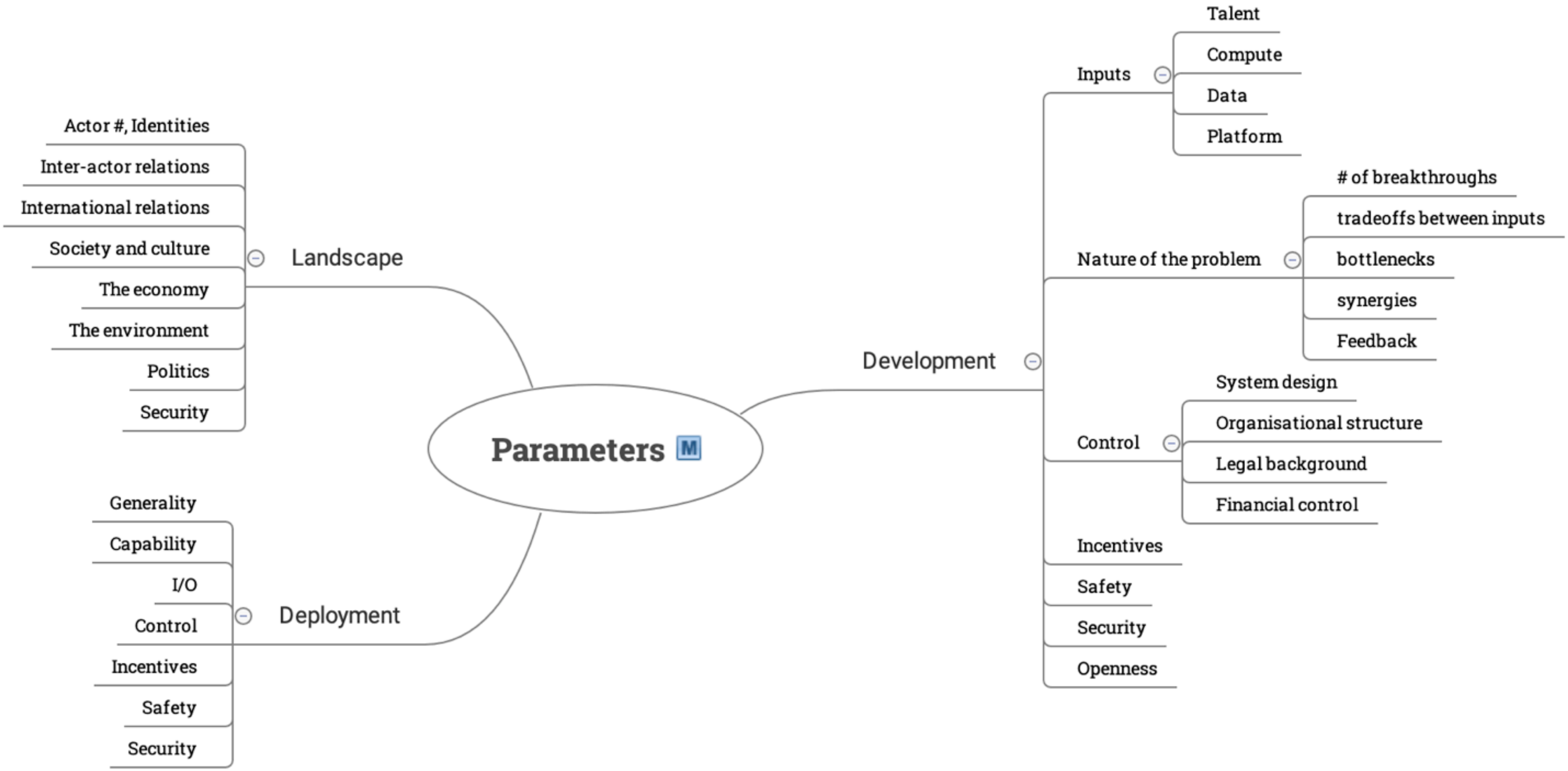
AI for Earth is a Microsoft program aimed at empowering people and organizations to solve global environmental challenges by increasing access to AI tools and educational opportunities, while accelerating innovation.



Landscape factors

Security




Parameters

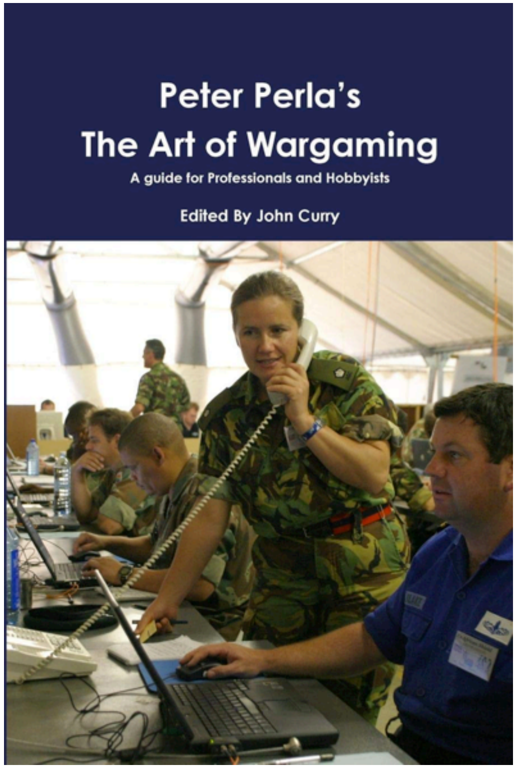
Landscape
- Actor #, Identities
- Inter-actor relations
- International relations
- Society and culture
- The economy
- The environment
- Politics
- Security

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


Peter Perla's

The Art of Wargaming

A guide for Professionals and Hobbyists

Edited By John Curry