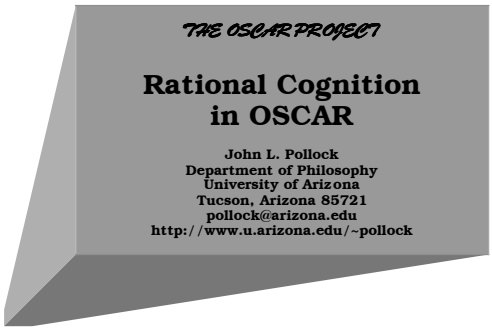


Part One: Evaluating Agent Architectures

- Stuart Russell: "rational agents are those that ***do the right thing***".
  - The problem of designing a rational agent then becomes the problem of figuring out what the right thing is.
  - There are two approaches to the latter problem, depending upon the kind of agent we want to build.
- *Anthropomorphic Agents* — those that can help human beings rather directly in their intellectual endeavors.
  - These endeavors consist of decision making and data processing.
  - An agent that can help humans in these enterprises must make decisions and draw conclusions that are rational by human standards of rationality.
- *Goal-Oriented Agents* — those that can carry out certain narrowly-defined tasks in the world.
  - Here the objective is to get the job done, and it makes little difference how the agent achieves its design goal.

Evaluating Goal-Oriented Agents

- If the design goal of a goal-oriented agent is sufficiently simple, it may be possible to construct a metric the measures how well an agent achieves it.
  - Then the natural way of evaluating an agent architecture is in terms of the expected-value of that metric.
  - An ideally rational goal-oriented agent would be one whose design maximizes that expected-value.
  - The recent work on *bounded-optimality* (Russell and Subramanian; Horvitz; Zilberstein and Russell, etc.) derives from this approach to evaluating agent architectures.
- This approach will only be applicable in cases in which it is possible to construct a metric of success.
- If the design goal is sufficiently complex, that will be at least difficult, and perhaps impossible.

Evaluating Anthropomorphic Agents

- Here it is the individual decisions and conclusions of the agent that we want to be rational.
- In principle, we could regard an anthropomorphic agent as a special case of a goal-oriented agent, where now the goal is to make rational decisions and draw rational conclusions, but it is doubtful that we can produce a metric that measures the degree to which such an agent is successful in achieving these goals.
- Even if we could construct such a metric, it would not provide an analysis of rationality for such an agent, because the metric itself must presuppose prior standards of rationality governing the individual cognitive acts of the agent being evaluated.

Evaluating Anthropomorphic Agents

- In AI it is often supposed that the standards of rationality that apply to individual cognitive acts are straightforward and unproblematic:
  - Bayesian probability theory provides the standards of rationality for beliefs;
  - classical decision theory provides the standards of rationality for practical decisions.
- It may come as a surprise then that most philosophers reject Bayesian epistemology, and I believe there are compelling reasons for rejecting classical decision theory.

Bayesian Epistemology

- Bayesian epistemology asserts that the degree to which a rational agent is justified in believing something can be identified with a subjective probability. Belief updating is governed by conditionalization on new inputs.
  - There is an immense literature on this. Some of the objections to it are summarized in Pollock and Cruz, *Contemporary Theories of Knowledge*, 2nd edition (Rowman and Littlefield, 1999).
- Perhaps the simplest objection to Bayesian epistemology is that it implies that an agent is always rational in believing any truth of logic, because any such truth has probability 1.
  - However, this conflicts with common sense. Consider a complex tautology like $[P \leftrightarrow (Q \,\&\, \sim P)] \rightarrow \sim Q$. If one of my logic students picks this out of the air and believes it for no reason, we do not regard that as rational.
  - He should only believe it if he has good reason to believe it. In other words, rational belief requires reasons, and that conflicts with Bayesian epistemology.

Classical Decision Theory

- Classical decision theory has us choose acts one at a time on the basis of their expected values.
- It is *courses of action*, or *plans*, that must be evaluated decision-theoretically, and individual acts become rational by being prescribed by rationally adopted plans. (See Pollock, *Cognitive Carpentry*, chapter 5, MIT Press, 1995.)
- Furthermore, I will argue below that we cannot just take classical decision theory intact and apply it to plans. A plan is not automatically superior to a competitor just because it has a higher expected value.

Anthopomorphic Agents and Procedural Rationality

- The design of an anthropomorphic agent requires a general theory of rational cognition.
- The agent's cognition must be rational by human standards. Cognition is a ***process***, so this generates an essentially procedural concept of rationality.
  - Many AI researchers have followed Herbert Simon in rejecting such a procedural account, endorsing instead a satisficing account based on goal-satisfaction, but that is not applicable to anthropomorphic agents.

Procedural Rationality

- We do not necessarily want an anthropomorphic agent to model human cognition exactly.
  - We want it to draw rational conclusions and make rational decisions, but it need not do so in exactly the same way humans do it. How can we make sense of this?
- Stuart Russell (following Herbert Simon) suggests that the appropriate concept of rationality should only apply to the *ultimate results* of cognition, and not the course of cognition.

Procedural Rationality

- A conclusion or decision is *warranted* (relative to a system of cognition) iff it is endorsed "in the limit".
  - i.e., there is some stage of cognition at which it becomes endorsed and beyond which the endorsement is never retracted.
- We might require an agent architecture to have the same theory of warrant as human rational cognition.
  - This is to evaluate its behavior in the limit.
  - An agent that drew conclusions and made decisions at random for the first ten million years, and then started over again reasoning just like human beings would have the same theory of warrant, but it would not be a good agent design.
  - This is a problem for any assessment of agents in terms of the results of cognition in the limit.

Procedural Rationality

- It looks like the best we can do is require that the agent's reasoning never strays very far from the course of human reasoning.
  - If humans will draw a conclusion within a certain number of steps, the agent will do so within a "comparable" number of steps, and if a human will retract the conclusion within a certain number of further steps, the agent will do so within a "comparable" number of further steps. This is admittedly vague.
  - We might require that the worst-case difference be polynomial in the number of steps, or something like that. However, this proposal does not handle the case of the agent that draws conclusions randomly for the first ten million years.
- I am not going to endorse a solution to this problem. I just want to call attention to it, and urge that whatever the solution is, it seems reasonable to think that the kind of architecture I am about to describe satisfies the requisite constraints.

Goal-Oriented Agents

- Simple goal-oriented agents might be designed without any attention to how human beings achieve the same cognitive goals.
- I will argue that for sophisticated goal-oriented agents that aim at achieving complex goals in complicated environments, it will often be necessary to build in many of the same capabilities as are required for anthropomorphic agents.
  - I must wait to argue for this after I have said more about what is required for anthropomorphic agents.

Part Two
Two Concepts of Rationality

- Philosophers have traditionally investigated rationality by relying upon their "philosophical intuitions".
  - Example: The *Nicod Principle* proposes that sets of premises of the form "(Ac & Bc)" confirm the generalization "All A's are B's", for any choice of A and B.
  - There was a time when virtually all epistemologists endorsed the Nicod Principle in just this form.
  - Goodman [1955] startled the philosophical world by constructing an example that was taken to conclusively refute this version of the Nicod Principle.
  - x is grue if and only if either (1) x is green and first examined before the year 2000, or (2) x is blue and not first examined before the year 2000.
  - If we now (prior to the year 2000) examine lots of emeralds and find that they are all green, that gives us an inductive reason for thinking that all emeralds are green. Our sample of green emeralds is also a sample of grue emeralds, so by the Nicod principle our observations would also give us a reason for thinking that all emeralds are grue. These two conclusions together would entail the absurd consequence that there will be no emeralds first examined after the year 2000.

Human Rational Norms

- Goodman's refutation of the unrestricted Nicod Principle is conclusive, because everyone who looks at the example agrees that it would not be rational to accept the conclusions drawn in accordance with the Nicod Principle.
- This illustrates the use of thought experiments in investigating rationality.
  - Thought experiments provide data about what is rational or irrational in particular contexts, and then the philosopher constructs a general theory that is intended to capture that data.
  - The theory consists of a set of rational norms compliance with which is supposed to constitute rationality.

Procedural Knowledge

- Thought experiments, and the resulting theories of rationality, are based upon "philosophical intuitions", but what are they?
- My proposal is that they are analogous to the "linguistic intuitions" used by linguists in studying language.
  - The competence/performance distinction (Chomsky)
    - » Performance theories describe actual human performance
    - » Competence theories describe how we do it when we are doing it the way we "know how to do it" (procedural knowledge)
    - » Procedural knowledge carries with it the ability to detect divergences from the norms prescribing our procedural knowledge.
  - Similarly, we know how to cognize. Our philosophical intuitions are just an instance of the more general ability to detect divergences from the norms prescribing our procedural knowledge — in this case our procedural knowledge governing how to cognize.

Human Rationality

- The best way to understand the standard philosophical methodology for investigating rationality is to take it as an attempt to elicit the procedural norms comprising a competence theory of human cognition.
- I will refer to this concept of rationality as ***human rationality***.
  - Theories of human rationality are specifically about human cognition.
  - They describe contingent psychological features of the human cognitive architecture. However, upon reflection this seems to be a rather parochial view of rationality.
  - Some features of human rationality may be strongly motivated by general constraints imposed by the design problem rationality is intended to solve, but other features may be quite idiosyncratic, reflecting rather arbitrary design choices.
    - » Example: modus tollens: from ~Q and (P → Q) infer ~P.
  - An agent whose procedural knowledge for how to cognize differed from that of human beings would not automatically be irrational.
  - This seems to indicate the need for a more general "generic" concept of rationality.

The Design Stance

- We can think of rationality as the solution to certain design problems.
  - I will argue below that quite general features of these problems suffice to generate much of the structure of rationality.
  - General logical and feasibility constraints have the consequence that there is often only one obvious way of solving certain design problems, and this is the course taken by human rationality.
  - Much of the general structure of human rationality can be explained in this way.
  - However, the details of the design of human cognition are not similarly forthcoming.
  - It may often be the case that the details could be filled out in several different ways, and human cognition represents just one choice for how to do that.
  - There could be other rational agents, either natural or artificial, whose cognitive architectures represent somewhat different solutions to the same design problems.

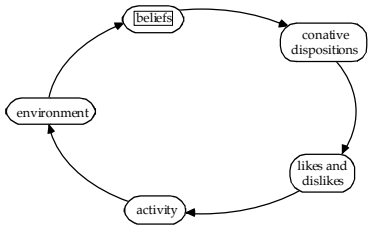
The Doxastic-Conative Loop

- I propose to understand generic rationality as attaching to any agent that constitutes a solution to the design problem that also generates the human cognitive architecture.
- However, if we are to explain rationality in this way, we must say what design problem rationality is designed to solve.
- I will begin by exploring one possibility, and then generalize the account.

The Doxastic-Conative Loop

- A simplistic view of evolution has it that evolutionary pressures select for traits that enhance the survivability of the creature.
- So we might regard the design problem for rationality to be that of creating an agent that can survive in a hostile world by virtue of its cognitive capabilities.
- This presupposes a prior understanding of what cognition is and how it might contribute to survivability.
- I take it as characteristic of rational cognition that a cognitive agent has doxastic states ("beliefs", broadly construed) reflecting the state of its environment, and conative states evaluating the environment as represented by the agent's beliefs.
- It is also equipped with cognitive mechanisms whereby it uses it beliefs and conations to select actions aimed at making the world more to its liking.

Schematic Rational Cognition



The Doxastic-Conative Loop

Generic Rationality

- Let us define a ***cognitive agent*** to be one implementing the doxastic-conative loop.
  - I take it that rationality pertains to cognitive agents.
  - We can assess a cognitive architecture in terms of how well it achieves its design goal.
  - However, different design goals are possible.
  - A natural design goal is to construct an agent capable of using its cognitive capabilties to survive in an uncooperative environment. This is motivated by considerations of evolution and natural selection, but it is based upon a simplistic view of evolution. For biological agents, a more natural design goal is propagation of the genome.
  - For artificial agents, we may have many different design goals.
    - » For example, in designing a Kamakaze robot warrior, the objective is neither survival nor propogation of the genome. This suggests that there is no privileged design goal in terms of which to evaluate cognitive architectures.

Generic Rationality

- It is striking, however, that for agents that implement the doxastic-conative loop, the same cognitive features seem to contribute to achievement of a wide range of design goals.
  - For example, for many different choices of design goal, an agent operating in a complex and uncooperative environment and intended to achieve its design goal by implementing the doxastic-conative loop will probably work better if it has a rather sophisticated system of mental representation, is capable of both deductive and defeasible reasoning, and can engage in long-range planning.
  - This is equally true of biological agents and Kamakaze robot warriors.
  - This suggests that there is a relatively neutral perspective from which we can evaluate cognitive architectures, and this in turn generates a generic concept of rationality.

Part Three: The OSCAR Architecture

- OSCAR is an architecture for rational agents based upon an evolving philosophical theory of rational cognition.
  - The general architecture is described in *Cognitive Carpentry* (MIT Press, 1995).
  - Related papers can be downloaded from http://www.u.arizona.edu/~pollock

An Architecture for Rational Cognition

- Epistemic cognition — about what to believe.
- Practical cognition — about what to do.
  - Epistemic cognition is skeptical; practical cognition is credulous.

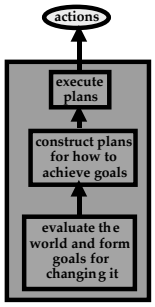
The Pre-eminence of Practical Cognition

- Most work on rational agents in AI has focussed on practical cognition rather than epistemic cognition, and for good reason.
- The whole point of an agent is *to do something*, to interact with the world, and such interaction is driven by practical cognition.
- From this perspective, epistemic cognition is subservient to practical cognition.

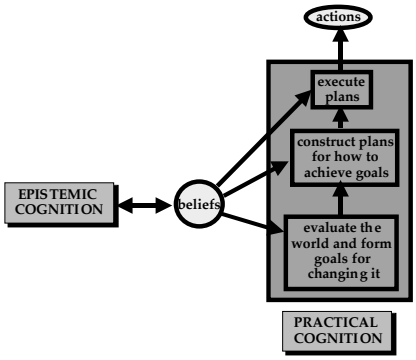
The Importance of Epistemic Cognition

- The OSCAR architecture differs from most agent architectures in that, although it is still practical cognition that directs the agent's interaction with the world, most of the work in rational cognition is performed by epistemic cognition.
  - Practical cognition evaluates the world (as represented by the agent's beliefs), and then poses queries concerning how to make it better.
  - These queries are passed to epistemic cognition, which tries to answer them by proposing plans.
  - Competing plans are evaluated and selected on the basis of their expected utilities, but those expected utilities are again computed by epistemic cognition.
  - Finally, plan execution generally requires a certain amount of monitoring to verify that things are going as planned, and that monitoring is again carried out by epistemic cognition.
  - In general, choices are made by practical cognition, but the information on which the choices are based is the product of epistemic cognition, and the bulk of the work in rational cognition goes into providing that information.

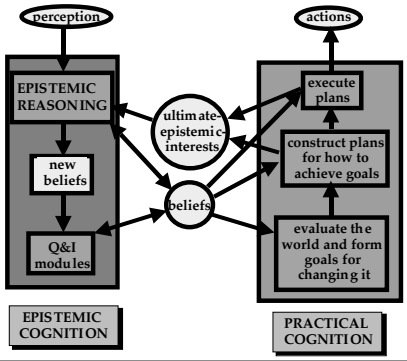
Practical Cognition



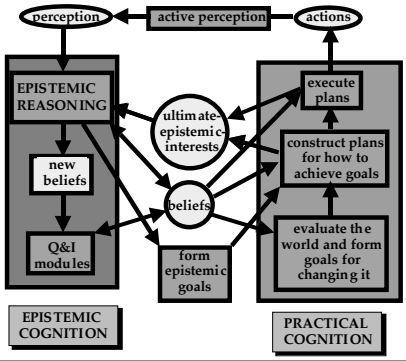
The Subservient Role of Epistemic Cognition

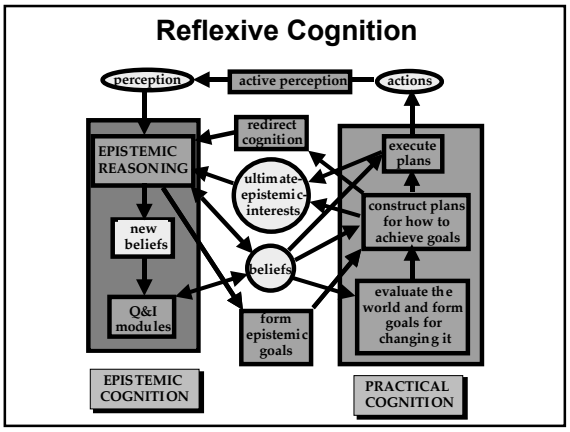


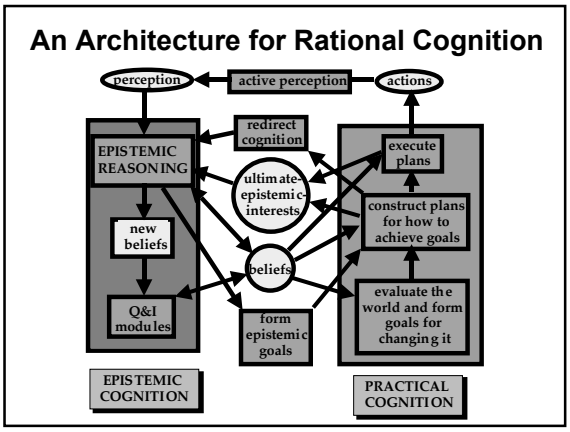
The Basic Interface



Empirical Investigation




Reflexive Cognition
perception
active perception
actions
EPISTEMIC REASONING
redirect cognition
execute plans
ultimate-epistemic-interests
construct plans for how to achieve goals
new beliefs
beliefs
Q&I modules
form epistemic goals
evaluate the world and form goals for changing it
EPISTEMIC COGNITION
PRACTICAL COGNITION


An Architecture for Rational Cognition
perception
active perception
actions
EPISTEMIC REASONING
redirect cognition
execute plans
ultimate-epistemic-interests
construct plans for how to achieve goals
new beliefs
beliefs
Q&I modules
form epistemic goals
evaluate the world and form goals for changing it
EPISTEMIC COGNITION
PRACTICAL COGNITION