
ARTIFICIAL INTELLIGENCE: OVERVIEW

Marcus Hutter

Canberra, ACT, 0200, Australia

<http://cs.anu.edu.au/student/comp3620/>



ANU



RSISE



NICTA

What is (Artificial) Intelligence?

Intelligence can have many faces:

creativity, solving problems, pattern recognition, classification, learning, induction, deduction, building analogies, optimization, surviving in an environment, language processing, planning, and knowledge.

⇒ formal definition incorporating every aspect of intelligence is difficult.

Recurring themes in intelligence definitions: Intelligence

- is a property that an individual agent has as it interacts with its environment or environments.
- is related to the agent's ability to succeed or profit with respect to some goal or objective.
- depends on how able the agent is to adapt to different objectives and environments.

<http://www.idsia.ch/~shane/intelligence.html>

Informal Definition of (Artificial) Intelligence?

Putting these key attributes together produces the informal definition of intelligence:

Intelligence measures an agent's ability to achieve goals in a wide range of environments. [S. Legg and M. Hutter]

Emergent: Features such as the ability to learn and adapt, or to understand, are implicit in the above definition as these capacities enable an agent to succeed in a wide range of environments.

The science of **Artificial Intelligence** is concerned with the construction of intelligent systems/artifacts/agents and their analysis.

What next? Substantiate all terms above: agent, ability, utility, goal, success, learn, adapt, environment, ...

INTRODUCTION TO ARTIFICIAL INTELLIGENCE

MARCUS HUTTER

CANBERRA, ACT, 0200, AUSTRALIA

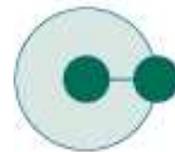
[HTTP://WWW.HUTTER1.NET](http://www.hutter1.net)



ANU



RSISE



NICTA

SLIDES ESSENTIALLY FROM RUSSELL&NORVIG, CHAPTER 1

Outline

- ◇ What is AI?
- ◇ A brief history
- ◇ The state of the art

What is AI?

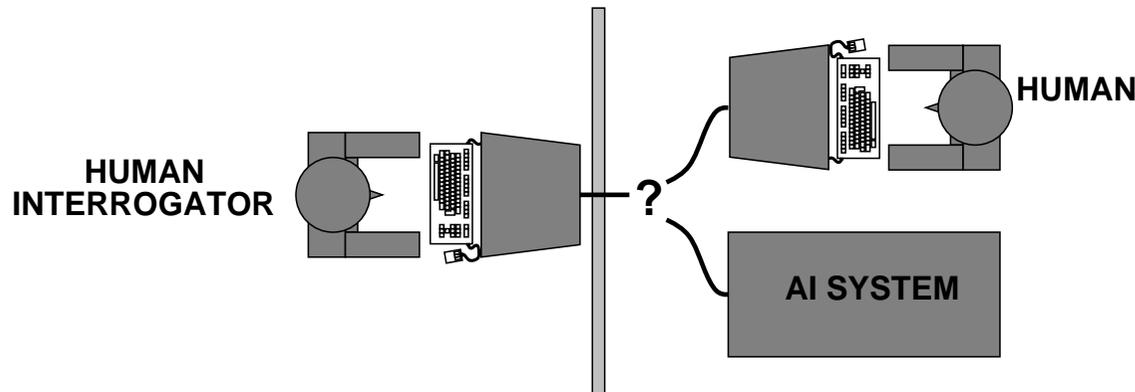
Four possible definitions of AI

Systems that think like humans	Systems that think rationally
Systems that act like humans	Systems that act rationally

Acting humanly: The Turing test

Turing (1950) “Computing machinery and intelligence”:

- ◇ “Can machines think?” → “Can machines behave intelligently?”
- ◇ Operational test for intelligent behavior: the **Imitation Game**



- ◇ Predicted that by 2000, a machine might have a 30% chance of fooling a lay person for 5 minutes
- ◇ Anticipated all major arguments against AI in following 50 years
- ◇ Suggested major components of AI: knowledge, reasoning, language understanding, learning

Problem: Turing test is not **reproducible**, **constructive**, or amenable to **mathematical analysis**

Thinking humanly: Cognitive Science

1960s “cognitive revolution”: information-processing psychology replaced prevailing orthodoxy of behaviorism

Requires scientific theories of internal activities of the brain

– What level of abstraction? “Knowledge” or “circuits”?

– How to validate? Requires

1) Predicting and testing behavior of human subjects (top-down)

or 2) Direct identification from neurological data (bottom-up)

Both approaches (roughly, Cognitive Science and Cognitive Neuroscience) are now distinct from AI

Both share with AI the following characteristic:

**the available theories do not explain (or engender)
anything resembling human-level general intelligence**

Hence, all three fields share one principal direction!

Thinking rationally: Laws of Thought

Normative (or prescriptive) rather than descriptive

Aristotle: what are correct arguments/thought processes?

Several Greek schools developed various forms of logic:

notation and **rules of derivation** for thoughts;
may or may not have proceeded to the idea of mechanization

Direct line through mathematics and philosophy to modern AI

Problems:

- 1) Not all intelligent behavior is mediated by logical deliberation
- 2) **What is the purpose of thinking?** What thoughts **should** I have out of all the thoughts (logical or otherwise) that I **could** have?

Acting rationally

Rational behavior: doing the right thing

The right thing: that which is expected to maximize goal achievement, given the available information

Doesn't necessarily involve thinking—e.g., blinking reflex—but thinking should be in the service of rational action

Aristotle (Nicomachean Ethics):

Every art and every inquiry, and similarly every action and pursuit, is thought to aim at some good

Rational agents

An **agent** is an entity that perceives and acts

This course is about designing **rational agents**

Abstractly, an agent is a function from percept histories to actions:

$$f : \mathcal{P}^* \rightarrow \mathcal{A}$$

For any given class of environments and tasks, we seek the class of agents with the best performance

Problem 1: **computational limitations make perfect rationality unachievable**

⇒ design best **program** for given machine resources

Problem 2: **we neither know nor have a formal description of most real-world env. our agent might operate in**

⇒ design a single agent that works well in a wide range of environments that includes reality

AI prehistory

Philosophy	logic, methods of reasoning mind as physical system foundations of learning, language, rationality
Mathematics	formal representation and proof algorithms, computation, (un)decidability, (in)tractability probability
Psychology	adaptation phenomena of perception and motor control experimental techniques (psychophysics, etc.)
Economics	formal theory of rational decisions
Linguistics	knowledge representation grammar
Neuroscience	plastic physical substrate for mental activity
Control theory	homeostatic systems, stability simple optimal agent designs

Potted history of AI

- 1943 McCulloch & Pitts: Boolean circuit model of brain
- 1950 Turing's "Computing Machinery and Intelligence"
- 1952–69 Early enthusiasm and great expectations "A machine can (never) do X"
- 1950s Early AI programs, including Samuel's checkers program, Newell & Simon's Logic Theorist, Gelernter's Geometry Engine
- 1956 Dartmouth meeting: "Artificial Intelligence" adopted
- 1965 Robinson's complete algorithm for logical reasoning
- 1966–74 AI discovers computational complexity
Neural network research almost disappears
- 1969–79 Early development of knowledge-based systems
- 1980–88 Expert systems industry booms
- 1988–93 Expert systems industry busts: "AI Winter"
- 1985–95 Neural networks return to popularity
- 1988– Resurgence of probability; general increase in technical depth
"Nouvelle AI": ALife, GAs, soft computing
- 1995– Agents, agents, everywhere . . .
- 2003– Human-level AI back on the agenda

State of the art

Which of the following can be done at present?

- ◇ Play a decent game of table tennis
- ◇ Drive safely along a curving mountain road
- ◇ Drive safely along Telegraph Avenue
- ◇ Buy a week's worth of groceries on the web
- ◇ Buy a week's worth of groceries at Coles
- ◇ Play a decent game of bridge
- ◇ Discover and prove a new mathematical theorem
- ◇ Design and execute a research program in molecular biology
- ◇ Write an intentionally funny story
- ◇ Give competent legal advice in a specialized area of law
- ◇ Translate spoken English into spoken Swedish in real time
- ◇ Converse successfully with another person for an hour
- ◇ Perform a complex surgical operation
- ◇ Unload any dishwasher and put everything away

State of the art

- ◇ Deep Blue defeated the reigning world chess champion Garry Kasparov in 1997
- ◇ In 1997, EQP proved a mathematical (Robbins) conjecture unsolved for decades
- ◇ ALVINN in NAVLAB drives autonomously 98% of the time from Pittsburgh to San Diego)
- ◇ During the 1991 Gulf War, US forces deployed an AI logistics planning and scheduling program that involved up to 50'000 vehicles, cargo, and people
- ◇ NASA's on-board autonomous planning program controlled the scheduling of operations for a spacecraft
- ◇ Proverb solves crossword puzzles better than most humans in 1991

INTELLIGENT AGENTS

MARCUS HUTTER

CANBERRA, ACT, 0200, AUSTRALIA

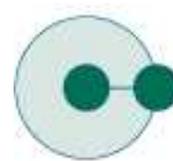
[HTTP://WWW.HUTTER1.NET](http://www.hutter1.net)



ANU



RSISE



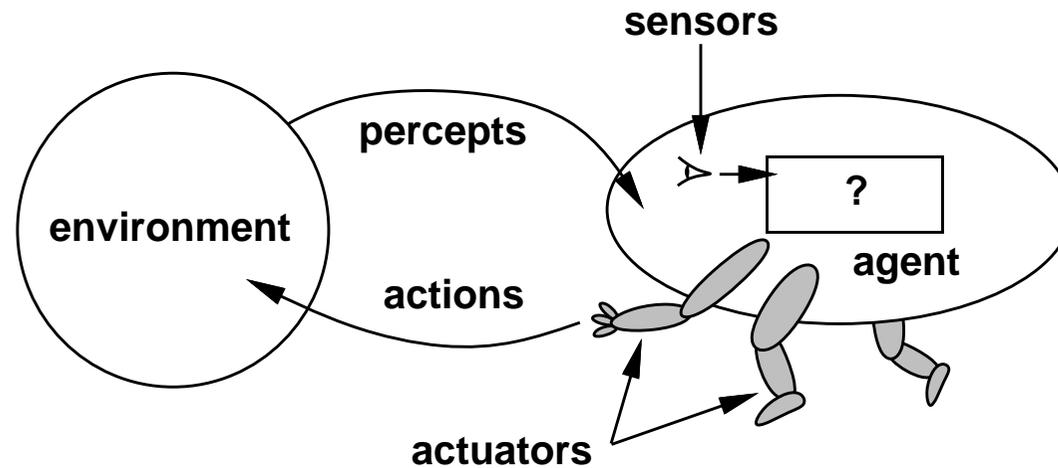
NICTA

SLIDES ESSENTIALLY FROM RUSSELL&NORVIG, CHAPTER 2

Outline

- ◇ Agents and environments
- ◇ Rationality
- ◇ PEAS (Performance measure, Environment, Actuators, Sensors)
- ◇ Environment types
- ◇ Agent types

Agents and environments



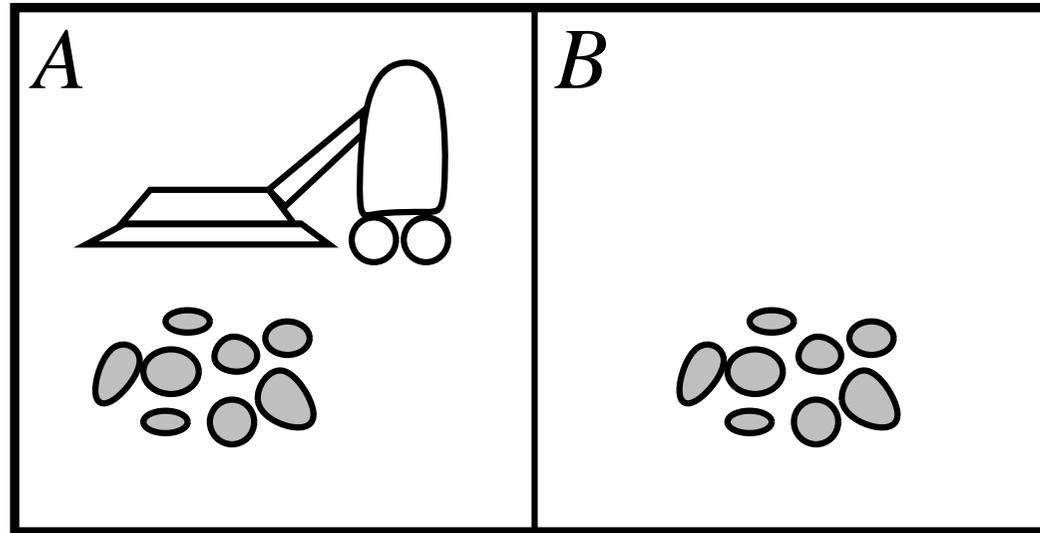
Agents include humans, robots, softbots, thermostats, etc.

The **agent function** maps from percept histories to actions:

$$f : \mathcal{P}^* \rightarrow \mathcal{A}$$

The **agent program** runs on the physical **architecture** to produce f

Vacuum-cleaner world



Percepts: location and contents, e.g., [A , $Dirty$]

Actions: $Left$, $Right$, $Suck$, $NoOp$

A vacuum-cleaner agent

Percept sequence	Action
<i>[A, Clean]</i>	<i>Right</i>
<i>[A, Dirty]</i>	<i>Suck</i>
<i>[B, Clean]</i>	<i>Left</i>
<i>[B, Dirty]</i>	<i>Suck</i>
<i>[A, Clean], [A, Clean]</i>	<i>Right</i>
<i>[A, Clean], [A, Dirty]</i>	<i>Suck</i>
⋮	⋮

function REFLEX-VACUUM-AGENT([*location, status*]) **returns** an action

if *status = Dirty* **then return** *Suck*

else if *location = A* **then return** *Right*

else if *location = B* **then return** *Left*

What is the **right** function?

Can it be implemented in a small agent program?

Rationality

Fixed performance measure evaluates the environment sequence

- one point per square cleaned up in time T ?
- one point per clean square per time step, minus one per move?
- penalize for $> k$ dirty squares?

A rational agent chooses whichever action maximizes the expected value of the performance measure given the percept sequence to date

Rational \neq omniscient

- percepts may not supply all relevant information

Rational \neq clairvoyant

- action outcomes may not be as expected

Hence, rational \neq successful

Rational \Rightarrow exploration, learning, autonomy

PEAS for Automated Taxi

To design a rational agent, we must specify the **task environment**

Consider, e.g., the task of designing an automated taxi:

Performance measure?? safety, destination, profits, legality, comfort, ...

Environment?? streets/freeways, traffic, pedestrians, weather, ...

Actuators?? steering, accelerator, brake, horn, speaker/display, ...

Sensors?? video, accelerometers, gauges, engine sensors, keyboard, GPS, ...

PEAS for Internet shopping agent

Performance measure?? price, quality, appropriateness, efficiency

Environment?? current and future WWW sites, vendors, shippers

Actuators?? display to user, follow URL, fill in form

Sensors?? HTML pages (text, graphics, scripts)

Environment types

	Solitaire	Backgammon	Internet shopping	Taxi
<u>Observable??</u>	Yes	Yes	No	No
<u>Deterministic??</u>	Yes	No	Partly	No
<u>Episodic??</u>	No	No	No	No
<u>Static??</u>	Yes	Semi	Semi	No
<u>Discrete??</u>	Yes	Yes	Yes	No
<u>Single-agent??</u>	Yes	No	Yes (except auctions)	No

The environment type largely determines the agent design

The real world is (of course) partially observable, stochastic, sequential, dynamic, continuous, multi-agent

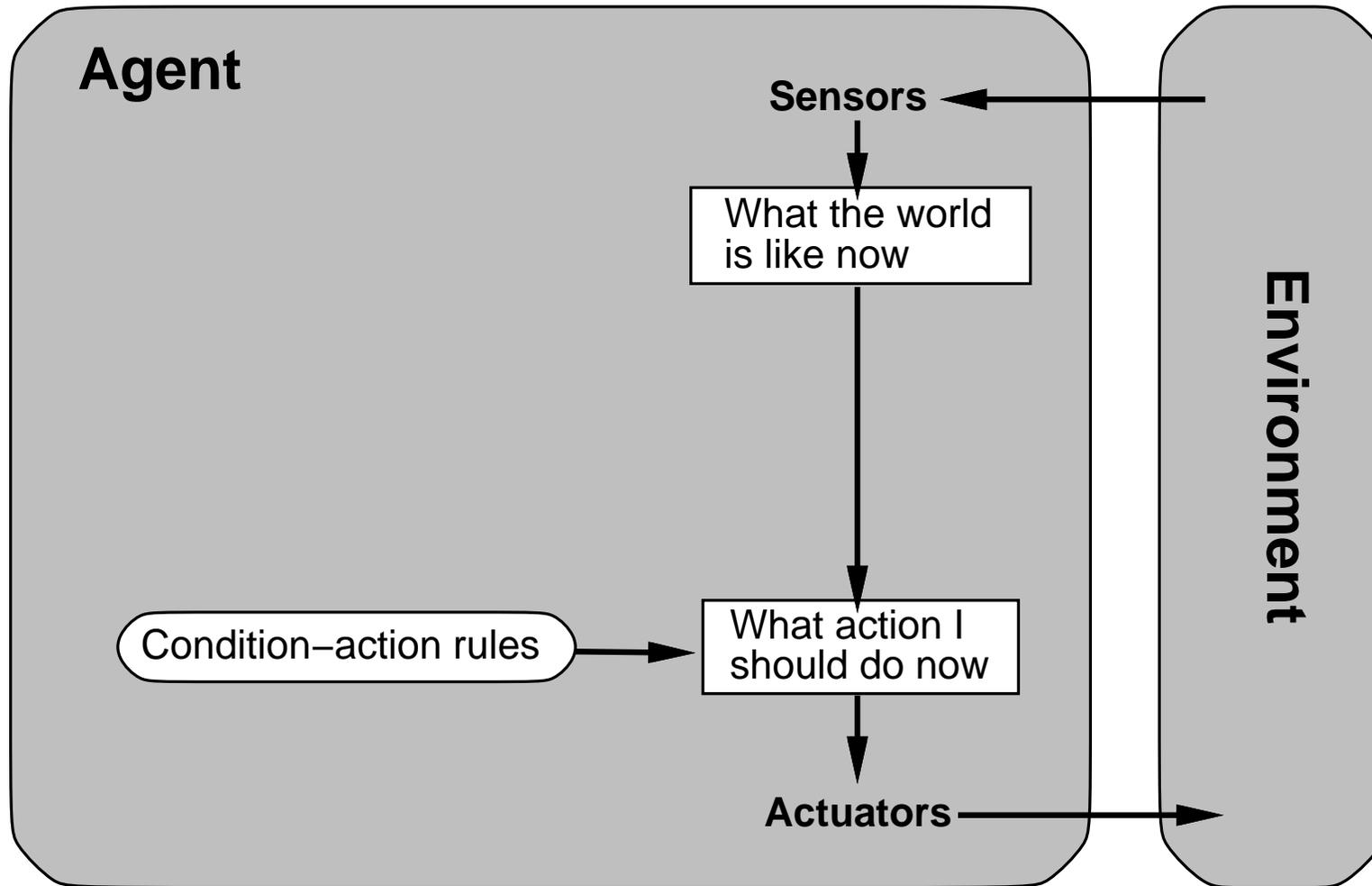
Agent types

Four basic types in order of increasing generality:

- simple reflex agents
- reflex agents with state
- goal-based agents
- utility-based agents

All these can be turned into learning agents

Simple reflex agents



Example

function REFLEX-VACUUM-AGENT(*[location,status]*) **returns** an action

if *status = Dirty* **then return** *Suck*

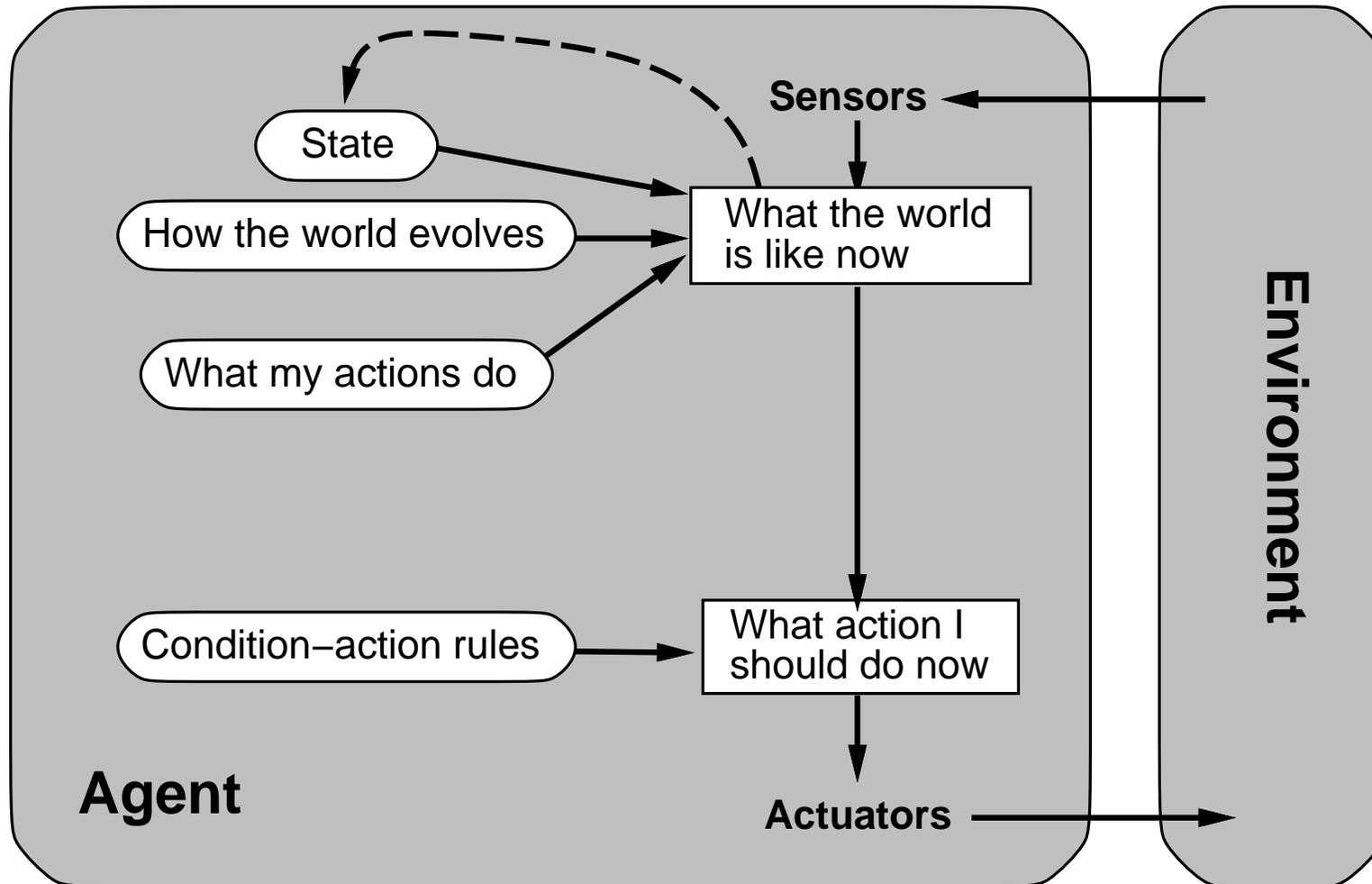
else if *location = A* **then return** *Right*

else if *location = B* **then return** *Left*

```
(setq joe (make-agent :name 'joe :body (make-agent-body)
                     :program (make-reflex-vacuum-agent-program)))
```

```
(defun make-reflex-vacuum-agent-program ()
  #'(lambda (percept)
      (let ((location (first percept)) (status (second percept)))
        (cond ((eq status 'dirty) 'Suck)
              ((eq location 'A) 'Right)
              ((eq location 'B) 'Left))))))
```

Reflex agents with state



Example

function REFLEX-VACUUM-AGENT(*[location,status]*) **returns** an action

static: *lastA, lastB*, numbers, initially ∞

lastA \leftarrow *lastA* + 1

lastB \leftarrow *lastB* + 1

if *location* = *A* **then**

if *status* = *Dirty* **then** *lastA* \leftarrow 0, **return** *Suck*

if *lastB* > 3 **then** **return** *Right*

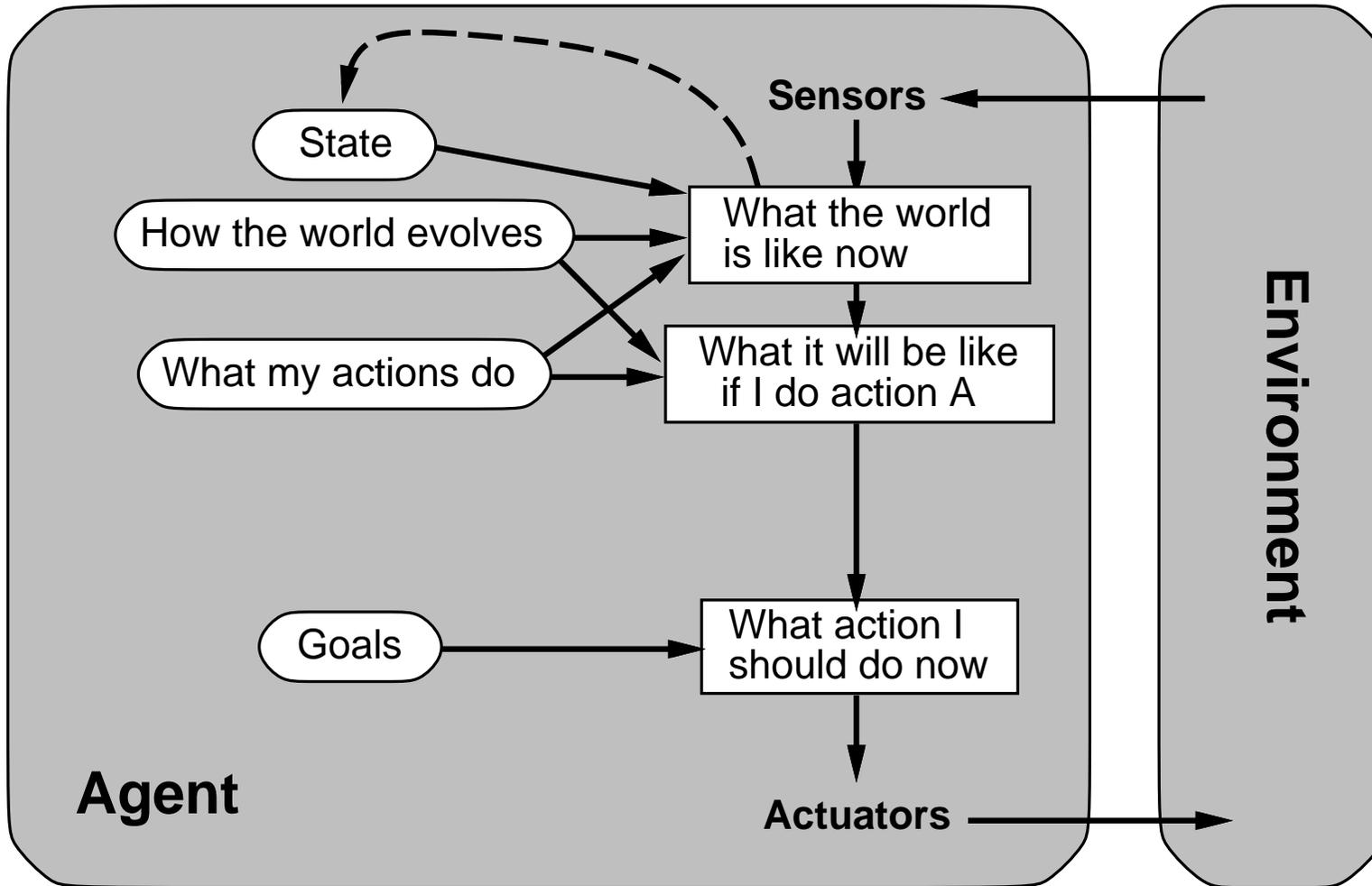
if *location* = *B* **then**

if *status* = *Dirty* **then** *lastB* \leftarrow 0, **return** *Suck*

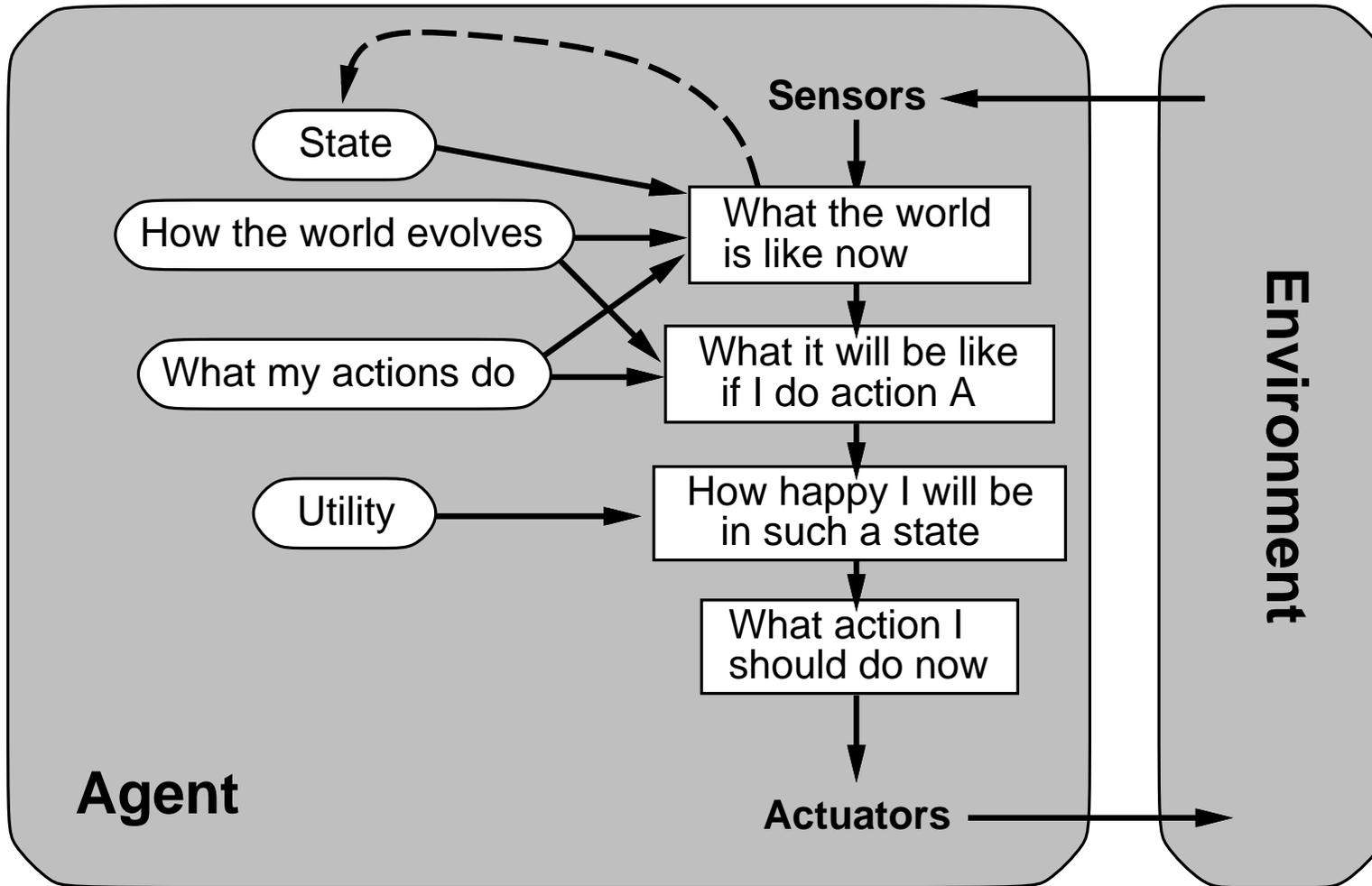
if *lastA* > 3 **then** **return** *Left*

return *NoOp*

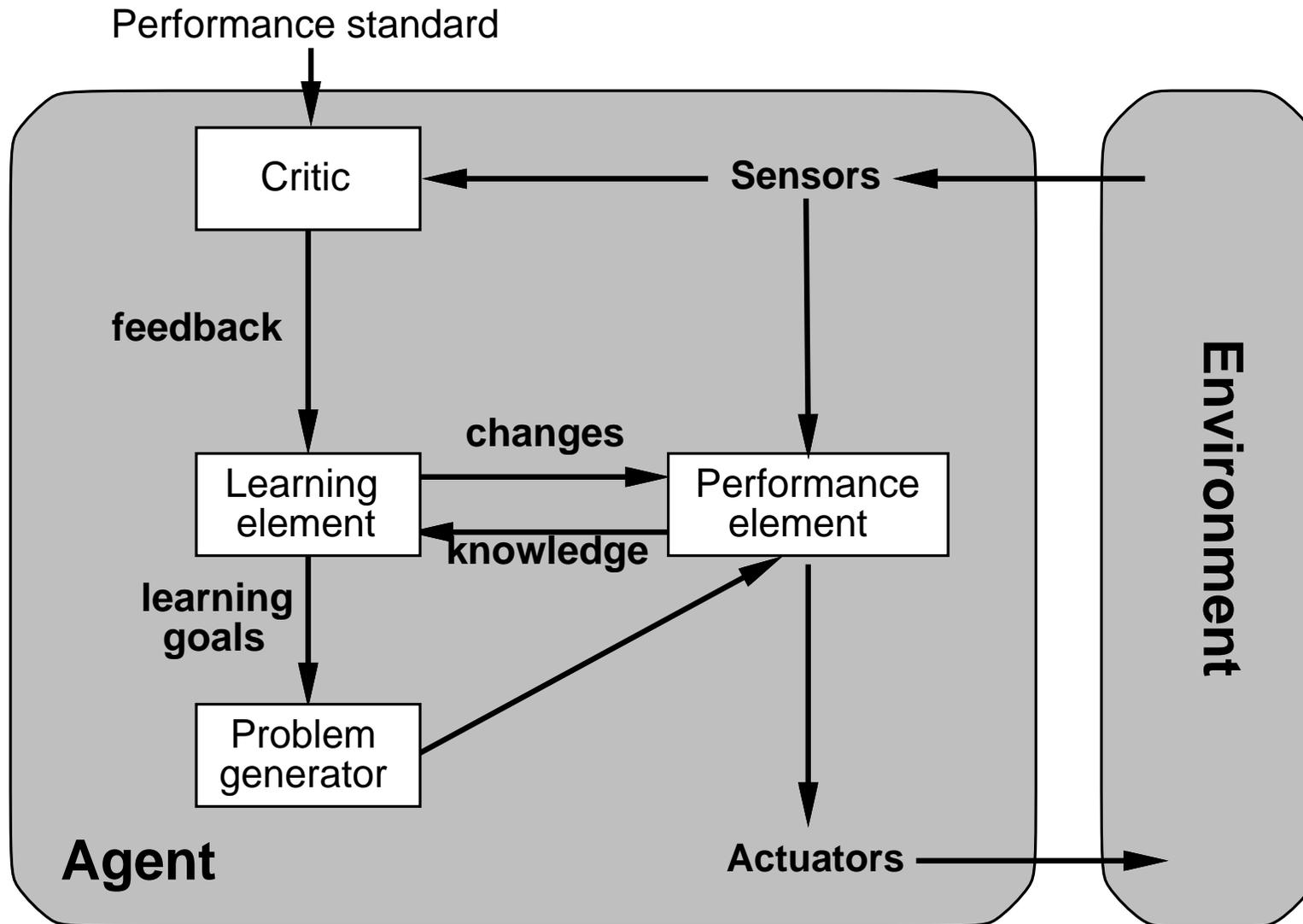
Goal-based agents



Utility-based agents



Learning agents



Exploitation vs. Exploration

An important issue for learning agents is *exploitation versus exploration*

- ◇ Exploitation: using what the agent has learned so far to select actions
- ◇ Exploration: trying actions just to see what happens in the hope of learning more successful behaviors
- ◇ In practice, agents must do some exploration otherwise they may be stuck in a subset of environment states having low(er) utility
- ◇ It even makes sense in some applications to choose actions randomly!
- ◇ Typically, agent explore more in the early stages of deployment and exploit more in later stages

Summary

- ◇ Agents interact with environments through actuators and sensors
- ◇ The agent function describes what the agent does in all circumstances
- ◇ The performance measure evaluates the environment sequence
- ◇ A perfectly rational agent maximizes expected performance
- ◇ Agent programs implement (some) agent functions
- ◇ PEAS descriptions define task environments
- ◇ Environments are categorized along several dimensions:
observable? deterministic? episodic? static? discrete? single-agent?
- ◇ Several basic agent architectures exist:
reflex, reflex with state, goal-based, utility-based, learning

Conclusion

- ◇ Rationality requires a learning component – it is necessary to know as much about the environment as possible before making a rational decision.
- ◇ When studying the various subfields of AI later, remember to keep in mind the *whole agent* view of AI.
- ◇ The individual subfields are interesting, but it's even more interesting to put them all together into an integrated system.

Table of Contents

- Introduction to AI
- Intelligent Agents
- Universal Artificial Intelligence
- AI: Philosophical Foundations

UNIVERSAL ARTIFICIAL INTELLIGENCE

Marcus Hutter

Canberra, ACT, 0200, Australia

<http://www.hutter1.net>



ANU



RSISE



NICTA

Abstract: Motivation

The dream of creating artificial devices that reach or outperform human intelligence is an old one, but a computationally efficient theory of true intelligence has not been found yet, despite considerable efforts in the last 50 years. Nowadays most research is more modest, focussing on solving more narrow, specific problems, associated with only some aspects of intelligence, like playing chess or natural language translation, either as a goal in itself or as a bottom-up approach. The dual, top down approach, is to find a mathematical (not computational) definition of general intelligence. Note that the AI problem remains non-trivial even when ignoring computational aspects.

Abstract: Contents

In this lecture we will outline such an elegant mathematical parameter-free theory of an optimal reinforcement learning agent embedded in an arbitrary unknown environment that possesses essentially all aspects of rational intelligence. Most of the course is devoted to giving an introduction to the key ingredients of this theory, which are important subjects in their own right: Occam's razor; Turing machines; Kolmogorov complexity; probability theory; Solomonoff induction; Bayesian sequence prediction; minimum description length principle; agents; sequential decision theory; adaptive control theory; reinforcement learning; Levin search and extensions.

Table of Contents

- Goal of Universal Artificial Intelligence
- Occam's Razor
- Information Theory & Kolmogorov Complexity
- Bayesian Probability Theory
- Algorithmic Probability Theory
- Inductive Inference & Universal Forecasting
- The Minimum Description Length Principle
- The Universal Similarity Metric
- Sequential Decision Theory
- Rational Agents in Known and Unknown Environment
- Computational Issues: Universal Search

Goal of Universal Artificial Intelligence

Goal: Construct a single universal agent that learns to act optimally in any environment.

State of the art: Formal (mathematical, non-comp.) definition of such an agent

Accomplishment: Well-defines AI, formalizes rational intelligence, formal “solution” of the AI problem

⇒ reduces the conceptual AI problem to a (pure) computational problem.

Occam's Razor

- Grue Emerald Paradox:

Hypothesis 1: All emeralds are green.

Hypothesis 2: All emeralds found till y2010 are green,
thereafter all emeralds are blue.

- Which hypothesis is more plausible? **H1!** Justification?
- **Occam's razor:** take simplest hypothesis consistent with data.
is the most important principle in science.

Information Theory & Kolmogorov Complexity

- Quantification/interpretation of Occam's razor:
- Shortest description of object is best explanation.
- Shortest program for a string on a Turing-machine T leads to best extrapolation=prediction.

$$K_T(x) = \min_p \{l(p) : T(p) = x\}$$

- Prediction is best for a universal Turing-machine U .

$$\text{Kolmogorov-complexity}(x) = K(x) = K_U(x) \leq K_T(x) + c_T$$

Bayesian Probability Theory

Given (1): Models $P(D|H_i)$ for probability of observing data D , when H_i is true.

Given (2): Prior probability over hypotheses $P(H_i)$.

Goal: Posterior probability $P(H_i|D)$ of H_i , after having seen data D .

Solution:

Bayes' rule:

$$P(H_i|D) = \frac{P(D|H_i) \cdot P(H_i)}{\sum_i P(D|H_i) \cdot P(H_i)}$$

(1) Models $P(D|H_i)$ usually easy to describe (objective probabilities)

(2) But Bayesian prob. theory does not tell us how to choose the prior $P(H_i)$ (subjective probabilities)

Algorithmic Probability Theory

- **Epicurus**: If more than one theory is consistent with the observations, keep all theories.
- \Rightarrow uniform prior over all H_i ?
- Refinement with **Occam's razor** quantified in terms of **Kolmogorov complexity**:

$$P(H_i) := 2^{-K_{T/U}(H_i)}$$

- **Fixing T** we have a complete theory for prediction.
Problem: How to choose T .
- **Choosing U** we have a universal theory for prediction.
Observation: Particular choice of U does not matter much.
Problem: Incomputable.

Inductive Inference & Universal Forecasting

- Solomonoff combined Occam, Epicurus, Bayes, and Turing in one formal theory of sequential prediction.
- $M(x)$ = probability that a universal Turing-machine outputs x when provided with fair coin flips on the input tape.
- A posteriori probability of y given x is $M(y|x) = M(xy)/M(x)$.
- Given $\dot{x}_1, \dots, \dot{x}_{t-1}$, the probability of x_t is $M(x_t|\dot{x}_1 \dots \dot{x}_{t-1})$.
- Immediate “applications”:
 - Weather forecasting: $x_t \in \{\text{sun, rain}\}$.
 - Stock-market prediction: $x_t \in \{\text{bear, bull}\}$.
 - Continuing number sequences in an IQ test: $x_t \in \mathbb{N}$.
- Works optimally for everything!

The Minimum Description Length Principle

- **Approximation** of Solomonoff, since M is incomputable:
- $M(x) \approx 2^{-K_U(x)}$ (quite good)
- $K_U(x) \approx K_T(x)$ (very crude)
- **Predict** y of highest $M(y|x)$ is approximately same as
- **MDL**: Predict y of smallest $K_T(xy)$.

The Universal Similarity Metric

- One example among many: Determination of composer of music.
- Let m_1, \dots, m_n be pieces of music of known composer $c = 1, \dots, n$.
- Let $m_?$ be (different!) piece of music of unknown composer.
- Concatenate each m_i with $m_?$
- Most similarity between pieces of music of same composer
 \Rightarrow maximal compression.
- Guess composer is
$$\hat{i} = \arg \max_i M(m_? | m_i) \approx \arg \min_i [K_T(m_i \circ m_?) - K_T(m_i)]$$
- For T choose Lempel-Ziv or bzip(2) compressor.
- No musical knowledge used in this method.

Sequential Decision Theory

Setup: For $t = 1, 2, 3, 4, \dots$

Given sequence x_1, x_2, \dots, x_{t-1}

(1) predict/make decision y_t ,

(2) observe x_t ,

(3) suffer loss $\text{Loss}(x_t, y_t)$,

(4) $t \rightarrow t + 1$, goto (1)

Goal: Minimize expected Loss.

Greedy minimization of expected loss **is optimal** if:

Important: Decision y_t does not influence env. (future observations).

Loss function is known.

Problem: Expectation w.r.t. what?

Solution: W.r.t. universal distribution M if true distr. is unknown.

Example: Weather Forecasting

Observation $x_t \in \mathcal{X} = \{\text{sunny, rainy}\}$

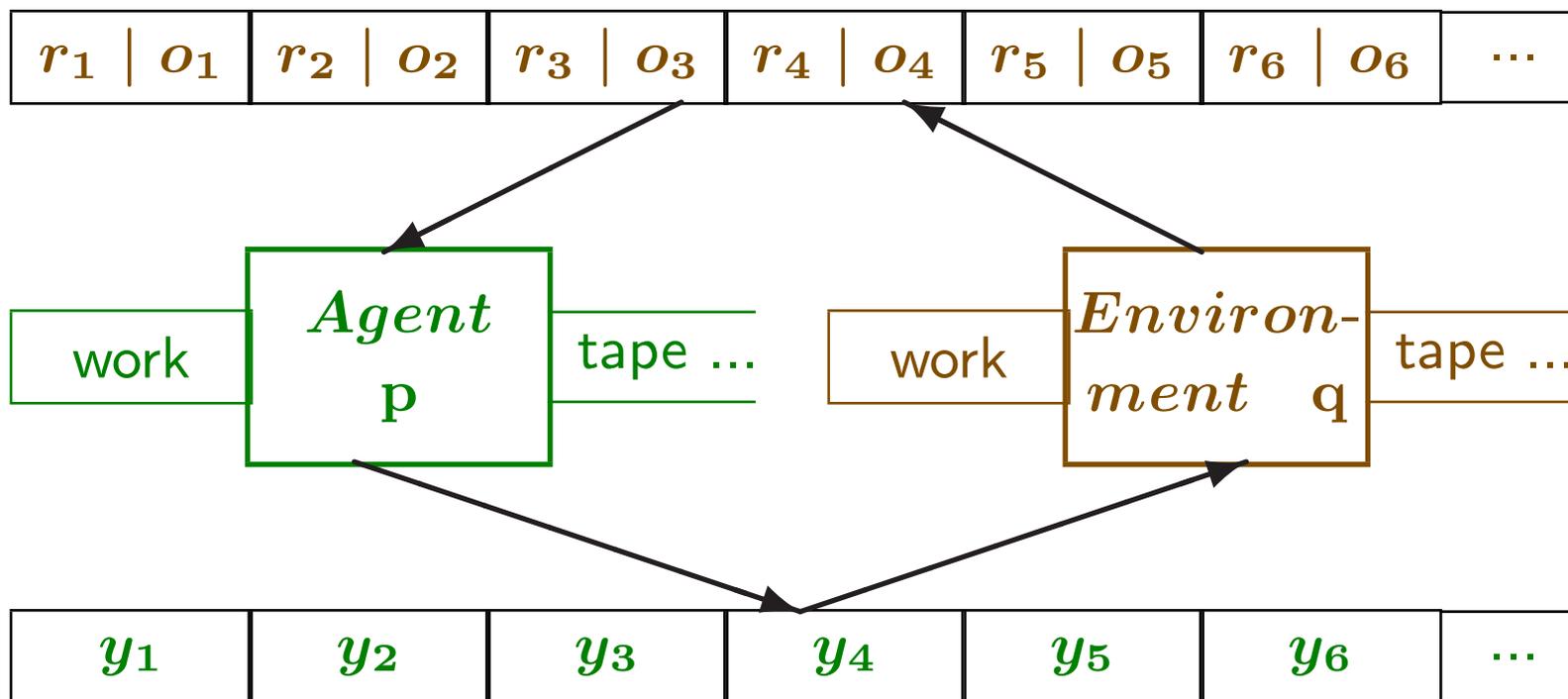
Decision $y_t \in \mathcal{Y} = \{\text{umbrella, sunglasses}\}$

Loss	sunny	rainy
umbrella	0.1	0.3
sunglasses	0.0	1.0

Taking umbrella/sunglasses does not influence future weather
(ignoring butterfly effect)

Agent Model with Reward

if actions/decisions y influence the environment q



Rational Agents in Known Environment

- **Setup:** Known deterministic or probabilistic environment
- **Greedy** maximization of reward r ($= -\text{Loss}$) **no longer optimal.**
Example: Chess
- **Exploration versus exploitation problem.**
 \Rightarrow **Agent has to be farsighted.**
- **Optimal solution:** Maximize future (expected) reward sum, called value.
- **Problem:** Things drastically change if environment is unknown

Rational Agents in Unknown Environment

Additional problem: (probabilistic) environment unknown.

Fields: reinforcement learning and adaptive control theory

Bayesian approach: Mixture distribution.

1. What performance does Bayes-optimal policy imply?
It does not necessarily imply self-optimization
(Heaven&Hell example).
2. Computationally very hard problem.
3. Choice of horizon? Immortal agents are lazy.

Universal Solomonoff mixture \Rightarrow universal agent AIXI.

Represents a formal (math., non-comp.) solution to the AI problem?

Most (all?) problems are easily phrased within AIXI.

Computational Issues: Universal Search

- **Levin search:**
Fastest algorithm for inversion and optimization problems.
- **Theoretical application:**
Assume somebody found a non-constructive proof of $P=NP$, then Levin-search is a polynomial time algorithm for every NP (complete) problem.
- **Practical applications** (J. Schmidhuber)
Maze, towers of hanoi, robotics, ...
- **FastPrg:** The asymptotically fastest and shortest algorithm for all well-defined problems.
- **AIXI $_{tl}$:** Computable variant of AIXI.

Exercises

1. What is the probability p that the sun will rise tomorrow,
2. Justify Laplace' rule ($p = \frac{n+1}{n+2}$, where $n = \#$ days sun rose in past)
3. Predict sequences:
2,3,5,7,11,13,17,19,23,29,31,37,41,43,47,53,59,?
3,1,4,1,5,9,2,6,5,3,?,
1,2,3,4,?
4. Argue in (1) and (3) for **different** continuations.

ARTIFICIAL INTELLIGENCE: PHILOSOPHICAL FOUNDATIONS

Marcus Hutter

Canberra, ACT, 0200, Australia

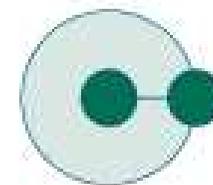
<http://www.hutter1.net>



ANU



RSISE



NICTA

Abstract

The dream of creating artificial devices that reach or outperform human intelligence is an old one, but has not been reached yet, despite considerable efforts in the last 50 years. I will discuss philosophical issues regarding what it means to think and whether artifacts could or should ever do so. Many arguments have been devised pro and against AI. Such philosophical considerations are important, since creating human level AI systems or beyond would have severe social, ethic, environmental, philosophical, and legal consequences.

Table of Contents

- Weak AI hypothesis:
machines can act as if they were intelligent
- Strong AI hypothesis:
machines can actually be intelligent
- Can a machine be conscious?

Can Weak AI Succeed?

The argument from disability:

- A machine can never do X.
- + These claims have been disproven for an increasing # of things X.

The mathematical objection (Lucas 1961, Penrose 1989,1994):

- No formal system incl. AIs, but only humans can “see” that Gödel’s unprovable sentence is true.
- + Lucas cannot consistently assert that this sentence is true.

The argument from informality of behavior:

- Human behavior is far too complex to be captured by any simple set of rules. Dreyfus (1972,1992) “What computers (still) can’t do”.
- + Computers already can generalize, can learn from experience, etc.

The Mathematical Objection to Weak AI

Applying Gödel's incompleteness theorem:

- $G(F) :=$ "This sentence cannot be proved in the formal axiomatic system F "
- We humans can easily see that $G(F)$ must be true.
- Lucas (1961), Penrose (1989,1994):
Since any AI is an F , no AI can prove $G(F)$.
- Therefore there are things humans, but no AI system can do.

Counter-argument:

- $L :=$ "J.R.Lucas cannot consistently assert that this sentence is true"
- Lucas cannot assert L , but now **we** can conclude that it is true.
- Lucas is in the same situation as an AI.

Strong AI versus Weak AI

Argument from consciousness:

- A machine passing the Turing test would not prove that it actually really thinks or is conscious about itself.
- + We do not know whether other humans are conscious about themselves, but it is a polite convention, which should be applied to AIs too.

Biological naturalism:

- Mental states can emerge from neural substrate only.

Functionalism:

- + Only the functionality/behavior matters.

Strong AI: Mind-Body and Free Will

Mind-body problem:

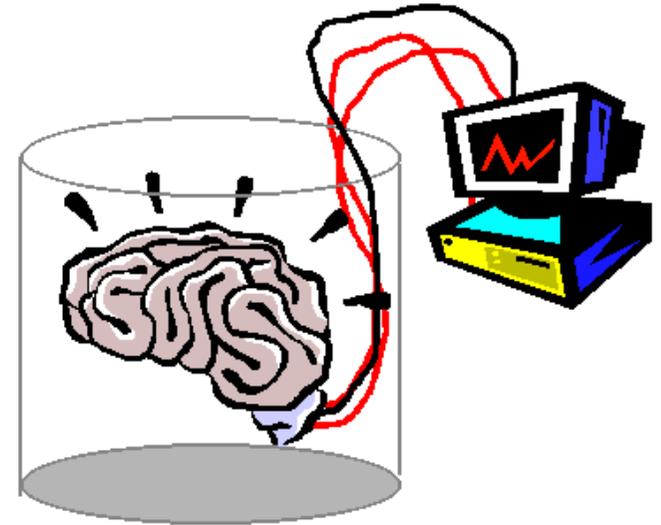
- + Materialist: There exists only the a mortal body.
- Dualist: There also exists an immortal soul.

Free will paradox:

- How can a purely physical mind governed by strictly by physical laws have free will?
- + By carefully reconstructing our naive notion of free will:
If it is impossible to predict and tell my next decision,
then I have effective free will.

Strong AI: Brain Dissection

The “brain in a vat” experiment:
(no) real experience:
(see Movie Matrix for details)



The brain prosthesis experiment:

Replacing some neurons in the brain by functionally identical electronic prostheses would neither effect external behavior nor internal experience of the subject.

Successively replace one neuron after the other until the whole brain is electronic.

Strong AI: Chinese Room & Lookup Table



Strong AI: Chinese Room & Lookup Table

Assume you have a huge table or rule book containing all answers to all potential questions in the Turing test (say in Chinese which you don't understand).

- You would pass the Turing test without understanding anything.
- + There is no big enough table.
- + The used rule book is conscious.
- + Analogy: Look, the brain just works according to physical rules without understanding anything.

Strong AI versus Weak AI: Does it Matter?

The phenomenon of consciousness is mysterious, but likely it is not too important whether a machine simulates intelligence or really *is* self aware. Maybe the whole distinction between strong and weak AI makes no sense.

Analogy:

- Natural ↔ artificial: urea, wine, paintings, thinking.
- Real ↔ virtual: flying an airplane versus simulator.

Is there a difference? Should we care?

Ethics and Risks of AI

- People might lose their jobs to automation.
- + So far automation (via AI technology) has created more jobs and wealth than it has eliminated.

- People might have too much (or too little) leisure time
- + AI frees us from boring routine jobs and leaves more time for pretentious and creative things.

- People might lose their sense of being unique.
- + We mastered similar degradations in the past (Galileo, Darwin, physical strength)
- + We will not feel so lonely anymore (cf. SETI)

- People might lose some of their privacy rights.

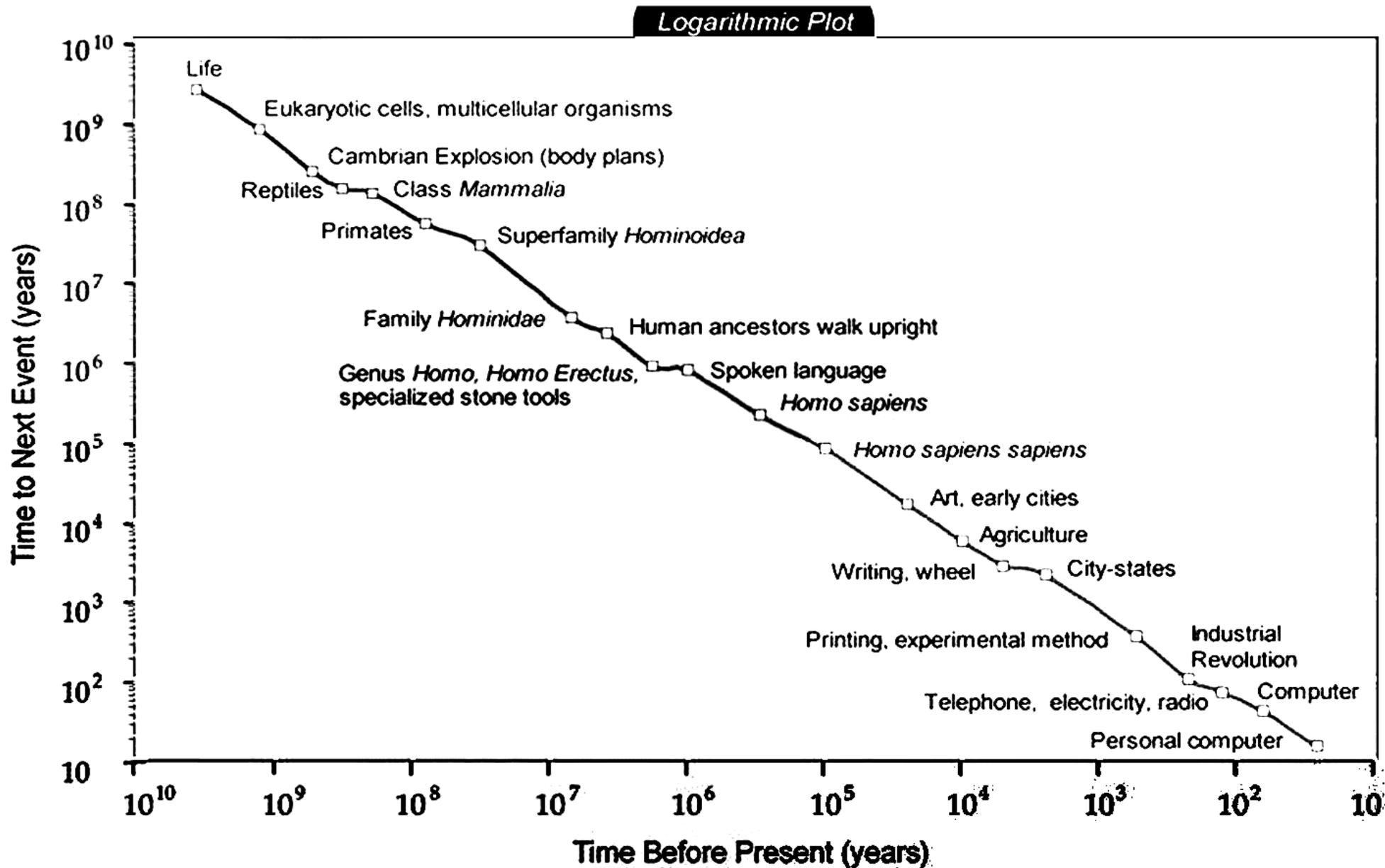
- The use of AI systems might result in a loss of accountability.
- ? Who is responsible if a physician follows the advice of a medical expert system, whose diagnosis turns out to be wrong?

What If We Do Succeed?

The success of AI might mean the end of the human race.

- Artificial evolution is replaced by natural selection.
AI systems will be our **mind children** (Moravec 2000)
- Once a machine surpasses the intelligence of a human it can design even smarter machines (I.J. Good 1965).
- This will lead to an **intelligence explosion** and a **technological singularity** at which the human era ends.
- Prediction beyond this **event horizon** will be impossible (Vernor Vinge 1993)
- Alternative 1: We keep the machines under control.
- Alternative 2: Humans merge with or extend their brain by AI.
Transhumanism (Ray Kurzweil 2000)

Countdown To Singularity



Three Laws of Robotics

Robots (should) have rights and moral duties

1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.



(Isaac Asimov 1942)

Recommended Literature

- [RN03] S. J. Russell and P. Norvig. *Artificial Intelligence. A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition, 2003.
- [HMU01] J. E. Hopcroft, R. Motwani, and J. D. Ullman. *Introduction to Automata Theory, Language, and Computation*. Addison-Wesley, 2nd edition, 2001.
- [LV97] M. Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications*. Springer, Berlin, 2nd edition, 1997.
- [SB98] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.
- [GP07] B. Goertzel and C. Pennachin, editors. *Artificial General Intelligence*. Springer, 2007.
- [Hut05] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005.
<http://www.hutter1.net/ai/uaibook.htm>.