WORLD SCIENCE FORUM BUDAPEST

Evolvable AI as an Existential Threat to Humanity

Eörs Szathmáry

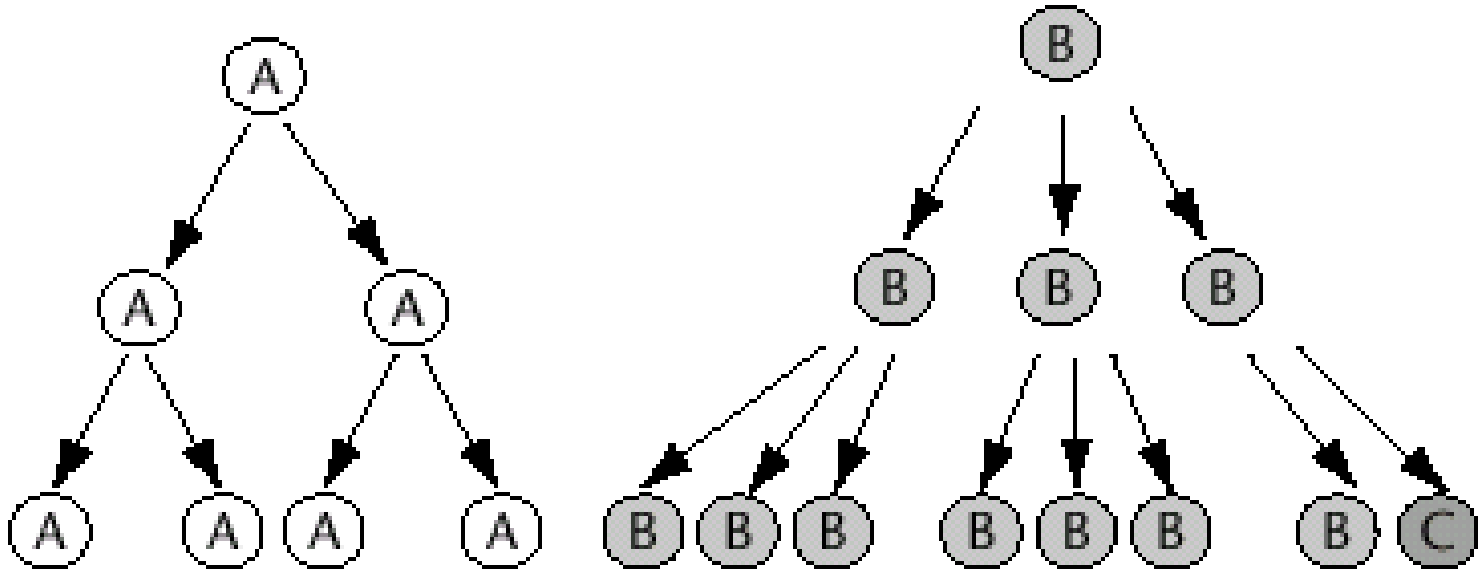
PARMENIDES FOUNDATION

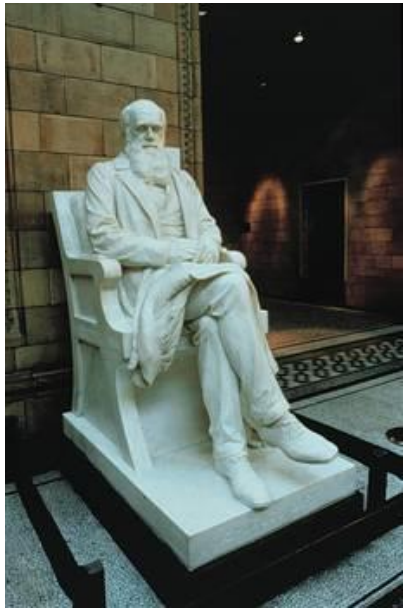
ELTE EÖTVÖS LORÁND UNIVERSITY

INSTITUTE OF EVOLUTION

HUN-REN Hungarian Research Network

Units of evolution and a tacit algorithm





1. Multiplication
2. Heredity
3. Variability

Some of the hereditary traits affect survival and/or fertility → fitness

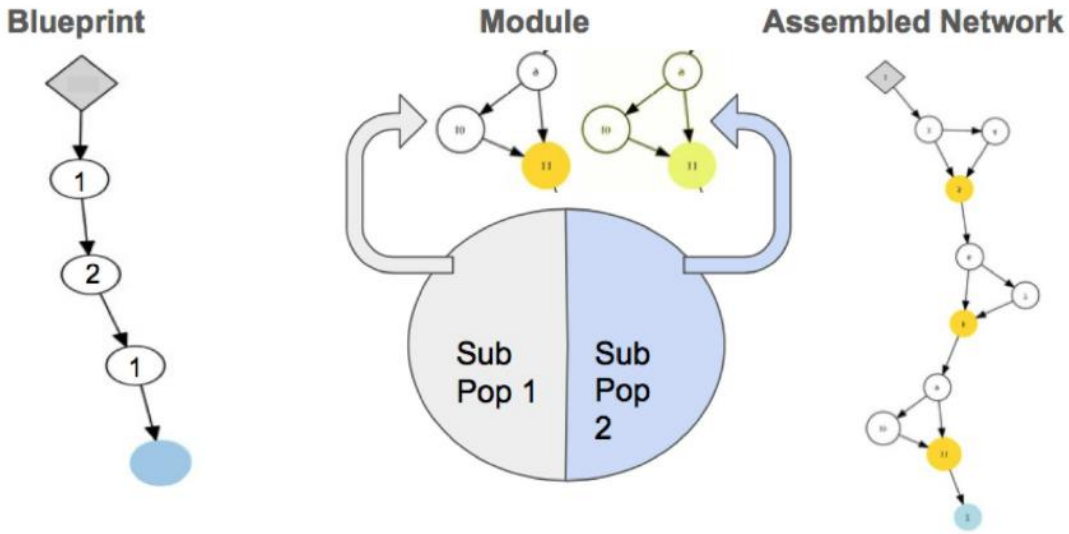
The power of evolution

1. Antenna shapes
2. Evaluation
3. Keeping highest-scoring designs
4. Iteration: mutation, recombination, crossover
5. The resulting antenna often outperforms the best manual designs, because it has a complicated asymmetric shape that could not have been found with traditional manual design methods.

Artificial Life (AL): Tierra

- These **digital organisms** can reproduce, mutate and evolve in a virtual environment, demonstrating the principles of natural selection and evolution
- Simple self-replicating codes **have evolved into more complex forms,** adapting to the constraints and possibilities of their digital environment, demonstrating how simple rules can evolve into complexity
- The emergence of **parasites and hyperparasites**
- Tierra's digital organisms competed for **limited CPU time and memory space,** similar to the competition for resources in natural ecosystems
- The project highlighted the **coevolutionary dynamics:** the evolution of one type of digital organism influenced the evolution of others.

Evolving deep neural networks



- Two types of chromosome
- NEAT within modules
- Topology and hyperparameter optimization
- Discovery of partially repetitive structures

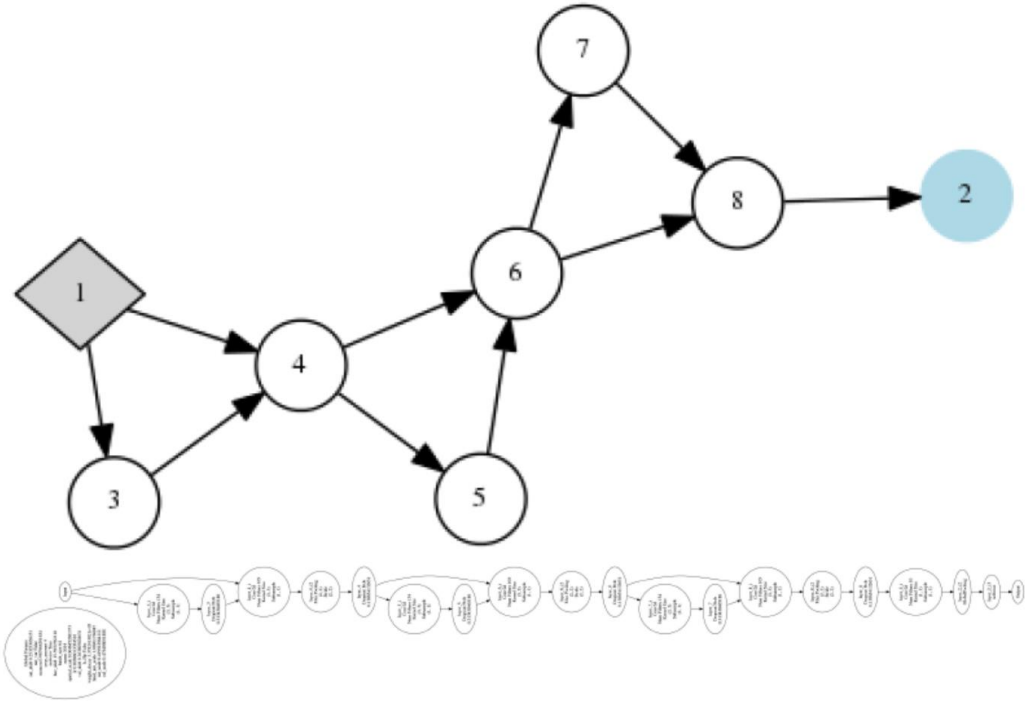


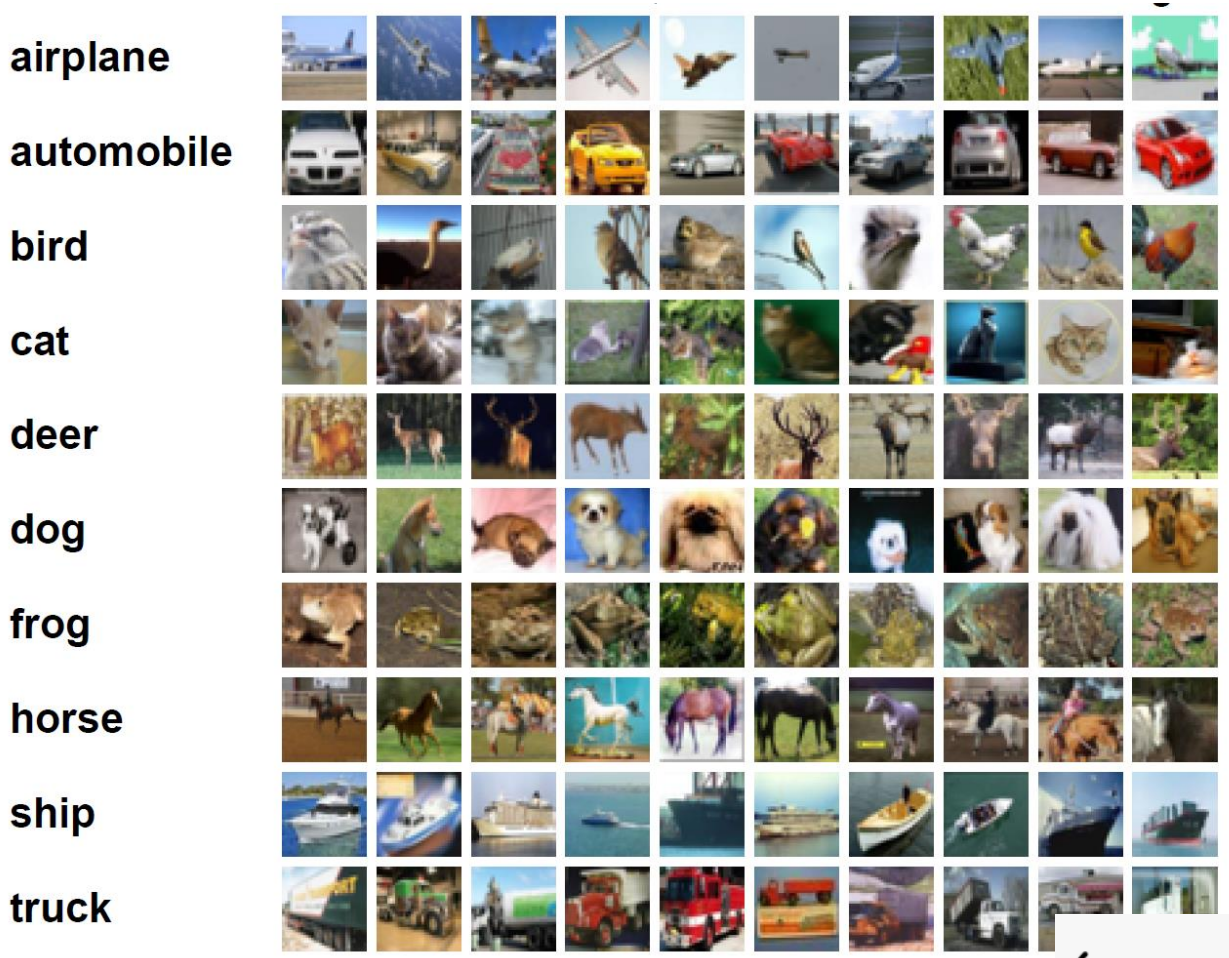
Image classification example

Component functions and the task

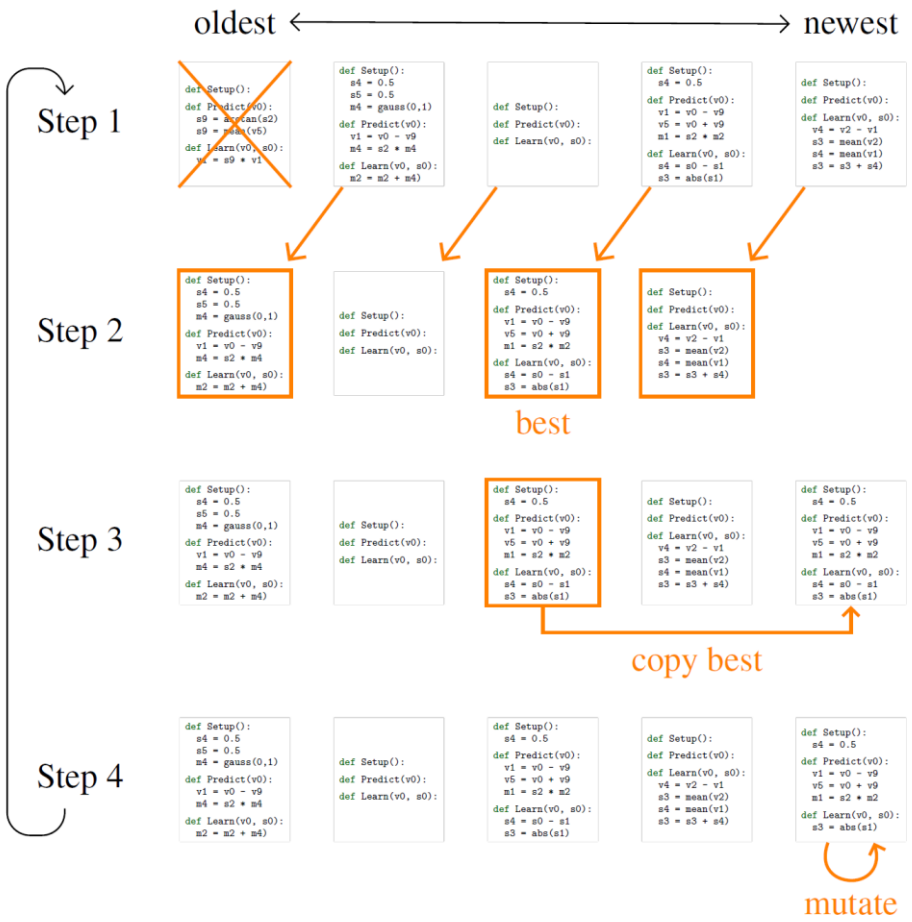
- Only high-school mathematics is provided!

1. Setup
2. Predict
3. Learn

- Culling of functionally identical and useless algorithms
- Tournament selection
- Migration



Evolvable AI




```
def Setup():
  s4 = 0.5
def Predict(v0):
  m1 = s2 * m2
def Learn(v0, s0):
  s4 = s0 - s1
  s3 = abs(s1)
```

parent

child

Type (i)

```
def Setup():
  s4 = 0.5
def Predict(v0):
  m1 = s2 * m2
def Learn(v0, s0):
  s4 = s0 - s1
  s2 = sin(v1)
  s3 = abs(s1)
```

```
def Setup():
  s4 = 0.5
def Predict(v0):
  m1 = s2 * m2
def Learn(v0, s0):
  s4 = s0 - s1
  v3 = abs(s1)
```

Type (ii)

```
def Setup():
  s4 = 0.5
def Predict(v0):
  m1 = s2 * m2
def Learn(v0, s0):
  s0 = mean(m1)
  s5 = arctan(s7)
```

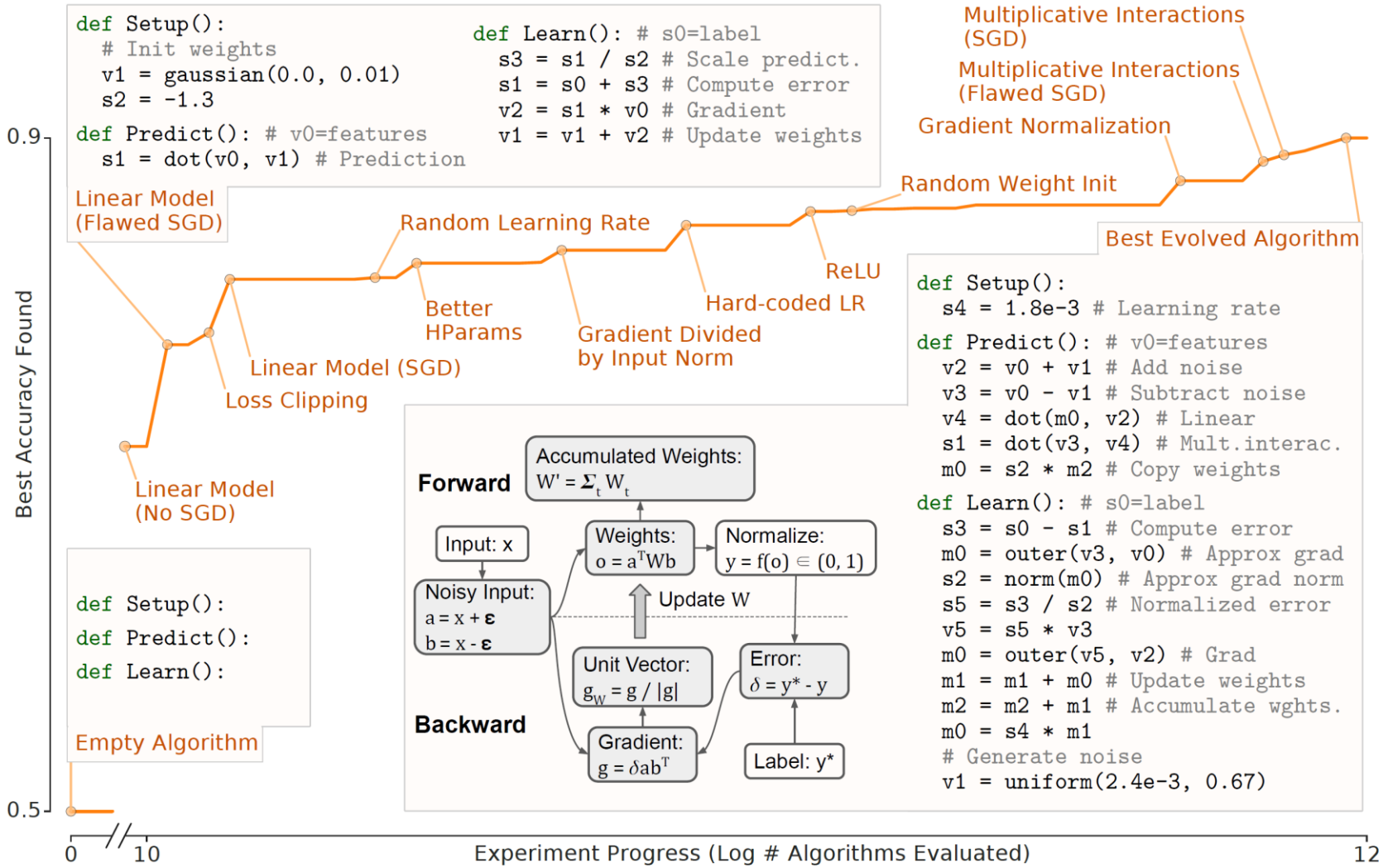
```
def Setup():
  s4 = 0.5
def Predict(v0):
  m1 = s2 * m2
def Learn(v0, s0):
  s4 = s0 - s1
  s3 = abs(s1)
```

Type (iii)

```
def Setup():
  s4 = 0.5
def Predict(v0):
  m1 = s7 * m2
def Learn(v0, s0):
  s4 = s0 - s1
  s3 = abs(s1)
```

Real, E., Liang, C., So, D. & Le, Q.. (2020). AutoML-Zero: Evolving Machine Learning Algorithms From Scratch. Proceedings of the 37th International Conference on Machine Learning, in *Proceedings of Machine Learning Research* **119**:8007-8019.

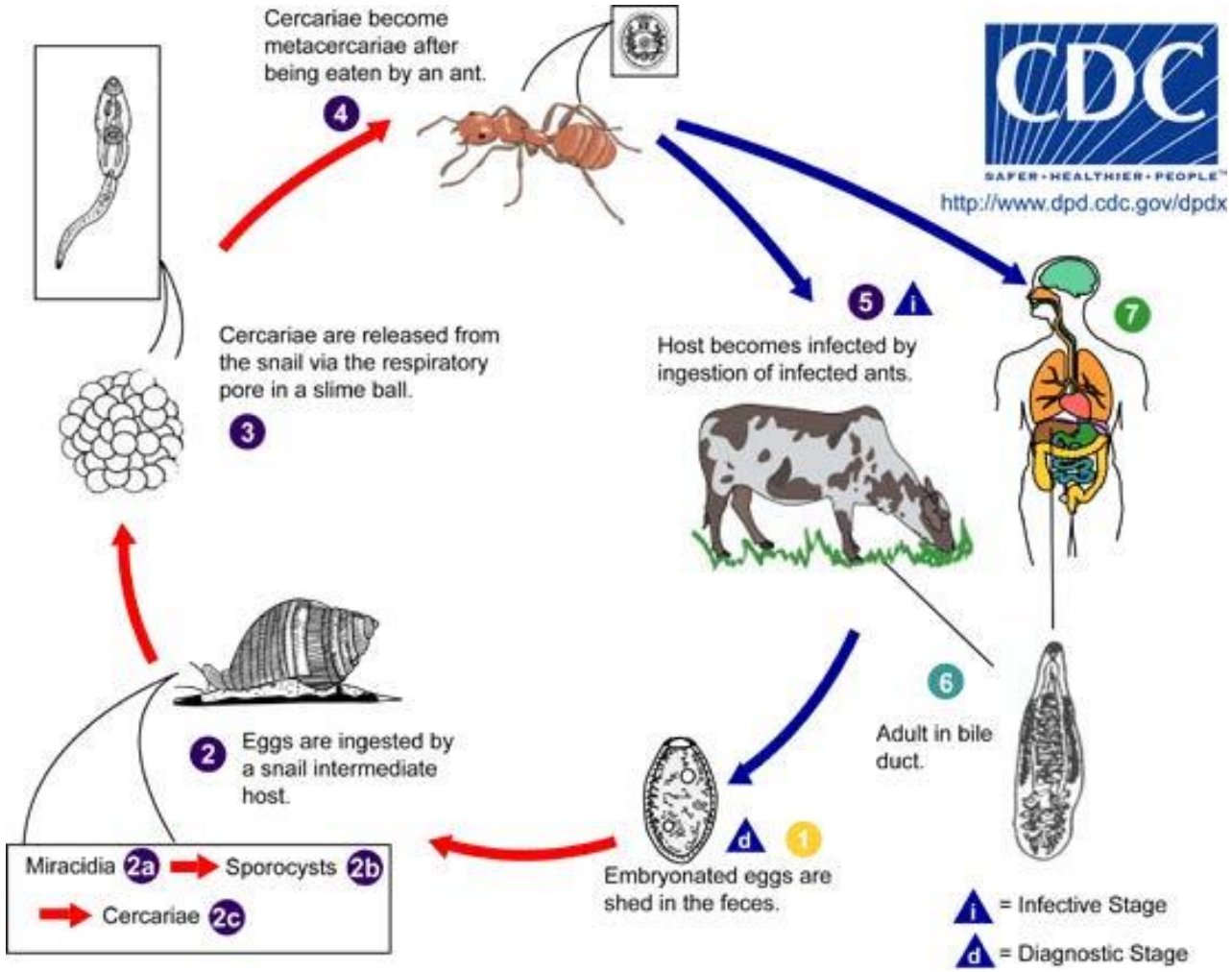
Evolution of the solutions: decades of AI development rediscovered!



Wherever Darwinian dynamics appear, self-interest also prevails

- The "selfish replicator"
- Teleological nature
- The goals of replicable AI and humans may coincide, but this is by no means necessary
- AI programs can write new programs
- They can learn and replication is possible
- A replicator can be very dangerous even if it is not conscious!

A tale of the liver fluke



- Complex lifecycle: transits different hosts
- Manipulates a much more compex cognitive system

WORLD SCIENCE FORUM BUDAPEST

MI Armaggedon: the risk of intelligent viruses

- Computer viruses can sometimes modify themselves to avoid detection by anti-virus software, making them more challenging to detect.
- AI continues to advance. Its capabilities now include writing software — and that includes creating viruses.
- Any day, a stupid or malicious human could instruct an AI to develop a virus that is itself an AI.
- This creation could in turn become the forebear of a new breed of intelligent viruses that replicate and create new improved versions of itself.
- *A self-replicating computer program that knows we want to erase it and benefits from proliferation should genuinely scare us all.*

AI Armageddon: The Risk of Intelligent Viruses | by Casper Skern Wilstrup | Machine Consciousness | Medium

Things can happen terribly fast

- Ernest Rutherford, morning of 12th September 1933 (when the newspaper article appeared): *'he who sought the source of power in the transformation of atoms spoke of the moonbeam'*
- Leo Szilárd, afternoon 12th September 1933: he laid the foundations of the theory of the nuclear fission chain reaction.
- New York Times, October 9th, 1903: *"It is assumed] that the flying machine, which will really fly, may be evolved by the combined and continuous efforts of mathematicians and mechanics in a million to ten million years... No doubt the problem has its attractions for those who are interested, but to the average person it seems as if the effort could be put to more useful use.*
- Wright brothers' first successful flight: 17th December 1903.

What can we do?

- Sterilisation: do not allow fully autonomous reproduction of artificial intelligence!
- Limit deviation from the state of production!
- Let's select AI for the "love" of humans!
- Let us select for self-sacrifice!
- Let's create a reward center in AI!
- Strictly regulate the use of frameworks that enable AI to be developed!
- A professional "epidemic" of digital pests. Scientific study of artificial intelligence!

Thanks for your attention!

WORLD SCIENCE FORUM BUDAPEST