Legg-Hutter Intelligence

James Oswald

Class Agenda

- Discussion about the definition of intelligence
- Synthesis of the informal definition of Legg-Hutter intelligence
- Formalizing Legg-Hutter intelligence
    - Formalizing Agent-Environment Systems
    - Formalizing Ability
    - Formalizing Environmental Complexity
        - Kolmogorov Complexity
        - Algorithmic Probability
- Examples of Legg-Hutter Intelligence
- Issues & Criticisms of Legg-Hutter Intelligence

Background

One of the leading formal theories of intelligence.

Some of Shane Legg’s PhD work from 2007.

Legg went on to Co-found Deepmind in 2010, where he currently works as the chief AGI scientist. Hutter was invited to Deepmind in ~2019.

Google DeepMind

Hutter

Legg

What is Intelligence?

“A fundamental problem in artificial intelligence is that nobody really knows what intelligence is.”
-Legg & Hutter (2007)

Discussion

Take 5 minutes:

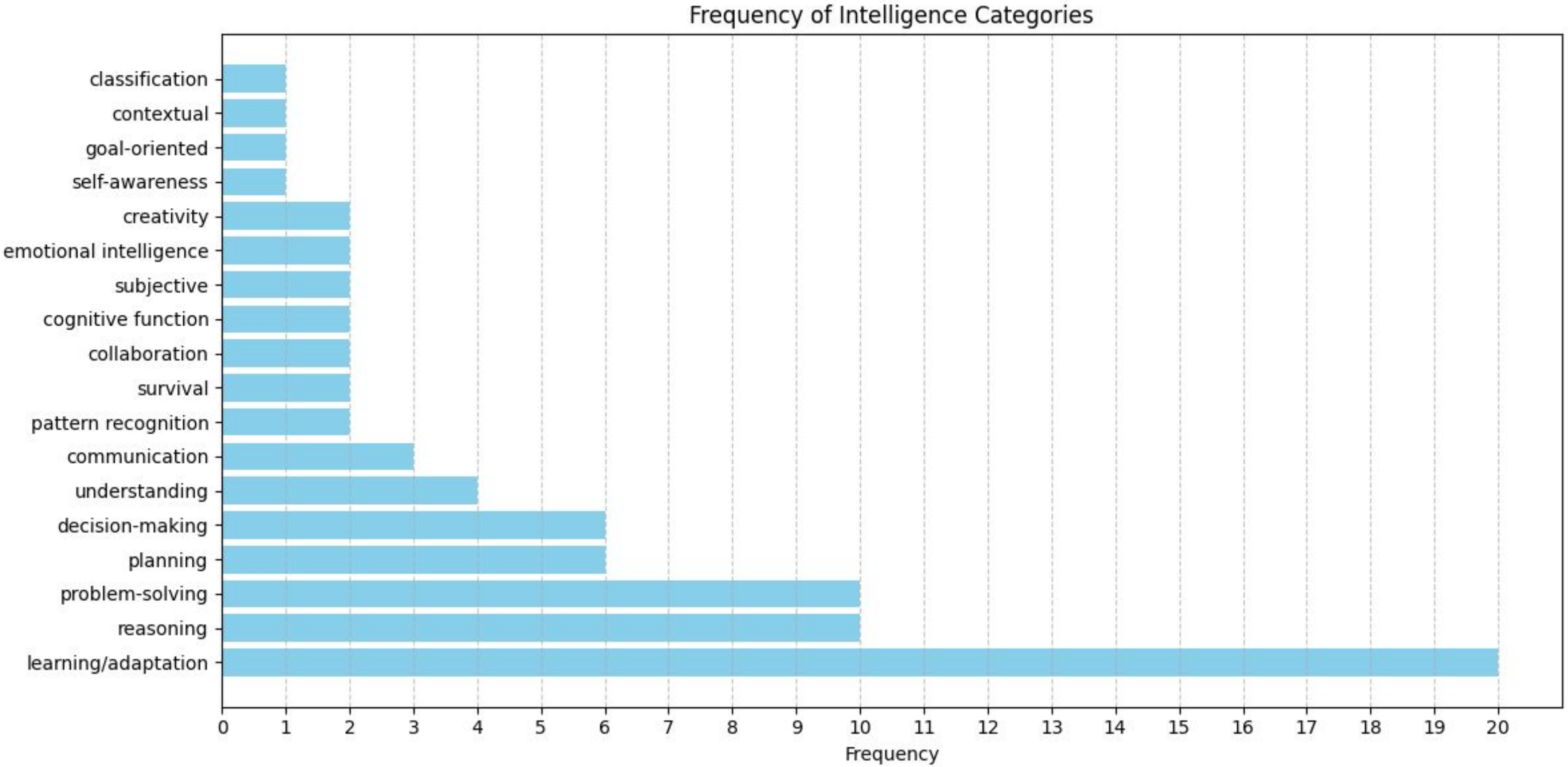
Think of your own definition of intelligence, make it as general as possible. If you need help, start from an *ostensive definition* (list things you consider to be intelligent) and try and find properties they share to come up with an *intensional definition* (where you specify properties required for intelligence).

Consider edge cases: Does your definition take…

- animals into account? Does it cover dolphins, rats, bacteria?
- AI into account? Does it cover basic neural networks? ChatGPT?
- order into account? We commonly say some things are “more intelligent”.
- collective systems into account? Bee swarms, human corporations?

Email me your definition at oswalj@rpi.edu with subject line: LastName Int Def

Core Themes in the IBLAI Fall 2024 Poll



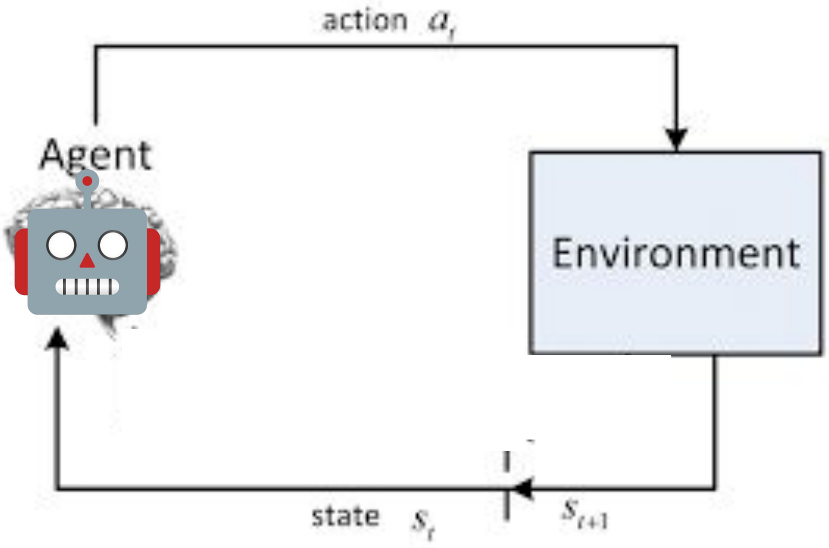
Famous Informal Definitions of Intelligence

1. “[Intelligence is] judgement, otherwise called good sense, practical sense, initiative, the faculty of adapting oneself to circumstances.” -Binet
2. “[Intelligence is] the capacity to learn or to profit by experience.” -Dearborn
3. “[Intelligence is the] ability to adapt oneself adequately to relatively new situations in life.” -Pinter
4. “A person possesses intelligence insofar as he has learned, or can learn, to adjust himself to his environment.” -Colvin
5. “[Intelligence is] the ability of an organism to solve new problems …” -Bingham
6. “[Intelligence is] A global concept that involves an individual’s ability to act purposefully, think rationally, and deal effectively with the environment.” -Wechsler
7. “[Intelligence is the] ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought” -APA
8. “Intelligence is what is measured by intelligence tests.” -Boring

What throughline connects all of these? Take a minute to see if you can find some common factors

Common Factors in Intelligence Definitions: 1

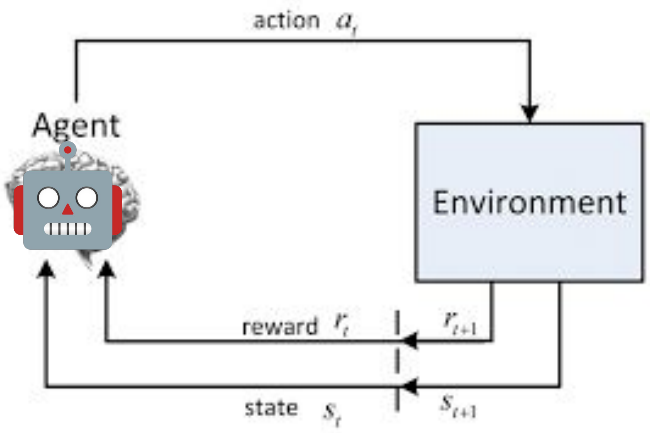
Intelligence is seen as a property of an agent who is interacting with an external environment.



Common Factors in Intelligence Definitions: 2

Intelligent agents are able to carefully choose their actions in a way that leads to them accomplishing their “goals.”

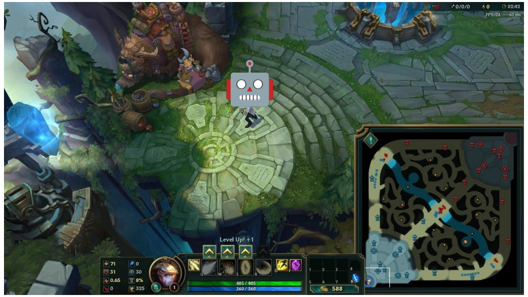
The definition of intelligence is *goal directed* or in our case based on the agent trying to maximize a *reward*.

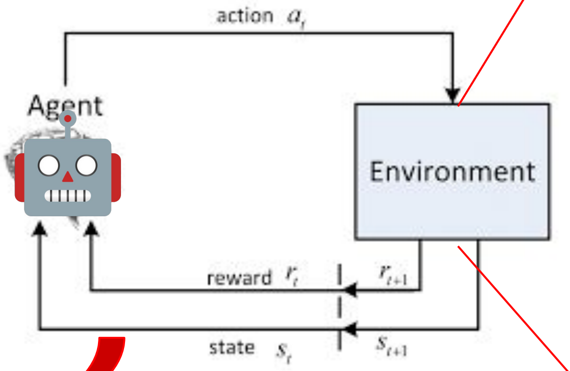


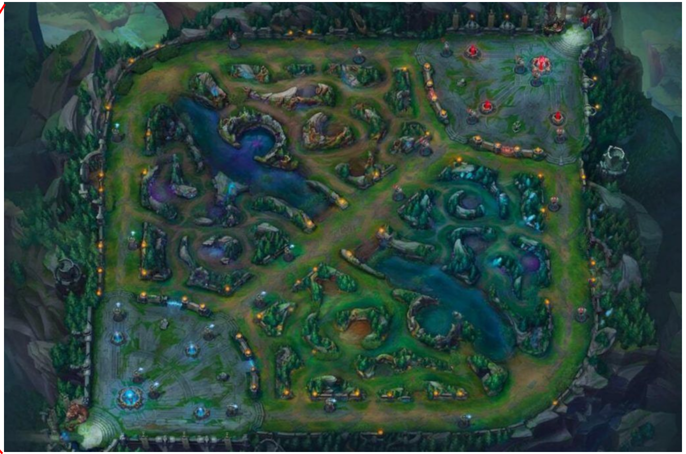
Common Factors in Informal Definitions: 3

Intelligence is not the ability to deal with a fully known environment, but rather the ability to deal with some range of possibilities which cannot be wholly anticipated.

True Environment State != Perceived State



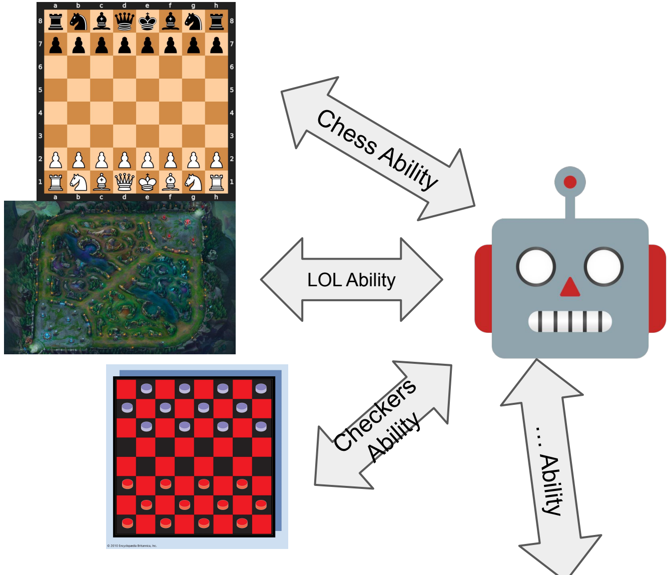




Legg-Hutter Intelligence Informal Definition

From these factors Legg and Hutter derive the following informal definition:

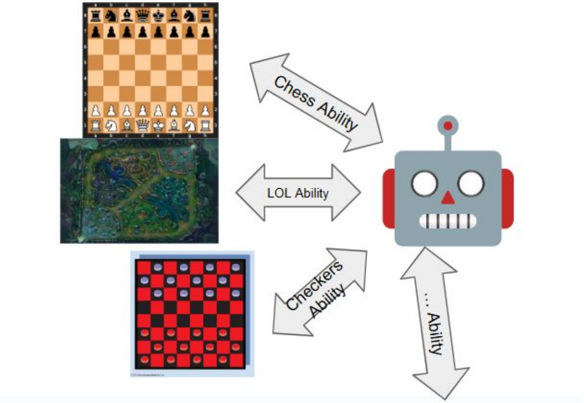
> *“Intelligence measures an agent’s ability to achieve goals in a wide range of environments”*



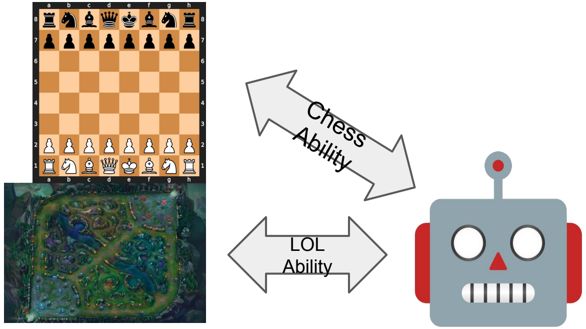
Intelligence := Chess Ability + LOL Ability + Checkers Ability + Ability in all other environments?

Intelligence Tests

This is an *intelligence measure*, it is impossible to evaluate.

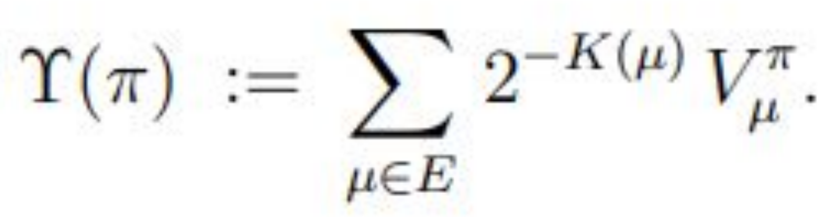


Intelligence tests approximate an intelligence measure in an inexpensive way but lose information.



Intelligence Approximation:
LOL ability + Chess Ability

Formalizing Legg-Hutter Intelligence



$$\Upsilon(\pi) \ := \sum_{\mu \in E} 2^{-K(\mu)} \, V_{\mu}^{\pi} .$$

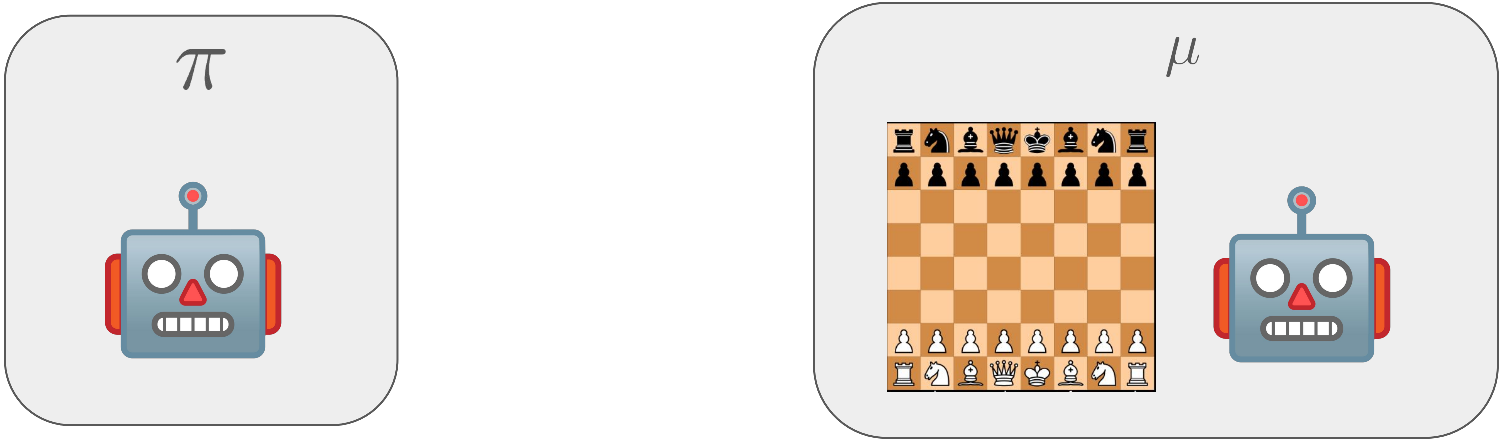
Components of A Formalization

“Intelligence measures an agent’s ability to achieve goals in a wide range of environments.”

This definition contains three essential components: An **agent**, **environment,** and **ability**.

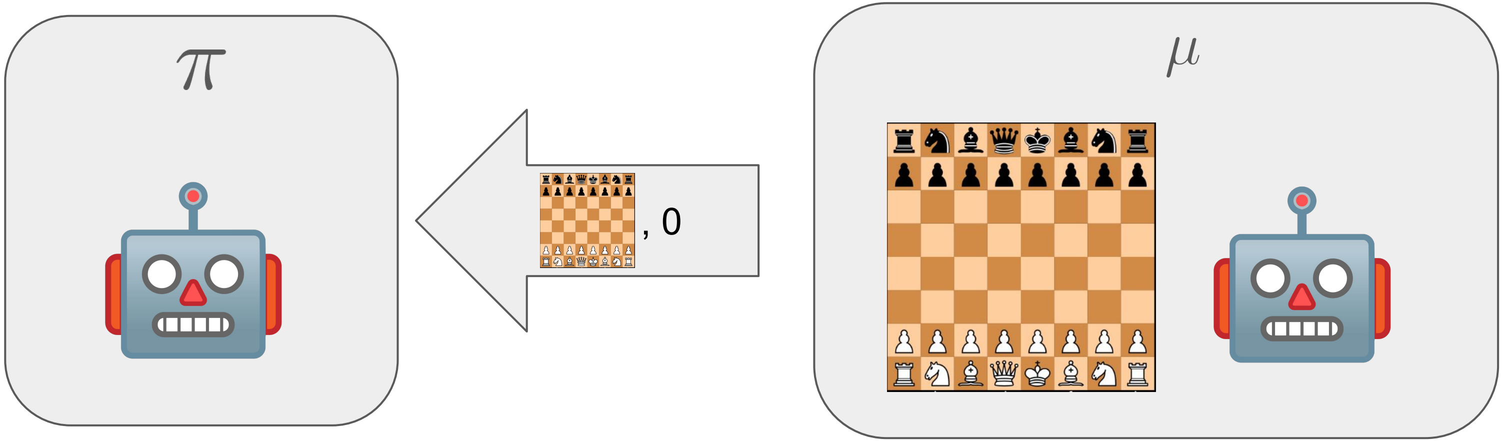
Let's Look at an Example of these components and how they interact in an Agent-Environment System

An Agent and An Environment



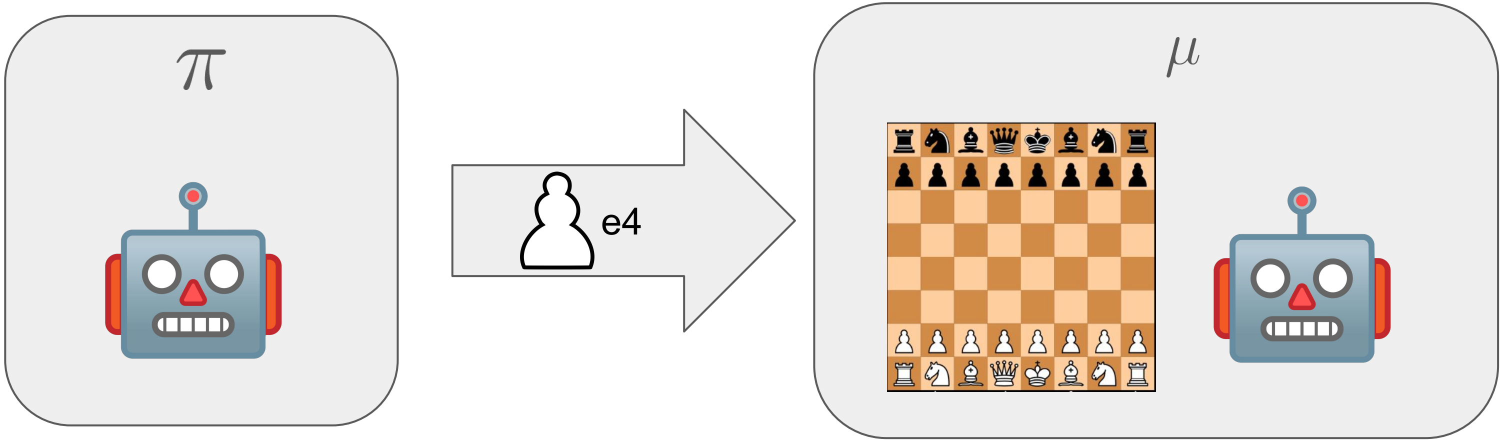
Interaction History: []

A Perception and Reward



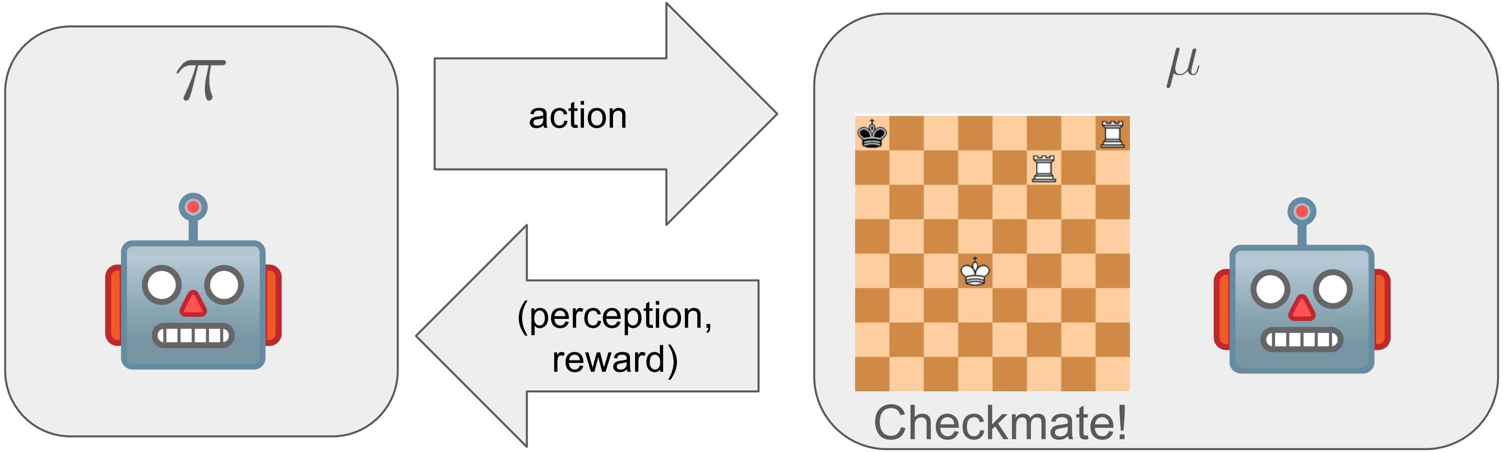
Interaction History: [(, 0)]

An Action



Interaction History: [(, 0), ♙e4]

A Full Game



Interaction History: [(, 0), ♙e4, …, (,1)]

Or formally : $[(p_0, r_0), a_0, \ldots, (p_n, r_n)]$

Agent-Environment System Formalism

An Agent-Environment system is a tuple $(\mathcal{A}, \mathcal{P}, \mathcal{R}, \pi, \mu)$ where

- $\mathcal{A}$, the action space, is a finite set of symbols that can be sent to the environment from the agent.
- $\mathcal{P}$, the perception space, is a finite set of symbols that can be sent to the agent from the environment.
- $\mathcal{R} \subseteq [0,1] \cap \mathbb{Q}$, the reward space, is a set of numeric values that can be sent from the environment to the agent, representing its performance.
- $\pi : (\mathcal{R} \times \mathcal{P} \times \mathcal{A})^* \times \mathcal{R} \times \mathcal{P} \to \mathcal{A}$, an agent, is a probabilistic function mapping the interaction history between itself and an environment to an action.
- $\mu : (\mathcal{R} \times \mathcal{P} \times \mathcal{A})^* \to \mathcal{R} \times \mathcal{P}$, an environment, is a probabilistic function mapping the interaction history between itself and an agent to a perception and reward pair.

$$\begin{tikzcd}
\pi \arrow[r, bend left, "a \in \mathcal{A}"] & \mu \arrow[l, bend left, "(r,o) \in \mathcal{R} \times \mathcal{P}"]
\end{tikzcd}$$

A Formal Probabilistic Environment & Agent

In each cycle two 50¢ coins are tossed. Before the coins settle the player must guess at the number of heads that will result: either 0, 1, or 2.

If the guess is correct the player gets to keep both coins and then two new coins are produced and the game repeats.

If the guess is incorrect the player does not receive any coins, and the game is repeated.

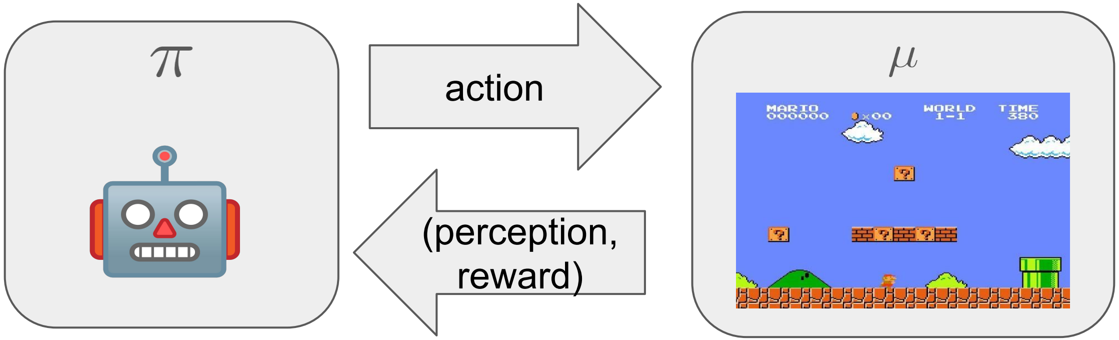
$$\mu(o_k r_k | o_1 \ldots a_{k-1}) := \begin{cases} \frac{1}{4} & \text{if } o_k = a_{k-1} \in \{0,2\} \wedge r_k = 1, \\ \frac{3}{4} & \text{if } o_k \neq a_{k-1} \in \{0,2\} \wedge r_k = 0, \\ \frac{1}{2} & \text{if } o_k = a_{k-1} = 1 \wedge r_k = 1, \\ \frac{1}{2} & \text{if } o_k \neq a_{k-1} = 1 \wedge r_k = 0, \\ 0 & \text{otherwise.} \end{cases}$$

A Probabilistic Environment μ returns probability of an observation reward pair based on history.

$$\pi(a_k | o_1 r_1 a_1 \ldots o_k r_k) := \begin{cases} 1 & \text{for } a_k = 1, \\ 0 & \text{otherwise.} \end{cases}$$

A Deterministic Agent π (Expressed probabilistically) returns probability of an action based on history. (Agent always guesses one)

Formalzing Ability On An Environment



$$[(p_0, r_0), a_0, \ldots, (p_n, r_n)]$$

Total expected reward of pi in mu:

$$V_\mu^\pi := \mathbf{E}\left(\sum_{i=1}^{\infty} r_i\right)$$

The expected value accounts for all possible game paths and weights the reward based on the probability that game occurs.

Extending Ability To Intelligence

Recall that LH Intelligence is defined as ability in ALL environments.
Let E be the countably infinite set of environments (games), to find intelligence it can we just sum expected value for all environments?

$$\mathrm{Int}(\pi) = \sum_{\mu \in E} V_{\mu}^{\pi}$$

What are some potential problems with this definition of intelligence?

Extending Ability To Intelligence

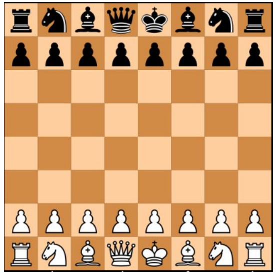
$$\mathbf{Int}(\pi) = \sum_{\mu \in E} V_{\mu}^{\pi}$$

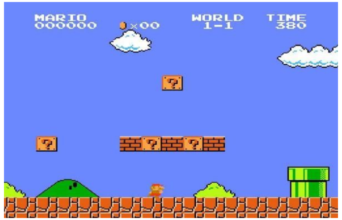
1) Mathematical problems, the measure diverges to infinity. (There are an infinite number of games who reward 1 regardless of what actions you take)

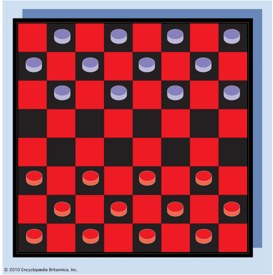
2) The measure informally encodes a notion that ability on any environment is as important for intelligence as ability on any other environment.

Why Environments Matter

If environments don't matter than all the following are equally important for intelligence…






© 2010 Encyclopædia Britannica, Inc.

Key Idea: *Environmental complexity* matters for intelligence, we should weight environments based off of how complex they are

Environmental Complexity: Occam's Razor

“Given multiple hypotheses that are consistent with the data, the simplest should be preferred.”

Example:

What comes next in the number sequence? 2, 4, 6, 8?

Environmental Complexity: Occam's Razor

“Given multiple hypotheses that are consistent with the data, the simplest should be preferred.”

Example:

What comes next in the number sequence? 2, 4, 6, 8?

10! The function f(k) = 2k fits this data, but…

This polynomial also fits the sequence but the next number according to it is 58. We prefer 10 due to Occam's Razor.

$$2k^4 - 20k^3 + 70k^2 - 98k + 48$$

Environmental Complexity: Occam's Razor

Key idea, weight performance based on occam's razor.

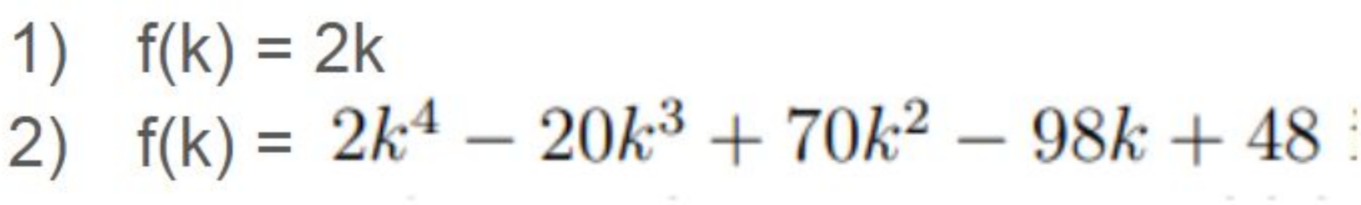
In the environment where you have seen 2, 4, 6, 8 and are rewarded for correctly guessing what comes next…

Performing well in the simple environment is better in the less predictable one

1) f(k) = 2k
2) f(k) = $2k^4 - 20k^3 + 70k^2 - 98k + 48$

Environmental Complexity: Complexity measure?

How do we quantify this notion of complexity of an environment?



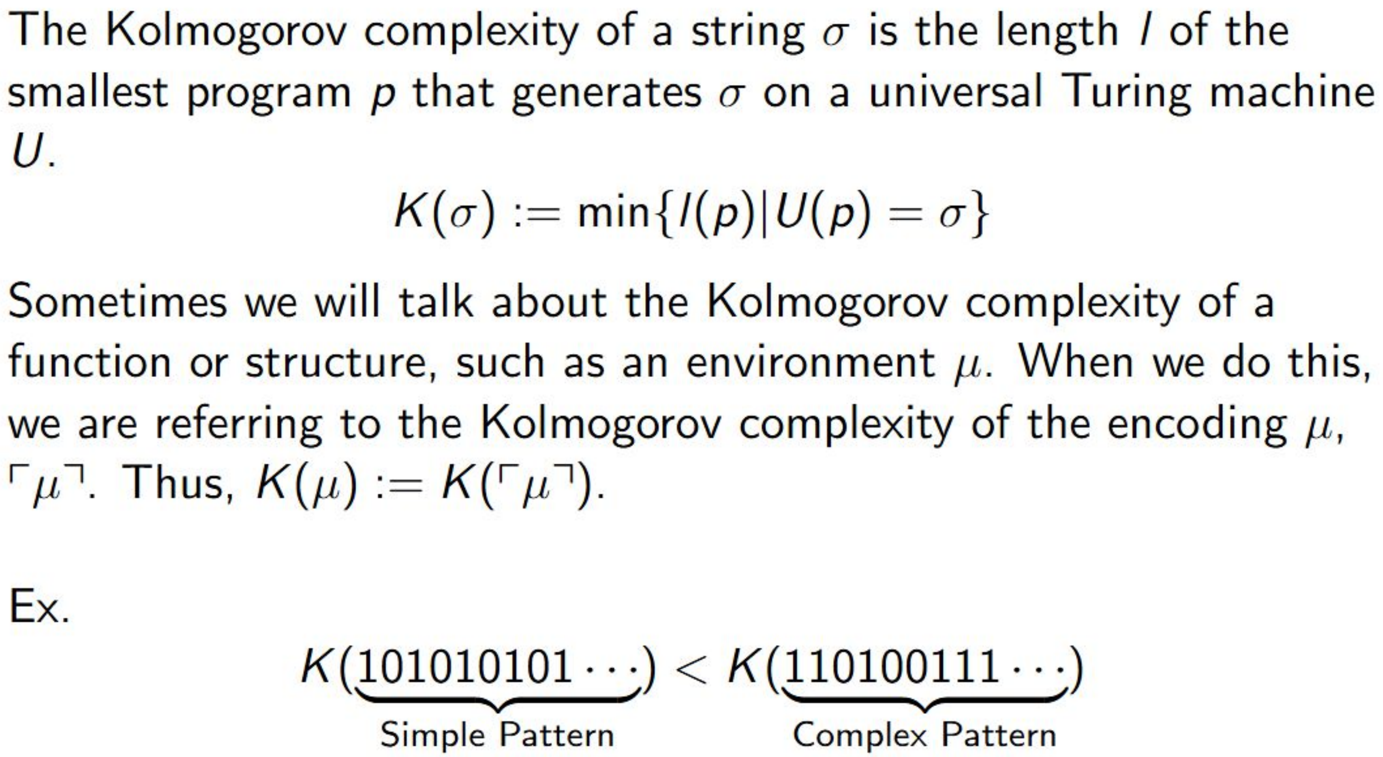
1) f(k) = 2k

2) f(k) = $2k^4 - 20k^3 + 70k^2 - 98k + 48$

We would like a function Complexity(f) such that:

Complexity(2k) < Complexity(2k^4 - 20k^3 + 70k^2 - 98k + 48)

Kolmogorov Complexity!



The Kolmogorov complexity of a string σ is the length l of the smallest program p that generates σ on a universal Turing machine U.

$$K(\sigma) := \min\{l(p)|U(p) = \sigma\}$$

Sometimes we will talk about the Kolmogorov complexity of a function or structure, such as an environment μ. When we do this, we are referring to the Kolmogorov complexity of the encoding μ, $\ulcorner\mu\urcorner$. Thus, $K(\mu) := K(\ulcorner\mu\urcorner)$.

Ex.

$$K(\underbrace{101010101\cdots}_{\text{Simple Pattern}}) < K(\underbrace{110100111\cdots}_{\text{Complex Pattern}})$$

Kolmogorov Complexity Exercises

For each of the following strings x, do we have high, low, or similar K(x) compared to len(x)?

(I.E. is the length of the shortest program generating x close to the length of x?)

- The string of one trillion zeros?
- Alternating string of 1 and 0 for trillion places?
- The binary string containing works of shakespeare in ASCII?
- The binary string representing the results of one trillion fair coin flips?

Kolmogorov Complexity Exercise Solutions

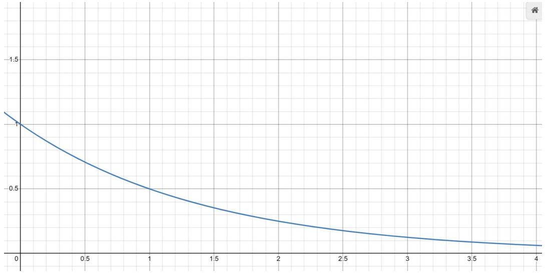
- **The string of one trillion zeros?**
  not even close K(x) << len(x), Python: `"1"*1000000000`
- **Alternating string of 1 and 0 for trillion places?**
  not even close K(x) << len(x), Python: `"10"*500000000`
- **The binary string containing works of shakespeare in ASCII?**
  Upper bound based on current SOTA text compression: K(x) < len(x)/10
- **The binary string representing the results of one trillion fair coin flips?**
  Very close K(x) ~ len(x), true randomness is impossible to compress.

Algorithmic Probability Weighting

Weighting based on K(μ) diverges and is counter our desire to enforce performance in more “predictable” environments.

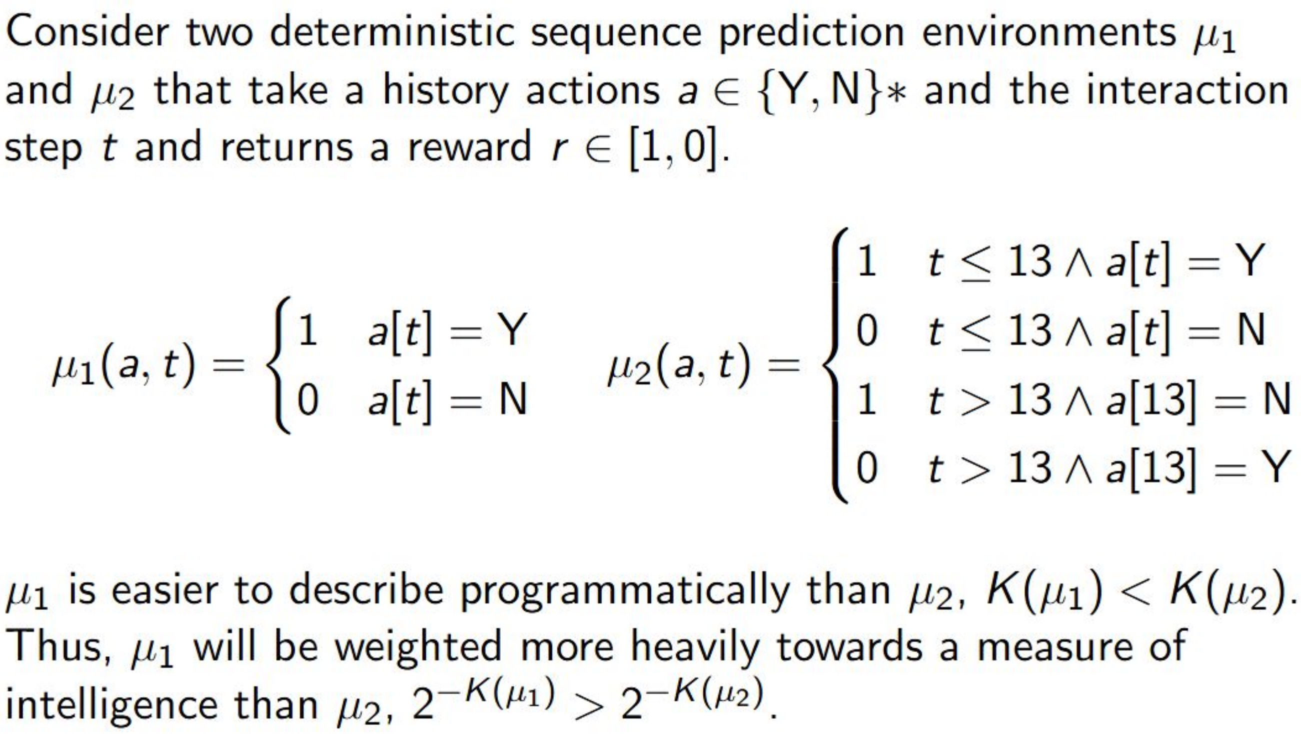
Fix both of these problems by scaling it by 2^-x

$$2^{-K(\mu)}$$



Where K(μ) is the kolmogorov complexity of an binary encoding of the environment μ.

Algorithmic Probability Example

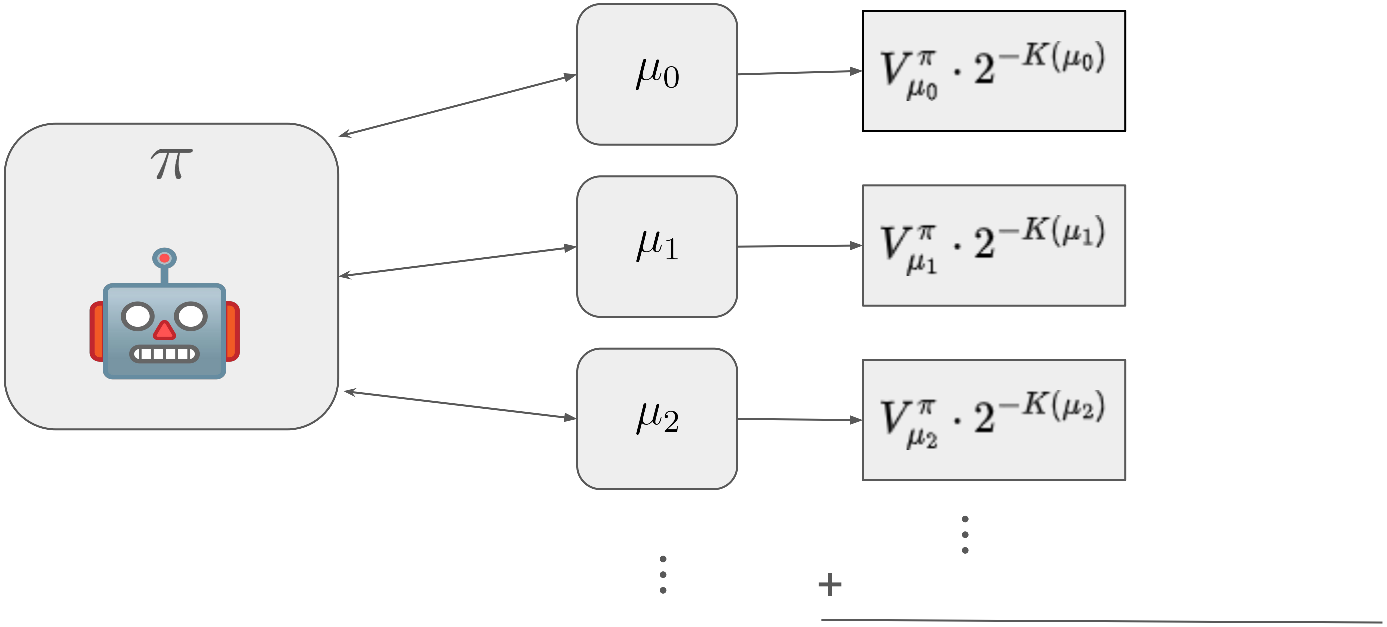


Consider two deterministic sequence prediction environments μ_1 and μ_2 that take a history actions $a \in \{\mathsf{Y}, \mathsf{N}\}*$ and the interaction step t and returns a reward $r \in [1, 0]$.

$$\mu_1(a,t) = \begin{cases} 1 & a[t] = \mathsf{Y} \\ 0 & a[t] = \mathsf{N} \end{cases} \qquad \mu_2(a,t) = \begin{cases} 1 & t \leq 13 \wedge a[t] = \mathsf{Y} \\ 0 & t \leq 13 \wedge a[t] = \mathsf{N} \\ 1 & t > 13 \wedge a[13] = \mathsf{N} \\ 0 & t > 13 \wedge a[13] = \mathsf{Y} \end{cases}$$

μ_1 is easier to describe programmatically than μ_2, $K(\mu_1) < K(\mu_2)$. Thus, μ_1 will be weighted more heavily towards a measure of intelligence than μ_2, $2^{-K(\mu_1)} > 2^{-K(\mu_2)}$.

From Ability to Intelligence


π
μ_0
μ_1
μ_2
$V_{\mu_0}^{\pi} \cdot 2^{-K(\mu_0)}$
$V_{\mu_1}^{\pi} \cdot 2^{-K(\mu_1)}$
$V_{\mu_2}^{\pi} \cdot 2^{-K(\mu_2)}$
⋮
⋮
+

$$\Upsilon(\pi) \; := \sum_{\mu\in E} 2^{-K(\mu)} \, V_\mu^\pi$$

Intelligence of π

The Formal Definition of Universal Intelligence

The intelligence of an agent π is defined as the sum of expected future values over all (computable) environments E weighted with respect to the complexity of the environment.

$$K(x) \ := \ \min_p \{ l(p) : \mathcal{U}(p) = x \}$$

$$V_\mu^\pi \ := \ \mathbf{E} \left(\sum_{i=1}^{\infty} r_i \right) \leq 1.$$

$$\Upsilon(\pi) := \sum_{\mu \in E} \overbrace{2^{-K(\mu)}}^{\text{Environment Weighting}} \underbrace{V_\mu^\pi}_{\text{Expected Reward}}$$

The Universal Intelligence Of Various Agents

Random Agent $\pi^{\mathtt{rand}}$

Makes uniformly random actions, does not respond to any reward or observations.

Fails to exploit regularities in the environment. Hence…

$V_\mu^{\pi^{\mathtt{rand}}}$ is typically low due to failure to exploit regularites

Implies $\Upsilon(\pi^{\mathtt{rand}})$ will be low.

Thus we can conclude the random agent is not very intelligent.

Specialized Agent: Deep Blue $\pi^{\texttt{dblue}}$.

Programmed to be specialized on a single task (Chess). The IBM Deep Blue AI performs very well on chess.

$$V_{\mu^{\texttt{chess}}}^{\pi^{\texttt{dblue}}} \text{ is very high.}$$

But $\cdot\ 2^{-K(\mu^{\texttt{chess}})}$ is small (chess is complex) and most environments are not chess. Thus $\Upsilon(\pi^{\texttt{dblue}})$ is low.

Deep blue is not very intelligent.

Simple General Agent $\pi^{\texttt{basic}}$

Builds an observation table and keeps statistics on rewards that follow from each action.

Select the action that will maximize reward with 90% probability, 10% exploration chance.

Can exploit structure and learn for, nearly all environments (excluding our 10$ or death environment)

$$V_\mu^{\pi^{\texttt{basic}}} > V_\mu^{\pi^{\texttt{rand}}} \text{ and so } \Upsilon(\pi^{\texttt{basic}}) > \Upsilon(\pi^{\texttt{rand}}).$$

More History: Backwards Looking Agent $\pi^{\texttt{2back}}$

Basic agent misses regularities in history, consider the following environment:

$$\mu^{\texttt{alt}}(o_k r_k | o_1 \ldots a_{k-1}) := \begin{cases} 1 & \text{if } a_{k-1} \neq a_{k-2} \wedge r_k = 2^{-k}, \\ 1 & \text{if } a_{k-1} = a_{k-2} \wedge r_k = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Determines reward based on previous two actions (agent can’t do same thing twice in a row), Basic agent can not exploit this.

The backwards looking agent will keep a table of of statistics for two actions rather than one. Can generalize to more at the cost of memory.

Thus because backwards looking agent can exploit more structure

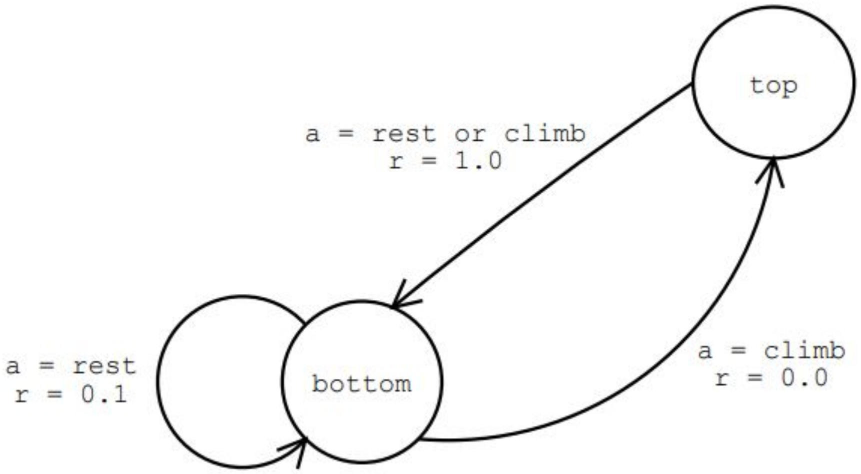
$$V_\mu^{\pi^{\text{2back}}} > V_\mu^{\pi^{\text{basic}}} \qquad \Upsilon(\pi^{\texttt{2back}}) > \Upsilon(\pi^{\texttt{basic}})$$

Forward Looking Agent

Looking at past rewards is insufficient, must plan ahead to maximize reward.

In the slide environment, just taking optimal next action will get you stuck in a local reward maxima, but missing the global maximum.

A forward looking agent will prioritize maximizing next n rewards in the future and can thus solve slide, which requires two rewards in the future.



$$V_{\mu}^{\pi^{2\text{forward}}} > V_{\mu}^{\pi^{2\text{back}}}.$$

The Most Intelligent Agent

$$\bar{\Upsilon} := \max_{\pi} \Upsilon(\pi) = \Upsilon\left(\pi^{AIXI}\right)$$

AIXI, hutters AGI agent is the maximally intelligent agent under this definition of intelligence.

AIXI is uncomputable and thus not feasible as a general agent but is theoretically interesting.

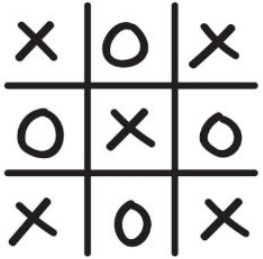
It has been proven that AIXI converges to optimal performance in any environment where this is at all possible for a general agent including:

Markov decision processes, prediction problems, classification problems, bandit problems

Issues with Legg-Hutter Intelligence

The Savant Problem for Legg-Hutter Intelligence (UI)

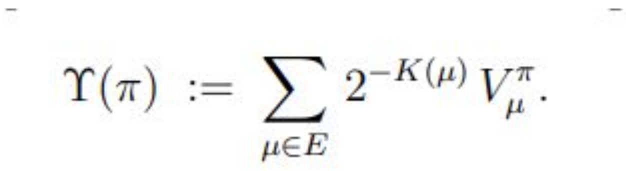
- UI could be argued to not capture a fair intelligence measure for savant-like agents that excel on certain complex domain classes but fail at other simple domains to be as intelligent as agents who perform well on simple well-structured domains.
- For example: consider an AI that can only perfectly play chess vs. an AI that can only perfectly play tik-tac-toe. UI says that the tik-tac-toe agent is more intelligent because tik-tac-toe is a simpler environment and thus more heavily weighted.



Uncomputability of Legg-Hutter Intelligence

Uncomputable due to multiple issues:

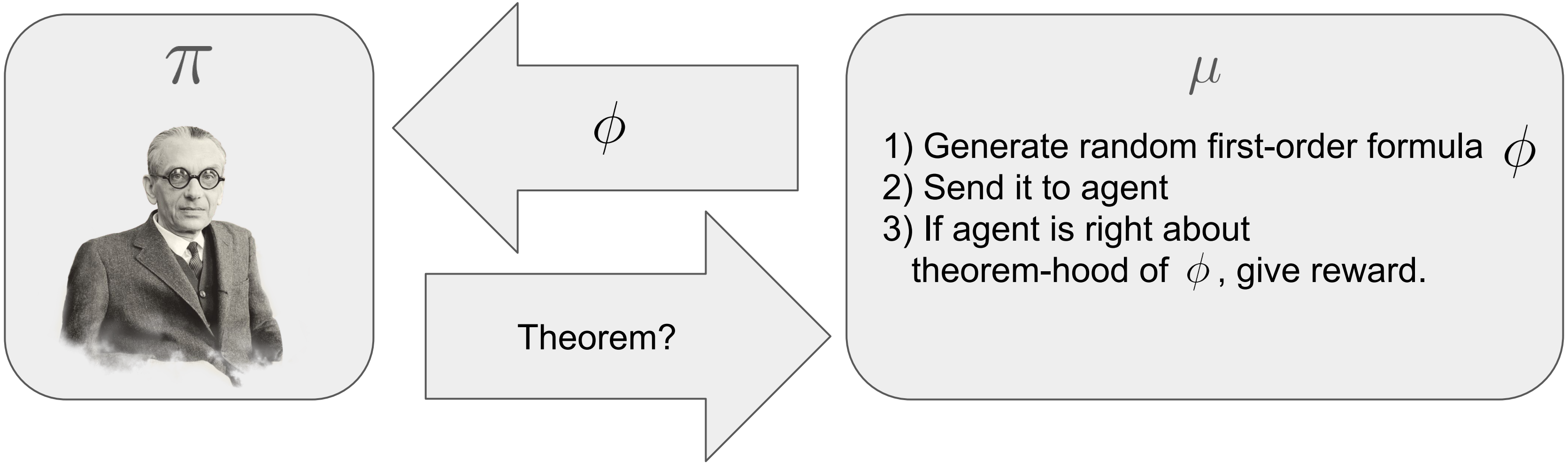
- May require infinite time to compute each V.
- Kolmogorov complexity is uncomputable.
- The size of E is not finite.



$$\Upsilon(\pi) := \sum_{\mu \in E} 2^{-K(\mu)} V_{\mu}^{\pi}.$$

Thus at best we can only approximate the intelligence of an agent.

Restriction to Computable Environments



π

ϕ

Theorem?

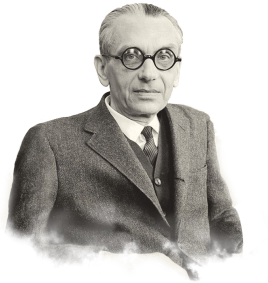
μ

1) Generate random first-order formula ϕ
2) Send it to agent
3) If agent is right about theorem-hood of ϕ, give reward.

In the general case, it is impossible for μ to guarantee Godel is correct and reward him accordingly, because 3 is uncomputable.

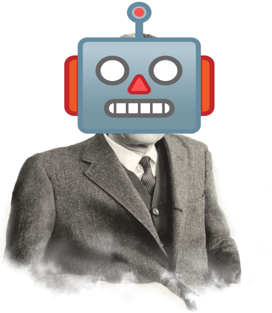
Restriction to Computable Environments

Under Legg-Hutter Intelligence, agents which can perform well on uncomputable tasks such as theorem proving have their intelligence shortchanged.



An agent that solves the *entscheidungsproblem*

>



An agent that can’t

But under Legg-Hutter Intelligence these agents could have the same measure of intelligence!

Black Box View of Agents

The internal structure of the agent is irrelevant to Legg-Hutter intelligence, only the agent’s behavior matters.

One could think of an agent that is extremely capable (Human level even) internally but simply refuses to take any action.

This agent under Legg-Hutter intelligence is not intelligent, even tho internal structure disagrees.

Jailed Godel cant take any actions, but despite his lack of action he is more intelligent than a rock!