Handout for honours seminar talk on AIXI[*]

Yuxi Liu[†]

Monday 20th May, 2019

Game of Life

Life is a game:

$$\begin{tikzcd}\text{me} \arrow[r, "\text{actions}"] & \text{the world} \arrow[l, bend left, "\text{events}"]\end{tikzcd}$$

There are two players: The world acts without desires. I act with desires.

See - Think - Act

See

- a is **action**.
- $e = (o, r)$ is **event** from **environment**, containing **observation** and **reward**.
- $æ = ae$ is one **round** of the game of life.
- $æ_{<t} = æ_{1:t-1} = a_1 e_1 \cdots a_{t-1} e_{t-1}$ is all **history** from round 1 to $t-1$.
- N is **horizon**, or length of the game.
- $R(æ_{1:N}) = r_1 + \cdots + r_N$ is **total reward** in life.

Beat the highscore, maximize $R(æ_{1:N})$.

Think

Metaphysics before physics.

Epicurus (300s BC): "Keep all hypotheses that are consistent with the facts."
Ptolemy (100s): "We consider it a good principle to explain the phenomena by the simplest hypothesis possible." (Occam's Razor)
Thomas Bayes (1760s):

$$P(H|E) = \frac{P(E,H)}{P(E)} = \frac{P(E|H)P(H)}{\sum_i P(E|H_i)P(H_i)}$$

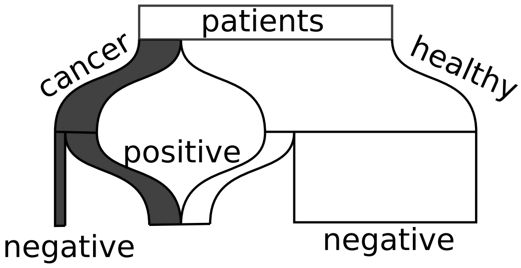

Figure 1: Bayes rule in cancer testing.

I like to interpret it as ”weighting the **multiverses**”.
Alan Turing (1930s): Everything calculable by a machine is calculable by a Turing machine.
Ray Solomonoff (1964): Predict using all consistent Turing machines, weighted by description length.

- p is the **program** run by the environment.
- $p(a_{1:t}) = e_{1:t}$ says that the program, given the action history $a_{1:t}$, replies with the environmental history $e_{1:t}$
- $\ell(p)$ is **length** of program.
- $$M(æ_{1:t}) = \sum_{p:p(a_{1:t})=e_{1:t}} 2^{-\ell(p)}$$

  is the probabilistic **mass** of all the multiverses where, given that I played $a_{1:t}$, the world replied with $e_{1:t}$.

[*] pdf at https://yuxiliu1995.github.io/notes/
[†] Honours student at ANU.

Act

John von Neumann, Oskar Morgenstern (1947): Maximize the expectation of reward.
Marcus Hutter (2000s): Intelligence measures an agent's general ability to achieve goals in a wide range of environments.

AIXI

Proposed by Marcus Hutter (professor at ANU, researcher at DeepMind), around 2000.
At final round: maximize expected $R(\text{æ}_{1:N})$:

$$\begin{aligned} a_N^* &= \operatorname*{argmax}_{a_N} \mathbb{E}[R(\text{æ}_{1:N}) | \text{æ}_{1:N-1} a_N] \\ &= \operatorname*{argmax}_{a_N} \sum_{e_N} R(\text{æ}_{1:N}) \frac{M(\text{æ}_{1:N})}{M(\text{æ}_{1:N-1})} \\ &= \operatorname*{argmax}_{a_N} \sum_{e_N} R(\text{æ}_{1:N}) M(\text{æ}_{1:N}) \end{aligned}$$

In general, at round t,

$$a_t^* = \arg \left(\max_{a_i} \sum_{e_i} \right)_{i=t}^{N} R(\text{æ}_{1:N}) M(\text{æ}_{1:N})$$

Why AIXI?

Artificial General Intelligence (AGI): The game of life is hard. Make someone who's better at the game.
AIXI is self-optimizing, Pareto-optimal, and has maximal intelligence. A mathematically precise **gold standard** for AGI.
It's not Turing computable, but it is approximately so.

Inspirational hyperboles(?)

John von Neumann (1950s): Accelerating progress of technology appears to approach an essential singularity in history, beyond which we cannot predict.
Irving Good (1964): The first ultraintelligent machine is the last invention that human need ever make.
Hugo de Garis (1990s): It would be a cosmic tragedy if humanity freezes evolution at the puny human level.
Nick Bostrom (2014): We are probably the stupidest possible biological species capable of starting a technological civilization.

Further reading

- [Bos14] Standard reference on super AI.

  New York Times bestseller.

- [Hut17] Online AI course by Marcus Hutter, archived at the Internet Archive.

- [Hut05] Standard reference on AIXI.

  Has online page `http://www.hutter1.net/ai/uaibook.htm`.

- [LH07] General definition of intelligence.

- [Leg08] PhD thesis on super AI, by Shane Legg, student of Marcus Hutter, cofounder of DeepMind.

References

[Bos14] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press, 2014.

[Hut05] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability.* Springer, Berlin, 2005.

[Hut17] Marcus Hutter. Advanced Topics in Artificial Intelligence COMP4620/COMP8620. `https://web.archive.org/web/20180821153654/https://cs.anu.edu.au/courses/comp4620/2017.html`, 2017.

[Leg08] Shane Legg. *Machine super intelligence.* PhD thesis, Università della Svizzera italiana, 2008.

[LH07] Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17(4):391–444, 2007.