Making Decisions

AI 109

Richard Kelley

Last Time

- Neural Representation
- Perceptrons

Today

- Making ***Decisions***
  - In Animals
  - In Humans
  - In Machines
- ***Expected Utility Theory***
- ***Exploration*** vs. ***Exploitation***

Natural things vs. Artifacts

What’s the difference?

Principles of Motion and Rest

- In his work the *Physics*, Aristotle was trying to understand the principles of ***change*** and ***motion*** and ***causality***.
  - The definition of Aristotle's "physics" is much broader than our definition today.
- In the *Physics*, Aristotle makes a distinction between natural things (think trees, people, etc.) and artifacts (manmade things like furniture):

Of things that exist, some exist by nature, some from other causes. By nature, the animals and their parts exist, and the plants and the simple bodies - for we say that these and the like exist 'by nature'.

All the things mentioned present a feature in which they differ from things which are not constituted by nature. ***Each of them has within itself a principle of motion and of stationariness.***

Motion and Rest (Continued)

- Aristotle continued:

On the other hand, a bed and a coat and anything else of that sort, qua receiving these designations i.e. in so far as they are products of art-have no innate impulse to change.

- There seems to be a fundamental distinction between things that move on their own and things that are moved by other things.
  - Natural substances (things) have *internal* sources of change.

Living Things vs Nonliving Things

What's the difference? The *most fundamental* difference?

Living Things vs. Nonliving Things (according to Aristotle)

- In his book *De Anima* (*On the Soul*), Aristotle gives the following:

> We say that a thing ***lives*** if it has *within itself* one or more of the following: intellect, perception, movement and rest in place, and also nourishment, growth, and decay.

- So living things are distinguished by having certain capacities for *acting*.
- Aristotle (and Aquinas following him) said that living things were distinguished by having ***souls***.
  - The soul is just what makes a living thing alive.

Souls

- Aristotle (and later Aquinas) distinguished three kinds of souls, corresponding to their capabilities:
  - **Plants (vegetative soul)**. The vegetative soul is the basic life principle that enables a living thing to take in nutrients, grow, and reproduce.
  - **Animals (sensitive soul)**. The sensitive soul includes the powers of sensation and appetite, allowing an organism to perceive its environment and move in response to what it perceives.
    - "**Sensitive**" means "having senses" rather than anything to do with emotions.
  - **Humans (rational soul)**. The rational soul includes the powers of intellect and reasoning, enabling a being to understand universal truths and deliberate about its actions.

Decision-making in Animals

- Animals can sense the world.
- Animals have *instincts*.
  - Your dog chases that squirrel by instinct, not because it understands anything about the universal concept of squirrels.
- Animals don't *reflect* on their choices. They follow their *appetites*.
- The process:
  - Perception -> appearance -> desire -> bodily movement.
  - ***Perception*** is the awareness of a thing through the senses.
  - ***Appearance*** is the internal image formed from perception. It remains even when the thing that caused it is not immediately present.
    - This is why the dog can search for the squirrel even after it climbs into the tree.

Decision-making in Humans

- Same basic process, with a (big) adjustment: in humans, the ***intellect*** can evaluate and reshape what appears good.
- Aquinas made this much more precise. The process:
  - ***External senses*** receive the object.
  - The ***internal senses*** (imagination, estimative power, memory) form and retain a “mental image”.
  - The intellect ***abstracts*** a universal concept from the phantasm.
  - The intellect ***judges*** something to be good (or not).
  - The will, as rational appetite, is moved by the good as presented by the intellect.
  - The will can command action.

Modern AI Approach

AIs vs Agents

- In modern AI, it is common to use the word ***Agent*** to refer to an AI program that can take actions in the world.
- Agent is (now) usually used in contrast with terms like LLM or chatbot.

Preferences

- Modern AI researchers ignore most of the complexity of what happens in humans and focus on *preferences*.
- This is probably because preferences are easier to describe mathematically than the process Aquinas described.
- A ***preference relation*** is a complete description of a set of things and for each pair, which is more preferred.
- Example. For Richard:
  - Vanilla bean ice cream > strawberry ice cream
  - Hamburgers > pork chops
  - Curly fries > pork chops
  - …

From Preferences to Utility

- In principle, preferences let us write down “what is good.”
- In AI, it’s common to require that preferences be ***rational***.
  - ***Transitive***: Preferences can’t have loops.
    - ice cream > fries
    - fries > pork chops
    - pork chops > ice cream.
  - ***Complete***: for everything (x and y) under consideration, x > y or y > x.
- In AI (and economics, and political science, and…), the word “rational” usually just means “complete and transitive.”
- This probably feels artificial, but it has one big advantage (to a computer scientist anyway): we can represent preferences *numbers*.

Utility Functions

- A ***utility function*** is an assignment of a number to everything under consideration, with the requirement that if one thing has a bigger number (called ***utility***) than another, it should be more preferred.
- We may write something like u(x) > u(y) if “x is more preferred than y.”
- A major assumption of modern AI is that we can build utility functions into our AI systems, and those utilities will capture everything we care about.
- ***Utilitarianism*** is the moral philosophy that humans have utility functions and we should always act to maximize “total utility” for society.
  - What’s good about this idea?
  - What’s bad?

Acting with utility functions

- In an AI system:
  - Perception -> utility calculation -> action.
    - The AI (or agent) “perceives” its options.
    - For each option a, the utility u(a) is calculated.
    - The agent does the action with the highest utility u(a).
- The notation for this is that a* = argmax u(a)
- The ***argmax*** is the input that maximizes an output.
- What’s good about this approach?
- What’s bad about this approach?

Probability

- One problem with the above approach is that life is uncertain: we don't know how to act because the effects of our actions aren't deterministic.
- ***Probability*** gives us a way to rigorously talk about uncertainty.
- Probability is *everywhere* in AI. We're not going to do a lot of math, but we'll need to get some key ideas to talk about more advanced AI systems.
- There are two interpretations of probability:
  - The ***frequentist*** interpretation. Probabilities are frequencies of occurrence.
  - The ***Bayesian*** interpretation. Probabilities are ***degrees of belief***.
- Which interpretation seems more reasonable?

Examples of probabilities

- What is the probability that
  - Fair coin comes up heads.
  - Biased coin comes up heads.
  - A standard 6-sided die comes up 1.
  - The sun comes up tomorrow.
  - Shakespeare wrote *Hamlet*.
  - The accused is guilty.
  - The optimal action for the self-driving car is to accelerate.
  - The next word in after “The quick brown” is “cat.”
    - This one is very important for AI.

Updating Beliefs

- In AI, a ***belief*** is just a variable with an attached probability.
  - I believe that there's a 70% chance of rain tomorrow.
- ***Conditional probabilities*** let us talk about probabilities given some kind of evidence or background information.
  - Given the forecast for tomorrow says rain, I believe there is a 90% chance of rain.
- Some terminology from probability gives us a way to talk about updating beliefs in response to ***evidence*** (learning):
  - **Priors**. Represent our beliefs before we see a piece of evidence.
  - **Likelihoods**. Represent a model of how evidence relates to beliefs.
  - **Posteriors**. Refer to a belief that been updated in response to evidence.
- ***Bayes' Rule*** is a mathematical rule that describes how to update beliefs in response to evidence (how to compute posteriors from likelihoods and priors).

Demo

Making Choices

- We can now use the above to say how AIs are programmed (more accurately *trained*) to act.
- Given a set of uncertain options for action:
  - Assign a probability to each action outcome.
  - Assign a utility to each action outcome.
  - Multiply each utility by its probability. This gives an ***expected utility*** for each outcome.
  - Choose the action that has the highest expected utility over its possible outcomes.

Exploration vs. Exploitation

- Whenever an agent has to act in the world, it wants to act optimally.
  - This requires knowing utilities or ***payoffs***.
- But it may have to take some actions to figure out what the utilities should be.
- The ***k-armed bandit*** is a model of this.

Demo