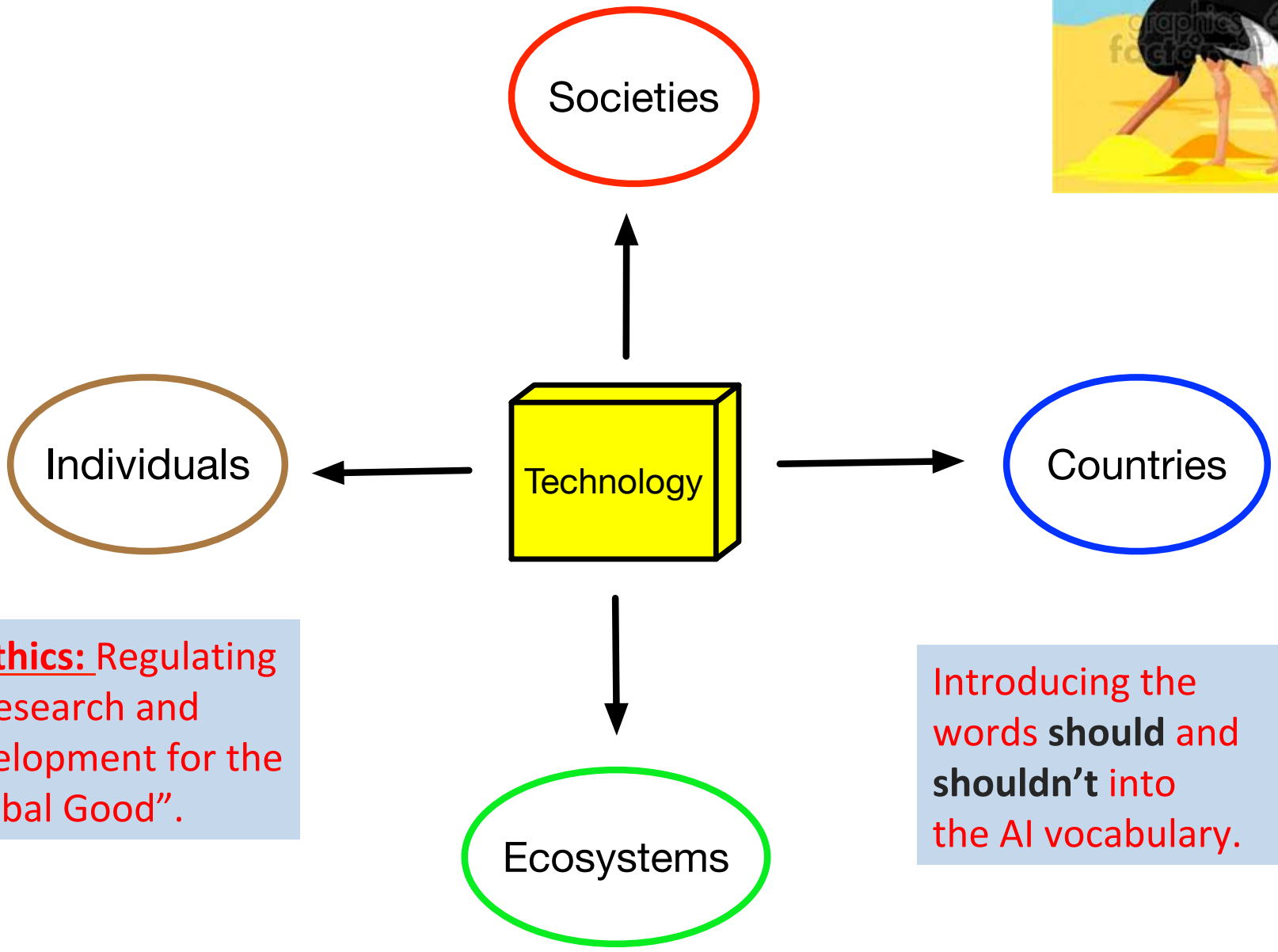
Can to *Should*: Ethics from the AI Research Perspective

Keith L. Downing
Department of Computer Science
&
The Norwegian Open AI Lab
NTNU

AI and Ethics

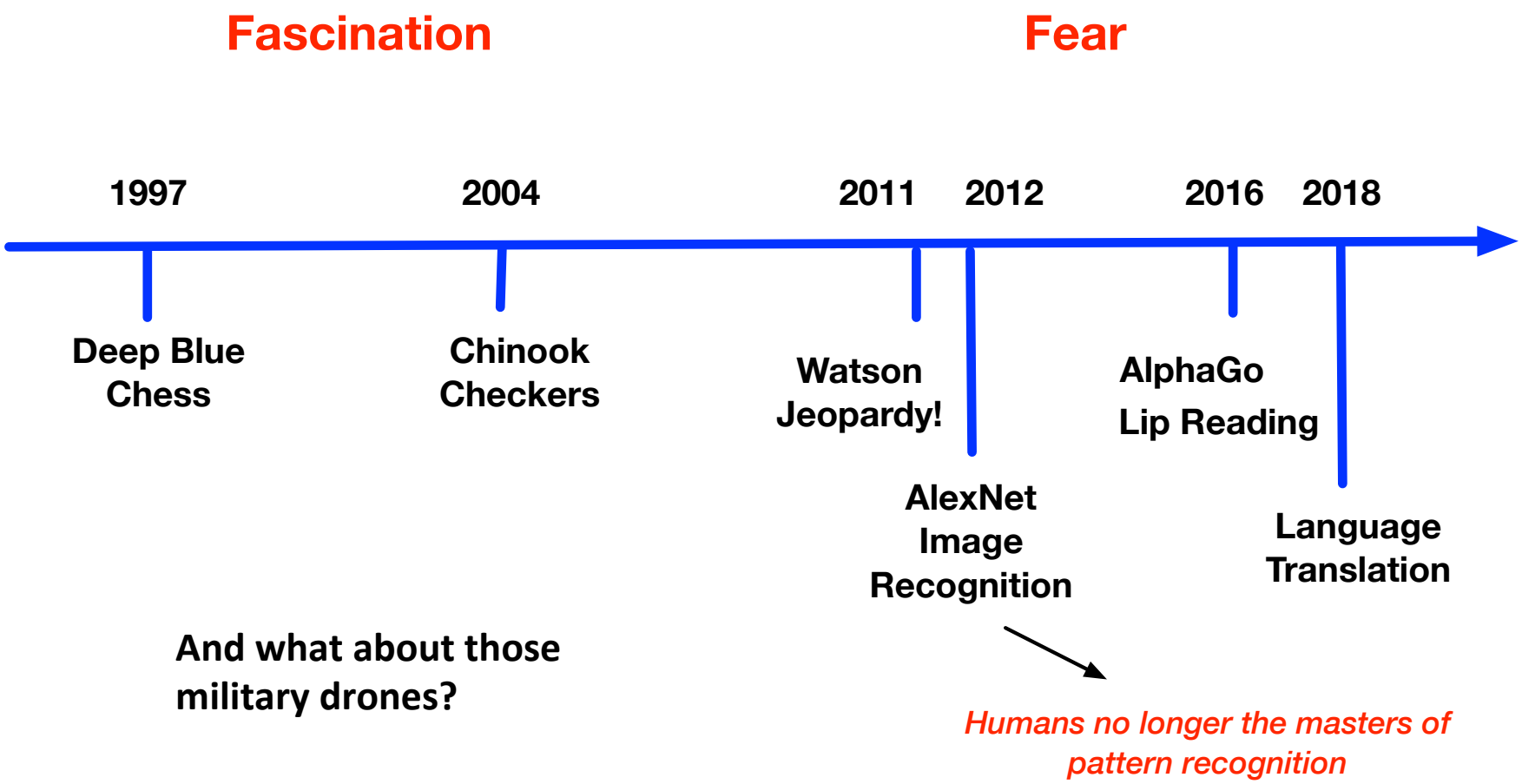


AI Ethics: Regulating AI Research and Development for the "Global Good".

Introducing the words **should** and **shouldn't** into the AI vocabulary.

Ethics Becomes an AI issue

AI Beats or Equals Human Experts

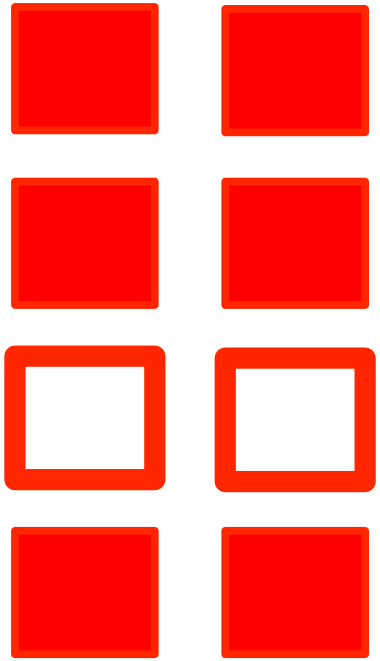


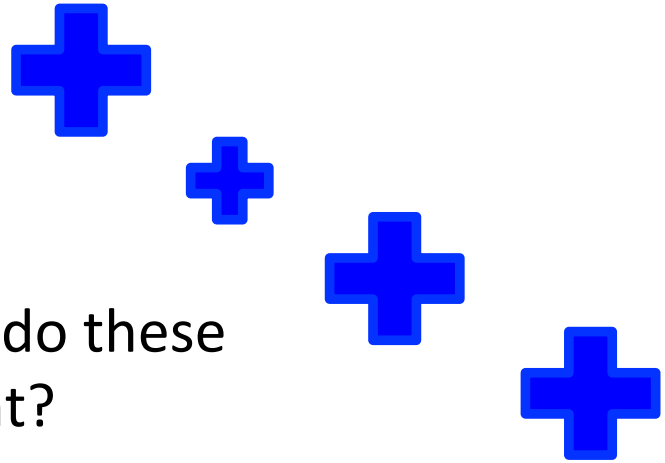
Why worry about AI?

- Complex Opaque Decision Making
- Adaptive => Surprising / Unpredictable
- Gives important advice to humans
- Monitors and *understands* humans
- Impersonates humans
- Manipulates humans
- Defeats humans*

*** In restricted domains**

Glaring Weaknesses





What concept do these both represent?

How much trust should we put in a technology with these types of deficiencies?

Humans are still the masters of **abstraction** & **common sense**

Ethics for Humans –vs- Machines

- **At this point in time**, we (as a species) have the power to control the development of AI and regulate how these machines behave.
- That regulation **can** be more complete and secure than society's attempts to regulate human behavior.
- Thus, we have an opportunity to **fully operationalize** ethical rules *.
- But only if we can agree on a) what they are, and b) how to implement them.

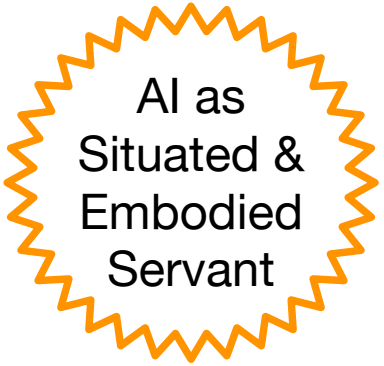
*** Implementation demands formalization demands deep understanding**

AI = Ethics with a Deadline

- Present
  - Privacy and security infringement by AI
  - Human life-altering decision-making by AI
  - Manipulation of humans by AI
- Immediate Future
  - Potential job loss
  - Warping of human intelligence
- Distant Future
  - Emergence of a robotic species
  - What happens to us?

Evolving Roles of AI

Today



Tomorrow

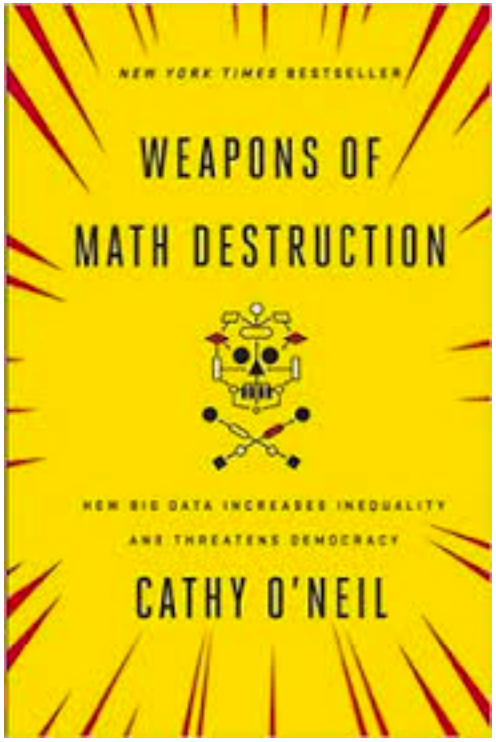
Someday ??

I. AI as Digital Tool

AI is Watching, Listening, Learning & Suggesting

- Who is here?
- When and where will we meet again?
- What do you like?
- What do you believe?
- Should you get a bank loan?
- Can we sell you our stuff?
- Can we influence your vote?
- Should we send you to jail?

How should humans handle the information produced by big-data mining, and the suggestions given by AI systems?

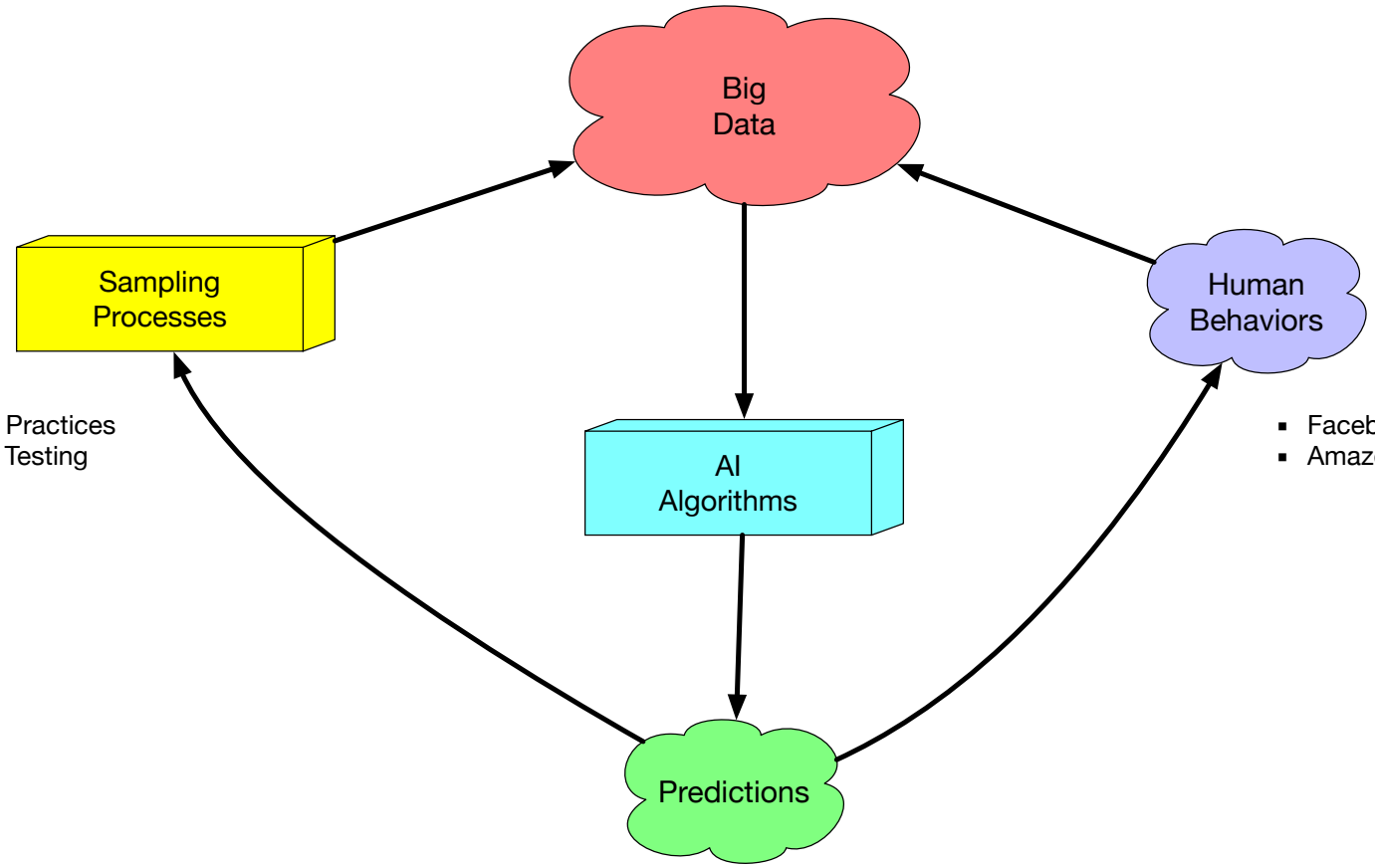


Algorithms = Opinions Embedded in Code

Properties of Algorithmic Weapons

- Opaque − Hard to understand
- Secret − The people who do understand them do not share that information.
- Socially influential − they affect people, often very asymmetrically.
- Questionable definitions of “success” (a.k.a. objective functions) such as e.g. maximizing profit, clicks, etc. These typically do not align with the user’s goals.
- Create dangerous feedback loops and self-fulfilling prophecies.

Dangerous (Positive) Feedbacks

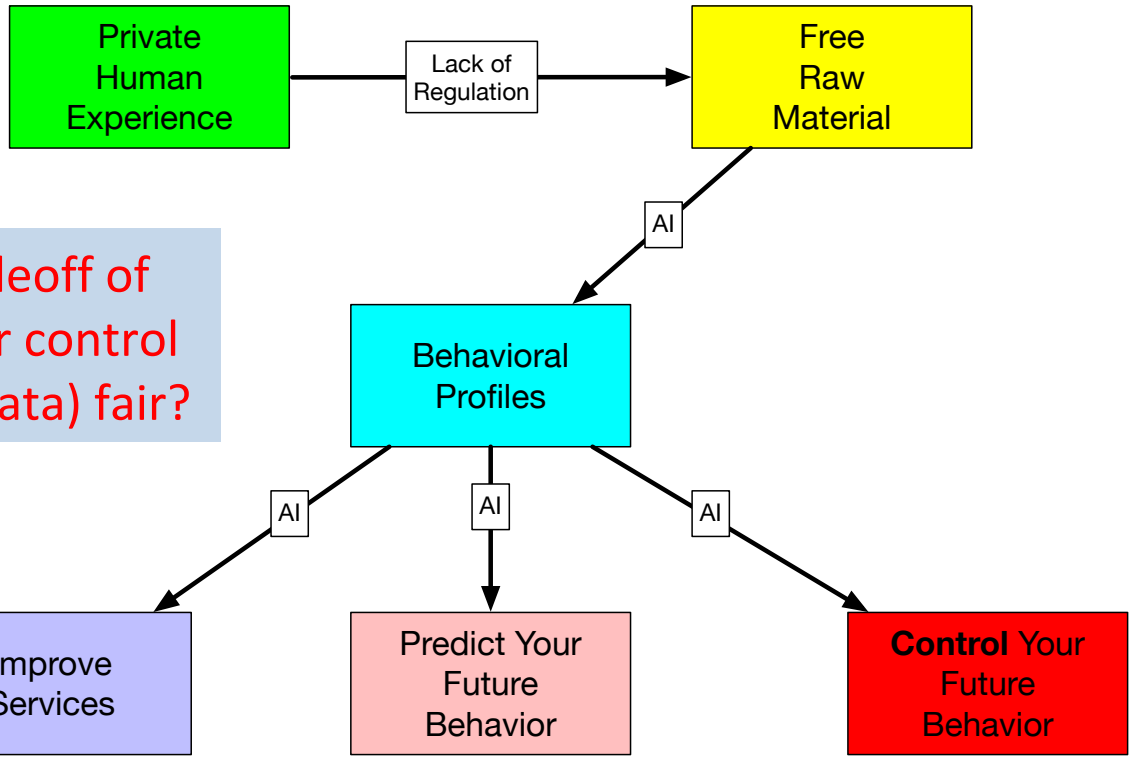


- Policing Practices
- Medical Testing

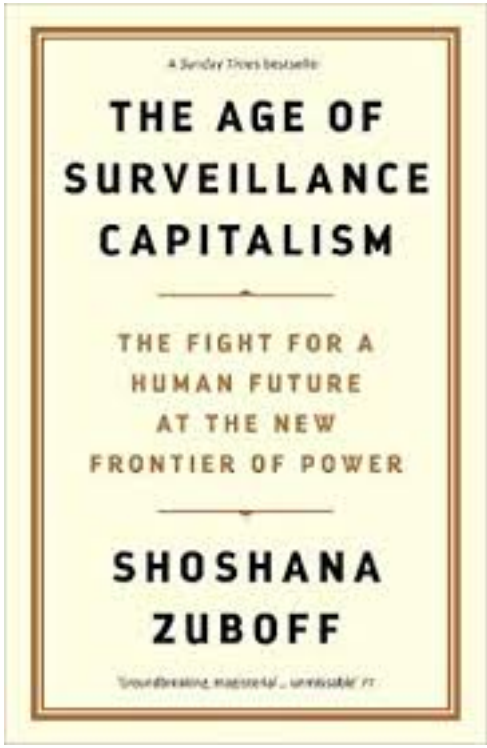
- Facebook News Feeds
- Amazon Book Purchases

Surveillance Capitalism

Money for nothing and your clicks for free.



Is the tradeoff of service for control (of your data) fair?



These predictions are traded in a new futures market, where surveillance capitalists sell certainty to businesses determined to know ***what we will do next****.*

In the competition for certainty, surveillance capitalists learned the most predictive data comes not just from monitoring but also from directing behavior.

Diminishing Transparency of AI

Showing Off

Nature

Military

Can we demand certain levels of transparency in AI R&D??

If so, who is privy to the info?

- Government
- UN
- Everyone

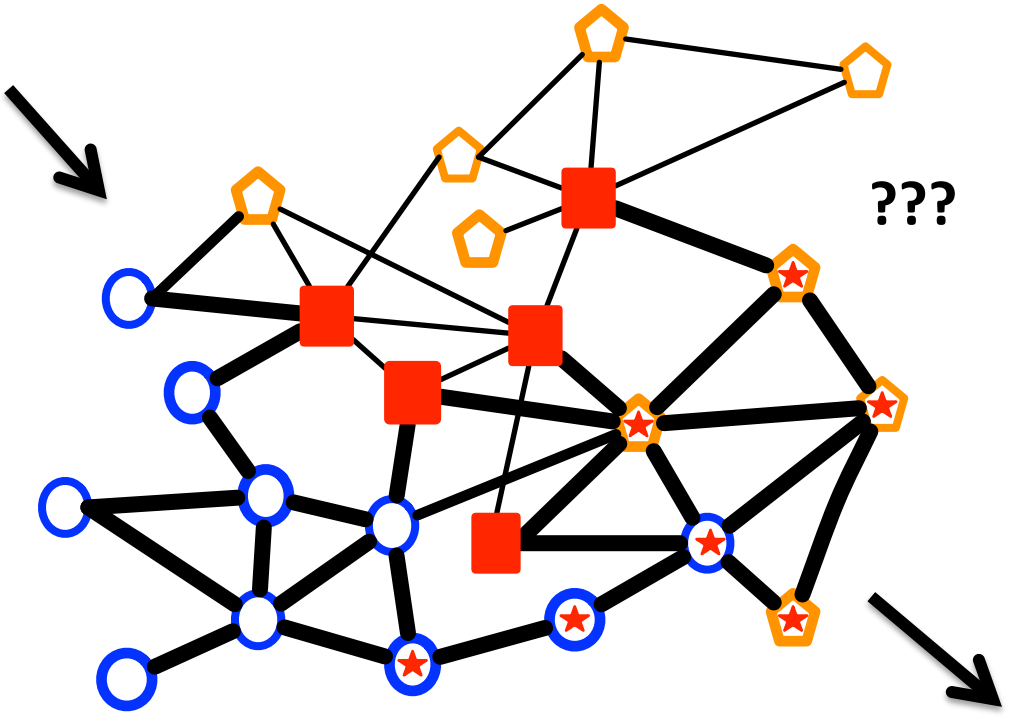
Academia

Financial Sector

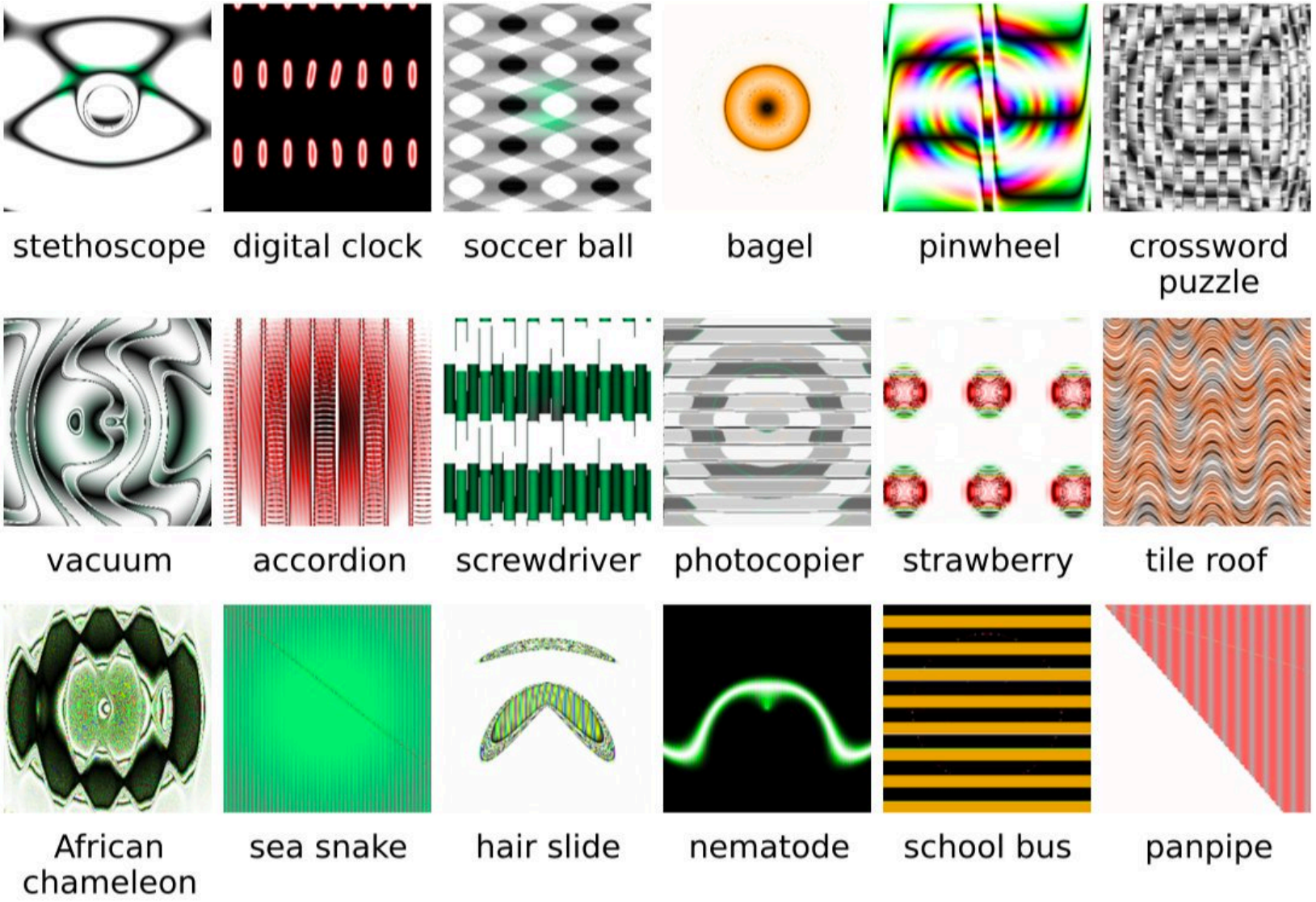
Translucent but not Transparent

Dissecting the Black Box

Competence without Comprehension (Dennett)



Evolution Deceives a Deep Learner

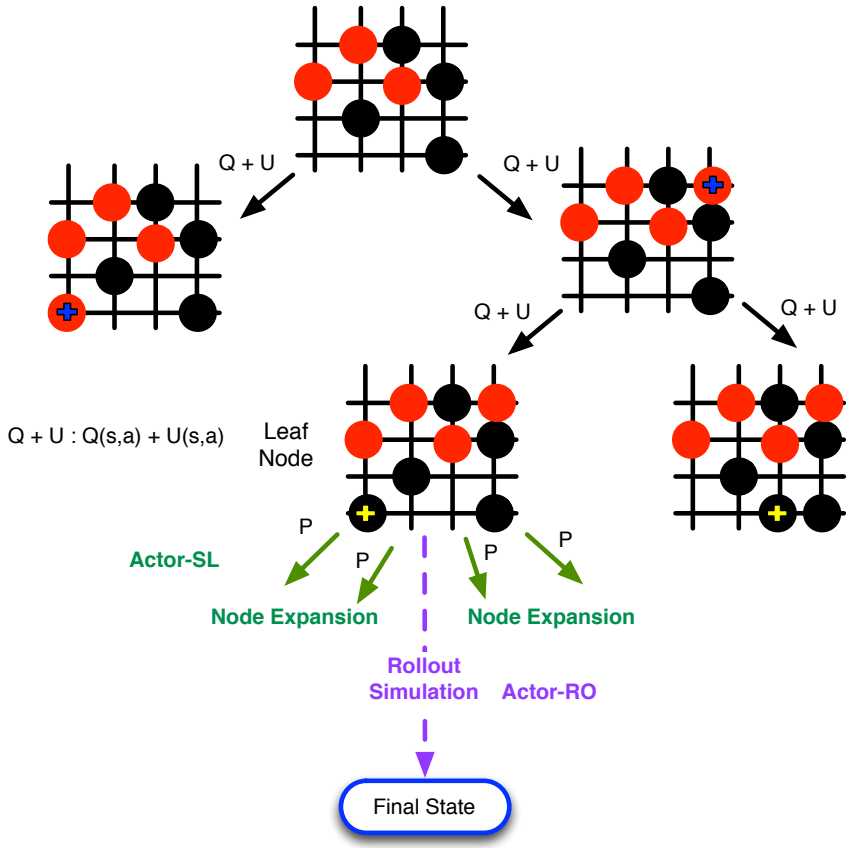


Deep Neural Networks are Easily Fooled. Nguyen, Yosinski & Clune (2015)

II. AI as Situated & Embodied Servant

AlphaGo

Silver et. al., *Mastering the game of Go with* ***deep neural networks*** *and* ***tree search***, Nature, 2016.



I guess I lost the game because I wasn't able to find any weaknesses…Lee Sedol (World # 2)

AlphaGo Zero & Alpha Zero

- No expert knowledge needed.
- Self-play is enough to become world champ

- Generalize AlphaGo Zero to other games.
- Becomes world champ at them.
- Another step toward AGI

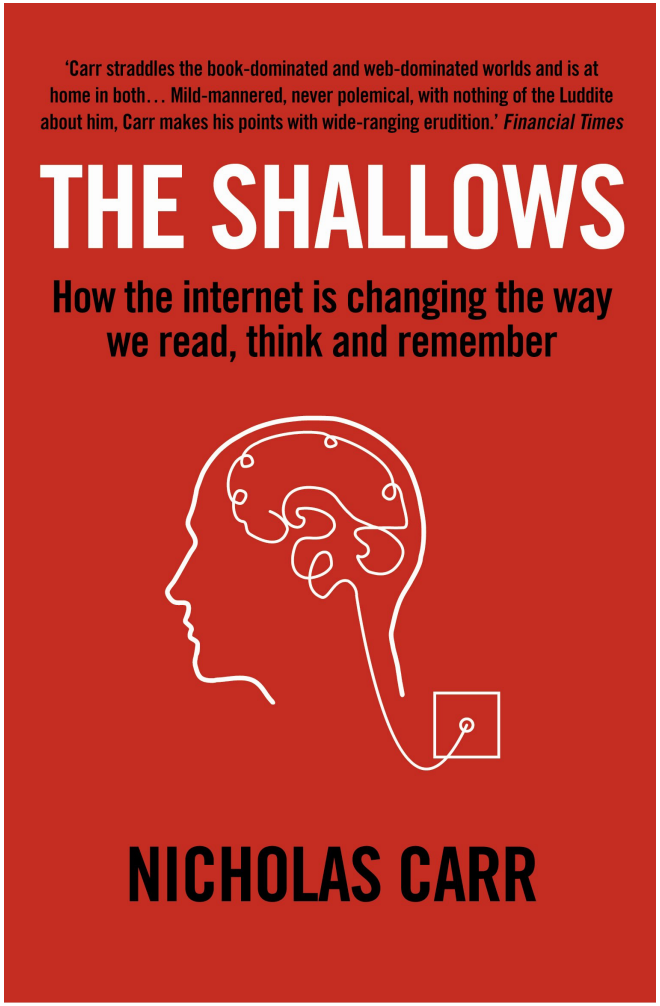
Humans Need Not Apply

- AI learns by itself. No need for human expertise.
- RL systems generate their own labeled datasets as they explore the world
- Humans are no longer the undisputed masters of pattern recognition
- Unbiased by humans => **Extremely Creative (Move 37)**

AI's Immediate Threats

- Job Loss
- Dumbing down of tech-dependent humanity

Robot-Run Warehouse



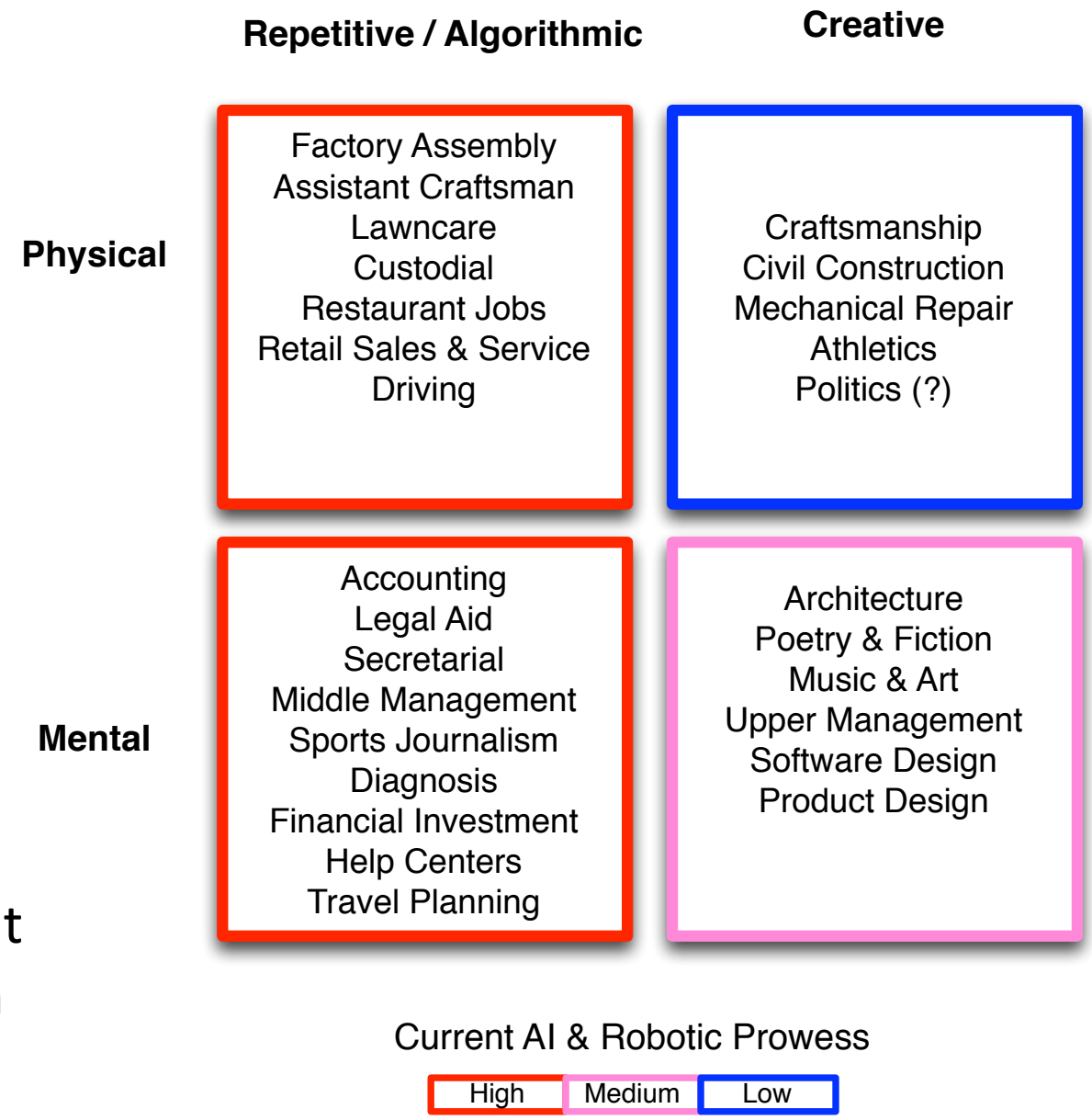
What is your value-add?

General advice:

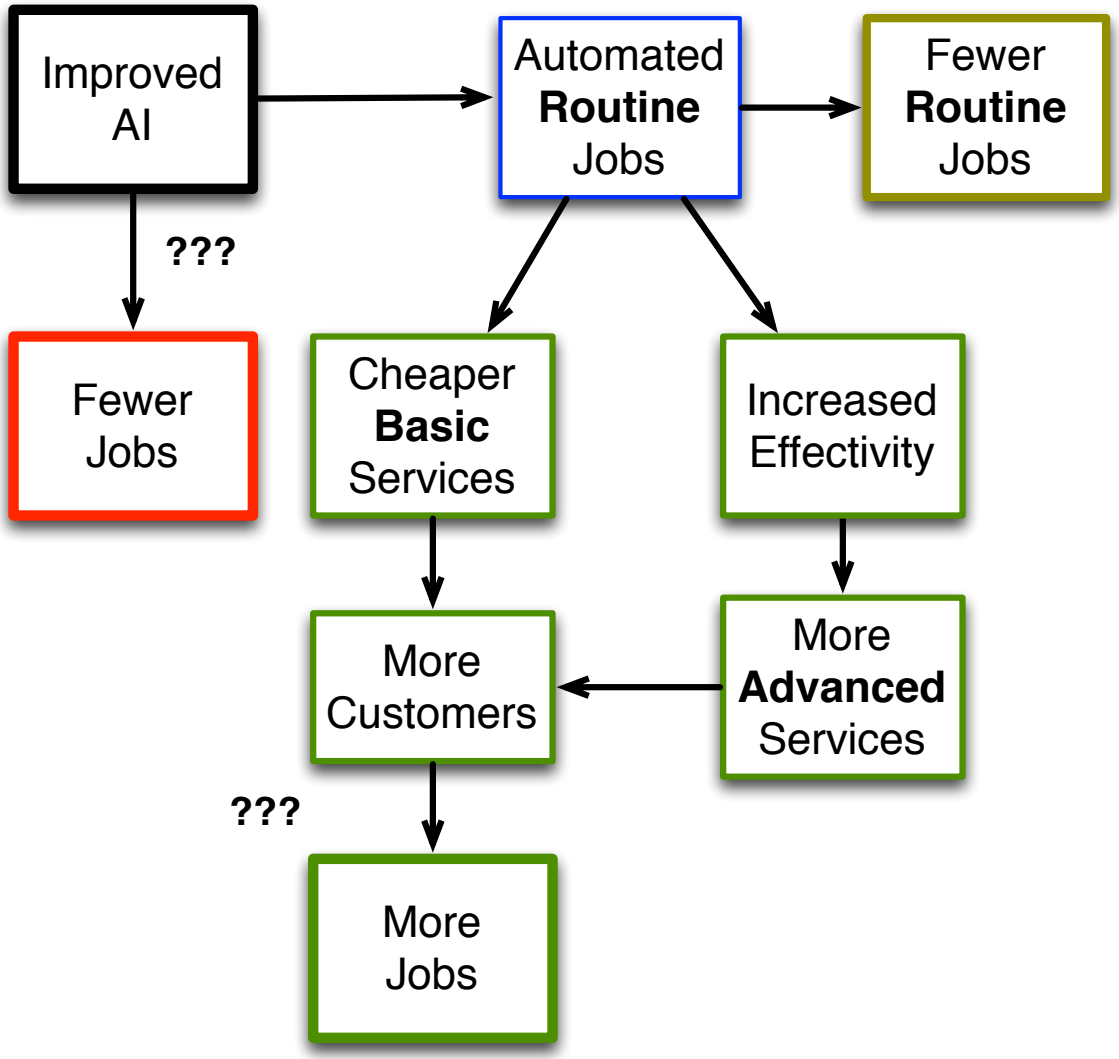
Know what AI can and cannot do!

Key personal traits:

- Deep, creative thought
- Empathy and emotion
- STEMpathy



Economics is not Rocket Science … Unfortunately



Plight of the Infovore

Constant distractions of cyberspace

"Shallows": Reduced depth of thought and emotion

Less Creative

Reduced value in an information society

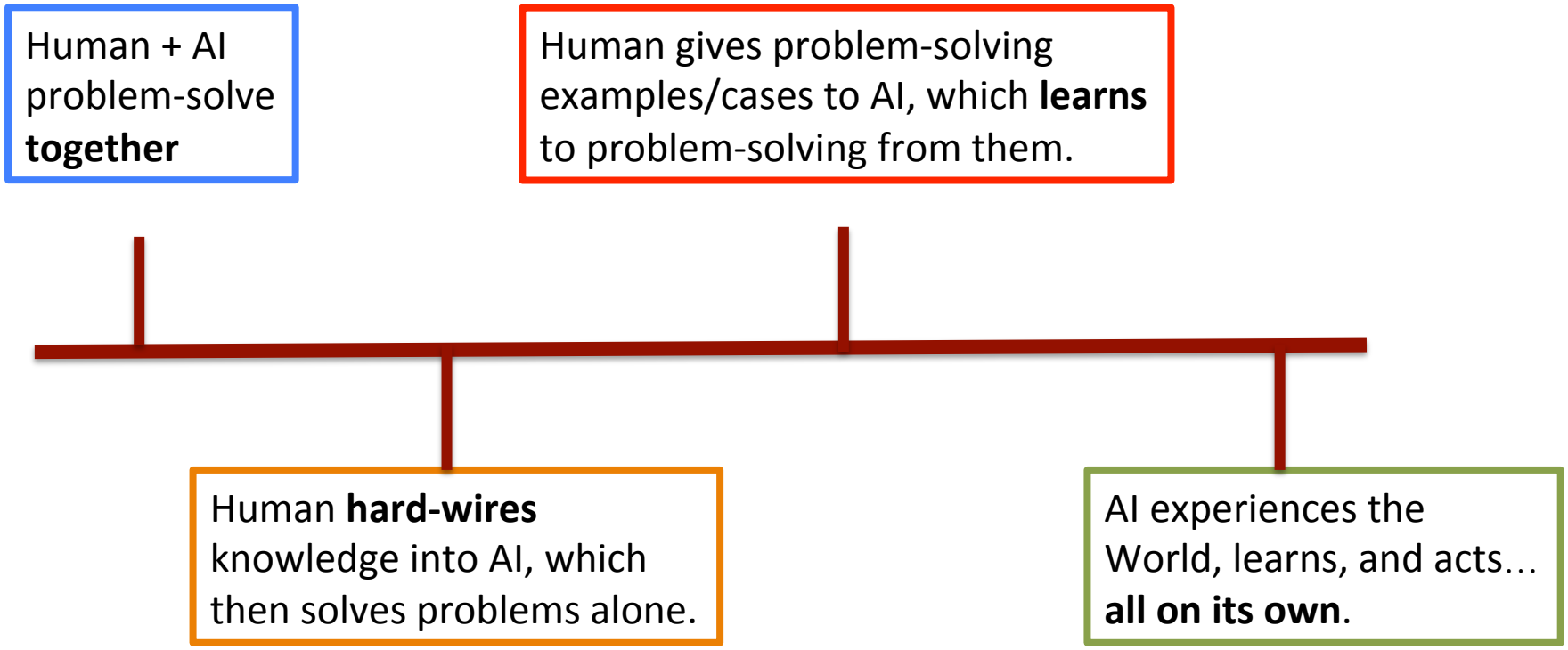
Rely on computers to understand the world

Rely on AI for **wisdom**

Automation Bias: Trust machines over ourselves.

Human intelligence becomes **artificial** … and thus more easily **predicted** and **controlled** by AI

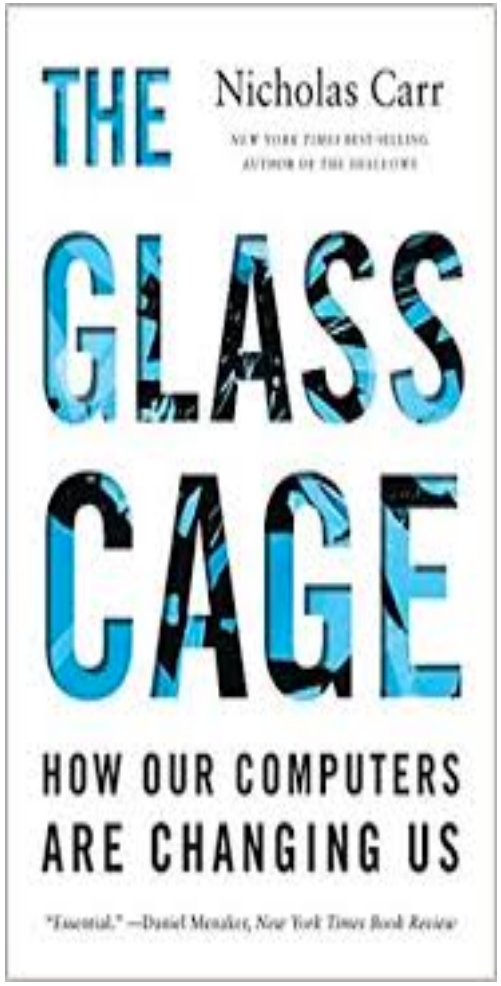
Human-AI Interaction Spectrum



- Should AI help us learn, not just do the job itself?
- For the good of humanity, shouldn't we **move back left** ?

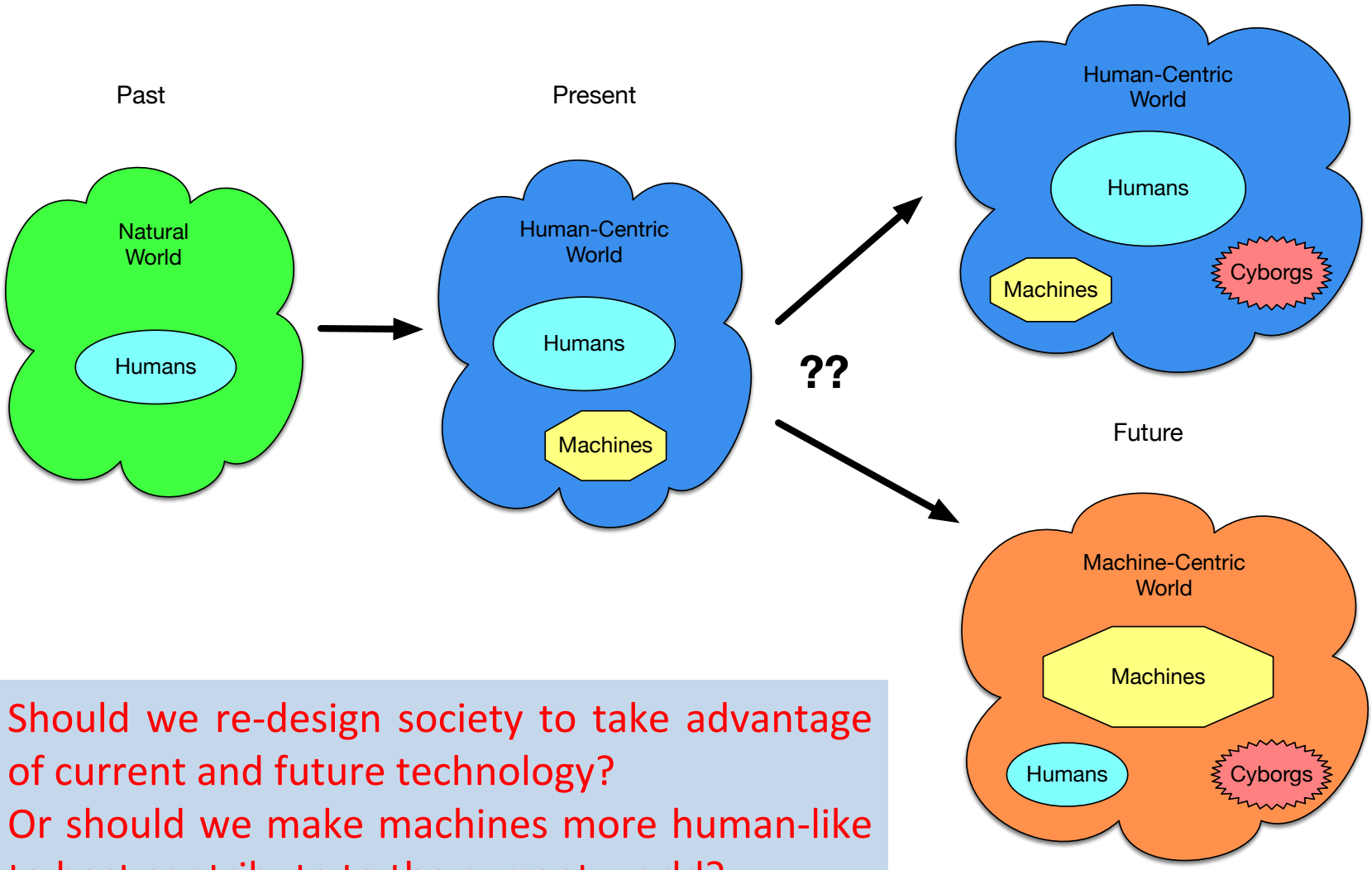
The Glass Cage (Nicholas Carr, 2014)

Reclaim our tools as instruments of ourselves, as instruments of ***experience*** *rather than just means of* ***production.***



Can (should) anything halt these general capitalistic forces that value production over the quality of human experience?

Re-Making our World

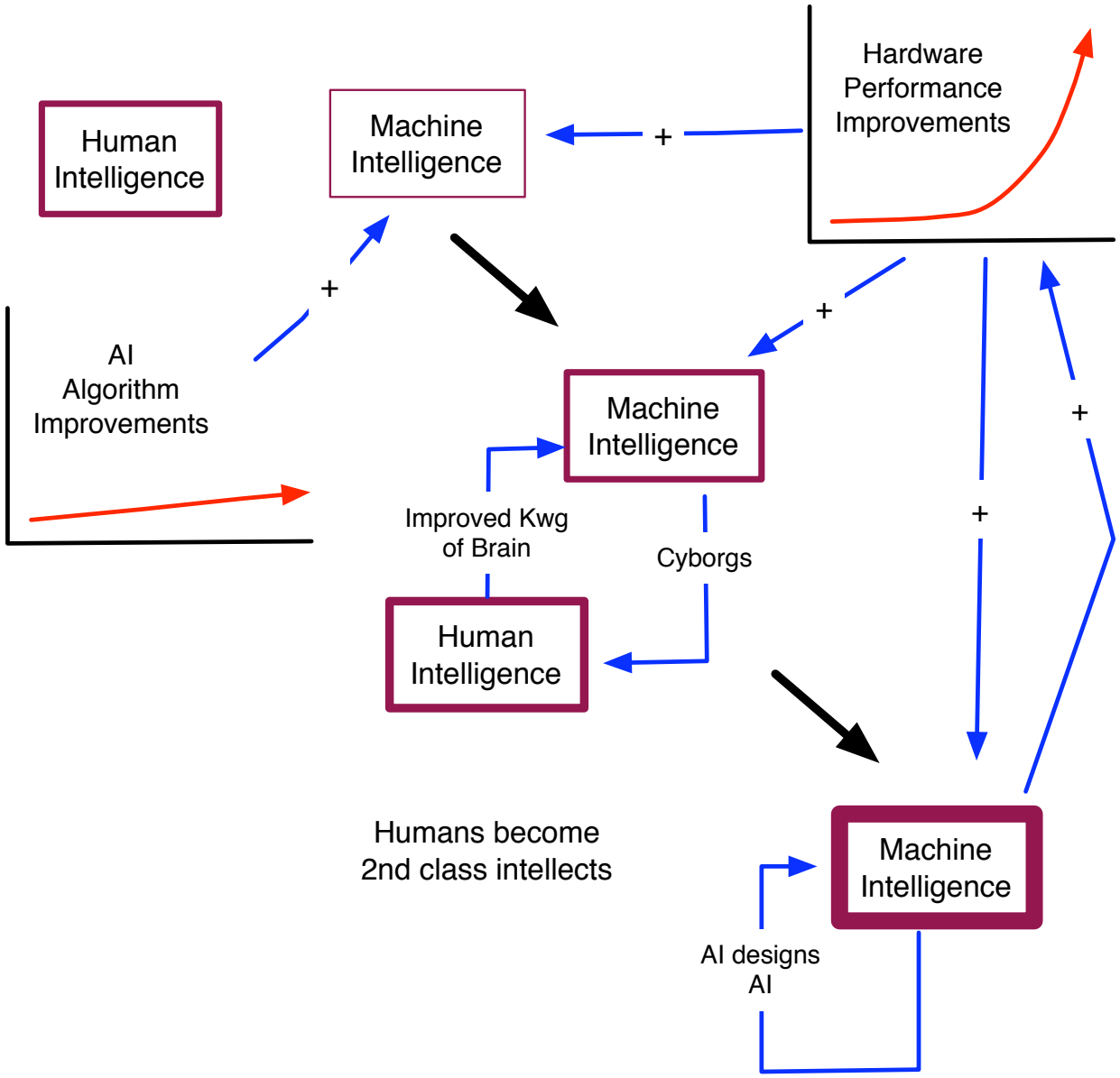


- Should we re-design society to take advantage of current and future technology?
- Or should we make machines more human-like to best contribute to the current world?

Re-designing environments to suit the current technology

III. AI as a Species

The Singularity (Vernor Vinge, Ray Kurzweil)



A point where normal expectations break down, where things become confusing, meaningless, and unpredictable (Joel Garreau, 2006).

4 Billion Years Ago (BYA)- Unicellular Life

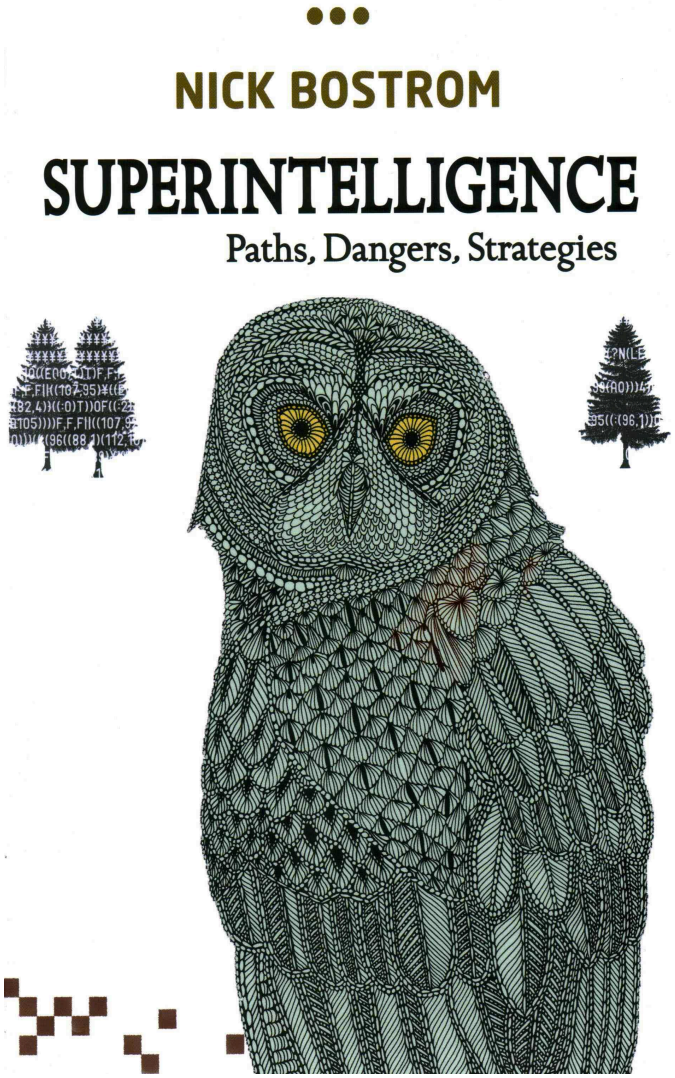
History of Life on Earth

2 BYA - Photosynthesis

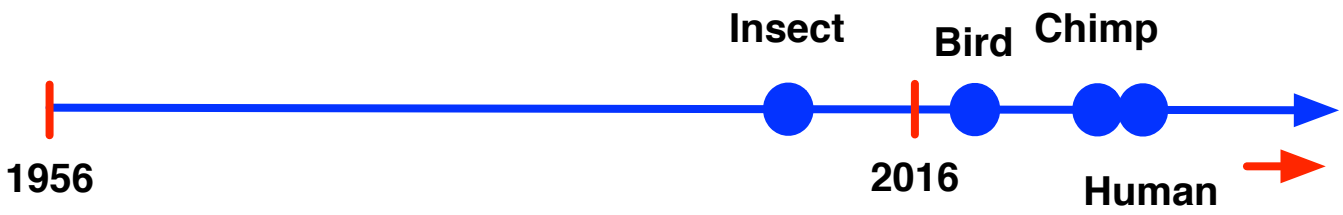
900 Million Years Ago (MYA) - Multicellular Life

500 MYA - Chordates
400 MYA - Insects
200 MYA - Birds

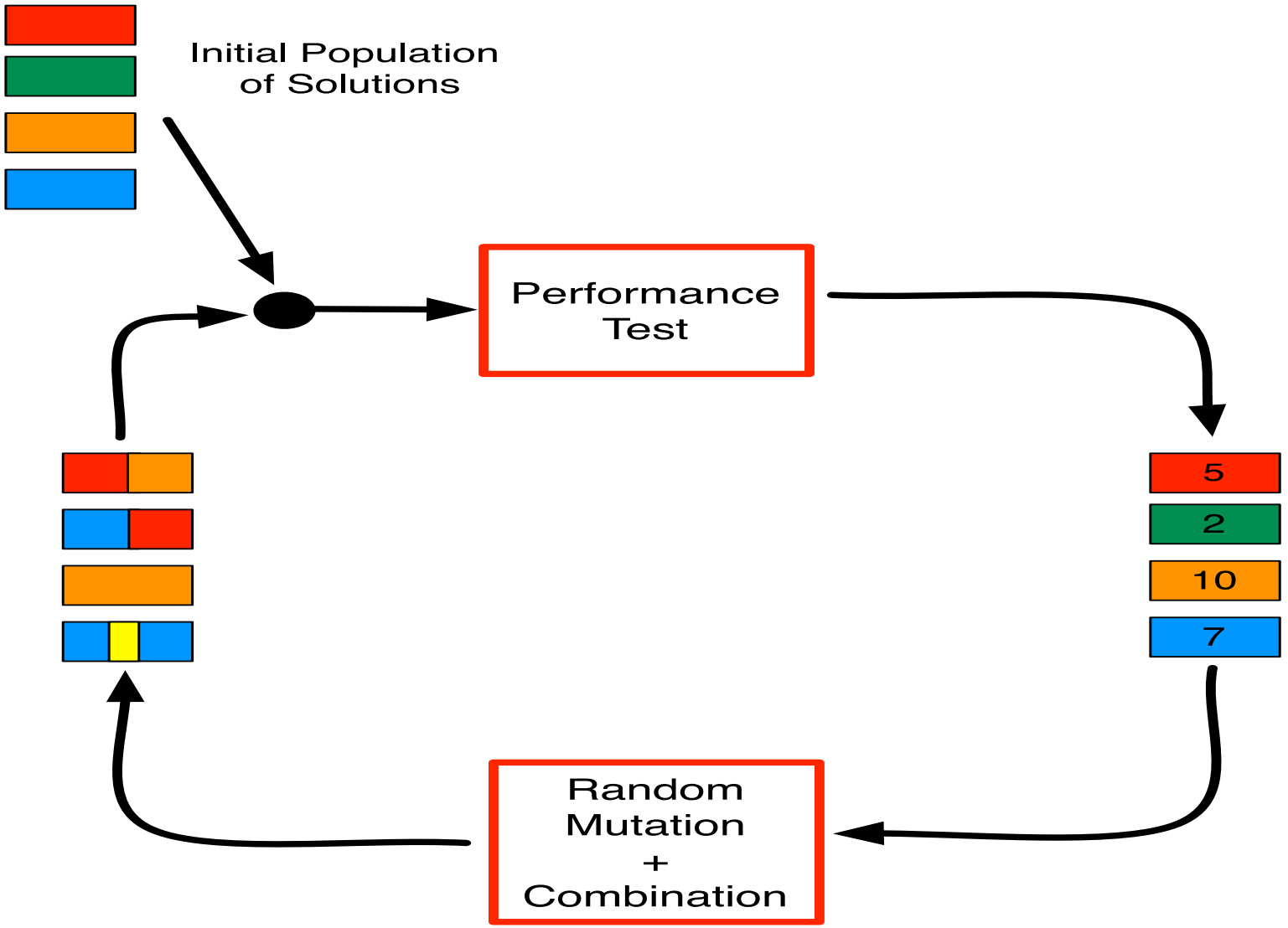
140 MYA - Mammals
75 MYA - Primates
6 MYA -Humans



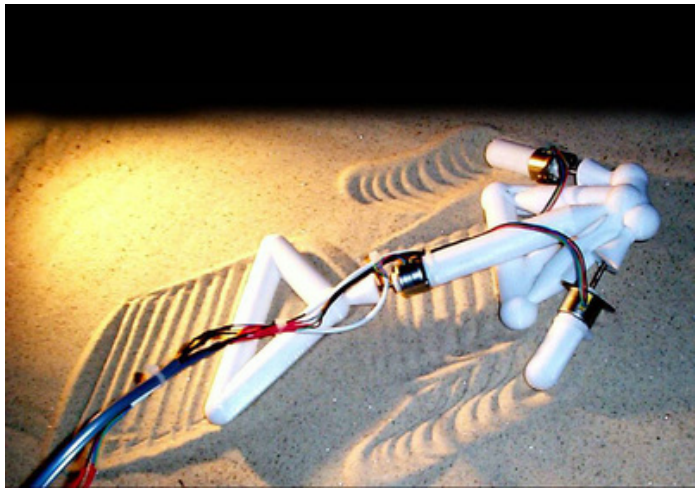
AI Timeline



Evolutionary Computation



Evolutionary Robotics



Evolving robotic
bodies + brains

Hod Lipson's
Creative Machines Lab
(Columbia University)

Survival of the Fittest

Artificial Selection of AI

- Competence
- Adaptability
- Obedience
- Focus on our goals

Natural Selection of AI

- Self-preservation
- Replication
- Focus on its goals

Problems Preserving Artificial Selection of AI

Our Goals

- Properly defining them (ethics)
- Risk-free embodiment in AIs
- Preventing AI from developing new goals that supersede ours.

We gradually give technologies more and more control over our decision-making.

Should we regulate AI's decision-making autonomy?

Birth Control

Who controls the resources needed for AI preservation and reproduction?

- Humans
- Nobody in particular
- AI

- Today humans have:
  - Good control of Hardware
  - Weaker control of Software

Should humans ever treat AI as a species…or always as just an extremely advanced digital tool?