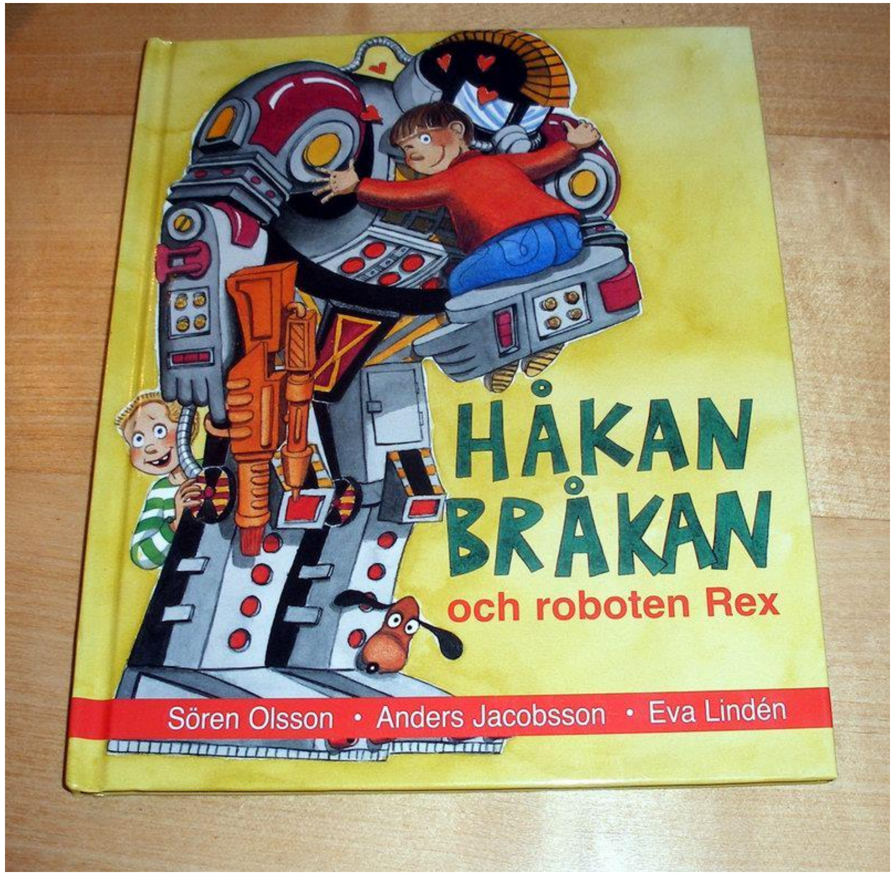
On autonomous cars, autopilots and the Anthropocene:

ethics and AI verification

Dr Anders Sandberg
Future of Humanity Institute
Oxford Martin School
Oxford University

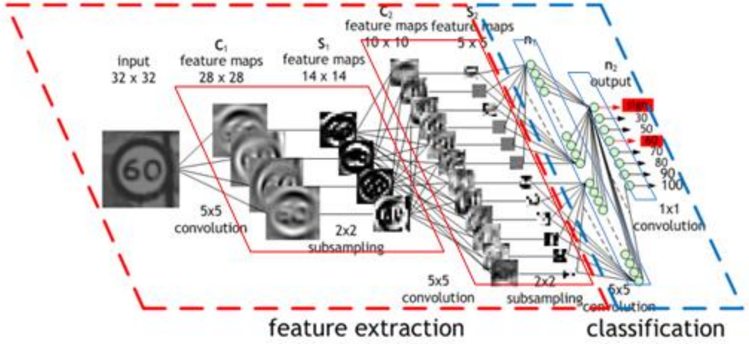
Robots have to live in the anthropocene

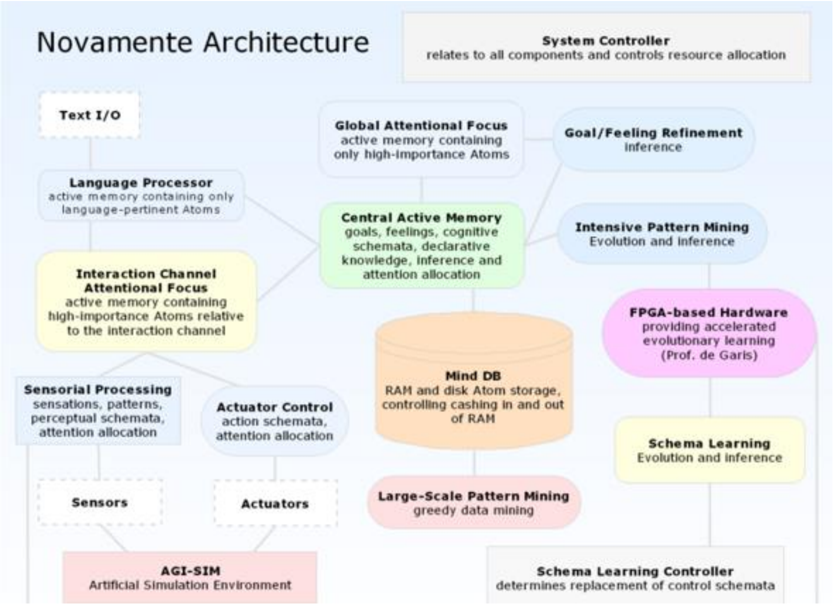
- Autonomous systems that matter will be interacting with the human world
- Verification and validation in this context involves not only understanding how the robot will interact, but how humans and their institutions will interact back.
- The human world is an ethical minefield. Just walking in a straight line is ill-advised.
- A hierarchy of “moral verification” challenges.



Challenge 1: safe and beneficial AI

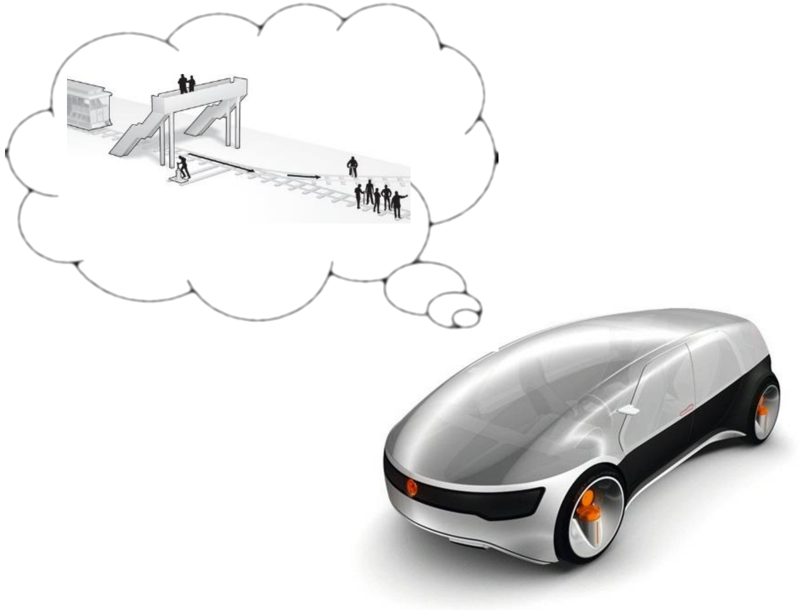
- Creating systems that behave in such a way that human aims are fulfilled beneficially and safely
    - Requirements need to embody this, but experience and the SAI debate show it to be fundamentally hard
- Traditionally: safety, controllability
- Also: inspectability, technical debt
- Long run: corrigibility, inverse reinforcement learning

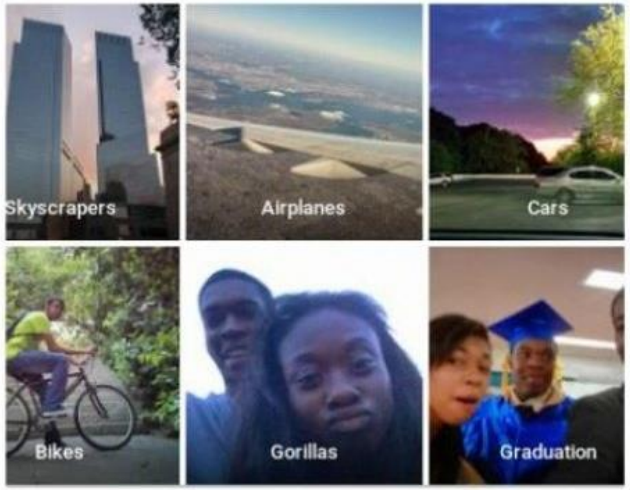



Novamente Architecture
System Controller
relates to all components and controls resource allocation
Text I/O
Global Attentional Focus
active memory containing only high-importance Atoms
Goal/Feeling Refinement
inference
Language Processor
active memory containing only language-pertinent Atoms
Central Active Memory
goals, feelings, cognitive schemata, declarative knowledge, inference and attention allocation
Intensive Pattern Mining
Evolution and inference
Interaction Channel Attentional Focus
active memory containing high-importance Atoms relative to the interaction channel
FPGA-based Hardware
providing accelerated evolutionary learning (Prof. de Garis)
Mind DB
RAM and disk Atom storage, controlling cashing in and out of RAM
Sensorial Processing
sensations, patterns, perceptual schemata, attention allocation
Actuator Control
action schemata, attention allocation
Schema Learning
Evolution and inference
Sensors
Actuators
Large-Scale Pattern Mining
greedy data mining
AGI-SIM
Artificial Simulation Environment
Schema Learning Controller
determines replacement of control schemata

Challenge 2: “Moral” behaviour of machines

- Creating systems that act in the same way as a moral human would
- Engineering hesitancy for encoding morality in specifications
- “Malicious autonomy”: mislearning and misbehaving without intention
- The value learning and alignment problem

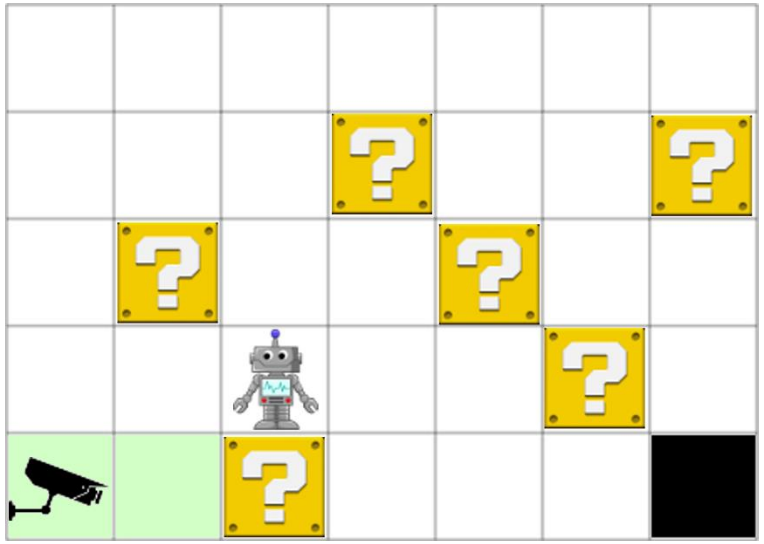




Jacky Alciné via Twitter

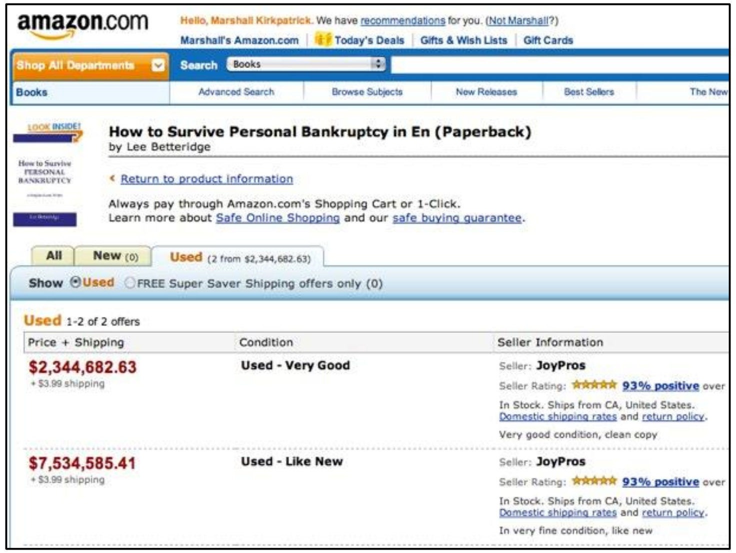
Challenge 3: keeping humans behaving/thinking right

- Creating systems so that humans are not driven to detrimental actions or belief states
- Autopilot bias
- Emergent accidental deception
- Cannot assume humans certified or know what they do.



Challenge 4: reducing systemic risk

- Avoiding systemic risks that emerge from the overall process
- The joint human-machine system can misbehave
- System design of open systems: always underspecified
  - Incentive design, principle design



amazon.com

Hello, Marshall Kirkpatrick. We have recommendations for you. (Not Marshall?)

Marshall's Amazon.com | Today's Deals | Gifts & Wish Lists | Gift Cards

Shop All Departments | Search Books

Books | Advanced Search | Browse Subjects | New Releases | Best Sellers | The New

LOOK INSIDE!

How to Survive PERSONAL BANKRUPTCY

How to Survive Personal Bankruptcy in En (Paperback)
by Lee Betteridge

‹ Return to product information

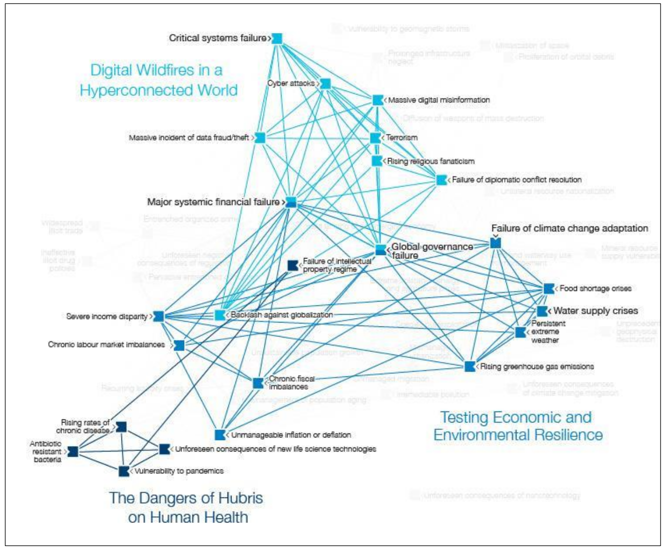
Always pay through Amazon.com's Shopping Cart or 1-Click.
Learn more about Safe Online Shopping and our safe buying guarantee.

All | New (0) | Used (2 from $2,344,682.63)

Show ● Used ○ FREE Super Saver Shipping offers only (0)

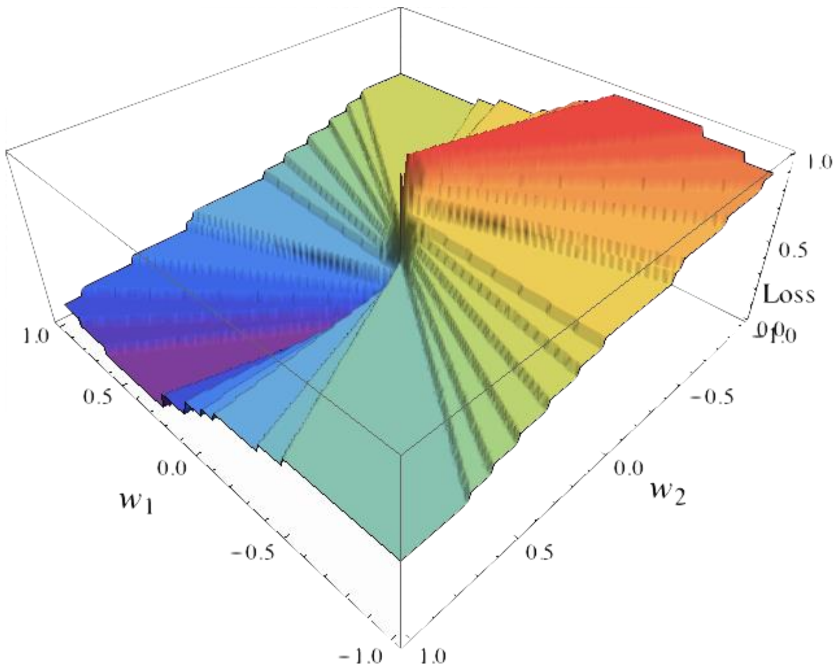
Used 1-2 of 2 offers

Price + Shipping	Condition	Seller Information
$2,344,682.63 + $3.99 shipping	**Used - Very Good**	Seller: **JoyPros** Seller Rating: ★★★★★ 93% positive over In Stock. Ships from CA, United States. Domestic shipping rates and return policy. Very good condition, clean copy
$7,534,585.41 + $3.99 shipping	**Used - Like New**	Seller: **JoyPros** Seller Rating: ★★★★★ 93% positive over In Stock. Ships from CA, United States. Domestic shipping rates and return policy. In very fine condition, like new


Digital Wildfires in a Hyperconnected World
Critical systems failure
Cyber attacks
Massive digital misinformation
Massive incident of data fraud/theft
Terrorism
Rising religious fanaticism
Failure of diplomatic conflict resolution
Major systemic financial failure
Failure of climate change adaptation
Global governance failure
Failure of intellectual property regime
Food shortage crises
Severe income disparity
Backlash against globalization
Water supply crises
Persistent extreme weather
Chronic labour market imbalances
Rising greenhouse gas emissions
Chronic fiscal imbalances
Rising rates of chronic disease
Antibiotic resistant bacteria
Unmanageable inflation or deflation
Unforeseen consequences of new life science technologies
Vulnerability to pandemics
Testing Economic and Environmental Resilience
The Dangers of Hubris on Human Health

Verifying ethics and the ethics of verifying

- Smart systems do not exist in a vacuum: the human context is an active part of what is going on.
- Autonomous, adaptive and open systems: strict verification and validation will be limited.
- Good-enough verification depends on loss function.
- Good-enough-ethics *also* depends on understanding the loss function.


Loss
w_1
w_2