Chapter 9: Decisions under Uncertainty

Making Decisions Under Uncertainty

What an agent should do depends on:

- The agent's ability — what options are available to it.
- The agent's beliefs — the ways the world could be, given the agent's knowledge.
  Sensing updates the agent's beliefs.
- The agent's preferences — what the agent wants and tradeoffs when there are risks.

Decision theory specifies how to trade off the desirability and probabilities of the possible outcomes for competing actions.

Making Decisions Under Uncertainty

An agent acts to

- affect the outside world
  - i.e. open a door
- change the relationship between the agent and the outside world
  - i.e. move to the kitchen
- aquire more information about the outside world (active sensing, communication)
  - i.e. looking behind the curtain, asking for help
- control its internal reasoning
  - i.e. selecting the next search state

Goals and Preferences

Alice . . . went on "Would you please tell me, please, which way I ought to go from here?"
"That depends a good deal on where you want to get to," said the Cat.
"I don't much care where —" said Alice.
"Then it doesn't matter which way you go," said the Cat.

Lewis Carroll, 1832–1898
Alice's Adventures in Wonderland, 1865
Chapter 6

Preferences

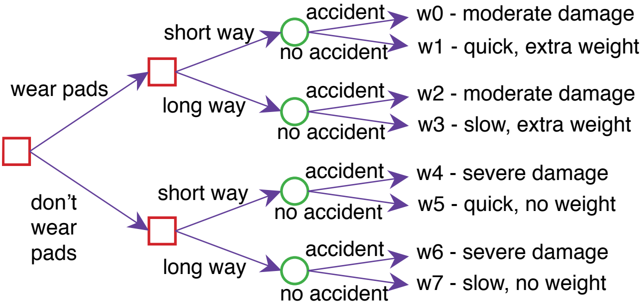
- Actions result in outcomes
- Agents have preferences over outcomes
- A rational agent will do the action that has the best outcome for them
- Sometimes agents don't know the outcomes of the actions, but they still need to compare actions
- Agents have to act.
  (Doing nothing is (often) an action).

Decision Variables

- Decision variables are like random variables that an agent gets to choose a value for.
- A possible world specifies a value for each decision variable and each random variable.
- For each assignment of values to all decision variables, the measure of the set of worlds satisfying that assignment sum to 1.
- The probability of a proposition is undefined unless the agent conditions on the values of all decision variables.

Decision Tree for Delivery Robot

The robot can choose to wear pads to protect itself or not.
The robot can choose to go the short way past the stairs or a long way that reduces the chance of an accident.
There is one random variable of whether there is an accident.



Expected Values

- The expected value of a function of possible worlds is its average value, weighting possible worlds by their probability.
- Suppose $f(\omega)$ is the value of function f on world ω.
  - The expected value of f is
    $$\mathcal{E}(f) = \sum_{\omega \in \Omega} P(\omega) \times f(\omega).$$
  - The conditional expected value of f given e is
    $$\mathcal{E}(f|e) = \sum_{\omega \models e} P(\omega|e) \times f(\omega).$$

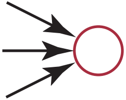
Utility

- Utility is a measure of desirability of worlds to an agent.
- Let $u(\omega)$ be the utility of world ω to the agent.
- Simple goals can be specified by: worlds that satisfy the goal have utility 1; other worlds have utility 0.
- Often utilities are more complicated: for example some function of the amount of damage to a robot, how much energy is left, what goals are achieved, and how much time it has taken.

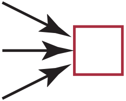
Decision Networks

- A decision network is a graphical representation of a finite (sequential) decision problem.
- Decision networks extend belief networks to include decision variables and utility.
- A decision network specifies what information is available when the agent has to act.
- A decision network specifies which variables the utility depends on.

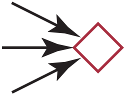
Decisions Networks



- A random variable is drawn as an ellipse. Arcs into the node represent probabilistic dependence.

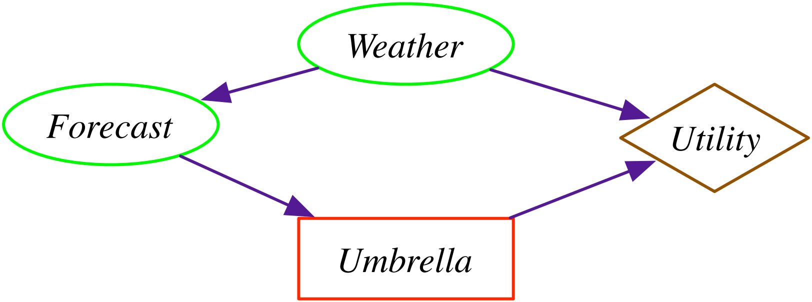


- A decision variable is drawn as an rectangle. Arcs into the node represent information available when the decision is taken.



- A utility node is drawn as a diamond. Arcs into the node represent variables that the utility depends on.

Umbrella Decision Network



You don't get to observe the weather when you have to decide whether to take your umbrella. You do get to observe the forecast.

Single decisions

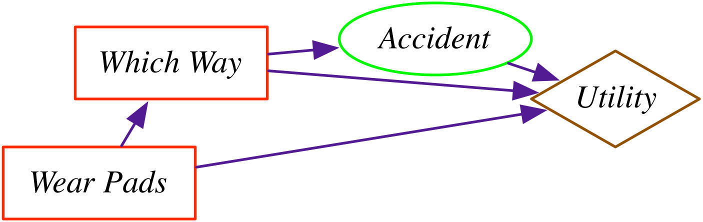
- In a single decision variable, the agent can choose $D = d_i$ for any $d_i \in dom(D)$.
- The expected utility of decision $D = d_i$ is $\mathcal{E}(u|D = d_i)$.
- An optimal single decision is the decision $D = d_{max}$ whose expected utility is maximal:

$$\mathcal{E}(u|D = d_{max}) = \max_{d_i \in dom(D)} \mathcal{E}(u|D = d_i).$$

Single-stage decision networks

Extend belief networks with:

- Decision nodes, that the agent chooses the value for. Domain is the set of possible actions. Drawn as rectangle.
- Utility nodes, the parents are the variables on which the utility depends. Drawn as a diamond.



This shows explicitly which nodes affect whether there is an accident and what affects the expected utility.

Finding the optimal decision

- Suppose the random variables are $X_1, \ldots, X_n$, and utility depends on $X_{i_1}, \ldots, X_{i_k}$

$$\begin{aligned}\mathcal{E}(u|D) &= \sum_{X_1,\ldots,X_n} P(X_1,\ldots,X_n|D) \times u(X_{i_1},\ldots,X_{i_k}) \\ &= \sum_{X_1,\ldots,X_n} \prod_{i=1}^{n} P(X_i|parents(X_i)) \times u(X_{i_1},\ldots,X_{i_k})\end{aligned}$$

  To find the optimal decision:
  - Create a factor for each conditional probability and for the utility
  - Sum out all of the random variables
  - This creates a factor on D that gives the expected utility for each D
  - Choose the D with the maximum value in the factor.

Example Initial Factors

Which Way	Accident	Value
long	true	0.01
long	false	0.99
short	true	0.2
short	false	0.8

Which Way	Accident	Wear Pads	Value
long	true	true	30
long	true	false	0
long	false	true	75
long	false	false	80
short	true	true	35
short	true	false	3
short	false	true	95
short	false	false	100

After summing out Accident

Which Way	Wear Pads	Value
long	true	74.55
long	false	79.2
short	true	83.0
short	false	80.6

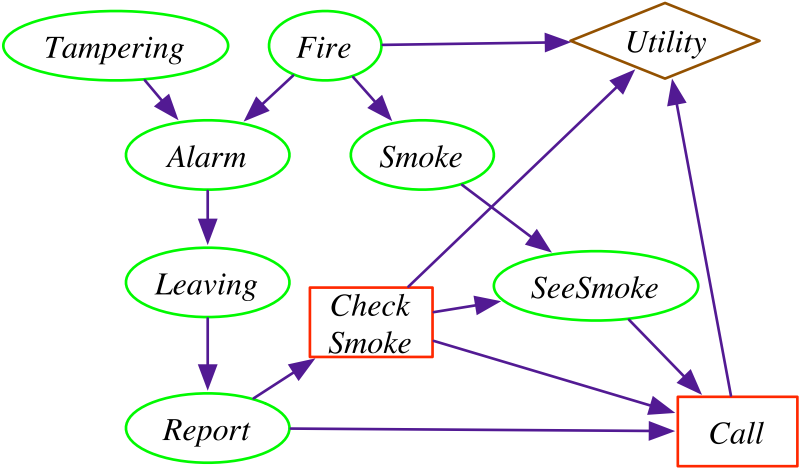
Sequential Decisions

- An intelligent agent doesn't carry out a multi-step plan ignoring information it receives in between steps.
- A more typical scenario is where the agent:
  observes, acts, observes, acts, . . .
- Subsequent actions can depend on what is observed.
  What is observed depends on previous actions.
- Often the sole reason for carrying out an action is to provide information for future actions.
  For example: diagnostic tests, spying.

Sequential decision problems

- A sequential decision problem consists of a sequence of decision variables $D_1, \ldots, D_n$.
- Each D_i has an information set of variables $parents(D_i)$, whose value will be known at the time decision D_i is made.

Decision Network for the Alarm Problem



No-forgetting

A No-forgetting decision network is a decision network where:

- The decision nodes are totally ordered. This is the order the actions will be taken.
- All decision nodes that come before D_i are parents of decision node D_i. Thus the agent remembers its previous actions.
- Any parent of a decision node is a parent of subsequent decision nodes. Thus the agent remembers its previous observations.

What should an agent do?

- What an agent should do at any time depends on what it will do in the future.
- What an agent does in the future depends on what it did before.

Policies

- A policy specifies what an agent should do under each circumstance.
- A policy is a sequence $\delta_1, \ldots, \delta_n$ of decision functions

$$\delta_i : dom(parents(D_i)) \to dom(D_i).$$

This policy means that when the agent has observed $O \in dom(parents(D_i))$, it will do $\delta_i(O)$.

Expected Utility of a Policy

- Possible world ω satisfies policy δ, written $\omega \models \delta$ if the world assigns the value to each decision node that the policy specifies.
- The expected utility of policy δ is

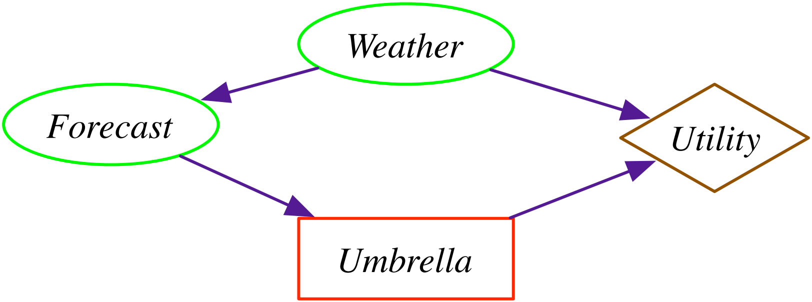
$$\mathcal{E}(u|\delta) = \sum_{\omega \models \delta} u(\omega) \times P(\omega),$$

- An optimal policy is one with the highest expected utility.

Finding the optimal policy

- Remove all variables that are not ancestors of the utility node
- Create a factor for each conditional probability table and a factor for the utility.
- Sum out variables that are not parents of a decision node.
- Select a variable D that is only in a factor f with (some of) its parents.
- Eliminate D by maximizing. This returns:
  - the optimal decision function for D, $\arg\max_D f$
  - a new factor to use in VE, $\max_D f$
- Repeat till there are no more decision nodes.
- Eliminate the remaining random variables. Multiply the factors: this is the expected utility of the optimal policy.

Umbrella Decision Network



You don't get to observe the weather when you have to decide whether to take your umbrella. You do get to observe the forecast.

Initial factors for the Umbrella Decision

Weather	Value
norain	0.7
rain	0.3

Weather	Fcast	Value
norain	sunny	0.7
norain	cloudy	0.2
norain	rainy	0.1
rain	sunny	0.15
rain	cloudy	0.25
rain	rainy	0.6

Weather	Umb	Value
norain	take	20
norain	leave	100
rain	take	70
rain	leave	0

Eliminating By Maximizing

f:

Fcast	Umb	Val
sunny	take	12.95
sunny	leave	49.0
cloudy	take	8.05
cloudy	leave	14.0
rainy	take	14.0
rainy	leave	7.0

$\max_{Umb} f$:

Fcast	Val
sunny	49.0
cloudy	14.0
rainy	14.0

$\arg\max_{Umb} f$:

Fcast	Umb
sunny	leave
cloudy	leave
rainy	take

Complexity of finding the optimal policy

- If there are k binary parents, to a decision D, there are 2^k assignments of values to the parents.
- If there are b possible actions, there are b^{2^k} different value assignments for the decision function.
- The number of policies is the product of the decision functions for the different decision nodes.

Complexity of finding the optimal policy

- If there are m decision nodes, and if there are k_i parents for decision node D_i, and D_i has domain size b_i the number of policies is:

$$\prod_{i=1}^{m} d_i^{2^{k_i}}$$

- The number of optimizations in the dynamic programming is the sum of the number of assignments of values to parents.

$$\sum_{i=1}^{m} 2^{k_i}$$

- The dynamic programming algorithm is much more efficient than searching through policy space.

Value of Information

- The value of information X for decision D is the utility of the network with an arc from X to D ($+$ no-forgetting arcs) minus the utility of the network without the arc.
- The value of information is always non-negative.
- It is positive only if the agent changes its action depending on X.
- The value of information provides a bound on how much an agent should be prepared to pay for a sensor. How much is a better weather forecast worth?
- We need to be careful when adding an arc would create a cycle. E.g., how much would it be worth knowing whether the fire truck will arrive quickly when deciding whether to call them?

Value of Control

- The value of control of a variable X is the value of the network when you make X a decision variable (and add no-forgetting arcs) minus the value of the network when X is a random variable.
- You need to be explicit about what information is available when you control X.
- If you control X without observing, controlling X can be worse than observing X. E.g., controlling a thermometer.
- If you keep the parents the same, the value of control is always non-negative.

Modelling Preferences

If o_1 and o_2 are outcomes of an action

- $o_1 \succeq o_2$ means o_1 is at least as desirable as o_2.
- $o_1 \sim o_2$ means $o_1 \succeq o_2$ and $o_2 \succeq o_1$.
- $o_1 \succ o_2$ means $o_1 \succeq o_2$ and $o_2 \not\succeq o_1$

Lotteries

- An agent may not know the outcomes of their actions, but only have a probability distribution of the outcomes.
- A lottery is a probability distribution over outcomes. It is written

  $$[p_1 : o_1, p_2 : o_2, \ldots, p_k : o_k]$$

  where the o_i are outcomes and $p_i \geq 0$ such that

  $$\sum_i p_i = 1$$

  The lottery specifies that outcome o_i occurs with probability p_i.
- When we talk about outcomes, we will include lotteries.

Properties of Preferences

- Completeness: Agents have to act, so they must have preferences:

  $$\forall o_1 \forall o_2 \;\; o_1 \succeq o_2 \text{ or } o_2 \succeq o_1$$

- Transitivity: Preferences must be transitive:

  $$\text{if } o_1 \succeq o_2 \text{ and } o_2 \succ o_3 \text{ then } o_1 \succ o_3$$

  (Similarly for other mixtures of $\succ$ and $\succeq$.)

Properties of Preferences

- Completeness: Agents have to act, so they must have preferences:

$$\forall o_1 \forall o_2 \; o_1 \succeq o_2 \text{ or } o_2 \succeq o_1$$

- Transitivity: Preferences must be transitive:

$$\text{if } o_1 \succeq o_2 \text{ and } o_2 \succ o_3 \text{ then } o_1 \succ o_3$$

(Similarly for other mixtures of $\succ$ and $\succeq$.)
Rationale: otherwise $o_1 \succeq o_2$ and $o_2 \succ o_3$ and $o_3 \succeq o_1$.
If they are prepared to pay to get o_2 instead of o_3,
and are happy to have o_1 instead of o_2,
and are happy to have o_3 instead of o_1
$\longrightarrow$ money pump.

Properties of Preferences (cont.)

Monotonicity: An agent prefers a larger chance of getting a better outcome than a smaller chance:

- If $o_1 \succ o_2$ and $p > q$ then

$$[p : o_1, 1 - p : o_2] \succ [q : o_1, 1 - q : o_2]$$

Consequence of axioms

- Suppose $o_1 \succ o_2$ and $o_2 \succ o_3$. Consider whether the agent would prefer
  - o_2
  - the lottery $[p : o_1, 1 - p : o_3]$

  for different values of $p \in [0, 1]$.
- Plot which one is preferred as a function of p:

Properties of Preferences (cont.)

Continuity: Suppose $o_1 \succ o_2$ and $o_2 \succ o_3$, then there exists a $p \in [0, 1]$ such that

$$o_2 \sim [p : o_1, 1 - p : o_3]$$

Properties of Preferences (cont.)

Decomposability: (no fun in gambling). An agent is indifferent between lotteries that have same probabilities and outcomes. This includes lotteries over lotteries. For example:

$$[p : o_1, 1 - p : [q : o_2, 1 - q : o_3]] \\ \sim \quad [p : o_1, (1 - p)q : o_2, (1 - p)(1 - q) : o_3]$$

Properties of Preferences (cont.)

Substitutability: if $o_1 \sim o_2$ then the agent is indifferent between lotteries that only differ by o_1 and o_2:

$$[p : o_1, 1 - p : o_3] \sim [p : o_2, 1 - p : o_3]$$

Alternative Axiom for Substitutability

Substitutability: if $o_1 \succeq o_2$ then the agent weakly prefers lotteries that contain o_1 instead of o_2, everything else being equal.
That is, for any number p and outcome o_3:

$$[p : o_1, (1-p) : o_3] \succeq [p : o_2, (1-p) : o_3]$$

What we would like

- We would like a measure of preference that can be combined with probabilities. So that

$$\begin{aligned} & value([p : o_1, 1-p : o_2]) \\ = \quad & p \times value(o_1) + (1-p) \times value(o_2) \end{aligned}$$

- Money does not act like this.
  What would you prefer

$$\$1,000,000 \text{ or } [0.5 : \$0, 0.5 : \$2,000,000]?$$

- It may seem that preferences are too complex and muti-faceted to be represented by single numbers.

Theorem

If preferences follow the preceding properties, then preferences can be measured by a function

$$utility : outcomes \to [0, 1]$$

such that

- $o_1 \succeq o_2$ if and only if $utility(o_1) \geq utility(o2)$.
- Utilities are linear with probabilities:

$$utility([p_1 : o_1, p_2 : o_2, \ldots, p_k : o_k]) = \sum_{i=1}^{k} p_i \times utility(o_i)$$

Proof

- If all outcomes are equally preferred, set $utility(o_i) = 0$ for all outcomes o_i.
- Otherwise, suppose the best outcome is *best* and the worst outcome is *worst*.
- For any outcome o_i, define $utility(o_i)$ to be the number u_i such that

$$o_i \sim [u_i : best, 1 - u_i : worst]$$

This exists by the Continuity property.

Proof (cont.)

- Suppose $o_1 \succeq o_2$ and $utility(o_i) = u_i$, then by Substitutability,

$$\begin{aligned} & [u_1 : best, 1 - u_1 : worst] \\ \succeq \ & [u_2 : best, 1 - u_2 : worst] \end{aligned}$$

  Which, by completeness and monotonicity implies $u_1 \geq u_2$.

Proof (cont.)

- Suppose $p = utility([p_1 : o_1, p_2 : o_2, \ldots, p_k : o_k])$.
- Suppose $utility(o_i) = u_i$. We know:

$$o_i \sim [u_i : best, 1 - u_i : worst]$$

- By substitutability, we can replace each o_i by $[u_i : best, 1 - u_i : worst]$, so

$$\begin{aligned} p = utility(\quad [\quad & p_1 : [u_1 : best, 1 - u_1 : worst] \\ & \ldots \\ & p_k : [u_k : best, 1 - u_k : worst]]) \end{aligned}$$

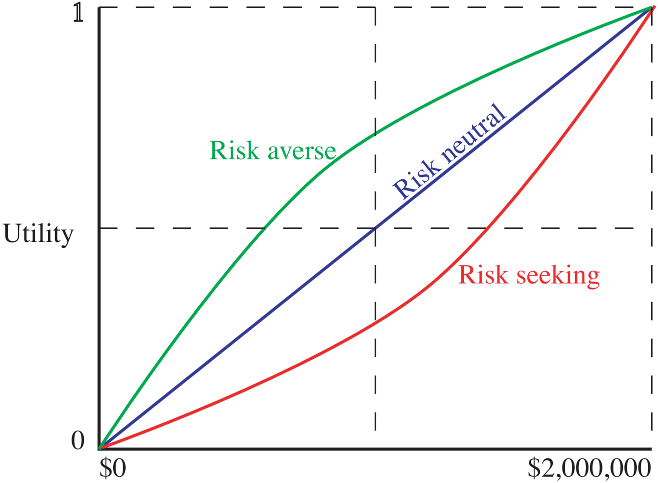
- By decomposability, this is equivalent to:

$$\begin{aligned} p = utility(\quad [\quad & p_1 u_1 + \cdots + p_k u_k \\ & \quad : best, \\ & p_1(1-u_1) + \cdots + p_k(1-u_k) \\ & \quad : worst]]) \end{aligned}$$

- Thus, by definition of utility,

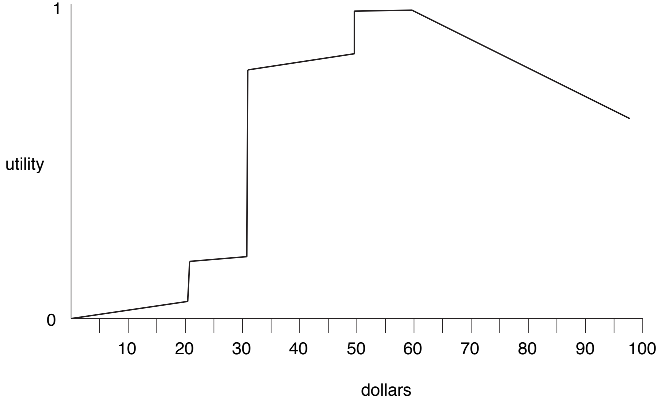
$$p = p_1 \times u_1 + \cdots + p_k \times u_k$$

Utility as a function of money



Possible utility as a function of money

Someone who really wants a toy worth \$30, but who would also like one worth \$20:



Allais Paradox (1953)

What would you prefer:

- A: $\$1m$ — one million dollars
- B: lottery $[0.10 : \$2.5m, 0.89 : \$1m, 0.01 : \$0]$

Allais Paradox (1953)

What would you prefer:

A: $\$1m$ — one million dollars

B: lottery $[0.10 : \$2.5m, 0.89 : \$1m, 0.01 : \$0]$

What would you prefer:

C: lottery $[0.11 : \$1m, 0.89 : \$0]$

D: lottery $[0.10 : \$2.5m, 0.9 : \$0]$

Allais Paradox (1953)

What would you prefer:

A: $\$1m$ — one million dollars

B: lottery $[0.10 : \$2.5m, 0.89 : \$1m, 0.01 : \$0]$

What would you prefer:

C: lottery $[0.11 : \$1m, 0.89 : \$0]$

D: lottery $[0.10 : \$2.5m, 0.9 : \$0]$

It is inconsistent with the axioms of preferences to have $A \succ B$ and $D \succ C$.

Allais Paradox (1953)

What would you prefer:

A: $\$1m$ — one million dollars

B: lottery $[0.10 : \$2.5m, 0.89 : \$1m, 0.01 : \$0]$

What would you prefer:

C: lottery $[0.11 : \$1m, 0.89 : \$0]$

D: lottery $[0.10 : \$2.5m, 0.9 : \$0]$

It is inconsistent with the axioms of preferences to have $A \succ B$ and $D \succ C$.

A,C: lottery $[0.11 : \$1m, 0.89 : X]$

B,D: lottery $[0.10 : \$2.5m, 0.01 : \$0, 0.89 : X]$

Framing Effects [Tversky and Kahneman]

- A disease is expected to kill 600 people. Two alternative programs have been proposed:

  Program A: 200 people will be saved
  Program B: probability $1/3$: 600 people will be saved
  probability $2/3$: no one will be saved

  Which program would you favor?

Framing Effects [Tversky and Kahneman]

- A disease is expected to kill 600 people. Two alternative programs have been proposed:

  Program C: 400 people will die

  Program D: probability $1/3$: no one will die
  probability $2/3$: 600 will die

  Which program would you favor?

Framing Effects [Tversky and Kahneman]

- A disease is expected to kill 600 people. Two alternative programs have been proposed:

  Program A: 200 people will be saved
  Program B: probability 1/3: 600 people will be saved
  probability 2/3: no one will be saved

  Which program would you favor?

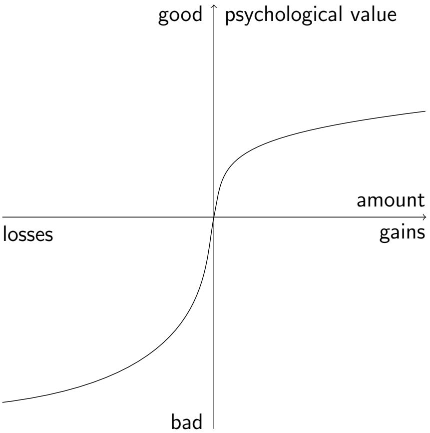
- A disease is expected to kill 600 people. Two alternative programs have been proposed:

  Program C: 400 people will die
  Program D: probability 1/3: no one will die
  probability 2/3: 600 will die

  Which program would you favor?

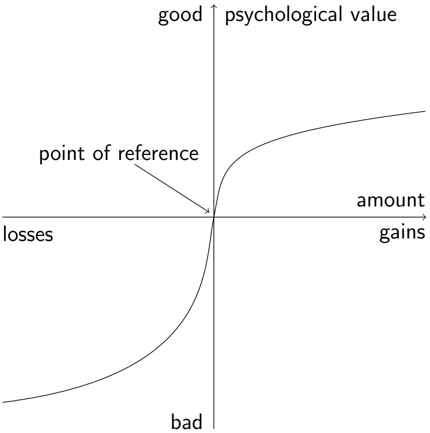
Tversky and Kahneman: 72% chose A over B.
22% chose C over D.

Prospect Theory



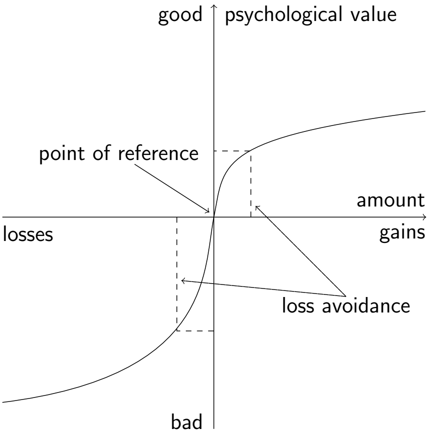
- In mixed gambles, loss aversion causes extreme risk-averse choices
- In bad choices, diminishing responsibility causes risk seeking.

Prospect Theory



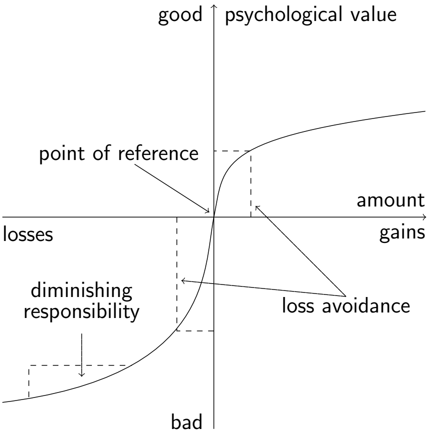
- In mixed gambles, loss aversion causes extreme risk-averse choices
- In bad choices, diminishing responsibility causes risk seeking.

Prospect Theory



- In mixed gambles, loss aversion causes extreme risk-averse choices
- In bad choices, diminishing responsibility causes risk seeking.

Prospect Theory



- In mixed gambles, loss aversion causes extreme risk-averse choices
- In bad choices, diminishing responsibility causes risk seeking.

Reference Points

Consider Anthony and Betty:

- Anthony's current wealth is \$1 million.
- Betty's current wealth is \$4 million.

They are both offered the choice between a gamble and a sure thing:

- Gamble: equal chance to end up owning \$1 million or \$4 million.
- Sure thing: own \$2 million

What does expected utility theory predict?

Reference Points

Consider Anthony and Betty:

- Anthony's current wealth is \$1 million.
- Betty's current wealth is \$4 million.

They are both offered the choice between a gamble and a sure thing:

- Gamble: equal chance to end up owning \$1 million or \$4 million.
- Sure thing: own \$2 million

What does expected utility theory predict?
What does prospect theory predict?

[From D. Kahneman, *Thinking, Fast and Slow*, 2011, pp. 275-276.]

Framing Effects

What do you think of Alan and Ben:

- Alan: intelligent—industrious—impulsive—critical—stubborn—envious

Framing Effects

What do you think of Alan and Ben:

- Ben: envious—stubborn—critical—impulsive—industrious—intelligent

Framing Effects

What do you think of Alan and Ben:

- Alan: intelligent—industrious—impulsive—critical—stubborn—envious
- Ben: envious—stubborn—critical—impulsive—industrious—intelligent

[From D. Kahneman, Thinking Fast and Slow, 2011, p. 82]

Framing Effects

- Suppose you had bought tickets for the theatre for \$50. When you got to the theatre, you had lost the tickets. You have your credit card and can buy equivalent tickets for \$50. Do you buy the replacement tickets on your credit card?

Framing Effects

- Suppose you had bought tickets for the theatre for \$50. When you got to the theatre, you had lost the tickets. You have your credit card and can buy equivalent tickets for \$50. Do you buy the replacement tickets on your credit card?
- Suppose you had \$50 in your pocket to buy tickets. When you got to the theatre, you had lost the \$50. You have your credit card and can buy equivalent tickets for \$50. Do you buy the tickets on your credit card?

[From R.M. Dawes, Rational Choice in an Uncertain World, 1988.]

The Ellsberg Paradox (1961)

There are 90 chips in a bag, 30 red, the other yellow or black.

What do you prefer?

A: You will win if a red chip is drawn, yellow and black are blanks

B: You will win if a yellow chip is drawn, red and black are blanks

The Ellsberg Paradox (1961)

There are 90 chips in a bag, 30 red, the other yellow or black.

What do you prefer?

A: You will win if a red chip is drawn, yellow and black are blanks

B: You will win if a yellow chip is drawn, red and black are blanks

C: You will win if a red or black chip is drawn, yellow is a blank

D: You will win if a yellow or black chip is drawn, red is a blank

The Ellsberg Paradox (1961)

There are 90 chips in a bag, 30 red, the other yellow or black.

What do you prefer?

A: You will win if a red chip is drawn,
yellow and black are blanks

B: You will win if a yellow chip is drawn,
red and black are blanks

C: You will win if a red or black chip is drawn,
yellow is a blank

D: You will win if a yellow or black chip is drawn,
red is a blank

People prefer A over B and D over C. Why?

The Ellsberg Paradox (1961)

There are 90 chips in a bag, 30 red, the other yellow or black.

What do you prefer?

- A: You will win if a red chip is drawn, yellow and black are blanks
- B: You will win if a yellow chip is drawn, red and black are blanks
- C: You will win if a red or black chip is drawn, yellow is a blank
- D: You will win if a yellow or black chip is drawn, red is a blank

People prefer A over B and D over C. Why?
Distinction between risk (probability distribution is known) and ambiguity (probability distribution is not known). If possible, people avoid ambiguity.

The limits of Rational Behavior

Daniel Kahnemann, Amon Tversky (1979): 12 reasons for irrational behavior

- overconfidence/over-confidentiality bias: overestimation of ones own capabilities, courage, and the ability to influence the future
- arrogance: underestimation of the abilities of opponents
- anchoring effect: opinions (even from unreliable sources) become a self-fulfilling prophecy
- stubbornness: positions, once adopted are kept
- familiarity bias: perception is biased in favour of the already known, alternatives will be ignored
- status quo bias: greater risks are taken to keep the status quo instead of changing it

The limits of Rational Behavior

- loss aversion: losses are feared more than gains are welcomed
- wrong priorities: unreasonably much effort is devoted to small decisions while too little attention is paid to the big ones
- inopportune regret: much time is spent on regretting a loss, without any benefit
- whitewashing: wrong decisions are euphemized
- manipulation: a decision is adopted more readily if motivated by a fear of loss, instead of potential gains
- priming: decisions are influenced by earlier, memorized and usually unconcious experiences and expectations
- foreboding: decisions are influenced by the ability to anticipate the future

If humans do not act rationally, should artificial agents do as well?

If humans do not act rationally, should artificial agents do as well?

No, but ...

If humans do not act rationally, should artificial agents do as well?

No, but ...

... they should be able to take the human deviation from rationality into consideration.

Factored Representation of Utility

- So far, utility has been described in terms of states.
- Usually, too many states have to be distinguished.
- Alternatively describing possible outcomes in terms of features $X_1, \ldots, X_n$.
- An additive utility is one that can be decomposed into set of factors:

$$u(X_1, \ldots, X_n) = f_1(X_1) + \cdots + f_n(X_n).$$

  This assumes additive independence.
- Strong assumption: contribution of each feature doesn't depend on other features.
- Many ways to represent the same utility:
  — a number can be added to one factor as long as it is subtracted from others.

Additive Utility

- An additive utility has a canonical representation:

$$u(X_1, \ldots, X_n) = w_1 \times u_1(X_1) + \cdots + w_n \times u_n(X_n).$$

- If $best_i$ is the best value of X_i, $u_i(X_i{=}best_i) = 1$.
  If $worst_i$ is the worst value of X_i, $u_i(X_i{=}worst_i) = 0$.
- w_i are weights, $\sum_i w_i = 1$.
  The weights reflect the relative importance of features.
- We can determine weights by comparing outcomes.

$$w_1 =$$

Additive Utility

- An additive utility has a canonical representation:

$$u(X_1, \ldots, X_n) = w_1 \times u_1(X_1) + \cdots + w_n \times u_n(X_n).$$

- If $best_i$ is the best value of X_i, $u_i(X_i{=}best_i) = 1$.
  If $worst_i$ is the worst value of X_i, $u_i(X_i{=}worst_i) = 0$.
- w_i are weights, $\sum_i w_i = 1$.
  The weights reflect the relative importance of features.
- We can determine weights by comparing outcomes.

$$w_1 = u(best_1, x_2, \ldots, x_n) - u(worst_1, x_2, \ldots, x_n).$$

for any values $x_2, \ldots, x_n$ of $X_2, \ldots, X_n$.

Complements and Substitutes

- Often additive independence is not a good assumption.
- Values x_1 of feature X_1 and x_2 of feature X_2 are **complements** if having both is better than the sum of the two.
  - E.g. booking a hotel room and an airplane ticket together.
  - E.g. two outings on the same day, if the locations are in close proximity.

Complements and Substitutes

- Often additive independence is not a good assumption.
- Values x_1 of feature X_1 and x_2 of feature X_2 are complements if having both is better than the sum of the two.
  - E.g. booking a hotel room and an airplane ticket together.
  - E.g. two outings on the same day, if the locations are in close proximity.
- Values x_1 of feature X_1 and x_2 of feature X_2 are substitutes if having both is worse than the sum of the two.
  - E.g. two lengthy outings into opposite directions on the same day.

Generalized Additive Utility

- A generalized additive utility can be written as a sum of factors:

$$u(X_1, \dots, X_n) = f_1(\overline{X_1}) + \dots + f_k(\overline{X_k})$$

  where $\overline{X_i} \subseteq \{X_1, \dots, X_n\}$.
- An intuitive canonical representation is difficult to find.
- It can represent complements and substitutes.

Utility and time

- Would you prefer \$1000 today or \$1000 next year?
- What price would you pay now to have an eternity of happiness?
- How can you trade off pleasures today with pleasures in the future?

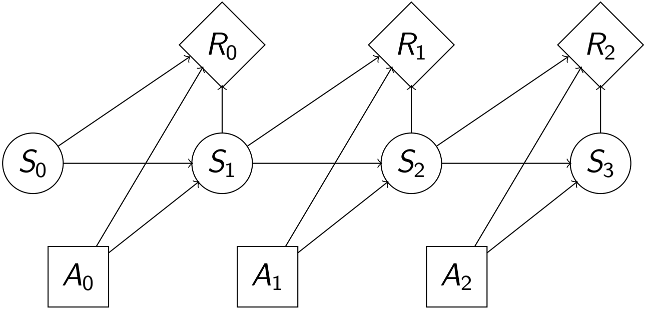
Utility and time

- How would you compare the following sequences of rewards (per week):

  A: \$1000000, \$0, \$0, \$0, \$0, \$0,...
  B: \$1000, \$1000, \$1000, \$1000, \$1000,...
  C: \$1000, \$0, \$0, \$0, \$0,...
  D: \$1, \$1, \$1, \$1, \$1,...
  E: \$1, \$2, \$3, \$4, \$5,...

Markov decision processes

- augmenting a Markov chain with actions



- fully or partially observable processes (MDP/POMDP)
- stationary models: state transitions and rewards do not depend on time

Markov decision processes

- Can the agent go on forever?
  - no: indefinite horizon problem
  - yes: infinite horizon problem
- utility has to be estimated continously, since the agent might never be able to reach an end state

Rewards and Values

Suppose the agent receives a sequence of rewards $r_1, r_2, r_3, r_4, \ldots$ in time. Three different possibilities to compute the utility

- total reward $V = \sum_{i=1}^{\infty} r_i$
- average reward $V = \lim_{n \to \infty} (r_1 + \cdots + r_n)/n$
- discounted return $V = r_1 + \gamma r_2 + \gamma^2 r_3 + \gamma^3 r_4 + \cdots$

  γ is the discount factor $0 \le \gamma \le 1$.

Properties of the Discounted Rewards

- The discounted return for rewards $r_1, r_2, r_3, r_4, \ldots$ is

$$\begin{aligned} V &= r_1 + \gamma r_2 + \gamma^2 r_3 + \gamma^3 r_4 + \cdots \\ &= \end{aligned}$$

Properties of the Discounted Rewards

- The discounted return for rewards $r_1, r_2, r_3, r_4, \ldots$ is

$$
\begin{aligned}
V &= r_1 + \gamma r_2 + \gamma^2 r_3 + \gamma^3 r_4 + \cdots \\
&= r_1 + \gamma(r_2 + \gamma(r_3 + \gamma(r_4 + \ldots)))
\end{aligned}
$$

- If $V(t)$ is the value obtained until time step t

$$V(t) =$$

Properties of the Discounted Rewards

- The discounted return for rewards $r_1, r_2, r_3, r_4, \ldots$ is
$$\begin{aligned} V &= r_1 + \gamma r_2 + \gamma^2 r_3 + \gamma^3 r_4 + \cdots \\ &= r_1 + \gamma(r_2 + \gamma(r_3 + \gamma(r_4 + \ldots))) \end{aligned}$$
- If $V(t)$ is the value obtained until time step t
$$V(t) = r_t + \gamma V(t-1)$$
- How is the infinite future valued compared to immediate rewards?

Properties of the Discounted Rewards

- The discounted return for rewards $r_1, r_2, r_3, r_4, \ldots$ is
$$\begin{aligned} V &= r_1 + \gamma r_2 + \gamma^2 r_3 + \gamma^3 r_4 + \cdots \\ &= r_1 + \gamma(r_2 + \gamma(r_3 + \gamma(r_4 + \ldots))) \end{aligned}$$
- If $V(t)$ is the value obtained until time step t
$$V(t) = r_t + \gamma V(t-1)$$
- How is the infinite future valued compared to immediate rewards?
$1 + \gamma + \gamma^2 + \gamma^3 + \cdots = 1/(1-\gamma)$
Therefore $\dfrac{\text{minimum reward}}{1-\gamma} \leq V(t) \leq \dfrac{\text{maximum reward}}{1-\gamma}$
- We can approximate V with the first k terms, with error:
$$V - (r_1 + \gamma r_2 + \cdots + \gamma^{k-1} r_k) = \gamma^k V(k+1)$$