Rationality and Intelligence

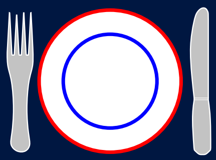
Stuart Russell
Computer Science Division
UC Berkeley

Joint work with Eric Wefald, Devika Subramanian, Shlomo Zilberstein, Othar Hansson, Andrew Mayer, Gary Ogasawara, Tim Huang, Ron Parr, Keiji Kanazawa, Daphne Koller, Jonathan Tash, Peter Norvig, and Jeff Forbes.

Includes ideas by Eric Horvitz, Michael Fehling, Jack Breese, Michael Bratman, Tom Dean, Martha Pollack, and others.

Outline

1. Constructive definitions of Intelligence
2. Some silly old definitions
3. A silly new definition



Three kinds of AI

Modelling human cognition

“Look! My model of humans is accurate!”

Building useful artifacts

“Look! PBTS made a small fortune!”

Creating Intelligence

“Look! My system is Intelligent!!”

“No it isn’t!” “Yes it is!” etc.

Why constructive definitions?

Avoid silly arguments, G & T.

Need a formal relationship between

input/structure/output and Intelligence

while avoiding overly narrow definitions that lead to sterile and irrelevant research!

Constructive definitions ...

Suppose a definition Int is proposed

"Look! My system is Int!"

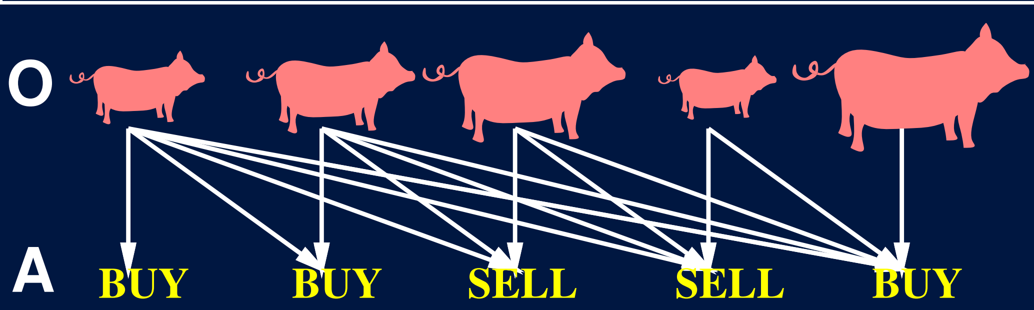
- Is the claim interesting?
- Is the claim sometimes true?
- What research do we do on Int?

Candidates for Int

And the candidates for Best Formal Definition of Intelligence are as follows:

- ♢ Int_1: Perfect rationality
- ♢ Int_2: Calculative rationality
- ♢ Int_3: Metalevel rationality
- ♢ Int_4: Bounded optimality

Agents and environments



Agents perceive **O** and act **A** in environment E

An agent function $f : \mathbf{O}^* \rightarrow \mathbf{A}$

specifies an act for any percept sequence

Global measure $V(f, E)$ evaluates f in E

$\text{Int}_1 = \textbf{perfect rationality}$

Agent f_{opt} is perfectly rational:

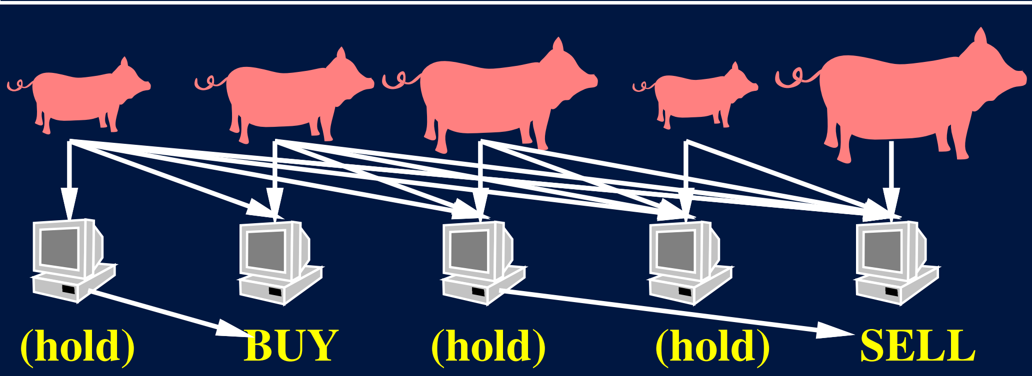
$$f_{opt} = \text{argmax}_f\, V(f, \mathbf{E})$$

i.e., the best possible behaviour

"Look! My system is perfectly rational!"

- Very interesting claim
- VERY seldom possible
- Research relates global measure to local constraints, e.g., maximizing utility

Machines and programs



Agent is a machine M running a program p

This defines an agent function f = Agent(p, M)

Int_2 = calculative rationality

p is calculatively rational if $\mathsf{Agent}(p, M) = f_{opt}$
when M is infinitely fast
i.e., p eventually computes the best action

"Look! My system is calculatively rational!"

- Useless in real-time* worlds
- Quite often true
- Research on calculative tools, e.g. logical planners, influence diagrams

The calculative toolbox

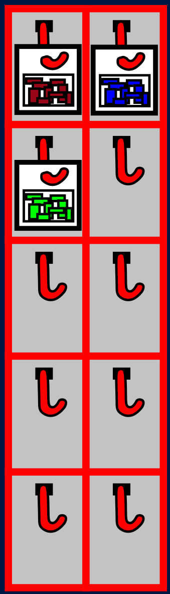
The toolbox is almost empty!!

Need tools for

- learning, modelling,
- deciding, compiling

in environments that are

- (non)deterministic,
- (partially) observable,
- discrete/continuous, static/dynamic



Complexity

Calculative rationality describes "in principle" capability

NP/PSPACE-completeness
 $\Rightarrow$ trade off decision quality for computation

Int$_3$: metalevel rationality

Agent(p, M) is metalevelly rational if it controls its computations optimally

"Look! My system is metalevelly rational!"

- Very interesting claim
- VERY seldom possible
- Research on rational metareasoning

Rational metareasoning

Do the Right Thinking:

- ♢ Computations are actions
- ♢ Cost=time Benefit=better decisions
- ♢ Value $\approx$ benefit minus cost

General agent program:

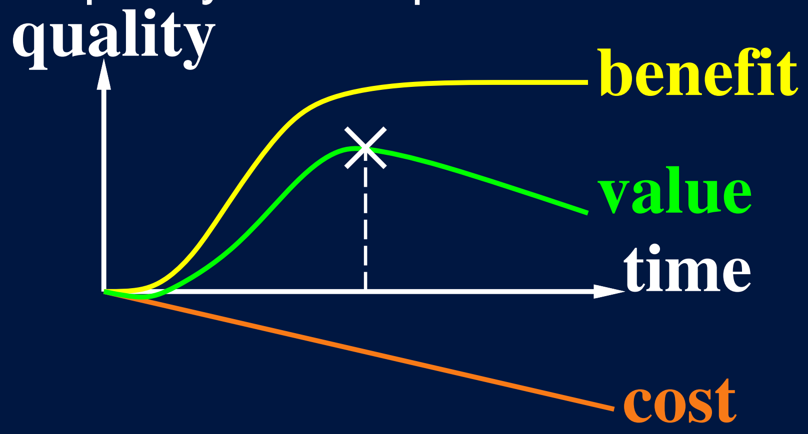
Repeat until no computation has value > 0:

Do the best computation

Do the current best action

Anytime algorithms

Decision quality that improves over time



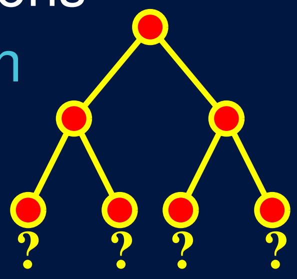
Rational metareasoning applies trivially

Anytime tools!

Fine-grained metareasoning

Explicit model of effects of computations

$\Rightarrow$ selection as well as termination



Compiled into efficient formula

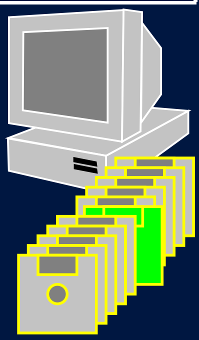
for value of computation

Applications in search, games, MDPs show improvement over standard algorithms

Algorithms in AI

Metareasoning replaces clever algorithms!

Int_4: bounded optimality



Agent(p_{opt}, M) is bounded-optimal iff

$$p_{opt} = \mathrm{argmax}_p V(\mathrm{Agent}(p, M), \mathbf{E})$$

i.e., the best program given M.

Look! My system is bounded-optimal!

- Very interesting claim
- Always possible
- Research on all sorts of things

Nonlocal constraints!

Translates into nonlocal constraints on action

$\Rightarrow$ Optimize over programs, not actions

Similar conclusions reached in other fields:

Economics: Herb Simon and others

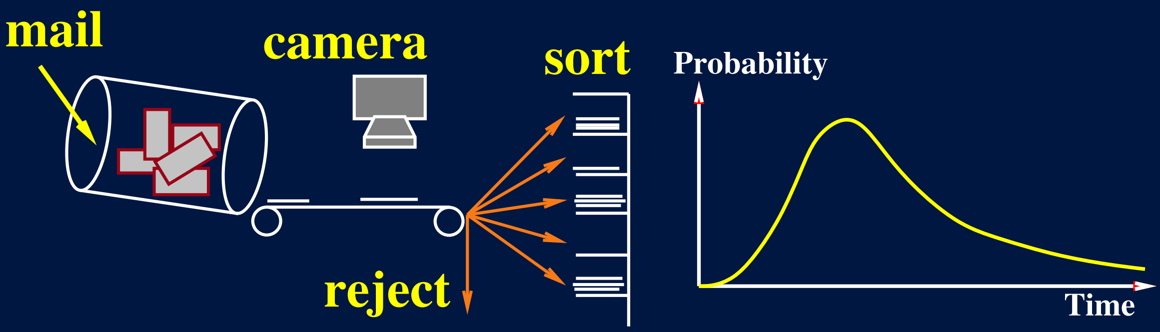
Game theory: Prisoners' Dilemma

Robert Aumann, Wed. 10.30 a.m.

Philosophy: Dennett's Moral First-Aid Manual

Politics: Toffler's* Creating a New Civilization

Example: Sorting mail



E: Letters arrive at random times

M: Runs one or more neural networks

p_{opt} is a sequence of networks computable from arrival distribution

Asymptotic bounded optimality

Strict bounded optimality is too fragile

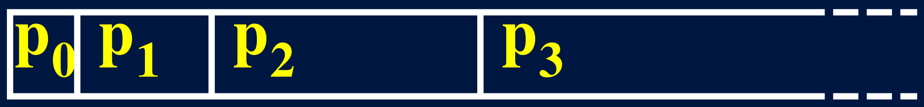
p is asymptotically bounded-optimal (ABO) iff
$\exists k\ V(\mathrm{Agent}(p, kM), \mathbf{E}) \geq V(\mathrm{Agent}(p_{opt}, M), \mathbf{E})$

I.e., speeding up M by k compensates
for p's inefficiency

Worst-case ABO and average-case ABO
generalize classical complexity

Complex real-time systems

Let p_i be ABO for a fixed deadline at $t = 2^i \epsilon$

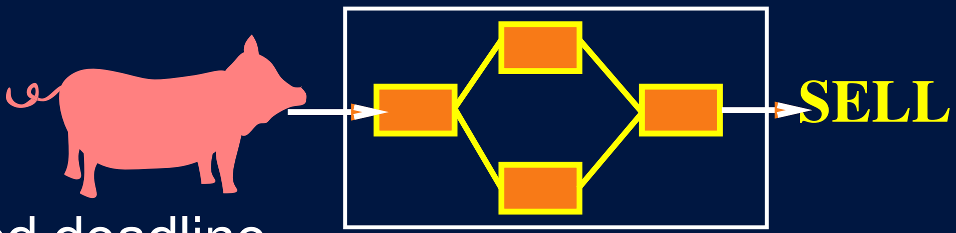


Sequence is ABO for any deadline distribution

As good as knowing the deadline in advance!

Complex systems contd.

Use the doubling construction to build composite anytime systems



Fixed deadline

⇒ allocation to components is easy

⇒ “compiler” for complex systems

Metalevel reinforcement learning

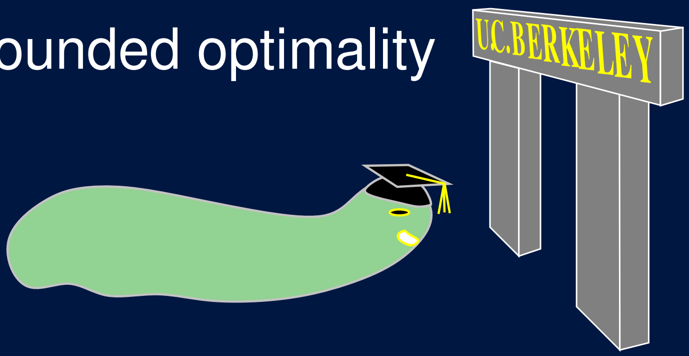
Object-level reinforcement learning:
learn long-term rewards for actions from short-term rewards

Metalevel reinforcement learning:
learn long-term rewards for computations

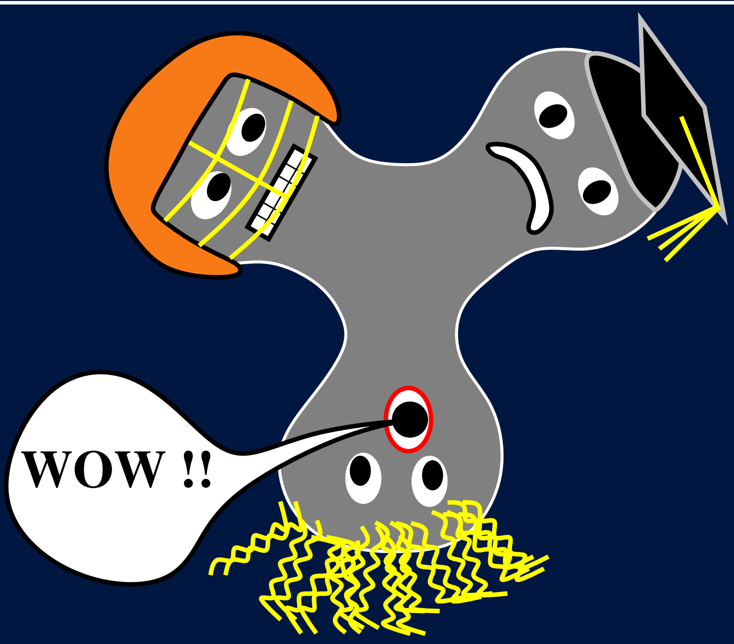
Criterion for "valid" update rules:
convergence to bounded optimality

What next?

- Prove convergence to bounded optimality within fixed software architectures
- Prove dominance between architectures
- Develop a “grammar” of AI architectures
- Learning and bounded optimality



Bounded optimal solutions



Conclusions

- Computational limitations ~~Brains~~ cause minds
- Tools in, algorithms out (eventually)
- Bounded optimality:
  - Fits intuitive idea of Intelligence
  - A bridge between theory and practice
- Crisis: LAP–FOPLBMLDTHTNPOPMEA