Self-Interested Agents and Utility Theory

What does it mean to say that an agent is self-interested?

- not that they want to harm other agents
- not that they only care about things that benefit them
- that the agent has its own description of states of the world that it likes, and that its actions are motivated by this description

Self-interested agents are individuals or entities that make decisions based on their personal preferences and goals, typically aiming to maximize their own utility, benefit, or payoff. In economics, game theory, and mechanism design, agents are often assumed to be rational and self-interested, meaning they will act to achieve the best possible outcome for themselves, regardless of the overall social welfare or the well-being of others.

Characteristics of Self-Interested Agents:

1. **Rationality**: Self-interested agents are rational, meaning they make decisions that maximize their personal utility based on the information available to them.
2. **Maximizing Personal Payoff**: They focus on maximizing their individual gain or minimizing their loss. Their actions are guided by a cost-benefit analysis of the options available.
3. **Strategic Behavior**: Agents anticipate the actions of other agents and take them into account when making decisions. They may act strategically to influence the outcome in their favor.
4. **Incentive-Driven**: Self-interested agents respond to incentives. Their behavior changes in response to rewards, punishments, or costs associated with different choices.
5. **No Altruism**: These agents are not necessarily concerned with the welfare of others unless it directly affects their own outcomes. They are motivated by self-interest rather than cooperative or altruistic behavior.

Role in Game Theory:

In **game theory**, self-interested agents (players) interact in strategic settings, where the outcome for each player depends not only on their own choices but also on the choices of others. Key concepts related to self-interested agents include:

- **Nash Equilibrium**: A situation in which no agent can improve their payoff by unilaterally changing their strategy, given the strategies of others.
- **Dominant Strategy**: A strategy that is best for an agent, regardless of what the other players do.
- **Prisoner's Dilemma**: A classic example of self-interested behavior, where two rational agents might not cooperate, even though cooperation would lead to a better outcome for both.

Key Challenges with Self-Interested Agents:

- **Misaligned Incentives**: If the agents' self-interest does not align with the overall goals of the system or society, it can lead to suboptimal outcomes (e.g., free-rider problems, tragedy of the commons).
- **Strategic Manipulation**: Agents may manipulate the system if the mechanism is not designed to prevent such behavior, resulting in inefficiencies or unfair outcomes.
- **Coordination Problems**: In certain situations, self-interested behavior can prevent cooperation or coordination that would otherwise benefit all agents (as seen in the Prisoner's Dilemma).

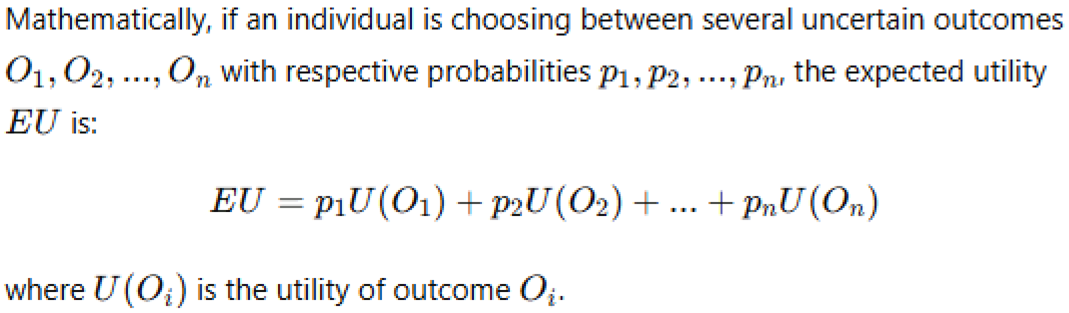
Utility theory:

- quantifies degree of preference across alternatives
- understand the impact of uncertainty on these preferences
- utility function: a mapping from states of the world to real numbers, indicating the agent's level of happiness with that state of the world
- Decision-theoretic rationality: take actions to maximize expected utility.

Utility theory is a fundamental concept in economics and decision theory that deals with how individuals or agents make choices under conditions of uncertainty or trade-offs. The theory posits that agents make decisions by comparing the expected utility (a measure of satisfaction or benefit) they derive from different options and choose the one that maximizes their utility.

Key Concepts of Utility Theory:

1. **Utility**: Utility is a numerical representation of an individual's preferences. It reflects the satisfaction or happiness an individual derives from consuming goods or services, or from different outcomes in decision-making. Utility can be thought of as a subjective measure of well-being.
   - **Cardinal Utility**: This assumes that utility can be measured and compared numerically, with meaningful differences (e.g., saying "option A gives me twice as much utility as option B").
   - **Ordinal Utility**: This assumes only that options can be ranked in terms of preference, but the differences in utility between options are not quantified (e.g., "I prefer A to B, but I don't quantify how much more I prefer it").
2. **Expected Utility**: In situations involving uncertainty or risk, individuals are assumed to maximize their **expected utility** rather than their utility. The **expected utility theorem** says that when faced with risky choices, rational agents will choose the option with the highest expected utility, calculated as the weighted sum of utilities over possible outcomes, with the weights being the probabilities of those outcomes.



Mathematically, if an individual is choosing between several uncertain outcomes $O_1, O_2, ..., O_n$ with respective probabilities $p_1, p_2, ..., p_n$, the expected utility EU is:

$$EU = p_1U(O_1) + p_2U(O_2) + ... + p_nU(O_n)$$

where $U(O_i)$ is the utility of outcome O_i.

Utility **Function**: A utility function represents an individual's preferences over different choices or outcomes. It assigns a number to each possible outcome, indicating the level of satisfaction (utility) the individual derives from it.

- For example, if x represents an amount of money, a utility function could be $U(x) = \sqrt{x}$,

  meaning that the individual derives utility from x in a diminishing manner (i.e., gaining $1000 when you already have $10,000 is less satisfying than when you have $100).

Risk Aversion, Risk Neutrality, and Risk Seeking:

- **Risk-Averse**: A person prefers certainty over a gamble with the same expected value. The utility function of a risk-averse individual is concave (e.g., $U(x) = \sqrt{x}$) , reflecting diminishing marginal utility of wealth.
- **Risk-Neutral**: A person is indifferent to risk; they care only about maximizing the expected value, not the risk. The utility function is linear (e.g., U(x)=x).
- **Risk-Seeking**: A person prefers a gamble to a certain outcome with the same expected value. The utility function is convex (e.g., U(x)=x^2).

Von Neumann-Morgenstern Utility Theorem: This theorem formalizes expected utility theory. It states that if an individual's preferences over uncertain outcomes satisfy certain rationality axioms (completeness, transitivity, independence, and continuity), then those preferences can be represented by an expected utility function.

- **Axioms**:
  - **Completeness**: An agent can compare any two options and has a preference or is indifferent between them.
  - **Transitivity**: If option A is preferred to B and B is preferred to C, then A is preferred to C.
  - **Independence**: If an agent is indifferent between two outcomes A and B, they should remain indifferent if the same probability of a third outcome is added to both A and B.
  - **Continuity**: There should exist some probabilities that make an agent indifferent between a certain outcome and a lottery between two other outcomes.

Utility Theorem in Practice:

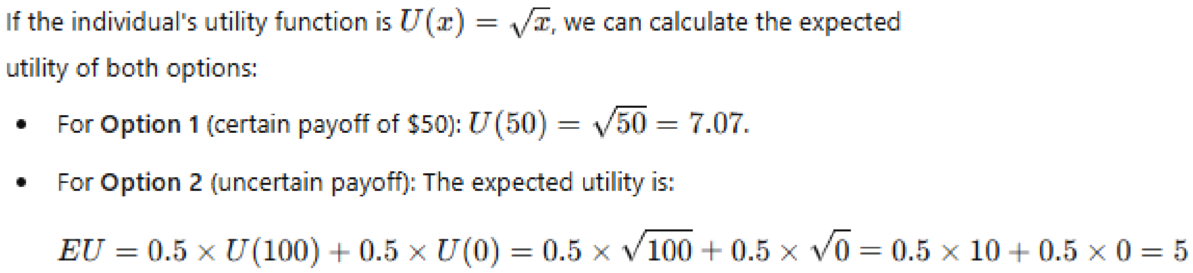
The **expected utility theorem** is widely used in fields such as economics, finance, and decision-making under uncertainty. Some real-world applications include:

- **Gambling and Insurance**: Risk-averse individuals prefer to avoid risk and thus buy insurance, while risk-seeking individuals may prefer gambling.
- **Investment Decisions**: Investors make decisions about asset allocation based on the expected utility of returns, balancing the trade-off between risk and reward.
- **Auction Design**: In auctions, bidders aim to maximize their expected utility by balancing the potential gain from winning an item with the risk of paying more than they are willing.

Example:

Consider an individual choosing between:

- **Option 1**: A certain payoff of $50.
- **Option 2**: A 50% chance of winning $100 and a 50% chance of winning $0.



If the individual's utility function is $U(x) = \sqrt{x}$, we can calculate the expected utility of both options:

- For **Option 1** (certain payoff of $50): $U(50) = \sqrt{50} = 7.07$.
- For **Option 2** (uncertain payoff): The expected utility is:

$$EU = 0.5 \times U(100) + 0.5 \times U(0) = 0.5 \times \sqrt{100} + 0.5 \times \sqrt{0} = 0.5 \times 10 + 0.5 \times 0 = 5$$

Since the expected utility of Option 1 (7.07) is greater than that of Option 2 (5), a risk-averse individual with this utility function would choose the certain payoff of $50.

Example: friends and enemies

- Alice has three options: club (c), movie (m), watching a video at home (h)
- On her own, her utility for these three outcomes is 100 for c, 50 for m and 50 for h
- However, Alice also cares about Bob (who she hates) and Carol (who she likes)
  - Bob is at the club 60% of the time, and at the movies otherwise
  - Carol is at the movies 75% of the time, and at the club otherwise
- If Alice runs into Bob at the movies, she suffers disutility of 40; if she sees him at the club she suffers disutility of 90.
- If Alice sees Carol, she enjoys whatever activity she's doing 1.5 times as much as she would have enjoyed it otherwise (taking into account the possible disutility caused by Bob)

What activity should Alice choose?

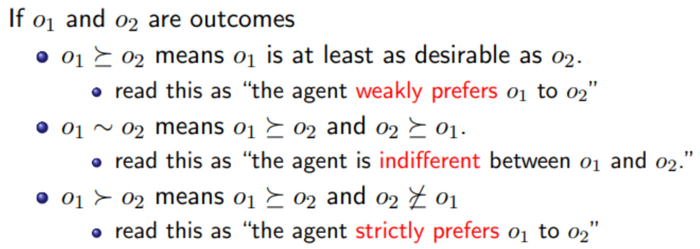
	$B = c$	$B = m$
$C = c$	15	150
$C = m$	10	100

$A = c$

	$B = c$	$B = m$
$C = c$	50	10
$C = m$	75	15

$A = m$

- Alice's expected utility for c:
  $0.25(0.6 \cdot 15 + 0.4 \cdot 150) + 0.75(0.6 \cdot 10 + 0.4 \cdot 100) = 51.75.$
- Alice's expected utility for m:
- $0.25(0.6 \cdot 50 + 0.4 \cdot 10) + 0.75(0.6(75) + 0.4(15)) = 46.75.$
- Alice's expected utility for h: 50.
  Alice prefers to go to the club (though Bob is often there and Carol rarely is), and prefers staying home to going to the movies (though Bob is usually not at the movies and Carol almost always is).
  Why utility?
  Why would anyone argue with the idea that an agent's preferences could be described using a utility function as we just did?
  Preferences Over Outcomes



If o_1 and o_2 are outcomes

- $o_1 \succeq o_2$ means o_1 is at least as desirable as o_2.
  - read this as "the agent weakly prefers o_1 to o_2"
- $o_1 \sim o_2$ means $o_1 \succeq o_2$ and $o_2 \succeq o_1$.
  - read this as "the agent is indifferent between o_1 and o_2."
- $o_1 \succ o_2$ means $o_1 \succeq o_2$ and $o_2 \not\succeq o_1$
  - read this as "the agent strictly prefers o_1 to o_2"

Lotteries

An agent may not know the outcomes of his actions, but may instead only have a probability distribution over the outcomes.

Definition (lottery)

A lottery is a probability distribution over outcomes. It is written [p1 : o1, p2 : o2, . . . , pk :ok] where the oi are outcomes and pi > 0 such that

$$\sum_i p_i = 1$$

The lottery specifies that outcome oi occurs with probability pi. We will consider lotteries to be outcomes

Completeness

A preference relationship must be defined between every pair of outcomes:

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

Transitivity

Preferences must be transitive:

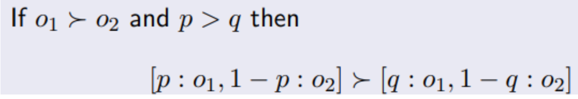
$$\text{if } o_1 \succeq o_2 \text{ and } o_2 \succeq o_3 \text{ then } o_1 \succeq o_3$$

- This makes good sense: otherwise
  $o_1 \succeq o_2$ and $o_2 \succeq o_3$ and $o_3 \succ o_1$

- An agent should be prepared to pay some amount to swap between an outcome they prefer less and an outcome they prefer more Intransitive preferences mean we can construct a "money pump"!

Monotonicity

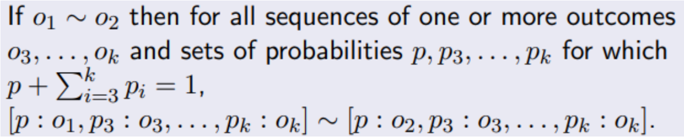
An agent prefers a larger chance of getting a better outcome to a smaller chance:



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

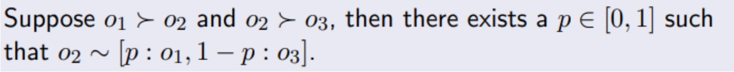
$$[p : o_1, 1-p : o_2] \succ [q : o_1, 1-q : o_2]$$

Substitutability



If $o_1 \sim o_2$ then for all sequences of one or more outcomes $o_3, \ldots, o_k$ and sets of probabilities $p, p_3, \ldots, p_k$ for which $p + \sum_{i=3}^{k} p_i = 1$, $[p : o_1, p_3 : o_3, \ldots, p_k : o_k] \sim [p : o_2, p_3 : o_3, \ldots, p_k : o_k]$.

Continuity



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

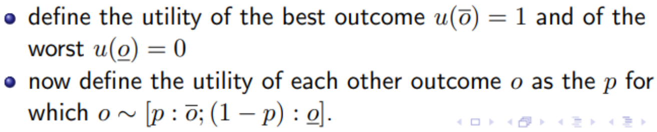
Preferences and utility functions

Theorem

If an agent's preference relation satisfies the axioms Completeness, Transitivity, Decomposability, Substitutability, Monotonicity and Continuity then there exists a function u : O → [0, 1] with the properties that:

- $u(o_1) \geq u(o_2)$ iff the agent prefers o_1 to o_2; and
- when faced about uncertainty about which outcomes he will receive, the agent prefers outcomes that maximize the expected value of u.

Proof idea:



- define the utility of the best outcome $u(\overline{o}) = 1$ and of the worst $u(\underline{o}) = 0$
- now define the utility of each other outcome o as the p for which $o \sim [p : \overline{o}; (1-p) : \underline{o}]$.