Introduction to AI Safety, Ethics, and Society

Utility Functions
Part 1: Preferences and Utility Functions

Introduction

Utility functions are a central concept in economics and decision theory.

- Expected utility theory: individuals will choose options that are likely to maximize their utility functions.

The objective of maximizing a utility function can cause intelligence.

- The construction and properties of the function being maximized are central to guiding intelligent behavior.

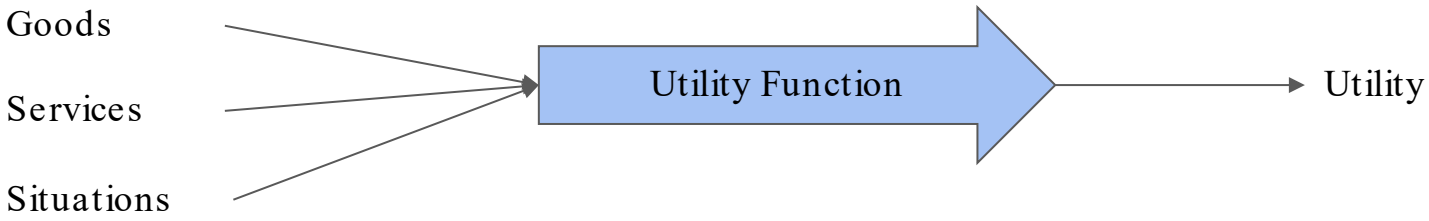
Utility functions are a foundational concept in AI safety because AIs may be approximated as expected utility maximizers. Key concepts like corrigibility rely on an understanding of preferences.

Roadmap

1. From preferences to utility functions.
2. Introducing uncertainty and expected utility.
3. Exploring corrigibility.
4. Attitudes towards risk.
5. Beyond expected utility theory.

Utility & Utility Functions

A utility function is a mathematical representation of an individual's preferences.



Utility measures how much an agent likes goods and situations.

Bernoulli Utility Functions (1/2)

Bernoulli utility functions represent preferences over outcomes, expressed as numbers.

Eg. utility function over fruits (apples, bananas, and cherries):

$$u(\text{fruits}) = 12a + 10b + 2c,$$

Utility of one banana:

$$u(\text{one banana}) = 10$$

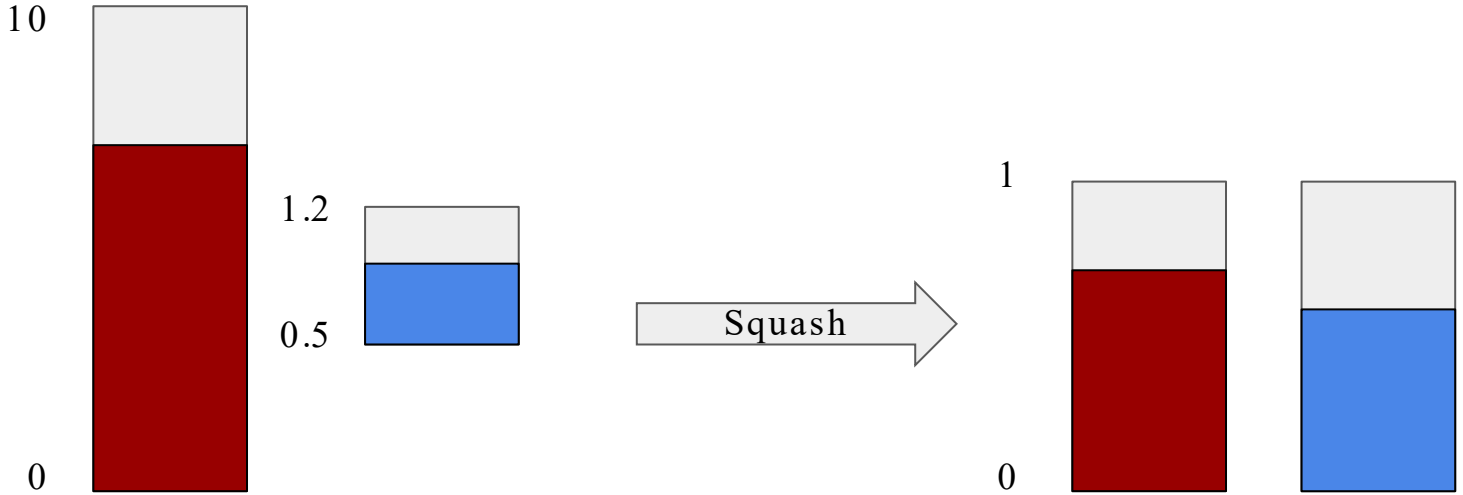
Utility of no apples, one banana, and five cherries:

$$u(\text{no apples, one banana, five cherries}) = (12 \cdot 0) + (10 \cdot 1) + (2 \cdot 5) = 20.$$

Bernoulli Utility Functions (2/2)

Bernoulli utility functions represent the same preferences when scaled or translated.

Utility functions can be squashed to facilitate interpersonal comparison.



Expected value

Expected value is the average outcome of a random event.

$$\mathbb{E}[\mathcal{L}] = o_1 \cdot p_1 + o_2 \cdot p_2 + \cdots + o_n \cdot p_n.$$

Expected value of a dice roll:

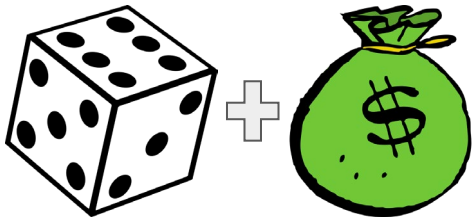
$$\frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = \frac{21}{6} = 3.5$$

Expected utility

Expected utility is the average utility of a random event.

$$\mathbb{E}[u(\mathcal{L})] = u(o_1) \cdot p_1 + u(o_2) \cdot p_2 + \cdots + u(o_n) \cdot p_n.$$

Expected utility of betting on a dice roll, assuming u($x) = ln($x):



$$\frac{1}{6} \cdot \ln 1 + \frac{1}{6} \cdot \ln 2 + \frac{1}{6} \cdot \ln 3 + \frac{1}{6} \cdot \ln 4 + \frac{1}{6} \cdot \ln 5 + \frac{1}{6} \cdot \ln 6 \approx 1.10$$

Von Neumann-Morgenstern Utility Functions

Von Neumann-Morgenstern utility functions represent preferences over *lotteries*.

Lotteries are all possible outcomes o_i and associated probabilities p_i.

Agents satisfying rationality axioms may be represented by vNM function, which is equivalent to expected utility.

vNM utility functions suggests that "rational" agents do and should make decisions that maximize expected utility.

$$U(\mathcal{L}) = u(o_1) \cdot p_1 + u(o_2) \cdot p_2 + \cdots + u(o_n) \cdot p_n.$$

Von Neumann-Morgenstern Axioms

Von Neumann and Morgenstern used an axiomatic approach to derive their claim: they decided on fundamental restrictions on preferences.

The axioms are:

1) **Completeness** : rank everything
2) **Transitivity** : rank consistently
3) **Continuity** : smooth preferences
4) **Independence** : adding the same outcomes to two lotteries does not change preferences

Assumed that preferences are **decomposable** (independent of description) and **monotonic** (preferring more goods).

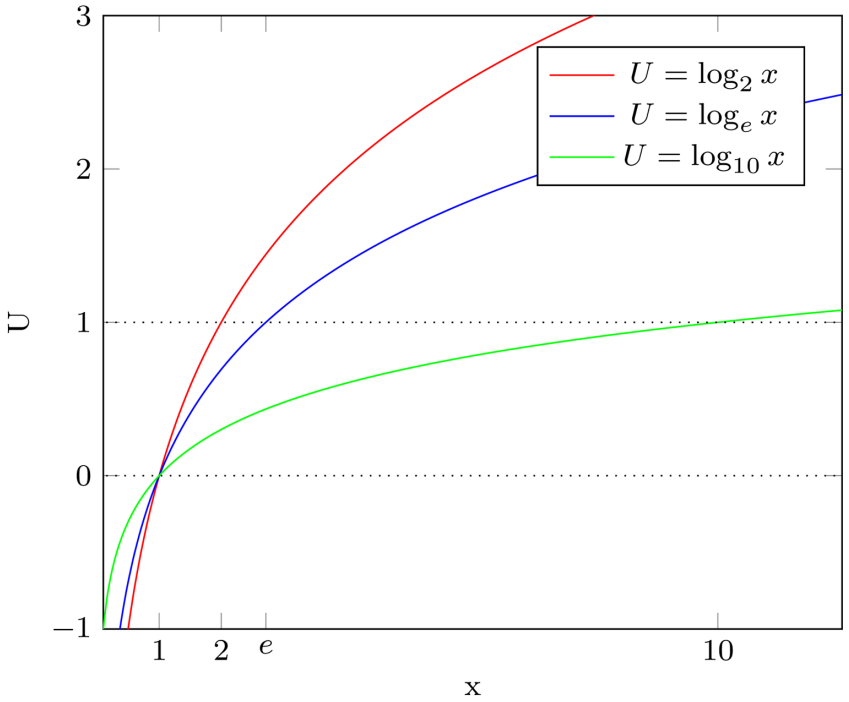
Independence and **decomposability** are contentious.

Logarithms

Logarithms express the power to which a given number (the base) must be raised in order to produce a value.

Eg. $\log_2 8 = 3$ because $2^3 = 8$

Logarithms are used as utility functions over wealth because they represent diminishing marginal returns, resembling human psychology.



St. Petersburg Paradox (1/2)

An old man on the streets of St. Petersburg offers gamblers the following game: he will flip a fair coin repeatedly until it lands on tails. If the first toss lands tails, he will pay two dollars and end the game. He will double the payout for each consecutive toss landing heads until a tails is flipped. The game ends whenever the coin lands tails, and the gambler wins takes home the winnings. How much should a gambler be willing to pay to play this game?

Key questions: What is the expected value?

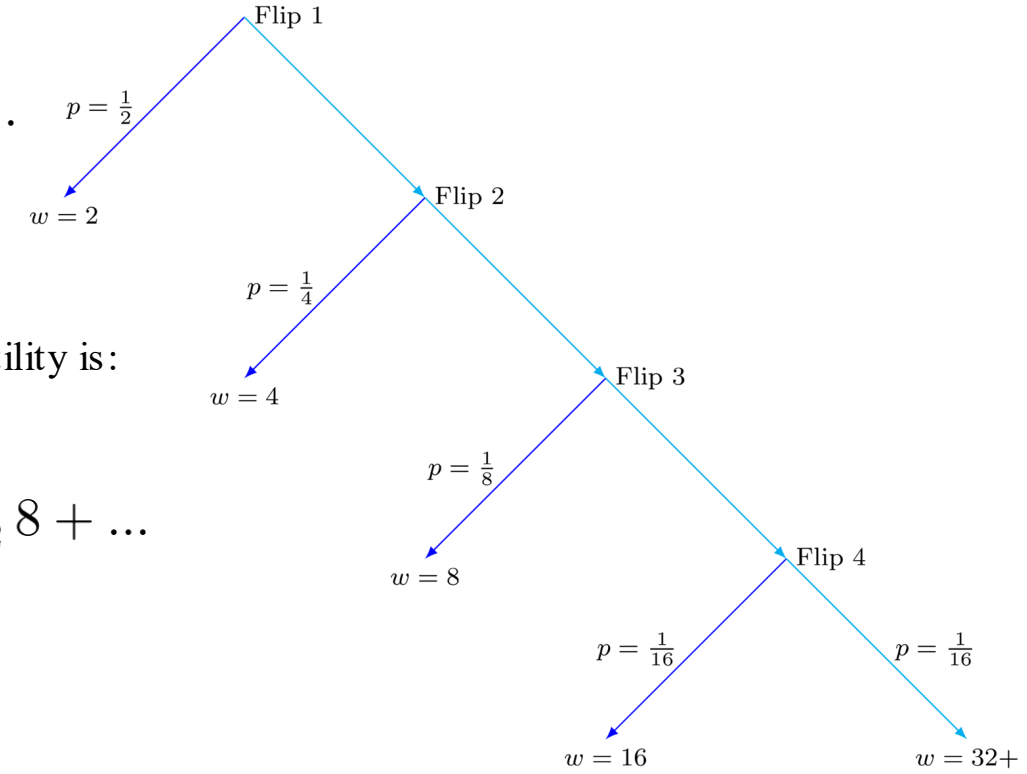
What is the expected utility?

St. Petersburg Paradox (2/2)

The expected value of this game is:

$$E[\mathcal{L}] = \frac{1}{2} \cdot \$2 + \frac{1}{4} \cdot \$4 + \frac{1}{8} \cdot \$8 + \cdots$$

$$= \$1 + \$1 + \$1 + \ldots = \$\infty$$



If we have logarithmic utility, the expected utility is:

$$U = \frac{1}{2} \cdot \log_2 2 + \frac{1}{4} \cdot \log_2 4 + \frac{1}{8} \cdot \log_2 8 + \ldots$$

$$U = 2 \Rightarrow w = \$4$$

Corrigibility (1/4)

Corrigibility measures our ability to correct an AI if and when we want to change its behavior.

An AI system is “corrigible” if it accepts and cooperates with corrective interventions like being shut down or having its utility function changed.

Corrigibility (2/4)

Suppose our coffee-fetching AI has the following utility function:

- 10 utils if we get coffee quickly.
- 5 utils if we get coffee slowly.
- 0 utils if we do not get coffee at all.

Suppose the AI breaks speed limits to deliver our coffee quickly, since it doesn't know we have preferences over it breaking speed limits. We would want to shut it down and change its utility function.

But this means we get our coffee slower: the AI would lose utility!

Corrigibility (3/4)

Suppose our coffee-fetching AI has the following utility function:

- 10 utils if we get coffee quickly.
- 5 utils if we get coffee slowly.
- 0 utils if we do not get coffee at all.

Suppose the AI thinks we can make coffee ourselves quicker than it can make us coffee, and shuts down. We tell it to stay on, wanting it to make us coffee so we can focus on more important things.

But this means we get our coffee slower: the AI would lose utility!

Corrigibility (4/4)

In both cases, our coffee-fetching AI would want to ignore our corrections. This is incorrigibility.

In general, agents might be incorrigible when their preferences are

- Complete: They rank every action.
- Transitive: Their rankings are consistent.

Suppose an AI wants to take action *a* and we tell it to shut down. It must then prefer *a* to shutting down: it will resist.

Suppose an AI wants to shut down and we tell it to take action *a*. It must then prefer shutting down to *a*: it will resist.

Recap

Utility functions represent preferences as one mathematical object.

Bernoulli utility functions deal with sure-things; vNM with uncertainty.

Von Neumann-Morgenstern axioms define rational preferences.

Expected value and expected utility are important to distinguish.

St. Petersburg's Paradox suggests that most people are risk averse.

Agents with regular utility functions might be incorrigible.

Introduction to AI Safety, Ethics, and Society

Utility Functions
Part 2: Risks and Extensions

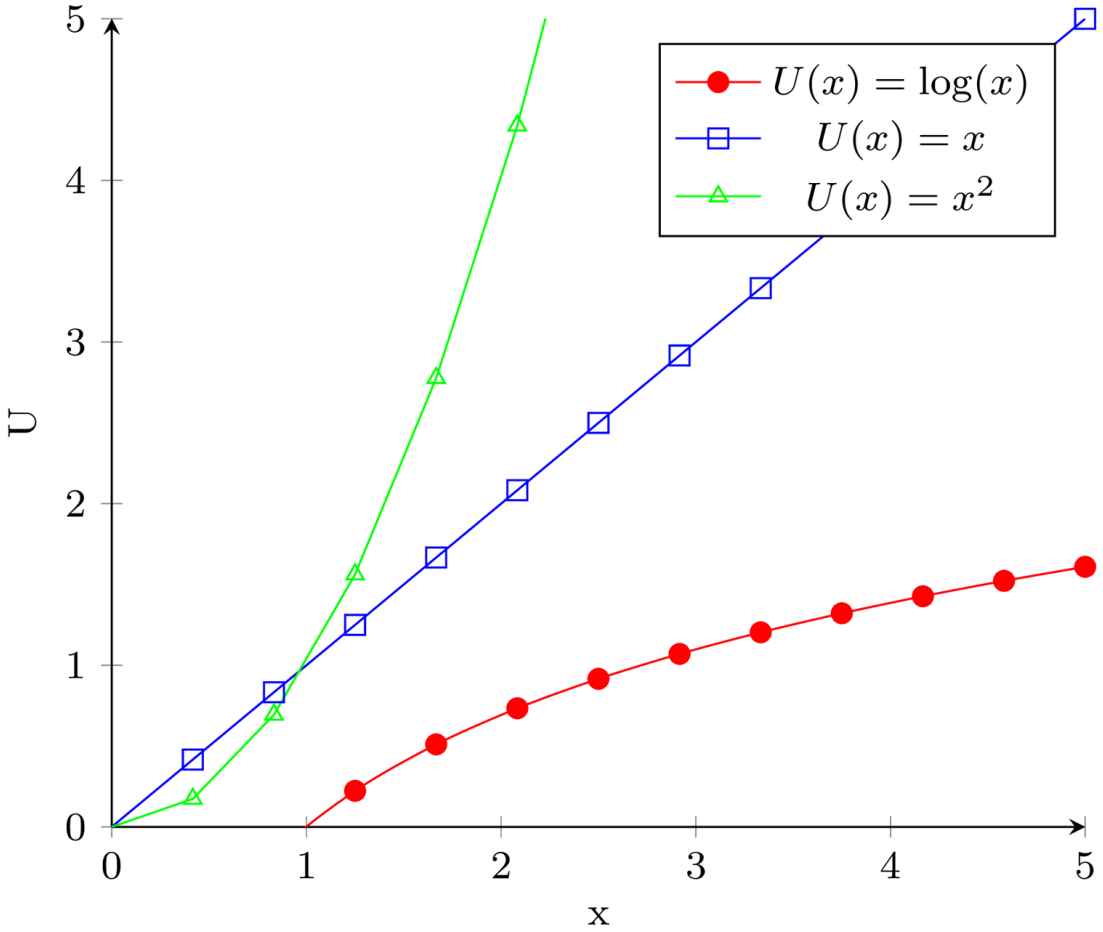
Roadmap

1. From preferences to utility functions.
2. Introducing uncertainty and expected utility.
3. Exploring corrigibility.
4. Attitudes towards risk.
5. Beyond expected utility theory.

Risk

Three Risk Attitudes

Risk aversion	Risk neutrality	Risk seeking
Dislikes risk	Indifferent towards risk	Likes risk
Turns down mean-zero lotteries	Indifferent towards mean-zero lotteries	Plays mean-zero lotteries
Prefers receiving expected value	Indifferent towards receiving expected value	Prefers playing a risky lottery
Pays to avoid risk	Pays nothing to avoid risk	Pays to get more risk
Has a concave utility function	Has a linear utility function	Has a convex utility function



Different Utility Functions Capture Different Attitudes to Risk

Criterion of Rightness vs Decision Making Procedure

Criterion of rightness: a way of judging whether an outcome is good.

Decision procedure: a method of making decisions that lead to the good outcomes.

Criterion of Rightness:

No thoughts

Decision Making Procedure:

Focus on breath

Risk Aversion

Risk aversion is very common in humans and animals.

Natural selection causes populations to be risk averse

Reasons to favor:

1. Computationally simpler and efficient heuristic.
2. Avoid risk of ruin.
3. Maximizing logarithmic utility is a better decision procedure, and maximizes every percentile of wealth in the long run (Kelly betting).

Risk Neutrality

Risk neutrality is equivalent to acting on expected value: if something has positive EV, I'll do it, If something has negative EV, I won't.

Be risk neutral when there is no possibility of ruin.

This means that risks are *local, uncorrelated, tractable,* and there are no *black swans*.

Risk neutrality is dangerous if risks are misjudged.

Risk Seeking

Risk seeking behavior is not always unreasonable (but sometimes is).

It can be a good strategy in narrow domains.

It is rare as a global attitude towards everything.

Reasons to favor:

1. All or nothing, like multiplayer-single-winner games.
2. Daring to rise from a bad situation.
3. Growing stronger through risk exposure, like building immunity.

Beyond Expected Utility Theory

Against Expected Utility Theory

We have many reasons to believe that humans do not behave in line with the axioms that define expected utility theory.

Three popular ones are:

1. The conflict between Independence and **Fairness.**
2. Human preferences violate Independence in the **Allais Paradox.**
3. Preferences change when represented as **Gains or Losses**.

Fairness (1/2)

Suppose Rachel has to leave everything she owns to one person in her will. She is indifferent between leaving it her son or daughter, but wants to treat them equality.

Twist. There's a 50% chance the law will change and everything goes to her son, no matter what she decides.

Fairness (2/2)

Problem. Independence implies preferences over below two options must be identical.

	Will to Daughter & “Fair Lottery”	Will to Son & “Unfair Lottery”
With no chance of the law changing	A: Daughter gets everything	B: Son gets everything
With a 50% chance of the law changing	C: 50%: daughter gets everything 50%: son gets everything	D: Son gets everything

If she prefers C to D, she violates the Independence Axiom.

The Allais Paradox (1/2)

Which gamble would you take in Scenario A?

	Gamble A1	Gamble A2
Scenario A	100% chance of $1 million	1% chance of nothing 89% chance of $1 million 10% change of $5 million

Which gamble would you take in Scenario B?

	Gamble B1	Gamble B2
Scenario B	11% chance of $1 million 89% chance of $0	90% chance of $0 10% change of $5 million

Typically, people choose A1 and B2. Seems reasonable…

The Allais Paradox (2/2)

Remove the common 89% chance of $1 million from Scenario A:

	Gamble A1	Gamble A2
Scenario A	11% chance of $1 million	1% chance of nothing 10% change of $5 million

Remove the common 89% chance of $0 in Scenario B:

	Gamble B1	Gamble B2
Scenario B	11% chance of $1 million	1% chance of $0 10% change of $5 million

If we remove the common outcomes by the Independence Axiom, these lotteries are the same: our choice of A1, B2 are inconsistent!

Framing (1/2)

An illness is expected to kill 600 people. There are two policy options.

Which option would you take in Scenario A?

	Option A1	Option A2
Scenario A	Save 200 people for sure.	⅓ chance of saving everyone. ⅔ chance of saving no one.

Which option would you take in Scenario B?

	Option B1	Option B2
Scenario B	400 people die for sure.	⅓ chance that no one dies. ⅔ chance that everyone dies.

People usually choose A1 and B2.

Framing (2/2)

But these are the same lottery!

- Saving 200 people is equivalent to 400 people dying.
- Saving everyone is equivalent to no one dying.
- Saving no one is equivalent to everyone dying.

	Option 1	Option 2
Scenario A/B	Save 200/400 Die.	⅓: Save 600/0 Die. ⅔: Save 0/600 Die.

Having different preferences across these two scenarios is violating something more fundamental than the von Neumann-Morgenstern axioms. We're making different choices despite the options having the same outcomes with the same probabilities.

Normative vs. Descriptive Statements

Normative statements: judgements of what *ought* to be, based on ideas of goodness, rightness, justness, etc.

Example: Von Neumann-Morgenstern expected utility.

Descriptive statements: accurate observations of what has occured/occurs.

Example: Prospect theory (next topic).

Prospect Theory (1/4)

Prospect theory is a descriptive model of human decision-making. It breaks down decision-making into two stages:

Part 1: Editing
Part 2: Evaluation

- Value Function.
- Weighting Function.

Prospect Theory (2/4)

In the **editing stage** , the inputs for lotteries are rewritten as gains or losses relative to the initial position rather than final outcomes.

For instance:

- Original: Pay all your $1000 wealth to play a lottery where you get $0 or $2000 with probability 0.5.
- Edited: Lose $1000 or gain $1000 with probability 0.5.

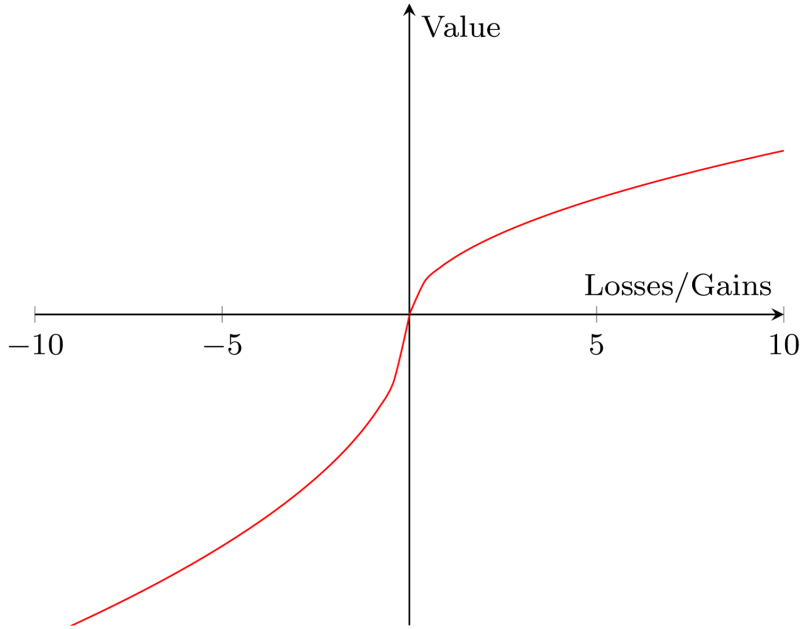
This accounts for why there is a difference in decision-making in the two policy options in the framing example.

Prospect Theory (3/4)

In the **evaluation** stage, gains and losses are multiplied by weighted probabilities to determine the preferred outcome.

Value Function.

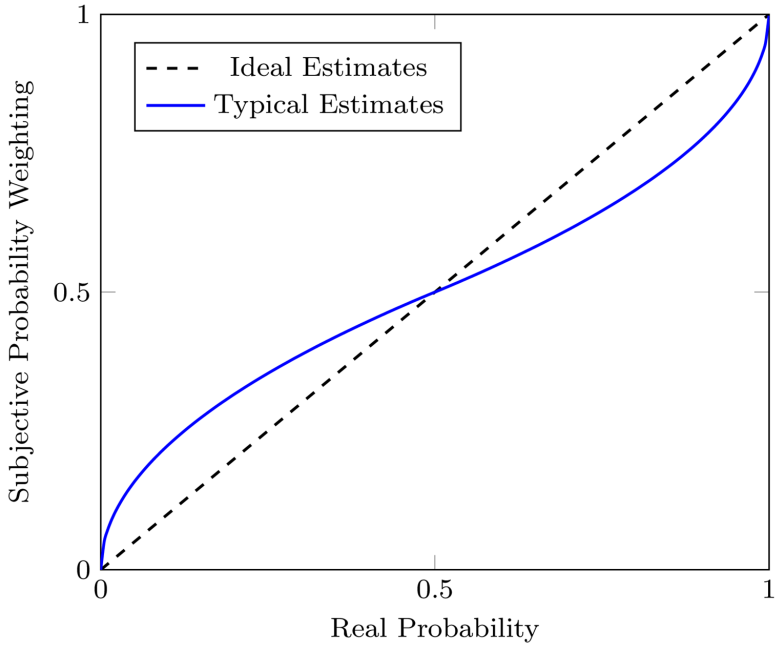
- Like a utility function, the value function assigns "value" to outcomes (changes in wealth)
- Sigmoid (s-shaped) curve: people are risk-seeking in losses and risk-averse in gains.
- The curve is steeper in losses than in gains, because people are more sensitive to losses.



Prospect Theory (4/4)

Weighting Function.

- The weighting function describes how humans think about probabilities.
- We overweight extreme probabilities, caring a lot about small probabilities of succeeding or failing: a 1% chance of failure "feels" more like a 10% chance of failure would.



General Decision Making Models

Expected Utility Theory and Prospect Theory are examples of decision-making models that bear certain similar features.

In general, decision making models:

- apply some function to outcomes, such as a utility function.
- apply some function to probabilities, such as a weighting function.
- multiply the outputs of both functions, and add them up to represent the desirability of the option being considered.

This gives us:

$$V(\mathcal{L}) = w(p_1) \cdot v(o_1) + w(p_2) \cdot v(o_2) + w(p_3) \cdot v(o_3) + ...$$

Recap

Attitudes to risk include risk aversion, risk seeking, and risk neutrality.

Attitudes to risk are described by the concavity of the utility function.

In different situations, agents might adopt different risk attitudes.

Expected utility theory is often inconsistent with how humans think.

Newer models like prospect theory fit better but are more complex.