

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?



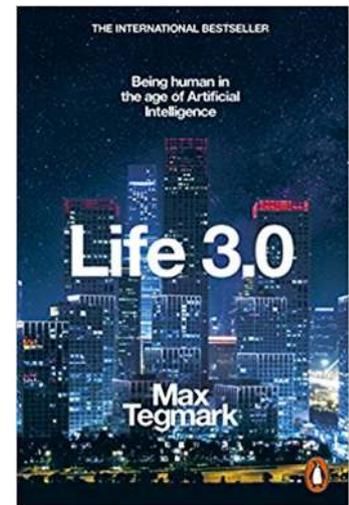
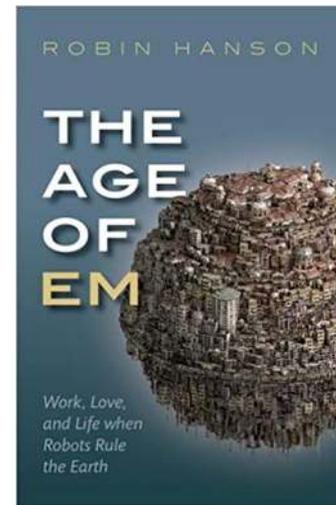
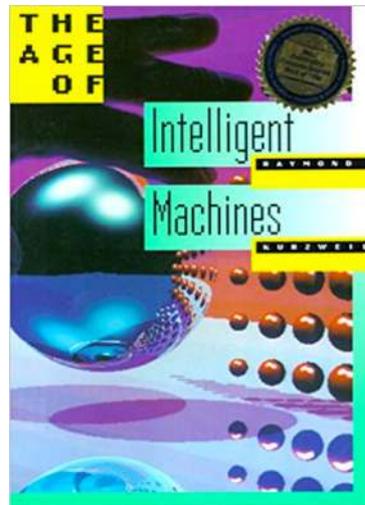
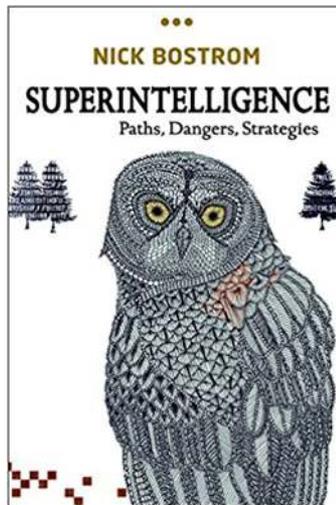
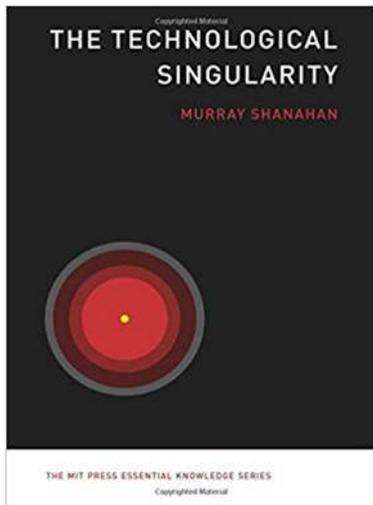
AGI futures narratives

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



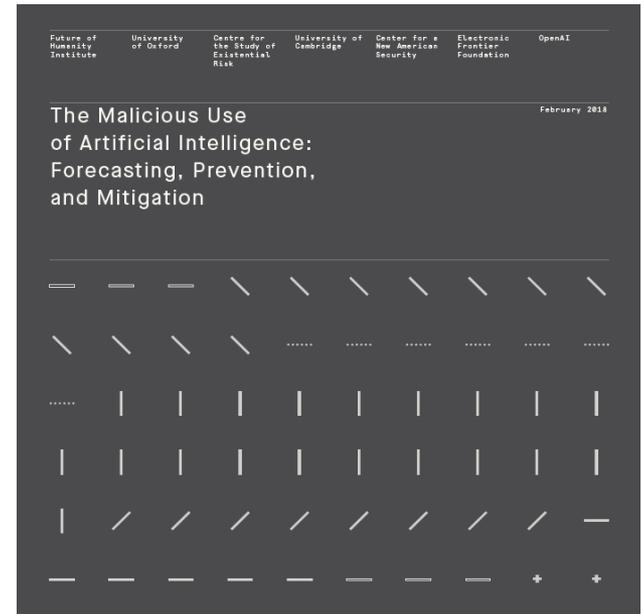
How do we explore and communicate these futures?

Single author exploration



How do we explore and communicate these futures?

Expert workshops, multi-authored reports

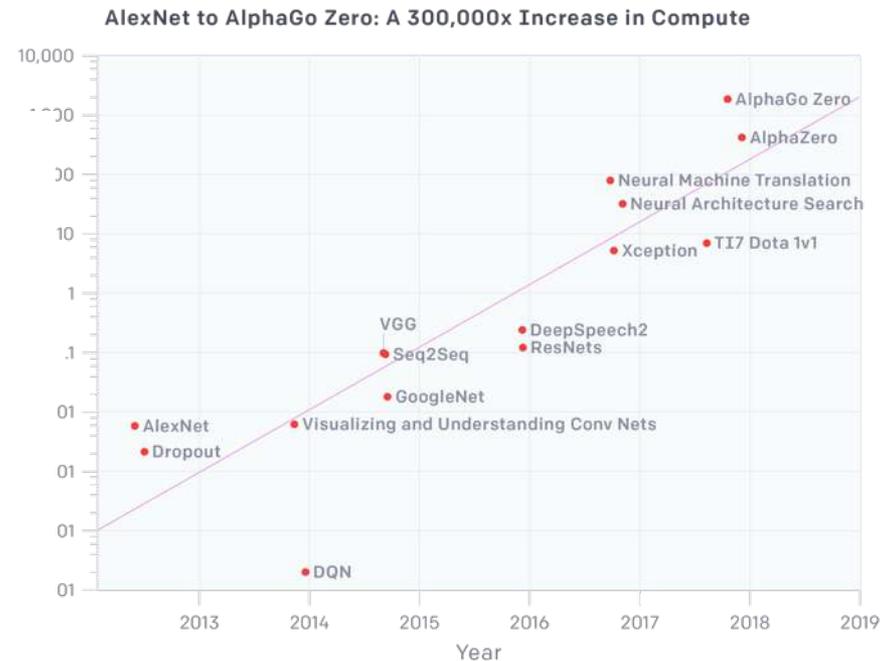
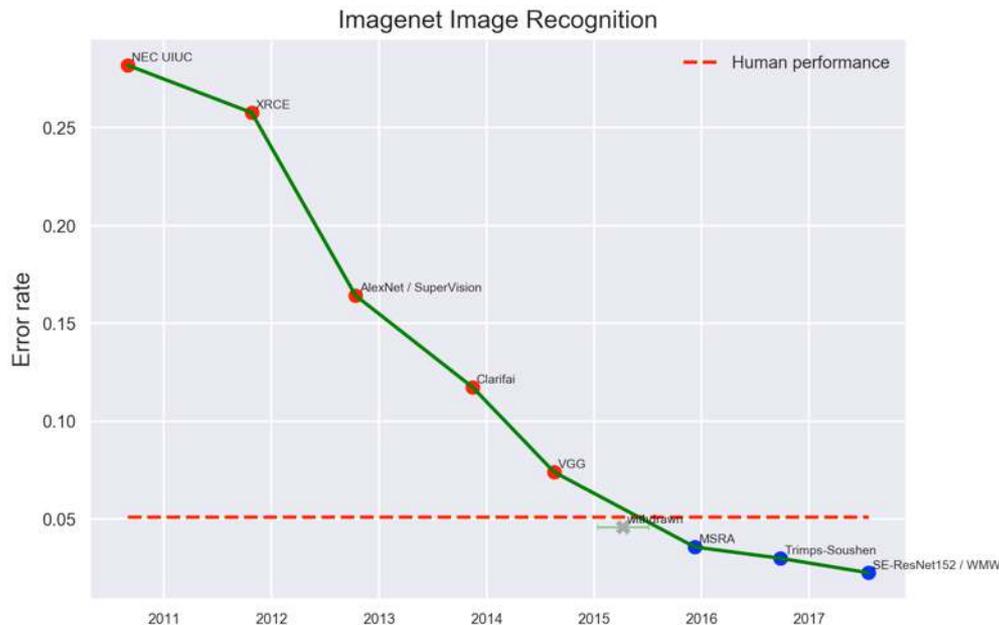


<http://maliciousaireport.com/>

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>



The screenshot shows the Metaculus website interface. The header includes the Metaculus logo and navigation links: "mapping the future", "crowdfounding critical contingencies", "modeling accurate insights", "predicting definite", "cumulative understanding", and "delivering definitive predictions". There are also buttons for "FIND QUESTIONS", "CATEGORIES", "CREATE A QUESTION", and "RANK".

The main content area displays a question: "Will the 'silver' Turing Test be passed by 2026?". The question was created by "nostradamus" and opened on Feb 1, 2016. It has 899 predictions, a 70% median, and 142 interested users. The question is currently "OPEN" and closes on Feb 29, 2020.

The question text reads: "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."

Buttons for "View Metaculus Prediction" and "Hide Community Prediction" are visible at the bottom of the question card.

<https://www.metaculus.com>

How do we explore and communicate these futures?

Video games



<https://bit.ly/2uqXNUn>

The screenshot displays the Decision Problem game interface. At the top, a black notification bar contains the text: "AutoClipper performance boosted by another 75%", "Manufacturing is now 5 times more effective", "Investment engine unlocked", "Lifetime investment revenue report: \$0", and "Using spectral froth annealing we now get 35,750 supply from every spool".

The main interface is divided into several panels:

- Paperclips: 128,768**
- 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%
- Manufacturing:** Clips per Second: 292.5
- Operations:** 10,126 / 10,000, Creativity: 17
- Quantum Computing:** qOps: 333
- Projects:** New Alogan (25 cost, 2,500 ops) - Improves marketing effectiveness by 50%; Strategic Modeling (12,000 ops) - Analyze strategy tournaments to generate Yomi; Photonix Chip (10000 ops) - Converts electromagnetic waves into quantum operations; HypersDrover (70,000 ops) - Autonomous aerial board ambassadors.
- Investments (Low Risk):** A table with columns: Stock, Amt., Price, Total, P/L. It shows a total investment of \$0.
- Trust:** 15, +1 Trust at: 144,000 clips.
- Time:** 00:19:15

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

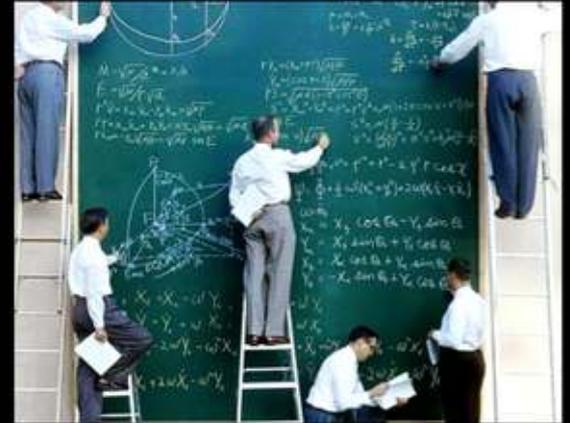
SPACECRAFT SCIENTIST/ENGINEER



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.



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](#).

Sharing and finding you on Facebook

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

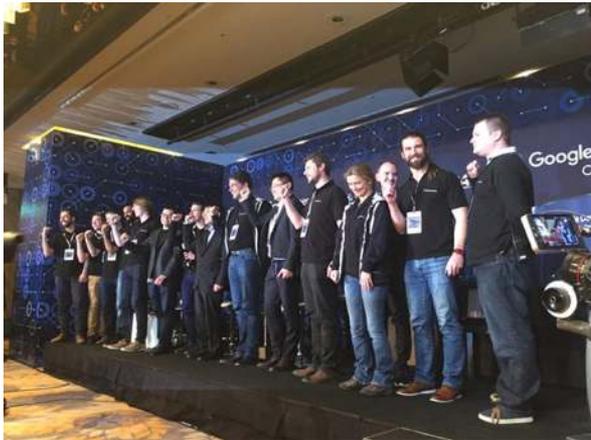
Sharing with other websites and applications

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

More resources

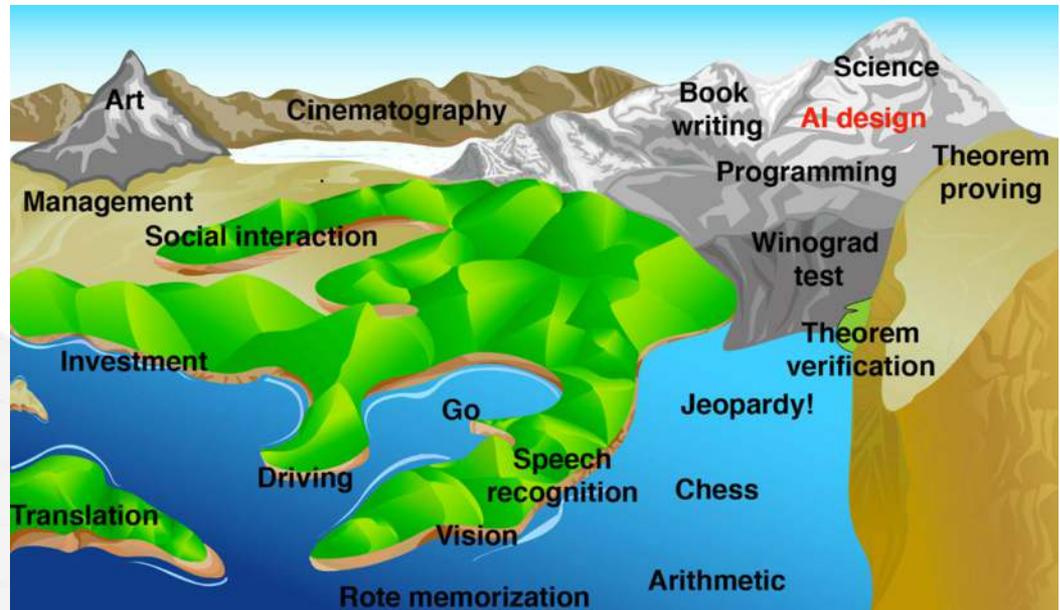
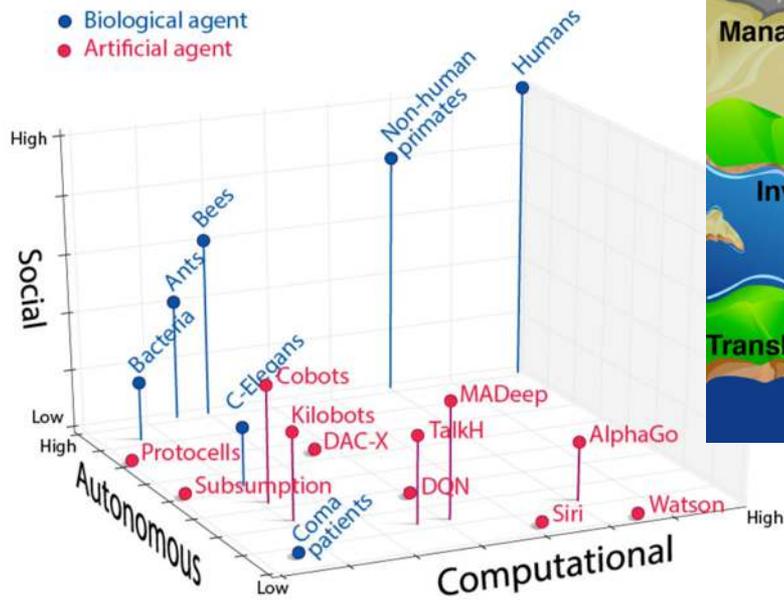
Interactive tools

[View the complete Data Use Policy](#)



Development factors

Nature of the problem



Development factors

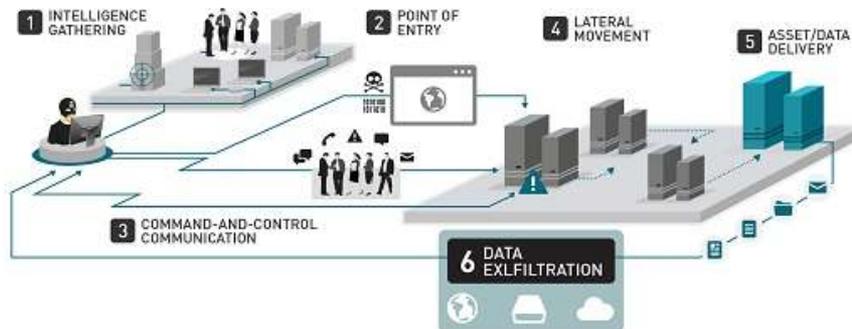
Control, incentives, openness



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 <ul style="list-style-type: none"> Bugs & inconsistencies Ambiguities Side-effects High-level specification languages Preference learning Design protocols 	Prevention and Risk <ul style="list-style-type: none"> Risk sensitivity Uncertainty estimates Safety margins Safe exploration Cautious generalisation Verification Adversaries 	Monitoring <ul style="list-style-type: none"> Interpretability Behavioural screening Activity traces Estimates of causal influence Machine theory of mind Tripwires & honeypots
Emergent <ul style="list-style-type: none"> Wireheading Delusions Metalearning and sub-agents Detecting emergent behaviour 	Recovery and Stability <ul style="list-style-type: none"> Instability Error-correction Failsafe mechanisms Distributional shift Graceful degradation 	Enforcement <ul style="list-style-type: none"> 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.

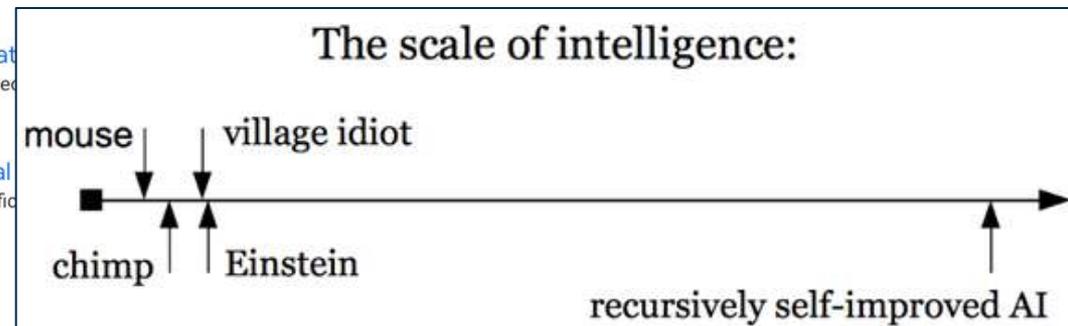
 **AutoML Translate**
Custom domain-specific translation.

 **AutoML Natural Language**
Custom text classification models.

CONVERSATION

 **Dialogflow Enterprise Edition**
Build conversational interfaces.

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



Landscape factors

Number and identity of actors

accenture

:) Affectiva

amazon



Baidu 百度

Cogitai

DeepMind



ELEMENT^{AI}

facebook

Google

IBM



McKinsey&Company

Microsoft

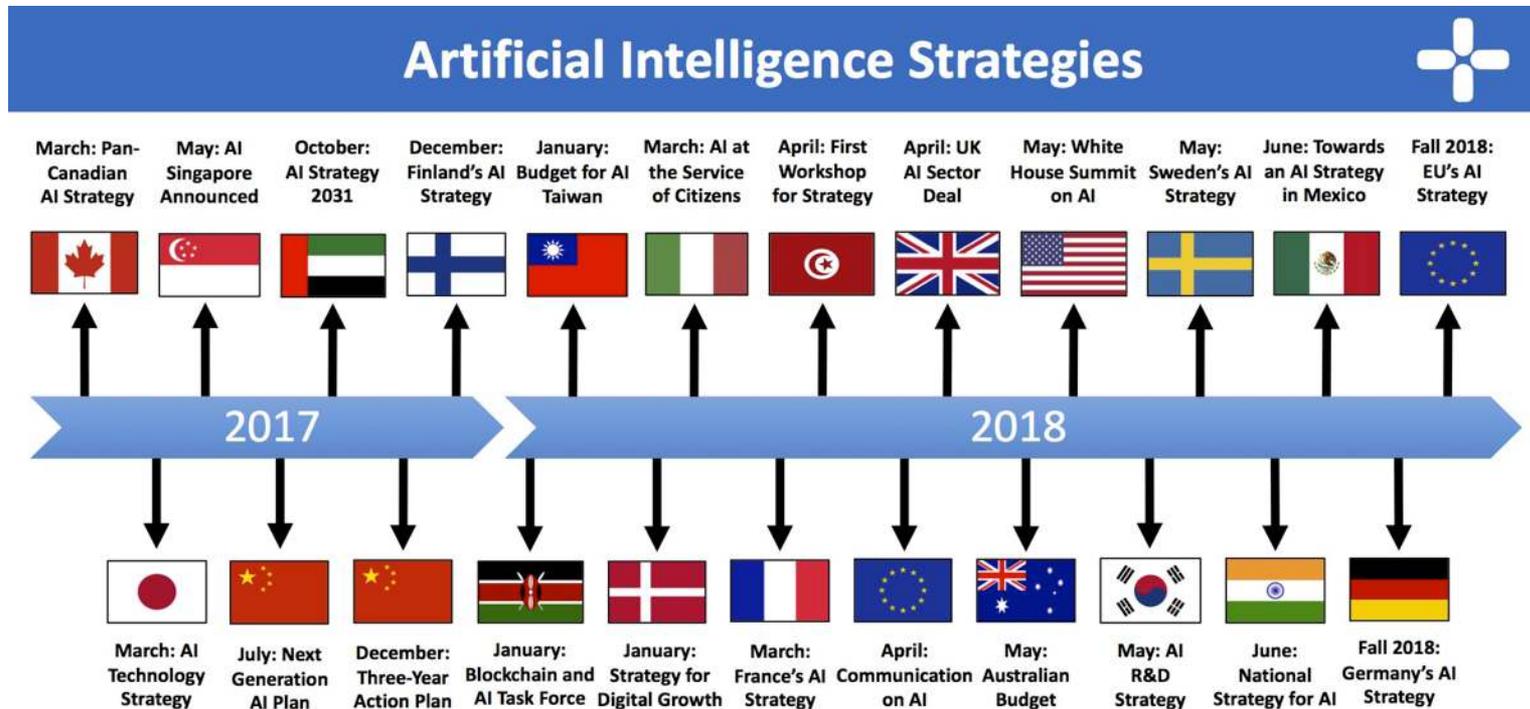
Landscape factors

Inter-actor relationships



Landscape factors

International relations



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

Landscape factors

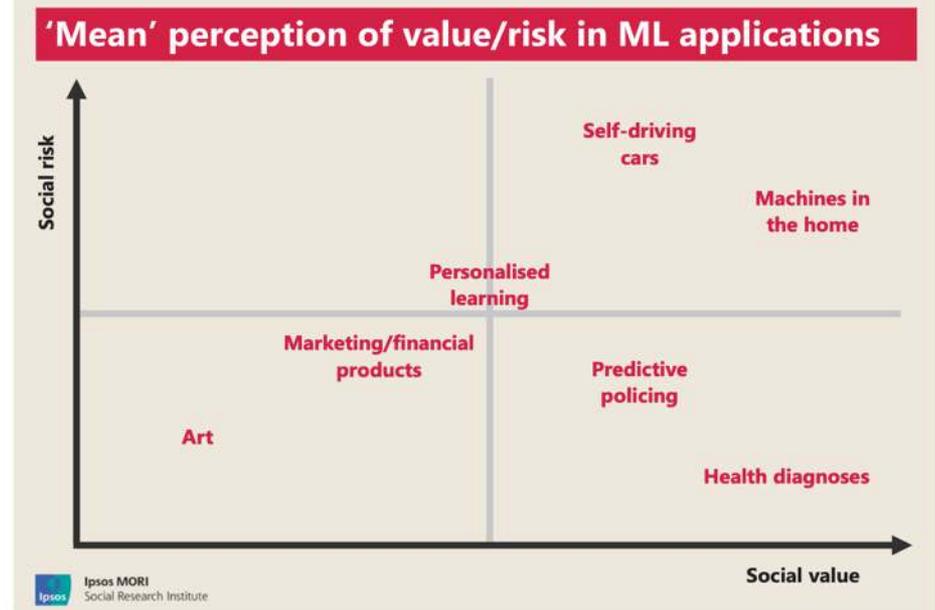
Society and culture



The New York Times

*Wielding Rocks and Knives,
Arizonans Attack Self-Driving Cars*

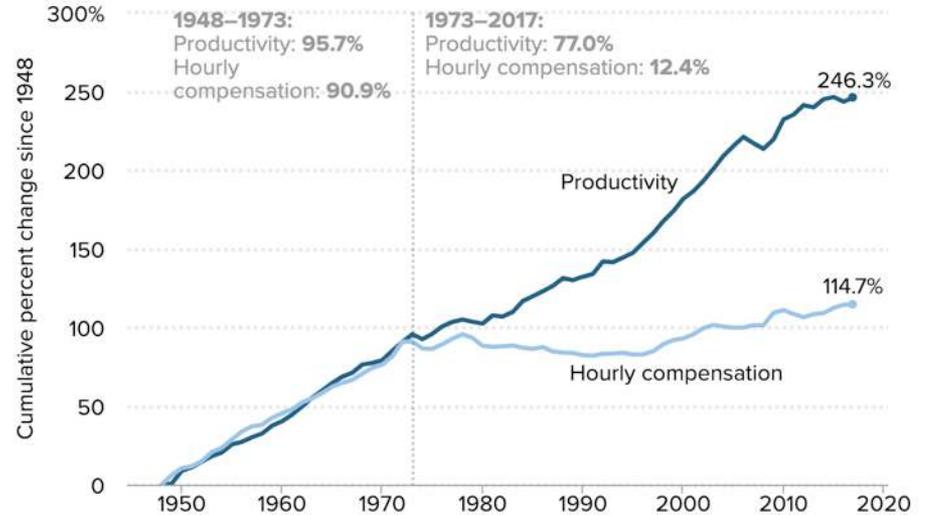
Figure 4.1: Overall social risk v. social value assessment



Landscape factors

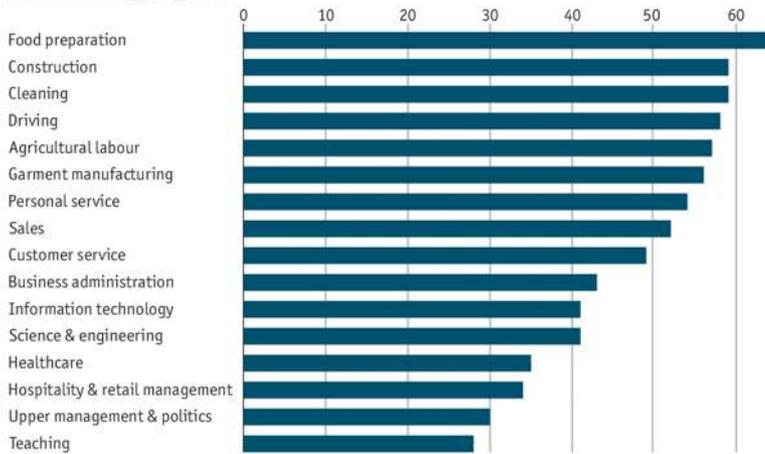
The economy

Productivity growth and hourly compensation growth, 1948–2017



Automated for the people

Automation risk by job type, %



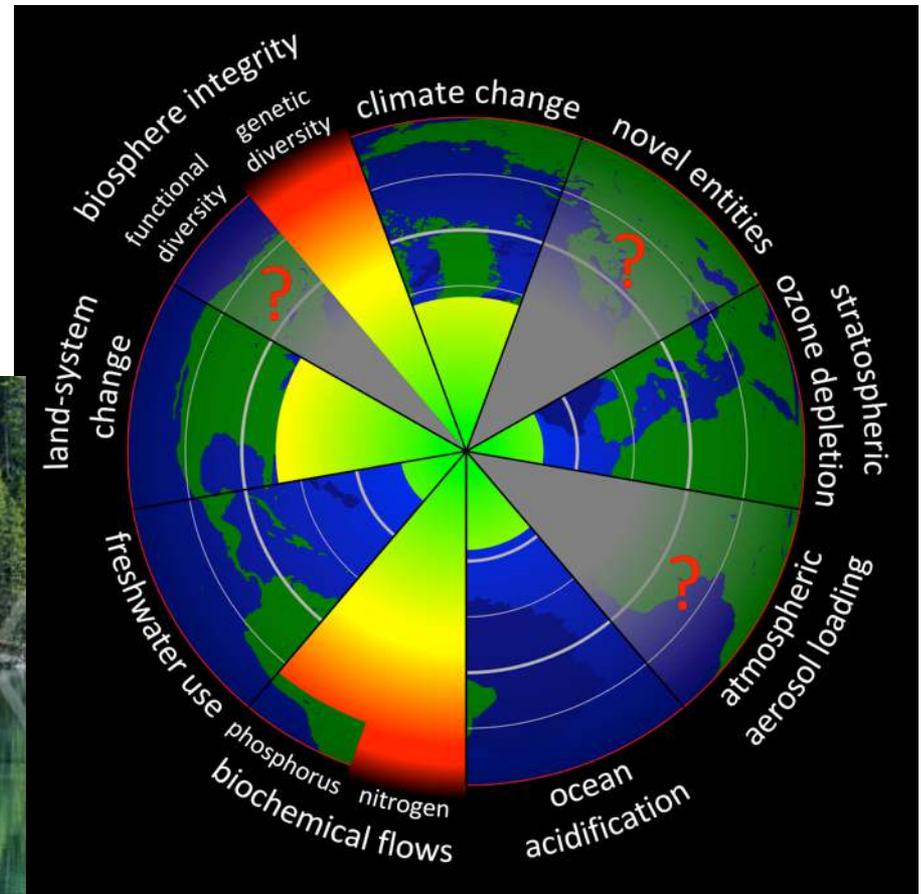
Source: OECD

Economist.com



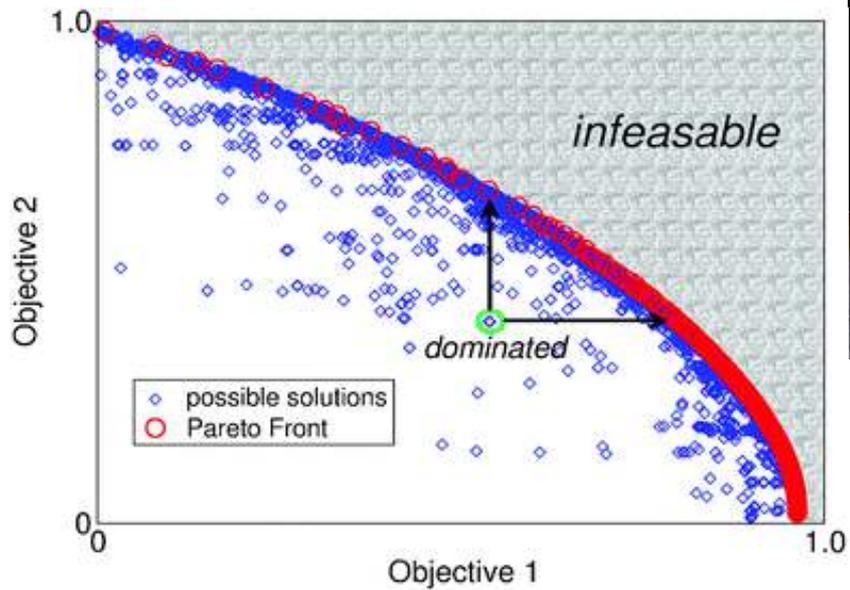
Landscape factors

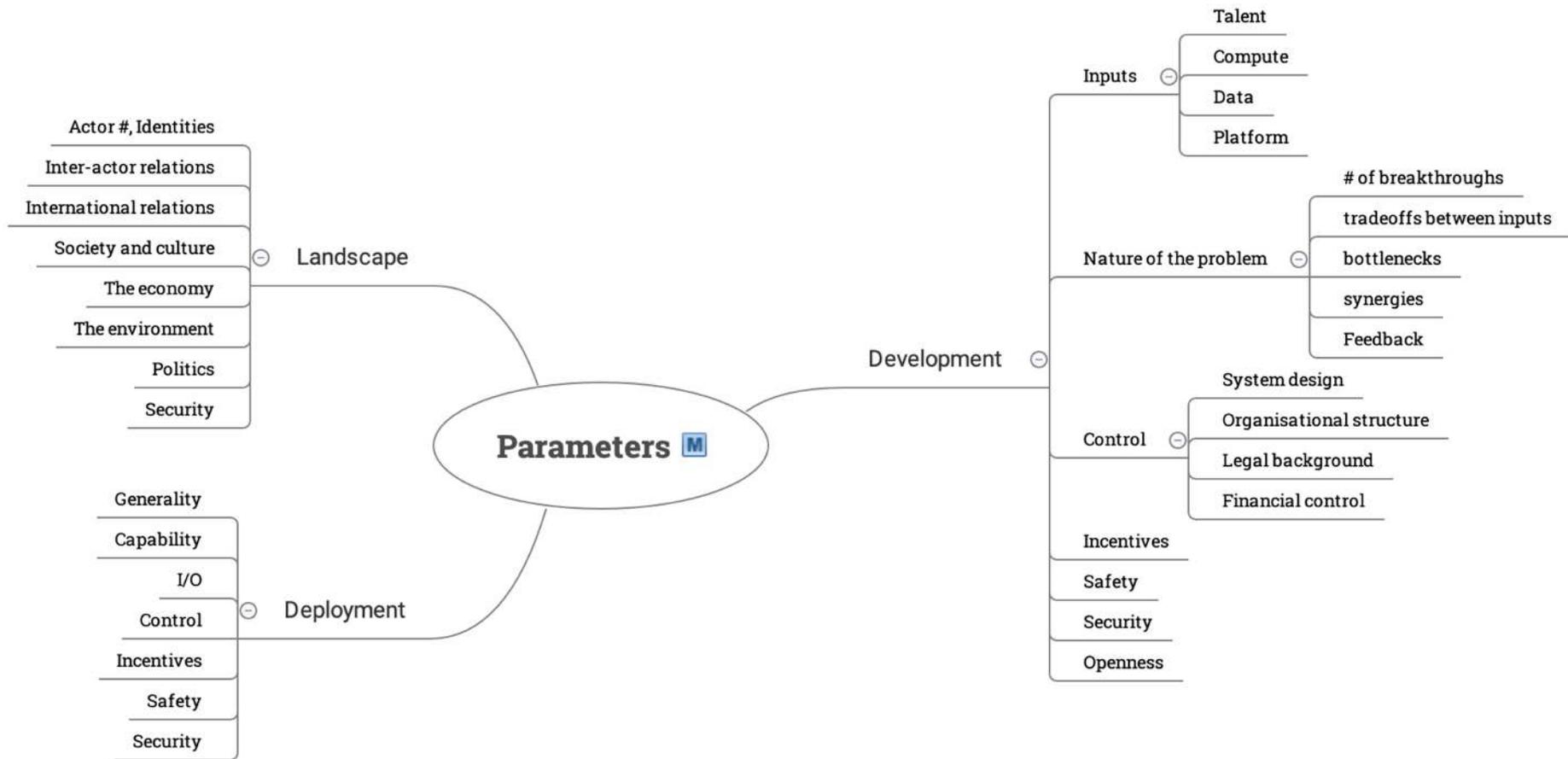
The environment



Landscape factors

Security





How do we explore and communicate these futures?

Scenario role-play

