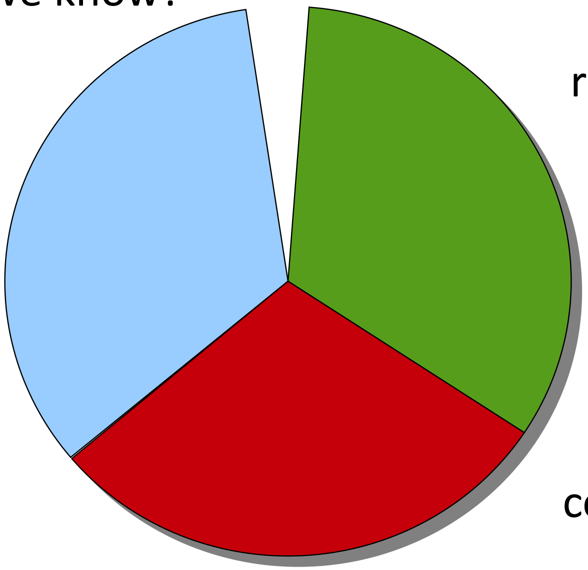
Behavioural Learning: Inspiration from Nature?

Chris Watkins
Dept of Computer Science
Royal Holloway, University of London

Plan of Talk



machine
learning

reinforcement
learning (RL)

RL as an honest
computational theory of
Skinner’s operant
conditioning

examples of
animal learning,
mis-learning, and
innate behaviour

learning and evolution:
why learn at all?

what don’t we know?

Machine learning: structure of the field

- Learning problems encountered either in technology or abstracted and simplified from biology/cognitive science

- Learning problems ***formalised*** as mathematical problems.
  - Formal definitions
  - Formal validation criteria so that different solutions can be compared
  - Public datasets, challenge problems

- Formalisation acts as an interface between engineers/computer scientists/mathematicians and … the study of learning. Research can proceed by considering the formalised problem.

- Study of formalised learning problems develops
  - algorithms + proofs of algorithm performance + technology
  - elaborations of problem definition occurs according to mathematical aesthetics/technological requirements

- Original cognitive science motivations can become a little forgotten…

Example: Supervised Learning

Given: an i.i.d. sample of examples x_1 , … x_N and a corresponding set y_1 , … , y_N of labels from some unknown probability distribution over (X,Y), and a loss function L:X,Y -> R^+

Find: a classification rule *f* : X -> Y with low expected loss on future samples from the same distribution.

Types of question:

- How does expected loss vary with size of sample (N), set of classification rules considered, loss function, etc.
- Efficient algorithms for finding *f*
- Variations of the problem (on-line mode, learning with privileged information, etc)

Is this related to cognitive science? Yes … but distantly

Reinforcement Learning

A family of learning algorithms that implement operant conditioning: learning from rewards and punishments.

RL algorithms learn to optimise cumulative reward over the medium term; can learn to take unpleasant preparatory actions that lead to rewards later.

RL can be viewed as a natural computational implementation of Thorndike’s ‘Law of Effect’ and of the folk psychology of learning from rewards and punishments.

Reinforcement Learning: motivating the model

Folk theory of training by reward and punishment: animals will learn to behave so as to seek rewards and avoid punishments.

Thorndike (1911) “Law of Effect”

“Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur.”

Reinforcement Learning: motivating the model

Early psychologists seemed so sure:

“Just as a sculptor carves a statue out of a block of marble, so does acquisition carve an activity out of a mass of random movements.”

Lloyd Morgan (1896)

“Formally the crab, fish, turtle, dog, cat, monkey, and baby have very similar intellects and characters. All are systems of connections subject to change by the laws of exercise and effort.”

Thorndike (1911)

“Learned behavior is constructed by a continual process of differential reinforcement from undifferentiated behavior.”

Skinner (1953)

RL: Intuitions

Formalise "learning from rewards and punishments":

"Situations" or **states:** a finite number of these $s_1 \ldots s_n$

Actions: finite number of actions possible in each state.

Rewards/costs: each time agent performs an action, the agent receives an immediate payoff that depends only on the state and the action performed.

State transitions: Performing an action takes agent to another state.

RL: Markov Decision Processes

States: s, observed by agent

Actions: a, chosen by agent

Payoffs: r, received by agent

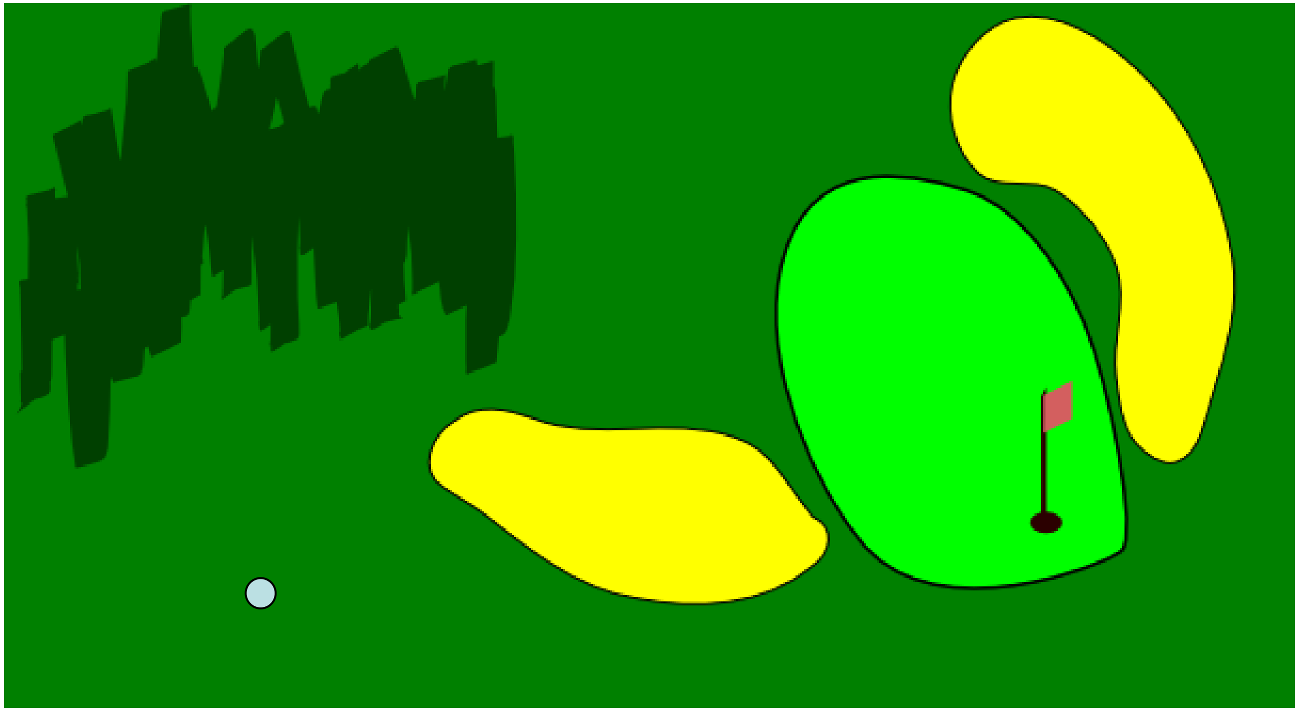
Transition probability distribution for each state,-action pair: P(s’ | s, a)

Markov property: Next state depends only on current state and action performed

Immediate payoff: r(s, a)

We suppose that the process terminates.

An example MDP: Robot Golf



State: Position of the ball

Actions: Settings for “stroke” by robot

Immediate payoff: each stroke costs 1

Aim: Minimise expected cost to get ball into hole.

Three types of control

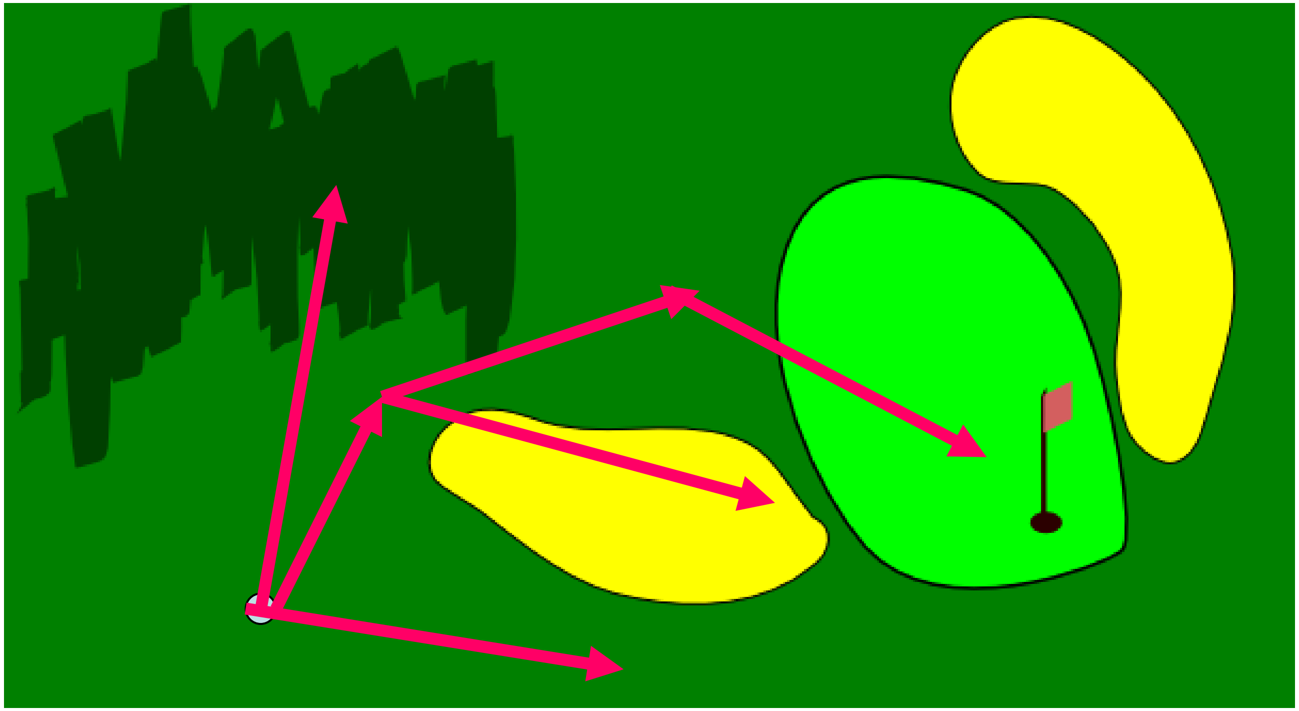
Look-ahead (multiple rollouts): plan a sequence of actions. Explore tree of possible sequences of actions to find a good route

Policy: For each location on golf course, have a stored action.

Value function (cost-to-go): Have a map of the golf course marked with the number of shots required to go to the hole.

Look ahead only one shot, to get to the next location with lowest cost-to-go.

Choosing actions: by look-ahead search

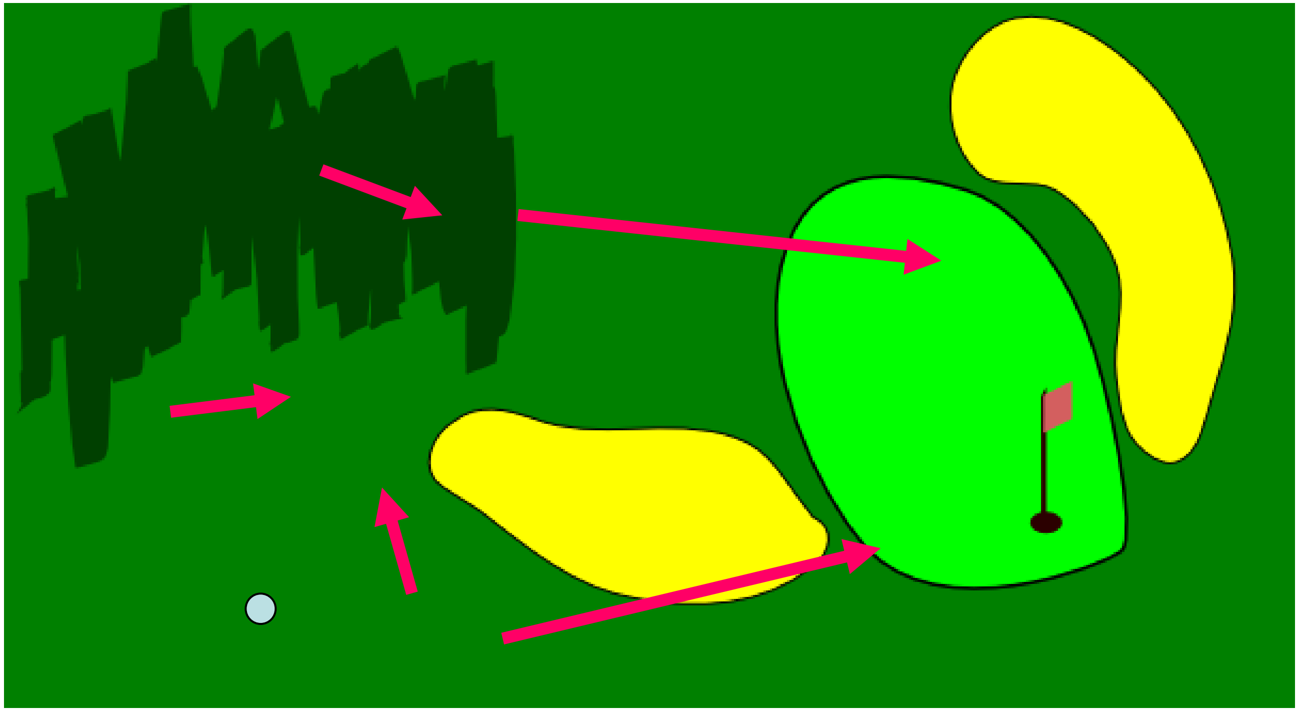


Computationally intensive

Needs a good model of effects of actions

No memory or learning

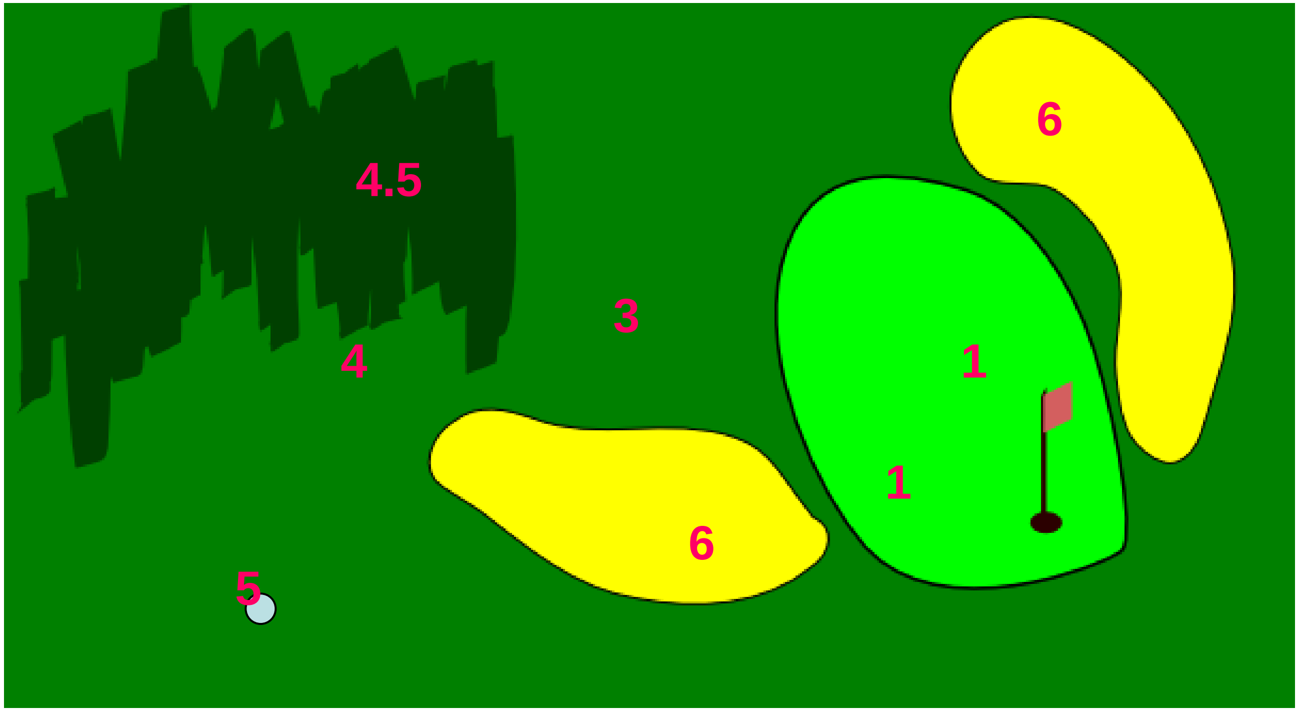
Choosing actions: by stored policy



Store a map of recommended shot for each position on course

In current state, look up the shot.

Choosing actions: by stored value function



Value of state is no. of shots needed from that state

Look-ahead **one shot** and pick state of lowest value

Optimising policy and values together

In current state:

Can I see a feasible shot that gives a lower value than the current state value?

- adjust the state value downwards

Can I see an action that leads to a better position than the current policy action?

- replace the policy action with a new action

Q-Learning

Even simpler, and needs no model of effects of actions.

Store a function Q(*s,a*) for each state-action pair.

Q(*s,a*) is (after learning) an estimate of the number of shots, starting in state *s*, taking shot *a,* then following policy thereafter.

“Policy” is: policy(s) = argmin_a Q(s,a)

“Value” is: value(s) = min_a Q(s,a)

Everything in one (rather large) stored table

Q-Learning

Learning:

Can use a model … but can also use the world as its own model

"Atom of experience" : [s a s' r]

Learning update:

$$Q(s,a) \leftarrow r + \min_b Q(s, b)$$

All operations on one stored table: no memory, no predictions.

Developments in RL

- Generalisation of value functions over large, continuous state spaces
- Partially observable Markov decision processes (POMDPs), in which current state is uncertain
- New families of algorithms: stochastic gradient and natural gradient
- Complexity analysis
- Analyses of exploration
- Hierarchical extensions (turns out to be hard)
- New probabilistic approaches for control…

Successes of RL

- Games
  - Checkers
  - Backgammon
  - Chess
  - Go
- Algorithms
  - New approaches to old problems in control
- New approaches to learning by imitation
  - Inverse reinforcement learning

Limitations of RL

- Hierarchical skills
- Competences to achieve multiple goals
- Forming new representations
- Slow

Plausibility of RL

- Satisfies same folk intuitions as radical behaviourism
- Robot designer specifies performance criterion: RL (hopefully) learns to achieve it
- Biologically motivated: an honest implementation of the “law of effect”, with improvements
  - Learning in continuous time
  - Whole family of alternative mental representations and learning algorithms
- A natural “upgrade path” ?
  - Simple organisms could use Q-learning, then get an evolutionary upgrade to predictive models and forward planning??

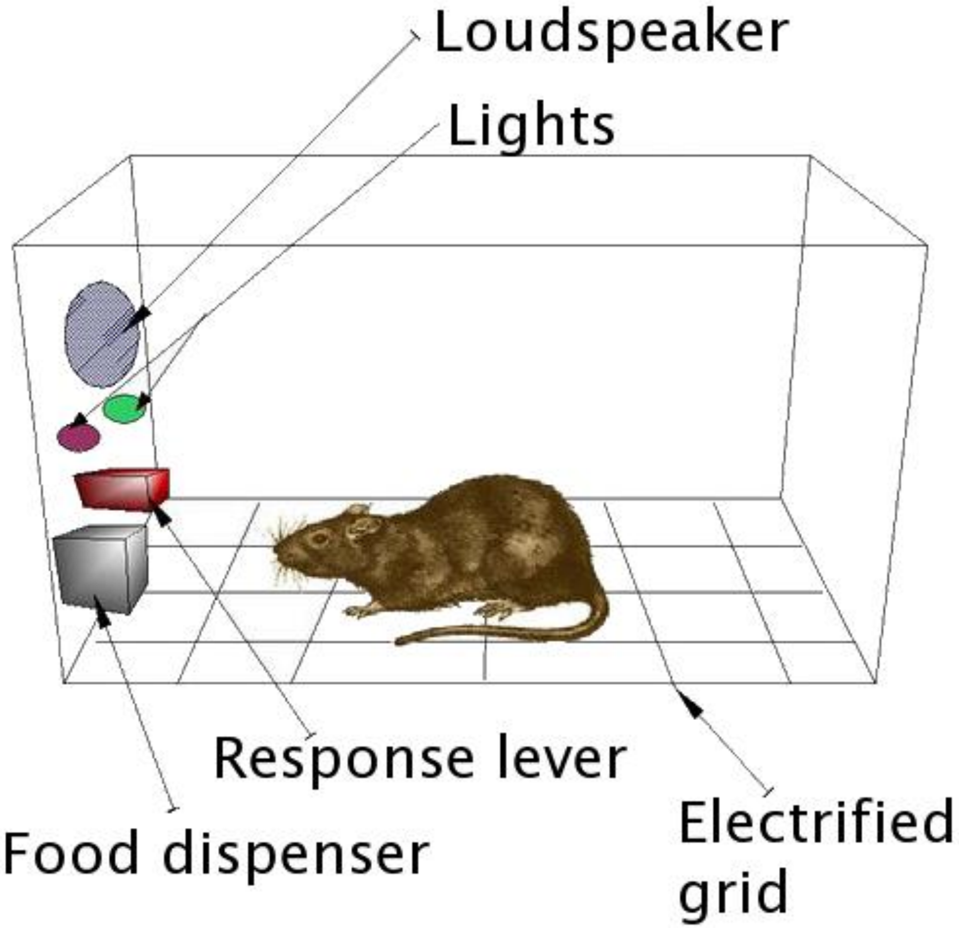
Part 2: RL in Animals

- Where animal reinforcement learning (conditioning) experiments go wrong: instinctive drift

- Extremes of innate behaviour: Cuckoos and their hosts

- Extremes of innate behaviour: the Megapodes

Conditioning Experiments



Skinner Box

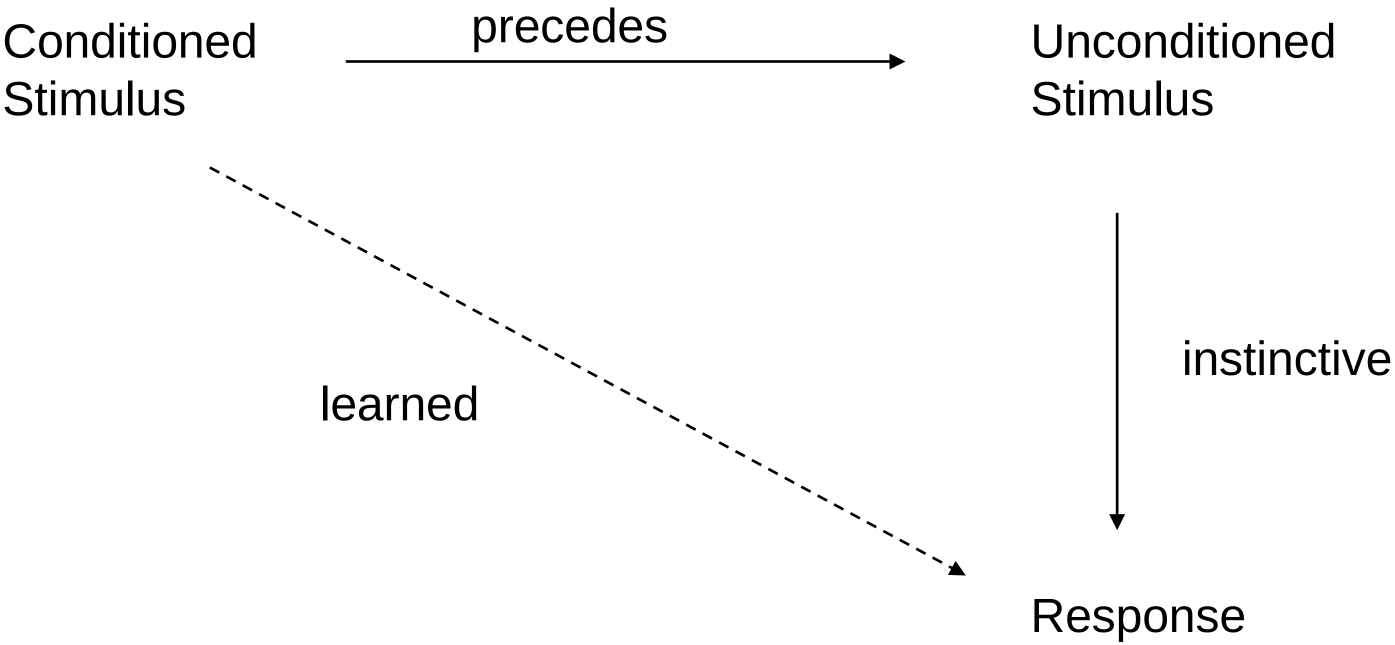
Simple environment

“Reinforcement schedule”

Simple responses

Reinforcement studied in its purest form?

Classical Conditioning



Conditioned Stimulus

precedes

Unconditioned Stimulus

instinctive

learned

Response

Keller and Marian Breland

Keller Breland with an otter

- Research students of B.F. Skinner in late 1940s
- Left before obtaining their PhDs to found an animal training company, using the new science of operant conditioning.
- Trained more than sixty species for novelty marketing displays, aquaria and zoos, movies and entertainment, military missions…
- After 10 years experience, they changed their views on operant conditioning theory.

Breland and Breland (1951)

Novelty animal acts:

“The success of these acts led to the development of a trained pig show, "Priscilla the Fastidious Pig," whose routine included turning on the radio, eating breakfast at a table, picking up the dirty clothes and putting them in a hamper, running the vacuum cleaner around, picking out her favorite feed from those of her competitors, and taking part in a quiz program, answering "Yes" or "No" to questions put by the audience, by lighting up the appropriate signs.”

‘The Misbehavior of Organisms’ (1961)

“The last instance we shall relate in detail is one of the most annoying and baffling for a good behaviorist. Here a pig was conditioned to pick up large wooden coins and deposit them in a large "piggy bank." The coins were placed several feet from the bank and the pig required to carry them to the bank and deposit them, usually four or five coins for one reinforcement. (Of course, we started out with one coin, near the bank.) “

The Misbehavior of Organisms

“Pigs condition very rapidly, they have no trouble taking ratios, they have ravenous appetites (naturally), and in many ways are among the most tractable animals we have worked with. However, this particular problem behavior developed in pig after pig, usually after a period of weeks or months, getting worse every day. “

“At first the pig would eagerly pick up one dollar, carry it to the bank, run back, get another; carry it rapidly and neatly, and so on, until the ratio was complete. “

The Misbehavior of Organisms

Thereafter, over a period of weeks the behavior would become slower and slower. He might run over eagerly for each dollar, but on the way back, instead of carrying the dollar and depositing it simply and cleanly, he would repeatedly drop it, root it, drop it again, root it along the way, pick it up, toss it up in the air, drop it, root it some more, and so on.

Rooting

Pigs find food by *rooting* – digging with their noses, and finding food by smell and touch.

To obtain a food reward by picking up a wooden object and then discarding it into a hole is contrary to their natural method of feeding.

The Misbehavior of Organisms

“We thought this behavior might simply be the dilly-dallying of an animal on a low drive. However, the behavior persisted and gained in strength in spite of a severely increased drive - he finally went through the ratios so slowly that he did not get enough to eat in the course of a day. Finally it would take the pig about 10 minutes to transport four coins a distance of about 6 feet. “

“This problem behavior developed repeatedly in successive pigs.”

Instinctive Drift

“The examples listed we feel represent a clear and utter failure of conditioning theory…..the animal simply does not do what he has been conditioned to do.”

“….wherever an animal has strong instinctive behaviors in the area of the conditioned response, after continued running the organism will drift toward the instinctive behavior to the detriment of the conditioned behavior… “

Questions

- Can we propose a more adequate theory of animal learning than one which is a combination of classical and operant conditioning?

- Why is instinctive drift so slow?
  - seems biologically useless if so slow

- Why hasn’t instinctive drift been more intensively studied?

Discussion

- RL can produce a powerful illusion of plausibility
  - paradigmatic example is to foraging for small morsels of food
  - in foraging, visible rewards correspond to animal’s internal subjective rewards
  - easy to forget that RL rewards are *subjective*

Example 2: Brood parasitism

The following example is an intricate behavioural repertoire in which rewards are hard to identify.

As evolutionarily aware observers, we can understand what is happening; but the animals don’t.

Point of example:

- Types of learning not naturally explainable with rewards
- Evolution puts strong constraints on learning

Brood parasitism: the Common Cuckoo

Cuckoo's egg in a reed-warbler's nest.

Cuckoo steals one egg in its beak, lays one to replace it, in less than 10 seconds, usually unobserved.

Cuckoo usually hatches first.

Immediately, while still blind, it ejects the host's eggs or chicks.

Hosts watch helplessly: they have no behavioural response.

Brood Parasitism

Hosts feed the cuckoo chick, which grows to 8 times their weight.

Cuckoo chicks take longer than host chicks to fledge: e.g. a pair of reed warblers may spend nearly 4 weeks longer rearing a cuckoo chick than they would for their own.

Hosts rarely desert the cuckoo chick: only one recent report…

95 bird species are obligate brood parasites

Sparrow feeding a shiny cowbird fledgling.

5 major groups of brood parasites worldwide.

Intense co-evolutionary `arms races’ between brood parasites and their hosts

Similar phenomena among the different groups

The Dangers of Theory

"It is wonderful to observe what great apparent delight the birds show when they see a female Cuckoo approach their abode....they seem quite beside themselves for joy. The little Wren...immediately quits its nest on the approach of the Cuckoo, as though to make room to enable her to lay her egg more commodiously. Meanwhile she hops round her with such expressions of delight that her husband at length joins her, and both seem lavish in their thanks for the honour which the great bird confers upon them by selecting their nest for its own use."

Bechstein, quoted in 1865

We consider only the hosts’ defences

Almost all hosts:

- evolve defences against parasitic eggs
- …but rear the parasitic chicks.

Why?

For host, more valuable to get rid of cuckoo egg than to reject the cuckoo chick…

But rejecting the chick at any stage enables hosts to conserve resources and possibly to re-nest.

An ability to reject a cuckoo chick would be advantageous enough to spread rapidly through host population.

Host defences against parasitic eggs

- Attack cuckoos on sight near nest.
- Defensive nests, with narrow entrance tunnels
- Recognise parasitic egg, and
  - eject it from nest
  - smash it
  - re-line nest, burying all the eggs
  - abandon nest

How does host recognise parasitic egg?

Surprisingly, host does *not* recognise the ‘odd one out’ among the eggs.

Classic experiments* show that hosts *imprint* on the appearance of the first egg they lay.

Host rejects subsequent eggs that are not sufficiently similar to the first (imprinted) egg.

* described in *Cuckoos, Cowbirds, and other Cheats* by N.B. Davies, 2000, Poyser

Why do hosts not imprint on their chicks also?

Lotem (1993)* noted that

- a host's first *egg* is her own

but

- a host's first hatched *chick* may well be a cuckoo, because cuckoo eggs hatch quickly (short incubation period).

If a host imprints on a cuckoo chick, then afterwards it would reject its own chicks and raise only cuckoos…

Host birds do not evolve ability to reject cuckoo chicks because "mental module" (imprinting) that is most available to use would cause them to reject their own chicks too frequently.

* Lotem, A., Nature 1993, **362,** pp743-744

Confirmation: Intra-species rejection of chicks*

Many birds drop eggs in each other’s nests: intra-specific brood parasitism.

Chicks of the same species look similar – yet parent birds can reject chicks that are not their own!

Key point: an intruder egg must be dropped into a nest that already has eggs in it.

Incubation period of intruder egg and host eggs is the same: host eggs hatch first.

American Coots

* Shizuka and Lyon, Nature 2009

Confirmation: Intra-species rejection of chicks

Parents imprint on first chicks hatched in each clutch.

Imprinting is precise enough that they can reject unrelated chicks that hatch afterwards.

Cross-fostering experiments showed that if unrelated chicks substituted for first hatched chicks, parents then reject their own chicks that hatch afterwards.

Not cost-free: mis-imprinting can occur in nature.

American Coots

Confirmation: Intra-species rejection of chicks

"... we observed rejection in action. … forms and intensities of parental aggression not seen in unparasitized broods, including actively seeking the chick from a distance to peck them vigorously and attempt to drown them, pecking chicks while brooding on the nest, and preventing chicks from access to the nest to be brooded

… [a] nest in which parents killed all of their own chicks after apparently mis-imprinting on chicks of a neighbouring pair.

These forms of parental aggression differed from the hostility that parent coots commonly use to control food allocation between surviving chicks."

Example 3: Megapodes

Australian brush turkey (+ 20 more species in Australasia)

Unique life history.

Eggs kept warm by being buried in huge mounds of manure

Chicks hatch with full feathers, burrow out of mound, run into bush, never see parents or siblings.

Fend for themselves.

Internal yolk gives them 48 hours to learn to survive in the outback.

Example 3: Megapodes

Needs more study!

Can stand, run fast, and soon fly (within 2 days)

Cautiously peck every salient object:

their feet, faeces, stones, seeds, anything moving

rapidly learn what is good to eat

Discover water by pecking salient objects floating on surface, then drink

Avoid objects moving towards them (either crouch or flee).

Example 3: A pure example of selection for innate competence

Low survival rate

Intense selection for rapid learning and complete innate behaviour

No cultural transmission or imitative learning

Significance: what are the limits of innate knowledge?

Not clear ! Some learning (including RL?) is needed – total adult behaviour is not immediate, but developed over a period of days/weeks.

Part 3: Learning and Evolution

1. Phenotypic plasticity
   - Genes cannot predict environment
2. Baldwin effects
3. Imitation, with and without culture
4. Learning as decoding of innate clues
5. Evolution is incremental optimisation

1. Why learn? Phenotypic Plasticity

Genes cannot predict the environment where animal finds itself

Different environments need different behaviour.

So animal should *learn* which environment it is in, and adopt the appropriate behaviour.

Continuum from moment-to-moment learning (perception) through to long-term skill learning.

Ideal form of learning is Bayesian inference to optimal behaviour.

2. Baldwin Effects

“Baldwin effect” is in two parts:

1. Learning allows behavioural plasticity: animals can learn new behaviours in new environments (eg on a new volcanic island).

2. If **learned behaviours are valuable** and **learning is costly/incomplete** then over evolutionary time, the learned behaviour may become innate.

Questions:

Can evolutionary learning and lifetime learning substitute for each other?

How important is behavioural plasticity in enabling behavioural evolution?

3. Cultural Transmission

- Parents can demonstrate their skills to children; children observe and imitate their parents.
  - A non-genetic channel for information to pass from one generation to the next.

- “Culture” in this sense is widespread in animals (chickens…)

- Questionable whether animal culture accumulates over many generations (imitation may not be accurate)

- Humans are distinct in that amount of information transmitted culturally seems enormously larger than for animals.

4. Imitation without culture

There is more genetic information in a population than in any single individual.

A naturalist observing a population might see innate adaptive behaviour that only some members show (eg fear of spiders).

But animals could make these observations too, and imitate salient behaviours that they see many other individuals doing.

This could enable *individuals* to exploit *population-level* genetic information.

No need to copy parents - can happen within a generation.

5. Does “Innate knowledge” require experience for its development?

Question: in what form is “innate behaviour” represented?

An animal’s whole behavioural repertoire is genetically specified: does process of cognitive development require experience?

Perhaps the most basic function of learning?

6. Evolution is incremental optimisation

Strong constraint on animal cognition.

Animal’s entire cognition is genetically specified.

Question: is one function “innate knowledge” actually to **limit** the nature and scope of what an animal can learn?

Animals of different sizes have brains of different sizes, but their behaviour may appear equally complex. (eg bats and cows)

Discussions: what don’t we know?

- “Cortex algorithm”? Does “innate knowledge” hold animals back?

- Do we need other paradigms of behavioural learning besides RL?

- What relevance does animal learning have for robots: should robots have “instincts” ?