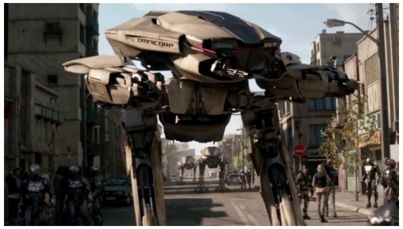
UNIVERSITY OF CAMBRIDGE

System 1 and System 2 Thinking Cycles

Shahar Avin
sa478@cam.ac.uk

The importance of AI strategy

t



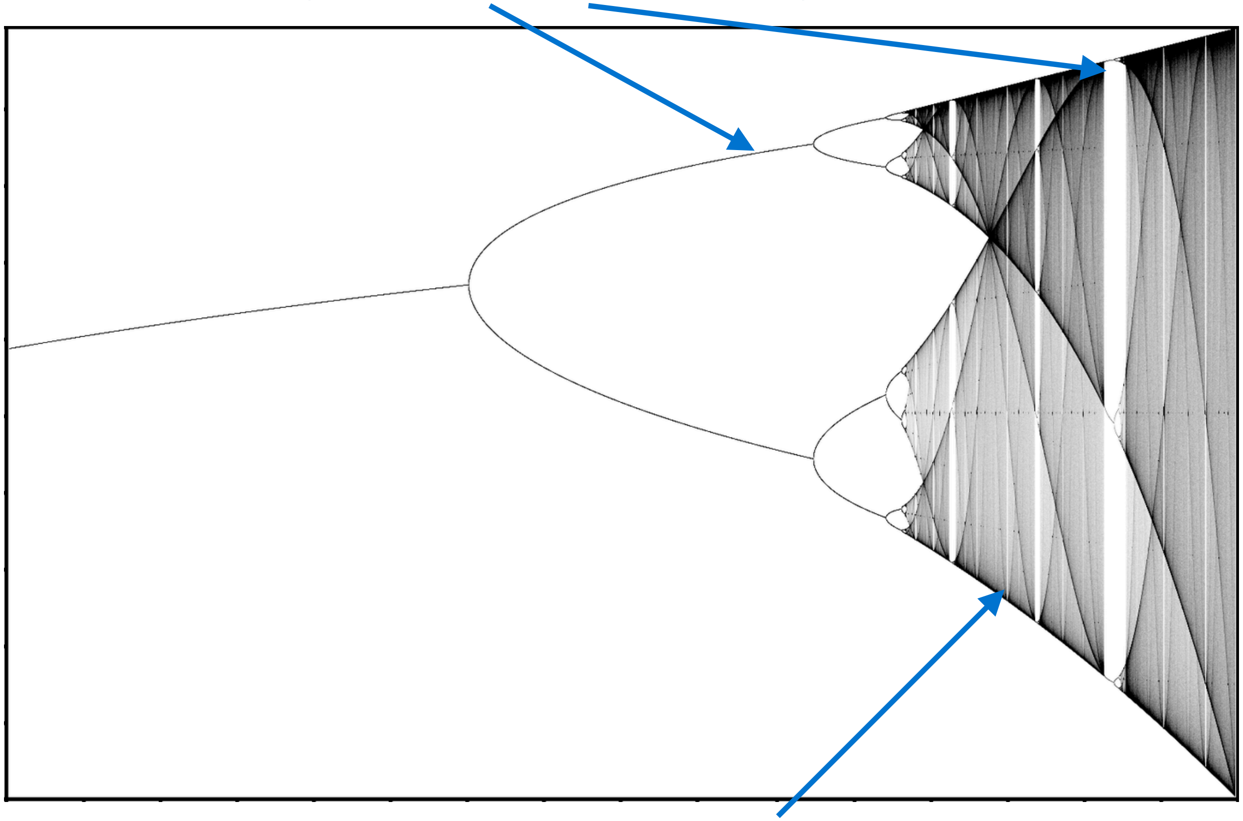
? ? ?

YOU ARE HERE

2019

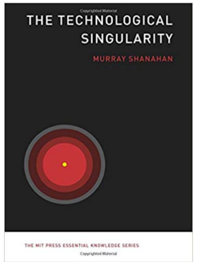
But history is messy

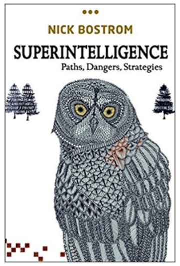
Requiring both mechanism design…

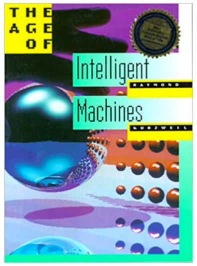


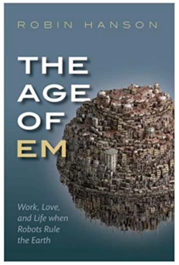
… and first-person perspective

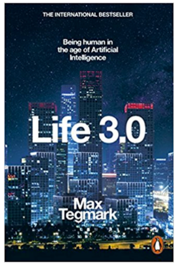
Mechanism design

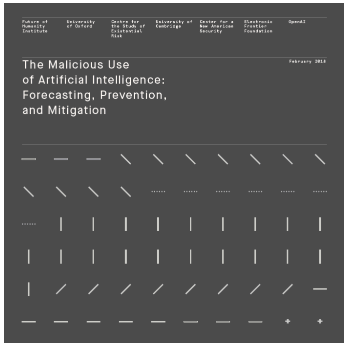

THE TECHNOLOGICAL SINGULARITY
MURRAY SHANAHAN
THE MIT PRESS ESSENTIAL KNOWLEDGE SERIES







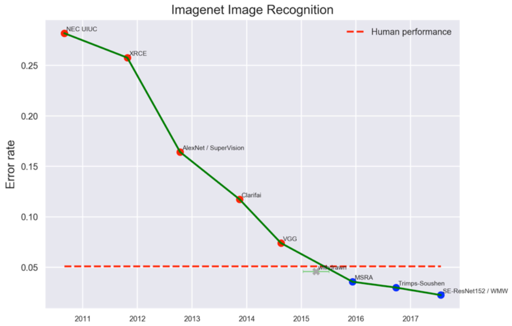

THE INTERNATIONAL BESTSELLER
Being human in the age of Artificial Intelligence
Life 3.0
Max Tegmark

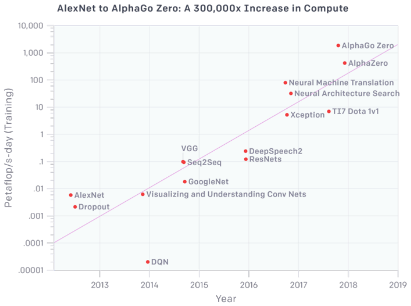


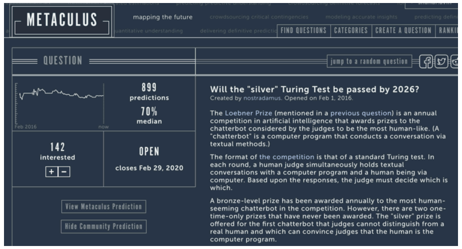
Future of Humanity Institute | University of Oxford | Centre for the Study of Existential Risk | University of Cambridge | Center for a New American Security | Electronic Frontier Foundation | OpenAI

February 2018

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






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

899 predictions

70% median

142 interested

OPEN

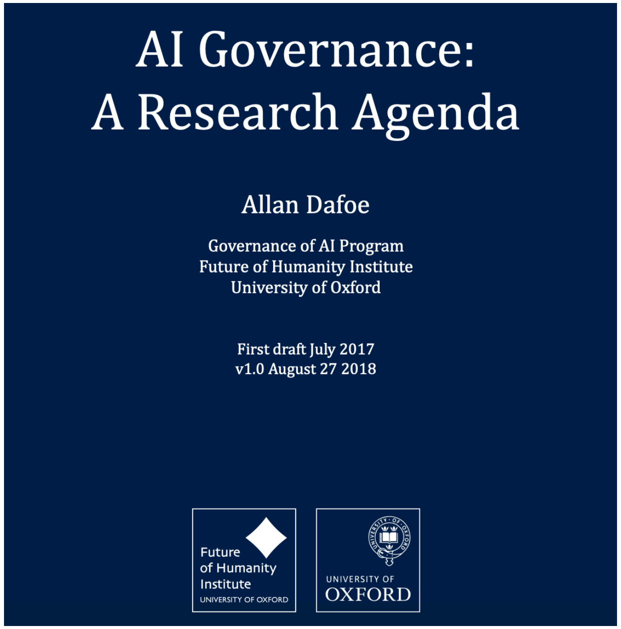
closes Feb 29, 2020

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.



AI Governance:
A Research Agenda

Allan Dafoe

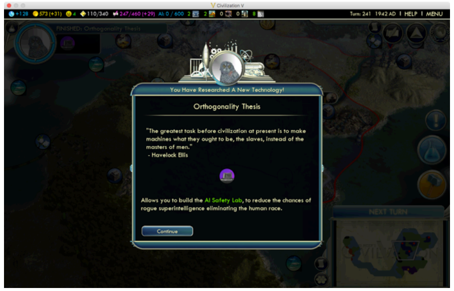
Governance of AI Program
Future of Humanity Institute
University of Oxford

First draft July 2017
v1.0 August 27 2018

Future of Humanity Institute
UNIVERSITY OF OXFORD

UNIVERSITY OF OXFORD

First-person perspective


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
NEXT TURN

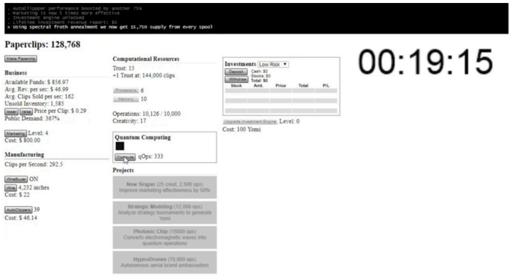
You Have Researched A New Technology!

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.

Continue



```
. 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

Make Paperclip

Business

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

Marketing Level: 4
Cost: $ 800.00

Manufacturing

Clips per Second: 292.5

WireBuyer ON
Wire 4,232 inches
Cost: $ 22

AutoClippers 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

Compute 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

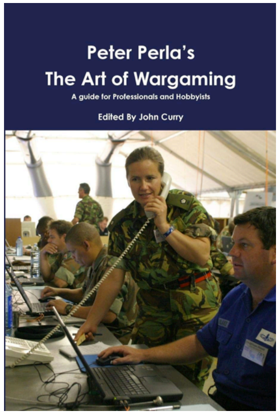
Investments Low Risk ▾

Deposit Cash: $0
Withdraw Stocks: $0
Total: $0

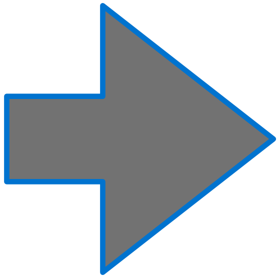
Stock	Amt.	Price	Total	P/L

Upgrade Investment Engine Level: 0
Cost: 100 Yomi

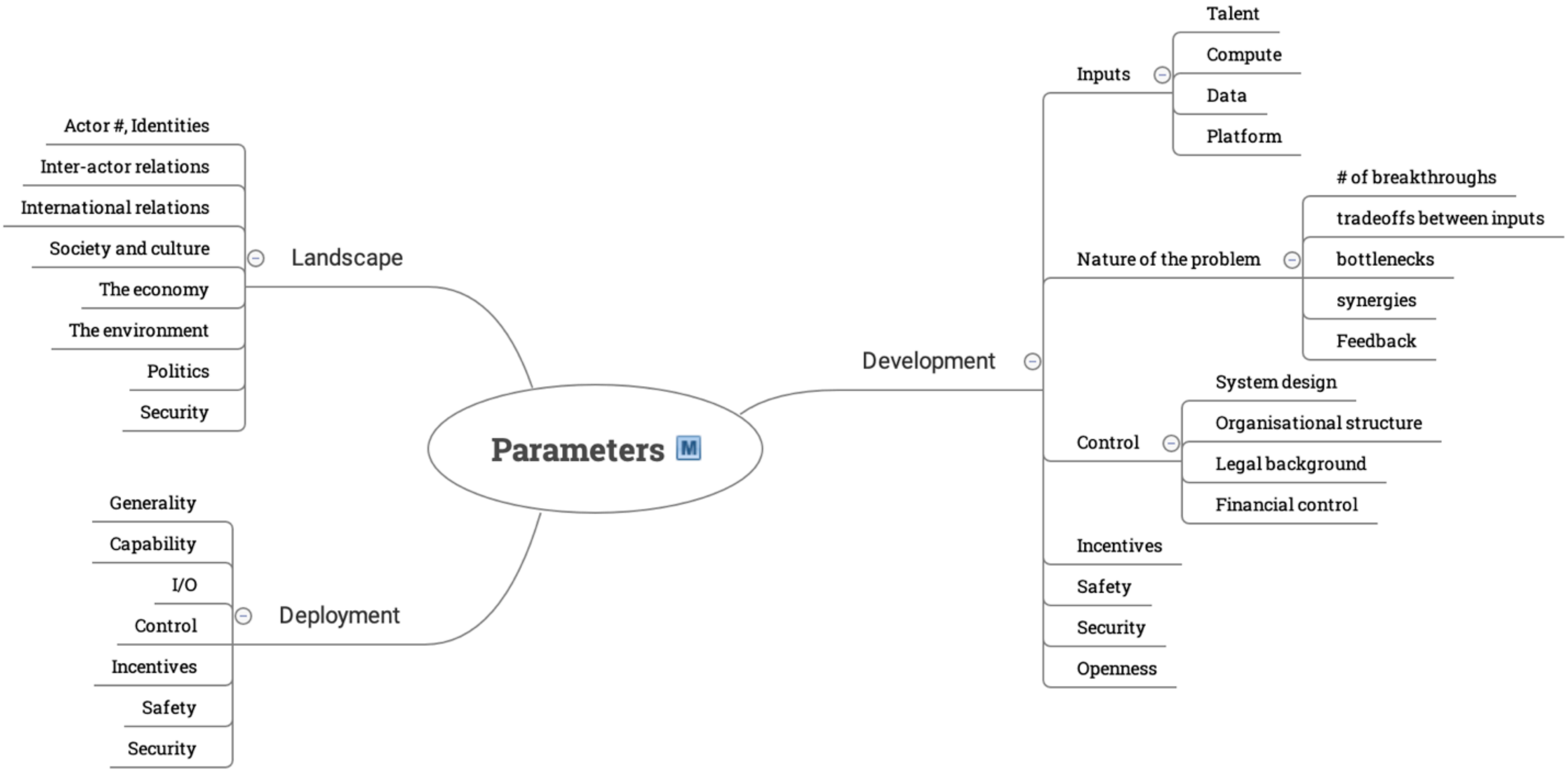
00:19:15



Peter Perla's
The Art of Wargaming
A guide for Professionals and Hobbyists
Edited By John Curry



Same system(s)…



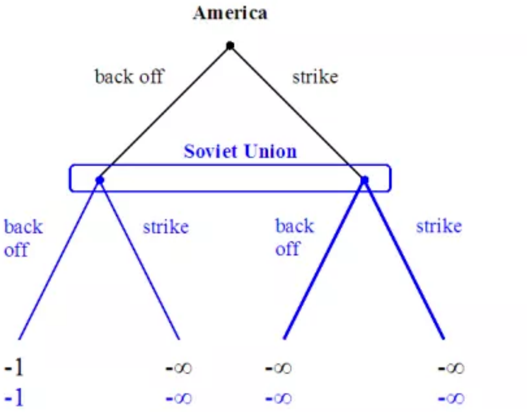
- Parameters
  - Landscape
    - Actor #, Identities
    - Inter-actor relations
    - International relations
    - Society and culture
    - The economy
    - The environment
    - Politics
    - Security
  - Deployment
    - Generality
    - Capability
    - I/O
    - Control
    - Incentives
    - 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

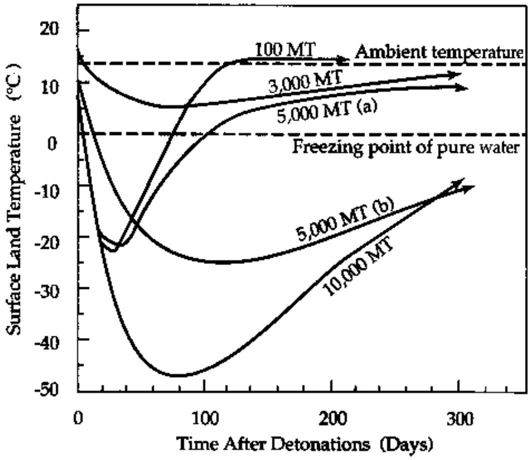
…different perspectives

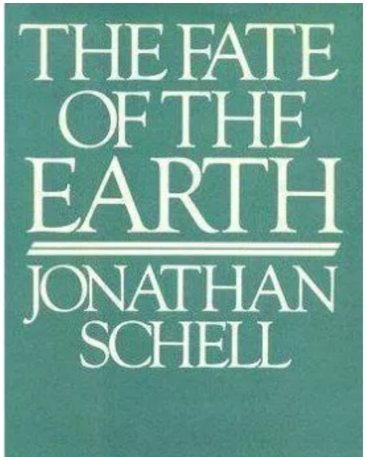
- incentives
- models
- generalisations
- attractors
- stability

- identities
- context
- world view
- salience
- narrative

Different, complementary outcomes







THE FATE OF THE EARTH

JONATHAN SCHELL

CSER AI Strategy scenario role-plays

- ~15 so far
- 3-60 participants
- AI researchers, policy researchers, effective altruists, general public
- Play as USG, PRC, EU, Tech company, Defence contractor, NGO
- Track talent, compute, technical breakthroughs, data, money
- Public and secret actions, negotiations, time pressure
- Fun, if depressing at times

- Get in touch if you’re interested

Earth 2045 scenario simulation

- Resolve a dramatic event
- Using a first-person perspective of a made-up character
- In collaboration with others
- While exploring an imaginary world
- Facilitated by an all-knowing story master

There is no rule saying changing the world can’t be (sometimes) fun