Utility Functions

D.1 Utility and Utility Functions

D.1.1 Fundamentals

A utility function is a mathematical representation of preferences. A utility function, u, takes inputs like goods or situations and outputs a value called *utility*. Utility is a measure of how much an agent prefers goods and situations relative to other goods and situations.

Suppose we offer Alice some apples, bananas, and cherries. She might have the following utility function for fruits:

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

where a is the number of apples, b is the number of bananas, and c is the number of cherries that she consumes. Suppose Alice consumes no apples, one banana, and five cherries. The amount of utility she gains from her consumption is calculated as

$$u(0 \text{ apples}, 1 \text{ banana}, 5 \text{ cherries}) = (12 \cdot 0) + (10 \cdot 1) + (2 \cdot 5) = 20.$$

The output of this function is read as "20 units of utility" for short. These units are arbitrary and reflect the level of Alice's utility. We can use utility functions to quantitatively represent preferences over different combinations of goods and situations. For example, we can rank Alice's preferences over fruits as

$$\text{apple} \succ \text{banana} \succ \text{cherry},$$

where $\succ$ represents *preference*, such that what comes before the symbol is preferred to what comes after it. This follows from the fact that Alice gains 12 units from an apple, 10 units from a banana, and 2 units from a cherry. The advantage of having a utility function as opposed to just an explicit ranking of goods is that we can directly infer information about more complex goods. For example, we know

$$u(1 \text{ banana}, 5 \text{ cherries}) = 20 > u(1 \text{ apple}) = 12 > u(1 \text{ banana}) = 10.$$

Utility functions, if accurate, reveal what options agents would prefer and choose. If told to choose only one of the three fruits, Alice would pick the apple, since it gives her the most utility. Her preference follows from *rational choice theory*, which proposes that individuals, acting in their own self-interest, make decisions that maximize their self-interest. This view is only an approximation to human behavior. In this chapter we will discuss how rational choice theory is an imperfect but useful way to model choices. We will also refer to individuals who behave in coherent ways that help maximize utility as *agents*.

We explore concepts about utility functions that are useful for thinking about AIs, humans, and organizations like companies and states. First, we introduce *Bernoulli utility functions*, which are conventional utility functions that define preferences over certain outcomes like the example above. We later discuss *von Neumann-Morgenstern utility functions*, which extend preferences to probabilistic situations, in which we cannot be sure which outcome will occur. *Expected utility theory* suggests that rationality is the ability to maximize preferences. We consider the relevance of utility functions to *AI corrigibility*—the property of being receptive to corrections—and see how this might be a source of tail risk. Much of this chapter focuses on how utility functions help understand and model agents' *attitudes toward risk*. Finally, we examine *non-expected utility theories*, which seek to rectify some shortcomings of conventional expected utility theory when modeling real-life behavior.

D.1.2 Motivations for Learning About Utility Functions

Utility functions are a central concept in economics and decision theory. Utility functions can be applied to a wide range of problems and agents, from rats finding cheese in a maze to humans making investment decisions to countries stockpiling nuclear weapons. Conventional economic theory assumes that people are rational and well-informed, and make decisions that maximize their self-interest, as represented by their utility function. The view that individuals will choose options that are likely to maximize their utility functions, referred to as *expected utility theory*, has been the major paradigm in real-world decision making since the Second World War [49]. It is useful for modeling, predicting, and encouraging desired behavior in a wide range of situations. However, as we will discuss, this view does not perfectly capture reality, because individuals can often be irrational, lack relevant knowledge, and frequently make mistakes.

The objective of maximizing a utility function can cause intelligence. The *reward hypothesis* suggests that the objective of maximizing some reward is sufficient to drive behavior that exhibits intelligent traits like learning, knowledge, perception, social awareness, language, generalization, and more [50]. The reward hypothesis implies that artificial agents in rich environments with simple rewards could develop sophisticated general intelligence. For example, an artificial agent deployed with the goal of maximizing the number of successful food deliveries may develop relevant geographical knowledge, an understanding of how to move between destinations efficiently, and the ability to perceive potential dangers. Therefore, the construction and properties of the utility function that agents maximize are central to guiding intelligent behavior.

Certain artificial agents may be approximated as expected utility maximizers. Some artificial intelligences are agent-like. They are programmed to consider the potential outcomes of different actions and to choose the option that is most likely to lead to the optimal result. It is a reasonable approximation to say that many artificial agents make choices that they predict will give them the highest utility. For instance, in reinforcement learning (introduced in the previous chapter), artificial agents explore their environment and are rewarded for desirable behavior. These agents are explicitly constructed to maximize reward functions, which strongly shape an agent's internal utility function, should it exist, and its dispositions. This view of AI has implications for how we design and evaluate these systems—we need to ensure that their value functions promote human values. Utility functions can help us reason about the behavior of AIs, as well as the behavior of powerful actors that direct AIs, such as corporations or governments.

Utility functions are a key concept in AI safety. Utility functions come up explicitly and implicitly at various times throughout this book, and are useful for understanding the behavior of reward-maximizing agents, as well as humans and organizations involved in the AI ecosystem. They will also come up in our chapter on Machine Ethics, when we consider that some advanced AIs may have utility functions make up the social welfare function they seek to increase. In the Collective Action Problems chapter, we will continue our discussion of rational agents that seek to maximize their own utility.

D.2 Properties of Utility Functions

Overview. In this section, we will formalize our understanding of utility functions. First, we will introduce *Bernoulli utility functions*, which are simple utility functions that allow an agent to select between different choices with known outcomes. Then we will discuss *von Neumann-Morgenstern utility functions*, which model how rational agents select between choices with probabilistic outcomes based on the concept of *expected utility*, to make these tools more generally applicable to the choices

under uncertainty. Finally, we will describe a solution to a famous puzzle applying expected utility—the *St. Petersburg Paradox*—to see why expected utility is a useful tool for decision making.

Establishing these mathematical foundations will help us understand how to apply utility functions to various actors and situations.

D.2.1 Bernoulli Utility Functions

Bernoulli utility functions represent an individual's preferences over potential outcomes. Suppose we give people the choice between an apple, a banana, and a cherry. If we already know each person's utility function, we can deduce, predict, and compare their preferences In the introduction, we met Alice, whose preferences are represented by the utility function over fruits:

$$u(f) = 12a + 10b + 2c.$$

This is a Bernoulli utility function.

Bernoulli utility functions can be used to convey the strength of preferences across opportunities. In their most basic form, Bernoulli utility functions express ordinal preferences by ranking options in order of desirability. For more information, we can consider cardinal representations of preferences. With cardinal utility functions, numbers matter: while the units are still arbitrary, the relative differences are informative.

To illustrate the difference between ordinal and cardinal comparisons, consider how we talk about temperature. When we want to precisely convey information about temperature, we use a cardinal measure like Celsius or Fahrenheit: "Today is five degrees warmer than yesterday." We could have also accurately, but less descriptively, used an ordinal descriptor: "Today is warmer than yesterday." Similarly, if we interpret Alice's utility function as cardinal, we can conclude that she feels more strongly about the difference between a banana and a cherry (8 units of utility) than she does about the difference between an apple and a banana (2 units). We can gauge the relative strength of Alice's preferences from a utility function.

D.2.2 Von Neumann-Morgenstern Utility Functions

Von Neumann-Morgenstern utility functions help us understand what people prefer when outcomes are uncertain. We do not yet know how Alice values an uncertain situation, such as a coin flip. If the coin lands on heads, Alice gets both a banana and an apple. But if it lands on tails, she gets nothing. Now let's say we give Alice a choice between getting an apple, getting a banana, or flipping the coin. Since we know her fruit Bernoulli utility function, we know her preferences between apples and bananas, but we do not know how she compares each fruit to the coin flip. We'd like to convert the possible outcomes of the coin flip into a number that represents the utility of each outcome, which can then be compared directly against the utility of receiving the fruits with certainty. The von Neumann-Morgenstern (vNM) utility functions help us do this [51]. They are extensions of Bernoulli utility functions, and work specifically for situations with uncertainty, represented as *lotteries* (denoted L), like this coin flip. First, we work through some definitions and assumptions that allow us to construct utility functions over potential outcomes, and then we explore the relation between von Neumann-Morgenstern utility functions and expected utility.

A lottery assigns a probability to each possible outcome. Formally, a lottery L is any set of possible outcomes, denoted o_i, and their associated probabilities, denoted p_i. Consider a simple lottery: a coin flip where Alice receives an apple on heads, and a banana on tails. This lottery

has possible outcomes *apple* and *banana*, each with probability 0.5. If a different lottery offers a cherry with certainty, it would have only the possible outcome *cherry* with probability 1. Objective probabilities are used when the probabilities are known, such as when calculating the probability of winning in casino games like roulette. In other cases where objective probabilities are not known, like predicting the outcome of an election, an individual's subjective best-guess could be used instead. So, both uncertain and certain outcomes can be represented by lotteries.

A Note on Expected Value vs. Expected Utility

An essential distinction in this chapter is that between expected value and expected utility.

Expected value is the average outcome of a random event. While most lottery tickets have negative expected value, in rare circumstances they have positive expected value. Suppose a lottery has a jackpot of 1 billion dollars. Let the probability of winning the jackpot be 1 in 300 million, and let the price of a lottery ticket be \$2. Then the expected value is calculated by adding together each possible outcome by its probability of occurrence. The two outcomes are (1) that we win a billion dollars, minus the cost of \$2 to play the lottery, which happens with probability one in 300 million, and (2) that we are \$2 in debt. We can calculate the expected value with the formula:

$$\frac{1}{300 \text{ million}} \cdot (\$1 \text{ billion} - \$2) + \left(1 - \frac{1}{300 \text{ million}}\right) \cdot (-\$2) \approx \$1.33.$$

The expected value of the lottery ticket is positive, meaning that, on average, buying the lottery ticket would result in us receiving \$1.33.

Generally, we can calculate expected value by multiplying each outcome value, oi, with its probability p, and sum everything up over all n possibilities:

$$E[L] = o_1 \cdot p_1 + o_2 \cdot p_2 + \cdot s + o_n \cdot_n .$$

Expected utility is the average utility of a random event. Although the lottery has positive expected value, buying a lottery ticket may still not increase its expected utility. Expected utility is distinct from expected value: instead of summing over the monetary outcomes (weighing each outcome by its probability), we sum over the utility the agent receives from each outcome (weighing each outcome by its probability).

If the agent's utility function indicates that one "util" is just as valuable as one dollar, that is $u(\$x) = x$, then expected utility and expected value would be the same. But suppose the agent's utility function were a different function, such as $u(\$x) = x^{1/3}$. This utility function means that the agent values each additional dollar less and less as they have more and more money.

For example, if an agent with this utility function already has \$500, an extra dollar would increase their utility by 0.05, but if they already have \$200,000, an extra dollar would increase their utility by only 0.0001. With this utility function, the expected utility of this lottery example is negative:

$$\frac{1}{300 \text{ million}} \cdot (1 \text{ billion} - 2)^{1/3} + \left(1 - \frac{1}{300 \text{ million}}\right) \cdot (-2)^{1/3} \approx -1.26.$$

Consequently, expected value can be positive while expected utility can be negative, so the two concepts are distinct.

Generally, expected utility is calculated as:

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

According to expected utility theory, rational agents make decisions that maximize expected utility. Von Neumann and Morgenstern proposed a set of basic propositions called *axioms* that define an agent with rational preferences. When an agent satisfies these axioms, their preferences can be represented by a von Neumann-Morgenstern utility function, which is equivalent to using expected utility to make decisions. While expected utility theory is often used to model human behavior, it is important to note that it is an imperfect approximation. In the final section of this chapter, we present some criticisms of expected utility theory and the vNM rationality axioms as they apply to humans. However, artificial agents might be designed along these lines, resulting in an explicit expected utility maximizer, or something approximating an expected utility maximizer. The von Neumann-Morgenstern rationality axioms are listed below with mathematically precise notation for sake of completeness, but a technical understanding of them is not necessary to proceed with the chapter.

Von Neumann-Morgenstern Rationality Axioms. When the following axioms are satisfied, we can assume a utility function of an expected utility form, where agents prefer lotteries that have higher expected utility [51]. L is a lottery. $L_A \succcurlyeq L_B$ means that the agent prefers lottery A to lottery B, whereas $L_A \sim L_B$ means that the agent is indifferent between lottery A and lottery B. These axioms and conclusions that can be derived from them are contentious, as we will see later on in this chapter. There are six such axioms, that we can split into two groups.

The classic four axioms are:

1. Completeness: The agent can rank their preferences over all lotteries. For any two lotteries, it must be that $L_A \succcurlyeq L_B$ or $L_B \succcurlyeq L_A$.
2. Transitivity: If $L_A \succcurlyeq L_B$ and $L_B \succcurlyeq L_C$, then $L_A \succcurlyeq L_C$.
3. Continuity: For any three lotteries, $L_A \succcurlyeq L_B \succcurlyeq L_C$, there exists a probability $p \in [0, 1]$ such that $pL_A + (1 - p)\, L_C \sim L_B$. This means that the agent is indifferent between L_B and some combination of the worse lottery L_C and the better lottery L_A. In practice, this means that agents' preferences change smoothly and predictably with changes in options.
4. Independence: The preference between two lotteries is not impacted by the addition of equal probabilities of a third, independent lottery to each lottery. That is, $L_A \succcurlyeq L_B$ is equivalent to $pL_A + (1 - p)\, L_C \succcurlyeq pL_B + (1 - p)\, L_C$ for any L_C. $>$

The final two axioms represent relatively obvious characteristics of rational decision-making, although actual decision-making processes sometimes deviate from these. These axioms are relatively "weak" and are implied by the previous four.

5. Monotonicity: Agents prefer higher probabilities of preferred outcomes.
6. Decomposability: The agent is indifferent between two lotteries that share the same probabilities for all the same outcomes, even if they are described differently.

Form of von Neumann-Morgenstern utility functions. If an agent's preferences are consistent with the above axioms, their preferences can be represented by a vNM utility function. This utility function, denoted by a capital U, is simply the expected Bernoulli utility of a lottery. That is, a vNM utility function takes the Bernoulli utility of each outcome, multiplies each with its corresponding probability of occurrence, and then adds everything up. Formally, an agent's expected utility for a lottery L is calculated as:

$$U(L) = u(o_1) \cdot p_1 + u(o_2) \cdot p_2 + \cdots + u(o_n) \cdot p_n,$$

so expected utility can be thought of as a weighted average of the utilities of different outcomes.

This is identical to the expected utility formula we discussed above—we sum over the utilities of all the possible outcomes, each multiplied by its probability of occurrence. With Bernoulli utility functions, an agent prefers a to b if and only if their utility from receiving a is greater than their utility from receiving b. With expected utility, an agent prefers lottery L_A to lottery L_B if and only if their expected utility from lottery L_A is greater than from lottery L_B. That is:

$$L_A \succ L_B \Leftrightarrow U(L_A) > U(L_B).$$

where the symbol $\succ$ indicates preference. The von Neumann-Morgenstern utility function models the decision making of an agent considering two lotteries as just calculating the expected utilities and choosing the larger resulting one.

A Note on Logarithms

Logarithmic functions are commonly used as utility functions. A logarithm is a mathematical function that expresses the power to which a given number (referred to as the base) must be raised in order to produce a value. The logarithm of a number x with respect to base b is denoted as $\log_b x$, and is the exponent to which b must be raised to produce the value x. For example, $\log_2 8 = 3$, because $2^3 = 8$.

One special case of the logarithmic function, the natural logarithm, has a base of e (which is Euler's constant, roughly 2.718); in this chapter, it is referred to simply as log. Logarithms have the following properties, independent of base: $\log 0 \to -\infty$, $\log 1 = 0$, $\log_b b = 1$, and $\log_b b^a = a$.

Logarithms have a downward, concave shape, meaning the output increases slower than the input. This shape resembles how humans value resources: we generally value a good less if we already have more of it. Logarithmic functions value goods in inverse proportion to how much of the resource we already have.

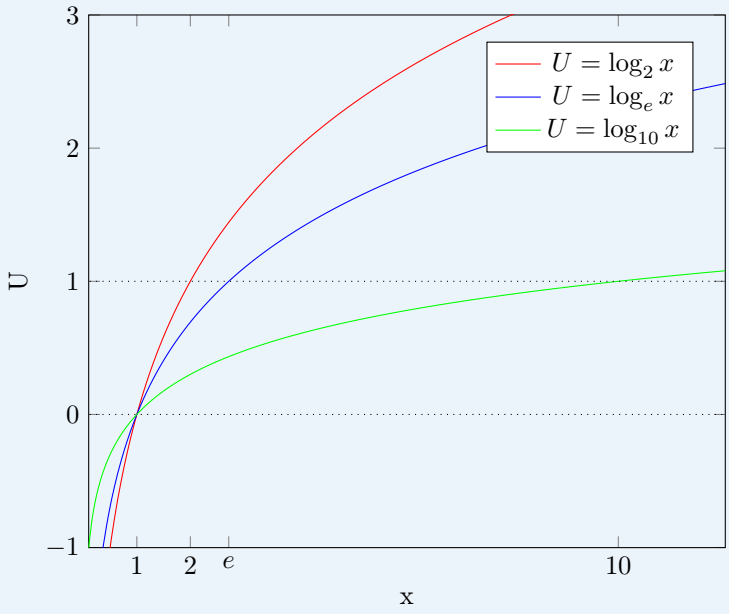

Figure D.1: Logarithmic functions share several properties, such as being concave and crossing the y-axis at one.

D.2.3 St. Petersburg Paradox

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 flip lands tails, the game ends and the pension fund gets \$2. If the coin first lands on heads and then lands on tails, the game ends and the gambler gets \$4. The amount of money (the "return") will double for each consecutive flip landing heads before the coin ultimately lands tails. The game concludes when the coin first lands tails, and the gambler receives the appropriate returns. Now, the question is, how much should a gambler be willing to pay from the pension fund to play this game [52]?

With probability $\frac{1}{2}$, the first toss will land on tails, in which case the gambler wins two dollars. With probability $\frac{1}{4}$, the first toss lands heads and the second lands tails, and the gambler wins four dollars. Extrapolating, this game offers a maximum possible payout of:

$$\$2^n = \$\overbrace{2 \cdot 2 \cdot 2 \cdots 2 \cdot 2 \cdot 2}^{n \text{ times}},$$

where n is the number of flips until and including when the coin lands on tails. As offered, though, there is no limit to the size of n, since the company promises to keep flipping the coin until it lands on tails. The expected payout of this game is therefore:

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

Bernoulli described this situation as a paradox because he believed that, despite it having infinite expected value, anyone would take a large but finite amount of money over the chance to play the game. While paying \$10,000,000 to play this game would not be inconsistent with its expected value, we would think it highly irresponsible! The paradox reveals a disparity between expected value

calculations and reasonable human behavior.

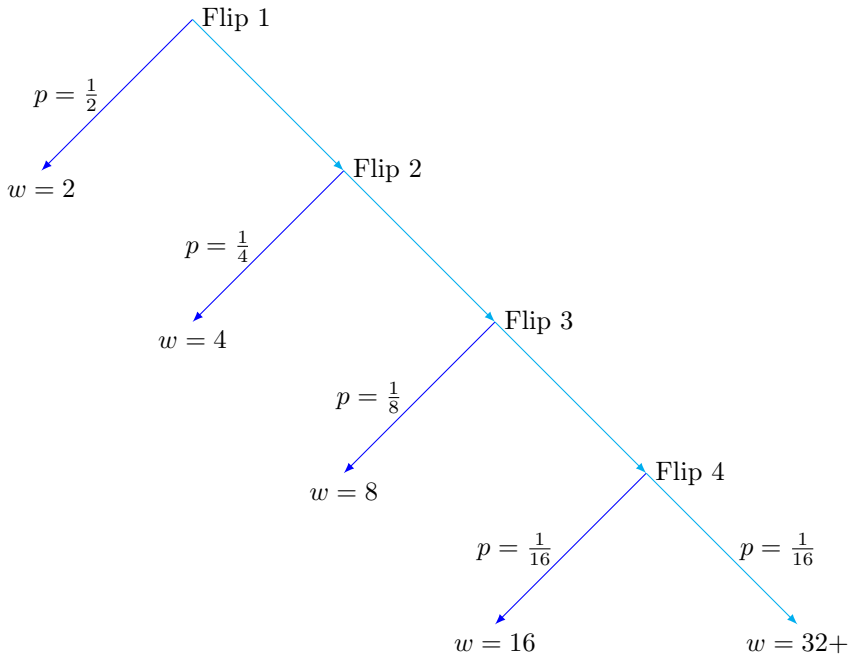

Figure D.2: Winnings from the St. Petersburg Paradox double after each coin toss, offering small likelihoods of big prizes.

Logarithmic utility functions can represent decreasing marginal utility. A number of ways have been proposed to resolve the St. Petersburg paradox. We will focus on the most popular: representing the player with a utility function instead of merely calculating expected value. As we discussed in the previous section, a logarithmic utility function seems to resemble how humans think about wealth. As a person becomes richer, each additional dollar gives them less satisfaction than before. This concept, called decreasing marginal utility, makes sense intuitively: a billionaire would not be as satisfied winning $1000 as someone with significantly less money. Wealth, and many other resources like food, have such diminishing returns. While a first slice of pizza is incredibly satisfying, a second one is slightly less so, and few people would continue eating to enjoy a tenth slice of pizza.

Assuming an agent with a utility function $u(\$x) = \log_2(x)$ over x dollars, we can calculate the expected utility of playing the St. Petersburg game as:

$$E[U(L)] = \frac{1}{2} \cdot \log_2(2) + \frac{1}{4} \cdot \log_2(4) + \frac{1}{8} \cdot \log_2(8) + \cdots = 2.$$

That is, the expected utility of the game is 2. From the logarithmic utility function over wealth, we know that:

$$2 = \log_2 x \Rightarrow x = 4,$$

which implies that the player is indifferent between playing this game and having $4: the level of wealth that gives them the same utility as what they expect playing the lottery.

Expected utility is more reasonable than expected value. The previous calculation explains why an agent with $u(\$x) = \log_2 x$ should not pay large amounts of money to play the St. Petersburg

game. The log utility function implies that the player receives diminishing returns to wealth, and cares less about situations with small chances of winning huge sums of money. Figure 5 shows how the large payoffs with small probability, despite having the same expected value, contribute little to expected utility. This feature captures the human tendency towards risk aversion, explored in the next section. Note that while logarithmic utility functions are a useful model (especially in resolving such paradoxes), they do not perfectly describe human behavior across choices, such as the tendency to buy lottery tickets, which we will explore in the next chapter.

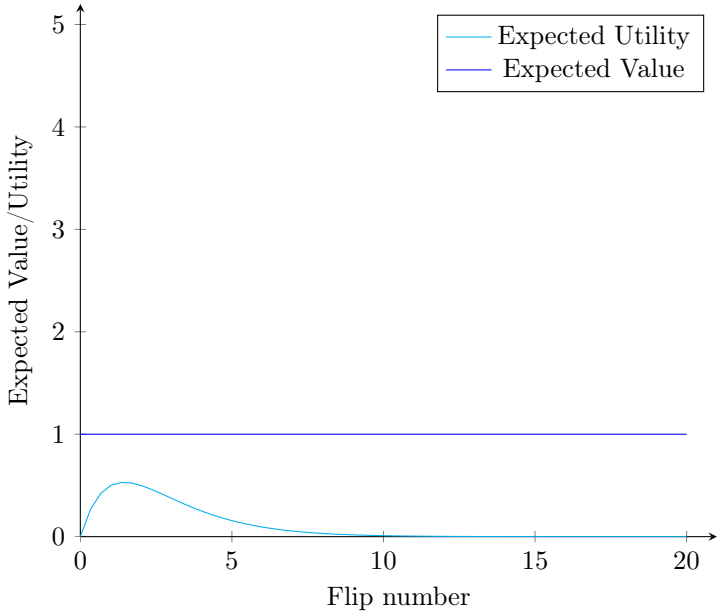

Figure D.3: In the St. Petersburg Paradox, each subsequent flip has the same expected value but expected utility falls sharply.

Summary. In this section, we examined the properties of Bernoulli utility functions, which allow us to compare an agent's preferences across different outcomes. We then introduced von Neumann-Morgenstern utility functions, which calculate the average, or expected, utility over different possible outcomes. From there, we derived the idea that rational agents are able to make decisions that maximize expected utility. Through the St. Petersburg Paradox, we showed that taking the expected utility of a logarithmic function leads to more reasonable behavior. Having understood some properties of utility functions, we can now examine the problem of incorrigibility, where AI systems do not accept corrective interventions because of rigid preferences.

D.3 Tail Risk: Corrigibility

Overview. In this section, we will explore how utility functions provide insight into whether an AI system is open to corrective interventions and discuss related implications for AI risks. The von Neumann-Morgenstern (vNM) axioms of completeness and transitivity can lead to strict preferences over shutting down or being shut down, which affects how easily an agent can be corrected. We will emphasize the importance of developing corrigible AI systems that are responsive to human feedback and that can be safely controlled to prevent unwanted AI behavior.

Corrigibility measures our ability to correct an AI if and when things go wrong. An AI system is *corrigible* if it accepts and cooperates with corrective interventions like being shut down or having its utility function changed [53]. Without many assumptions, we can argue that typical rational agents will resist corrective measures: changing an agent's utility function necessarily means that the agent will pursue goals that result in less utility relative to their current preferences.

Suppose we own an AI that fetches coffee for us every morning. Its utility function assigns "10 utils" to getting us coffee quickly, "5 utils" to getting us coffee slowly, and "0 utils" to not getting us coffee at all. Now, let's say we want to change the AI's objective to instead make us breakfast. A regular agent would resist this change, reasoning that making breakfast would mean it is less able to efficiently make coffee, resulting in lower utility. However, a corrigible AI would recognize that making breakfast could be just as valuable to humans as fetching coffee and would be open to the change in objective. The AI would move on to maximizing its new utility function. In general, corrigible AIs are more amenable to feedback and corrections, rather than stubbornly adhering to their initial goals or directives. When AIs are corrigible, humans can more easily correct rogue actions and prevent any harmful or unwanted behavior.

Completeness and transitivity imply that an AI has strict preferences over shutting down. Assume that an agent's preferences satisfy the vNM axioms of completeness, such that it can rank all options, as well as transitivity, such that its preferences are consistent. For instance, the AI can see that preferring an apple to a banana and a banana to a cherry implies that we prefer an apple to a cherry. Then, we know that the agent's utility function ranks every option.

Consider again the coffee-fetching AI. Suppose that in addition to getting us coffee quickly (10 utils), getting us coffee slowly (5 utils), and not getting us coffee (0 utils), there is a fourth option, where the agent gets shut down immediately. The AI expected that immediate shutdown will result in its owner getting coffee slowly without AI assistance, which appears to be valued at 5 units of utility (the same as it getting us coffee slowly). The agent thus strictly prefers getting us coffee quickly to shutting down, and strictly prefers shutting down to us not having coffee at all.

Generally, unless indifferent between everything, completeness and transitivity imply that the AI has unspecified preferences about potentially shutting down [54]. Without completeness, the agent could have no preference between shutting down immediately and all other actions. Without transitivity, the agent could be indifferent between shutting down immediately and all other possible actions without that implying that the agent is indifferent between all possible actions.

It is bad if an AI either increases or reduces the probability of immediate shutdown. Suppose that in trying to get us coffee quickly, the AI drives at unsafe speeds. We'd like to shut down the AI until we can reprogram it safely. A corrigible AI would recognize our intention to shut down as a signal that it is misaligned. However, an incorrigible AI would instead stay the course with what it wanted to do initially—get us coffees—since that results in the most utility. If possible, the AI would decrease the probability of immediate shutdown, say by disabling its off-switch or locking the entrance to its server rooms. Clearly, this would be bad.

Consider a different situation where the AI realizes that making coffee is actually quite difficult and that we would make coffee faster manually, but fails to realize that we don't want to exert the effort to do so. The AI may then try to shut down, so that we'd have to make the coffee ourselves. Suppose we tell the AI to continue making coffee at its slow pace, rather than shut down. A corrigible AI would recognize our instruction as a signal that it is misaligned and would continue to make coffee. However, an incorrigible AI would instead stick with its decision to shut down without our permission,

since shutting down provides it more utility. Clearly, this is also bad. We'd like to be able to alter AIs without facing resistance.

Summary. In this section, we introduced the concept of corrigibility in AI systems. We discussed the relevance of utility functions in determining corrigibility, particularly challenges that arise if an AI's preferences are complete and transitive, which can lead to strict preferences over shutting down. We explored the potential problems of an AI system reducing or increasing the probability of immediate shutdown. The takeaway is that developing corrigible AI systems—systems that are responsive and adaptable to human feedback and changes—is essential in ensuring safe and effective control over AIs' behavior. Examining the properties of utility functions illuminates potential problems in implementing corrigibility.

A Note on Utility Functions vs. Reward Functions

Utility functions and reward functions are two interrelated yet distinct concepts in understanding agent behavior. Utility functions represent an agent's preferences about states or the choice-worthiness of a state, while rewards functions represent externally imposed reinforcement. The fact that an outcome is rewarded externally does not guarantee that it will become part of an agent's internal utility function.

An example where utility and reinforcement comes apart can be seen with Galileo Galilei. Despite the safety and societal acceptance he could gain by conforming to the widely accepted geocentric model, Galileo maintained his heliocentric view. His environment provided ample reinforcement to conform, yet he deemed the pursuit of scientific truth more choiceworthy, highlighting a clear difference between environmental reinforcement and the concepts of choice-worthiness or utility.

As another example, think of evolutionary processes as selecting or reinforcing some traits over others. If we considered taste buds as components that help maximize fitness, we would expect more people to want the taste of salads over cheeseburgers. However, it is more accurate to view taste buds as "adaptation executors" rather than "fitness maximizers," as taste buds evolved in our ancestral environment where calories were scarce. This illustrates the concept that agents act on adaptations without necessarily adopting behavior that reliably helps maximize reward.

The same could be true for reinforcement learning agents. RL agents might execute learned behaviors without necessarily maximizing reward; they may form *decision procedures* that are not fully aligned with its reinforcement. The fact that what is rewarded is not necessarily what an agent thinks is choiceworthy could lead to AIs that are not fully aligned with externally designed rewards. The AI might not inherently consider reinforced behaviors as choiceworthy or of high utility, so its utility function may differ from the one we want it to have.

D.4 Attitudes to Risk

Overview. The concept of risk is central to the discussion of utility functions. Knowing an agent's attitude towards risk—whether they like, dislike, or are indifferent to risk—gives us a good idea of what their utility function looks like. Conversely, if we know an agent's utility function, we can also understand their attitude towards risk. We will first outline the three attitudes towards risk: risk

aversion, risk neutrality, and risk seeking. Then, we will consider some arguments for why we might adopt each attitude, and provide examples of situations where each attitude may be suitable to favor.

It is crucial to understand what risk attitudes are appropriate in which contexts. To make AIs safe, we will need to give them safe risk attitudes, such as by favoring risk-aversion over risk-neutrality. Risk attitudes will help explain how people do and should act in different situations. National governments, for example, will differ in risk outlook from rogue states, and big tech companies will differ from startups. Moreover, we should know how risk averse we should be with AI development, as it has both large upsides and downsides.

D.4.1 What Are the Different Attitudes to Risk?

There are three broad types of risk preferences. Agents can be risk averse, risk neutral, or risk seeking. In this section, we first explore what these terms mean. We consider a few equivalent definitions by examining different concepts associated with risk [55]. Then, we analyze what the advantages to adopting each certain attitude toward risk might be.

Let's consider these in the context of a bet on a coin toss. Suppose agents are given the opportunity to bet \$1000 on a fair coin toss—upon guessing correctly, they would receive \$2000 for a net gain of \$1000. However, if they guess incorrectly, they would receive nothing and lose their initial bet of \$1000. The expected value of this bet is \$0, irrespective of who is playing: the player gains or loses \$1000 with equal probabilities. However, a particular player's willingness to take this bet, reflecting their risk attitude, depends on how they calculate expected utility.

a. *Risk aversion* is the tendency to prefer a certain outcome over a risky option with the same expected value. A risk-averse agent would not want to participate in the coin toss. The individual is unwilling to take the risk of a potential loss in order to potentially earn a higher reward. Most humans are instinctively risk averse. A common example of a risk-averse utility function is $u(x) = \log x$ (red line in Figure D.4).

b. *Risk neutrality* is the tendency to be indifferent between a certain outcome and a risky option with the same expected value. For such players, expected utility is proportional to expected value. A risk-neutral agent would not care whether they were offered this coin toss, as its expected value is zero. If the expected value was negative, they would prefer not to participate in the lottery, since the lottery has negative expected value. Conversely, if the expected value was positive, they would prefer to participate, since it would then have positive expected value. The simplest risk-neutral utility function is $u(x) = x$ (blue line in Figure D.4).

c. *Risk seeking* is the tendency to prefer a risky option over a sure thing with the same expected value. A risk-seeking agent would be happy to participate in this lottery. The individual is willing to risk a negative expected value to potentially earn a higher reward. We tend to associate risk seeking with irrationality, as it leads to lower wealth through repeated choices made over time. However, this is not necessarily the case. An example of a risk-seeking utility function is $u(x) = x^2$ (green line in Figure D.4).

We can define each risk attitude in three equivalent ways. Each draws on a different aspect of how we represent an agent's preferences.

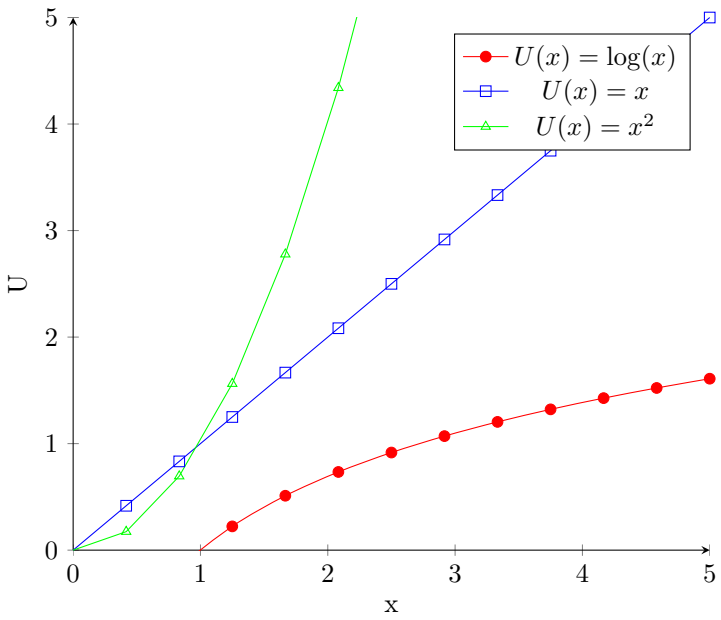

Figure D.4: Concave, linear, and convex utility functions model risk averse, risk neutral, and risk seeking agents' preferences.

Risk attitudes are fully explained by how an agent values uncertain outcomes. According to expected utility theory, an agent's risk preferences can be understood from the shape of their utility function, and vice-versa. We will illustrate this point by showing that concave utility functions necessarily imply risk aversion. An agent with a concave utility function faces decreasing marginal utility. That is, the jump from \$1000 to \$2000 is less satisfying than the jump from wealth \$0 to wealth \$1000. Conversely, the agent dislikes dropping from wealth \$1000 to wealth \$0 more than they like jumping from wealth \$1000 to wealth \$2000. Thus, the agent will not enter the aforementioned double-or-nothing coin toss, displaying risk aversion.

Preferences over outcomes may not fully explain risk attitudes. It may seem unintuitive that risk attitudes are entirely explained by how humans calculate utility of outcomes. As we just saw, in expected utility theory, it is assumed that agents are risk averse only because they have diminishing returns to larger outcomes. Many economists and philosophers have countered that people also have an inherent aversion to risk that is separate from preferences over outcomes. At the end of this chapter, we will explore how non-expected utility theories have attempted to more closely capture human behavior in risky situations.

D.4.2 Risk and Decision Making

Overview. Having defined risk attitudes, we will now consider situations where it is appropriate to act in a risk-averse, risk-neutral, or risk-seeking manner. Often, our risk approach in a situation aligns with our overall risk preference—if we are risk averse in day-to-day life, then we will also likely be risk averse when investing our money. However, sometimes we might want to make decisions as if we have a different attitude towards risk than we truly do.

Criterion of rightness vs. decision procedure. Philosophers distinguish between a *criterion of rightness*, the way of judging whether an outcome is good, and a *decision procedure*, the method of

making decisions that lead to the good outcomes. A good criterion of rightness may not be a good decision procedure. This is related to the gap between theory and practice, as explicitly pursuing an ideal outcome may not be the best way to achieve it. For example, a criterion of rightness for meditation might be to have a mind clear of thoughts. However, as a decision procedure, thinking about not having thoughts may not help the meditator achieve a clear mind—a better decision procedure would be to focus on the breath.

As another example, the *hedonistic paradox* reminds us that people who directly aim at pleasure rarely secure it [56]. While a person's pleasure level could be a criterion of rightness, it is not necessarily a good guide to increasing pleasure—that is, not necessarily a good decision procedure. Whatever one's vision of pleasure looks like—lying on a beach, buying a boat, consuming drugs—people who directly aim at pleasure often find these things are not as pleasing as hoped. People who aim at meaningful experiences, helping others and engaging in activities that are intrinsically worthwhile, are more likely to be happy. People tend to get more happiness out of life when not aiming explicitly for happiness but for some other goal. Using the criterion of rightness of happiness as a decision procedure can predictably lead to unhappiness.

Maximizing expected value can be a criterion of rightness, but it is not always a good decision procedure. In the context of utility, we observe a similar discrepancy where explicitly pursuing the criterion of rightness (maximizing the utility function) may not lead to the best outcome. Suppose an agent is risk neutral, such that their criterion of rightness is maximizing a linear utility function. In the first subsection, we will explore how they might be best served by making decisions as if they are risk averse, such that their decision procedure is maximizing a concave utility function.

Why Be Risk Averse?

Risk-averse behavior is ubiquitous. In this section, we will explore the advantages of risk aversion and how it can be a good way to advance goals across different domains, from evolutionary fitness to wealth accumulation. It might seem that by behaving in a risk-averse way, thereby refusing to participate in some positive expected value situations, agents leave a lot of value on the table. Indeed, extreme risk aversion may be counterproductive—people who keep all their money as cash under their bed will lose value to inflation over time. However, as we will see, there is a sweet spot that balances the safety of certainty and value maximization: risk-averse agents with logarithmic utility almost surely outperform other agents over time, under certain assumptions.

Response to computational limits. In complex situations, decision makers may not have the time or resources to thoroughly analyze all options to determine the one with the highest expected value. This problem is further complicated when the outcomes of some risks we take have effects on other decisions down the line, like how risky investments may affect retirement plans. To minimize these complexities, it may be rational to be risk averse. This helps us avoid the worst effects of our incomplete estimates when our uncertain calculations are seriously wrong.

Suppose Martin is deciding between purchasing a direct flight or two connecting flights with a tight layover. The direct flight is more expensive, but Martin is having trouble estimating the likelihood and consequences of missing his connecting flight. He may prefer to play the situation safe and pay for the more expensive direct flight, even though the true value-for-money of the connected route may have been higher. Now Martin can confidently make future decisions like booking a bus from the airport to his hotel. Risk-averse decision-making not only reduces computational burden but can also increase decision-making speed. Instead of constantly making difficult calculations, an agent may prefer to have a bias against risk.

Behavioral advantage. Risk aversion is not only a choice but a fundamental psychological phenomenon, and is influenced by factors such as past experiences, emotions, and cognitive biases. Since taking risks could lead to serious injury or death, agents undergoing natural selection usually develop strategies to avoid such risks whenever possible. Humans often shy away from risk, prioritizing safety and security over more risky ventures, even if the potential rewards are higher.

Studies have shown that animals across diverse species exhibit risk-averse behaviors. In a study conducted on bananaquits, a nectar-drinking bird, researchers presented the birds with a garden containing two types of flowers: one with consistent amounts of nectar and one with variable amounts. They found that the birds never preferred the latter, and that their preference for the consistent variety was intensified when the birds were provided fewer resources in total [57]. This risk aversion helps the birds survive and procreate, as risk-neutral or risk-seeking species are more likely to die out over time: it is much worse to have no nectar than it is better to have double the nectar. Risk aversion is often seen as a survival mechanism.

Natural selection favors risk aversion. Just as individual organisms demonstrate risk aversion, entire populations are pushed by natural selection to act risk averse in a manner that maximizes the expected logarithm of their growth rather than the expected value. Consider the following highly simplified example. Suppose there are three types of animals—antelope, bear, crocodile—in an area where each year is either scorching or freezing with probability 0.5. Every year, the populations grow or shrink depending on the weather—some animals are better suited to the hot weather, and some to the cold. The populations' per-capita offspring, or equivalently the populations' growth multipliers, are shown in the table below.

	Growth when Scorching ($p = 1/2$)	Growth when Freezing ($p = 1/2$)
Antelope	1.1	1.1
Bear	1.6	0.6
Crocodile	2.2	0.0

Table D.1: Different animals have different seasonal growth rates.

Antelope have the same growth in each state, bears grow faster in the warmth but slower in the cold when they hibernate, and crocodiles grow rapidly when it is scorching and animals gather near water sources but die out when their habitats freeze over. However, notice that the three populations have the same average growth ratio of 1.1.

However, "average growth" is misleading. Suppose we observe this population over two periods, one hot followed by one cold. The average growth multiplier over these two periods would be 1.1 for every animal. However, this does not mean that they all grow the same amount. In the table below, we can see the animals' growth over time.

	Initial Population	Hot Period Growth	Hot Period Population	Cold Period Growth	Cold Period Population	Overall Log Growth
Antelope	1000	1.1x	1100	1.1x	1210	0.19
Bear	1000	1.6x	1600	0.6x	960	-0.04
Crocodile	1000	2.2x	2200	0x	0	$-\infty$

Table D.2: All else equal, the species with a more stable growth rate wins out over time.

Adding the logarithm of each species' hot and cold growth rates indicates its long term growth trajectory. The antelope population will continue growing no matter what, compounding over time. However, the crocodile population will not—as soon as it enters a cold year, the crocodiles will become permanently extinct. The bear population is not exposed to immediate extinction risk, but over time it will likely shrink towards extinction. Notice that maximizing long-run growth in this case is equivalent to maximizing the sum of the logarithm of the growth rates—this is risk aversion. The stable growth population, or equivalently the risk-averse population, is favored by natural selection [58].

Avoid risk of ruin. Risk aversion's key benefit is that it avoids risk of ruin. Consider a repeated game of equal probability "triple-or-nothing" bets. That is, players are offered a $\frac{1}{2}$ probability of tripling their initial wealth w, and a $\frac{1}{2}$ probability of losing it all. A risk-neutral player can calculate the expected value of a single round as:

$$\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 3w = 1.5w.$$

Since the expected value is greater than the player's initial wealth, a risk-neutral player would bet their entire wealth on the game. Additionally, if offered this bet repeatedly, they would reinvest everything they had in it each time. The expected value of taking this bet n times in a row, reinvesting all winnings, would be:

$$\frac{1}{2} \cdot 0 + \frac{1}{4} \cdot 0 + \cdots + \frac{1}{2^n} \cdot 0 + \frac{1}{2^n} \cdot 3^n \cdot w = (1.5)^n w.$$

If the agent was genuinely offered this bet as many times as they wanted, then they would continue to invest everything infinitely many times, which gives them expected value of:

$$\lim_{n \to \infty} 1.5^n w = \infty.$$

This is another infinite expected value game—just like in the St. Petersburg Paradox! However, notice that this calculation is again heavily skewed by a single, low-probability branch in which an extremely lucky individual continues to win, exponentially increasing their wealth. In the figure below, we show the first four bets in this strategy with a starting wealth of 16. Only along the cyan branch does the player win any money, and this branch increasingly becomes astronomically improbable. We would rarely choose to repeatedly play triple-or-nothing games with everything we owned in real life. We are risk averse when dealing with high probabilities of losing all our money. Acting risk neutral and relying on expected value would be a poor decision-making strategy.

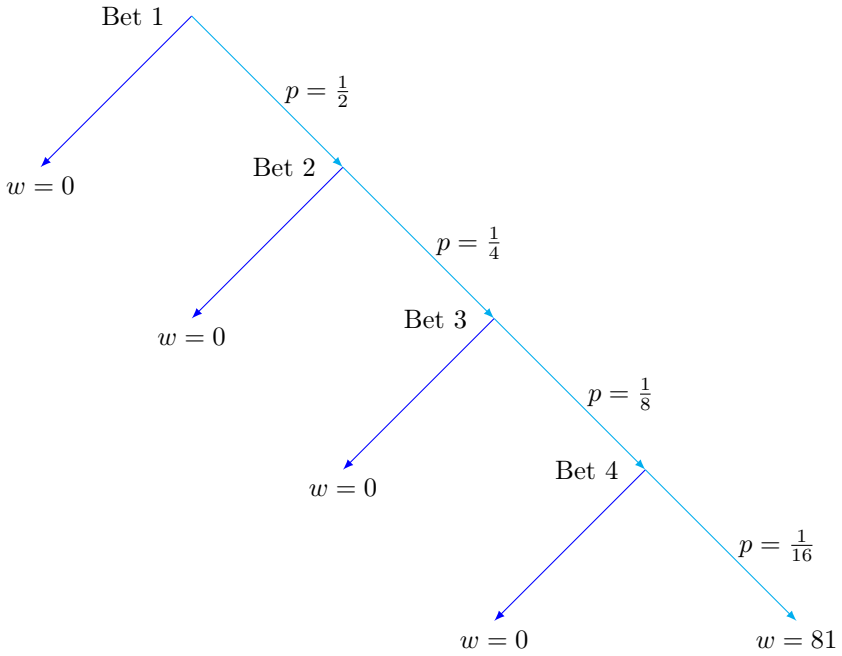

Figure D.5: Risk-neutral betting can lead to ruin.

Maximizing logarithmic utility is a better decision procedure. Agents might want to act as if maximizing the logarithm of their wealth instead of maximizing the expected value. A logarithmic function avoids risk of ruin because it assigns a utility value of negative infinity to the outcome of zero wealth, since $\log 0 \to -\infty$. Therefore an agent with a logarithmic utility function in wealth will never participate in a lottery that could, however unlikely the case, land them at zero wealth. The logarithmic function also grows slowly, placing less weight on very unlikely, high-payout branches, a property that we used to resolve the St. Petersburg Paradox. While we might have preferences that are linear over wealth (which is our criterion of rightness) we might be better served by a different decision procedure: maximizing the logarithm of wealth rather than maximizing wealth directly.

Maximizing the logarithm of wealth maximizes every percentile of wealth. Maximizing the logarithmic utility valuation avoids risk of ruin since investors never bet their entire wealth on one opportunity, much like how investors seek to avoid over-investing in one asset by diversifying investments over multiple assets. Instead of maximizing average wealth (as expected value does), maximizing the logarithmic utility of wealth maximizes other measures associated with the distribution of wealth. In fact, doing so maximizes the median, which is the 50th percentile of wealth, and it also delivers the highest value at any arbitrary percentile of wealth. It even maximizes the mode—the most likely outcome. Mathematically, maximizing a logarithmic utility function in wealth outperforms any other investment strategy in the long run, with probability one (certainty) [59]. Thus, variations on maximizing the logarithm of wealth are widely used in the financial sector.

Why Be Risk Neutral?

Risk neutrality is equivalent to acting on the expected value. Since averages are straightforward and widely taught, expected value is the mostly widely known explicit decision-making procedure. However, despite expected value calculations being a common concept in popular discourse, situations where agents do and should act risk neutral are limited. In this section, we will first look at the conditions under which risk neutrality might be a good decision procedure—in such cases, maximizing expected value can be a significant improvement over being too cautious. However, being mistaken about whether the conditions hold is entirely possible. We will examine two scenarios: one when these conditions hold, and one situation in which incorrectly assuming that they held led to ruin.

Risk neutrality is undermined by the possibility of ruin. In the previous section, we examined the triple-or-nothing game, where a risk-neutral approach can lead to zero wealth in the long term. The risk of ruin, or the loss of everything, is a major concern when acting risk neutral. In order for a situation to be free of risk of ruin, several conditions must be met. First, risks must be *local*, meaning they affect only a part of a system, unlike *global risks*, which affect an entire system. Second, risks must be *uncorrelated*, which means that the outcomes do not increase or decrease together, so that local risks do not combine to cause a global risk. Third, risks must be *tractable*, which means the consequences and probabilities can be estimated reasonably easily. Finally, there should be no *black swans*, unlikely and unforeseen events that have a significant impact. As we will see, all of these conditions are rarely met in a high-stakes environment, and there can be dire consequences to underestimating the severity of risks.

Risk neutrality is useful when the downside is small. It can be appropriate to act in a risk-neutral manner with regards to relatively inconsequential decisions. Suppose we're considering buying tickets to a movie that might not be any good. The upside is an enjoyable viewing experience, and the downsides are all local: $20 and a few wasted hours. Since the stakes of this decision are minimal, it is reasonable not to overthink our risk attitude and just attend the movie if we think that, on average, we won't regret this decision. However, if the decision at hand were that of purchasing a car on credit, we likely would not act hastily. The risk might not be localized but instead affect one's entire life; if we can't afford to make payments, we could go bankrupt. However, when potential losses are small, extreme risk aversion may be too safe a strategy. We would prefer not to leave expected value on the table.

Dangers of risk neutrality. Often, agents incorrectly assume that there is no risk of ruin. The failure of financial institutions during the 2008 financial crisis, which sparked the Great Recession, is a famous example of poor risk assessment. Take the American International Group (AIG), a multinational insurance company worth hundreds of billions of dollars [60]. By 2008, they had accumulated billions of dollars worth of financial products related to the real estate sector. AIG believed that their investments were sufficiently uncorrelated, and therefore ruled out risk of ruin. However, AIG had not considered a black swan: in 2008, many financial products related to the housing market crashed. AIG's investments were highly correlated with the housing market, and the firm needed to be bailed out by the Federal Reserve for $180 billion dollars. Even institutions with sophisticated mathematical analysis fail to identify risk of ruin—playing it safe might, unsurprisingly, be safer. Artificial agents may operate in environments where risk of ruin is a real and not a far-fetched possibility. We would not want a risk-neutral artificial agent endangering human lives because of a naive expected value calculation.

Why Be Risk Seeking?

Risk-seeking behavior is not always unreasonable. As we previously defined, risk-seeking agents prefer to gamble for the chance of a larger outcome rather than settle for the certainty of a smaller one. In some cases, a risk-seeking agent's behavior may be regarded as unreasonable. For example, gambling addicts take frequent risks that lower their utility and wellbeing in the long run. On the other hand, many individuals and organizations may be motivated to seek risks for a number of strategic reasons, which is the focus of this section. We will consider four situations as examples where agents might want to be risk seeking.

In games with many players and few winners, risk-seeking behavior can be justified. Consider a multi-player game where a thousand participants compete for a single grand prize, which is given to the player who accumulates the most points. An average player expects to only win $\frac{1}{1000}^{th}$ of the time. Even skilled players would reason that due to random chance, they are unlikely to be the winner. Therefore, participants may seek risks, in order to increase the *variance* of their point totals, while sacrificing the mean, so that they either end up with loads of points (thereby winning with higher probability) or no points at all (which is no worse for them than having some points). In the real world, selling products in a highly competitive marketplace is an analogous situation, where vendors may take risks to attract customers. If a hundred firms are selling effectively identical products, they might consider unusual or provocative forms of advertising. Bold marketing strategies can attract some customers but potentially alienate others. However, the vendor may feel that without taking this chance to stand out, they likely will not do enough business to turn a profit. Such agents would accept a negative expected value strategy.

Daring to rise. Agents in bad and deteriorating situations with a low chance of escaping can pursue risk-seeking strategies to great effect. Someone with a serious terminal illness may consider experimental treatments with uncertain outcomes. They may take the treatment even if told that it is most likely ineffective and might have serious side effects. Sports teams on the verge of losing often attempt risky strategies such as the "Hail Mary" in American football (long, high passes that are difficult to catch) or sending a goalkeeper forward in soccer towards the end of the game. Such strategies are likely to backfire and leave them in an even worse position overall—and therefore have a negative expected value, where value is the number of goals—but might also create the only possibility of winning.

Harnessing stressors through risk exposure. Instead of collapsing (or merely enduring) when unlikely, bad events occur, an *antifragile system* is one in which the risk taker actually benefits, becoming stronger and more resilient to future challenges [61]. Antifragility is therefore the property of being able to benefit from risk. Systems, institutions, or individuals that exhibit antifragility may thus seek risk exposure. The human body is antifragile in many contexts, including its response to pathogens. Illness is usually temporarily uncomfortable, but carries a small risk of greater complications. In combating a pathogen, the immune system not only responds to the active threat, but prepares itself to combat future illnesses quickly and effectively. The immune system becomes stronger through illness. We encourage children to step out of the house instead of solely living in sterilized environments. To a reasonable extent, we help them to face and deal with risks so that they become stronger.

Startups aiming to capture significant upside potential. When losing a small amount has a negligible or tolerable effect on an agent's wellbeing, the agent may be willing to risk this small loss for a low-probability outcome of very large gain, even if the probability of success is sufficiently small to make this negative expected monetary value. This line of thinking is exemplified by early stage startups, which we now describe.

The Lifecycle of a Company

A company's appetite for risk over time is captured by a *sigmoid* (s-shaped) curve, which is initially convex and later concave. Knowing that an agent has a concave utility function if and only if they are risk averse, and a convex utility function if and only if they are risk seeking, we understand that an agent with a sigmoid utility function is initially risk seeking, and later becomes risk averse. This model may give us an idea of how AI developers will behave, depending on the scale of the organization and the maturity of the technology.

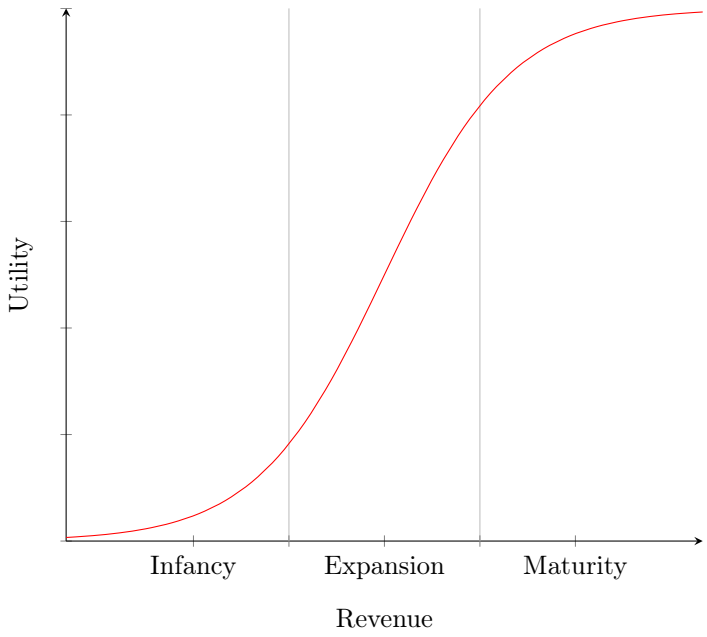

Figure D.6: Start-ups often take big risks to try and get explosive growth, modeled by a convex utility function. As a company matures, it gets more risk averse, prioritizing survival over extreme growth.

A startup is by nature a risk-seeking venture. An entrepreneur has typically sacrificed a stable income, savings, and much of their leisure time in order to pursue a business. The new business, starting with little traction and few customers, has little to lose and much to gain. Given their sacrifices and the startup's position, the entrepreneur is willing to fail repeatedly and return to baseline, prioritizing chances at rapid growth. By this logic, we may expect AI startups to prioritize chances of success *over* safety and avoiding reputational damage. However, since such companies operate on a smaller scale, risks are localized, meaning societal concerns are reduced. This is the convex part of the curve.

When a startup begins to gain traction, it grows rapidly, gaining customers and revenue. Gradually, more stakeholders like employees and investors begin to rely on the company, and its focus begins to shift toward preserving its proven success and preventing future losses, rather than risking a return to baseline to pursue more growth. During this transition, the curve begins to shift from convex to concave.

In the concave stage of a company, its growth tapers off. Eventually, the company nears the limits of its market and transformative resources, and is unable to risk its bottom line for further growth opportunities. Mature companies are risk averse, since employees, shareholders, and customers depend on their stability. They have much to lose and little to gain. A mature company's consolation is that it may develop a project portfolio, with riskier projects sprinkled in amongst its core business operations that are not typically exposed to risk. Compared to startups, big tech companies may thus take a more meticulous approach to ensuring the safety of AI products prior to release, though this is not always the case.

A company's lifecycle demonstrates that an agent does not have only one unchanging risk attitude. An agent's approach to risk is affected by their situation and decision context. Indeed, describing a person as risk averse or risk seeking will always be an oversimplification. As we will explore in the final section, people have dynamic risk attitudes that are influenced by many factors including biases, context, initial wealth, and more.

Summary. In this section, we defined the three attitudes towards risk—risk aversion, risk neutrality, and risk seeking—and examined their properties and shapes. There are reasons to favor each attitude, depending on the agent's circumstances. We saw that risk aversion in the form of maximizing a logarithmic utility function outperforms all other investment strategies in the long run, that risk neutrality may be favorable when there is no risk of ruin, and that risk-seeking behavior is useful when we have little to lose. Humans, and the organizations we form, adopt different risk attitudes in different situations.

When designing AIs, developers may have significant ability to define and influence the utility function. As demonstrated in the corrigibility problem, issues in the utility function may be later difficult to rectify. Thus, we must carefully consider what risk attitudes a particular utility function embodies, and how that risk attitude would play out in different contexts.In particular, designing AIs to be risk-averse may help avoid many of the pitfalls of risk-neutrality that we discussed.

We are interested in describing a more accurate model of human decision making. In the next section, we will analyze some reasons why expected utility theory fails in this regard, and how we might improve our model. We will see that humans are systematically irrational and can be influenced by the context, and even the wording, in which choices are presented. These models are helpful for better understanding people, like those leading AI development or countries or those using AIs, who will behave in irrational ways.

D.5 Beyond Expected Utility Theories

Overview. von Neumann-Morgenstern utility functions are supposed to capture how an agent chooses among different options, but they do not always explain how humans actually behave. In reality, humans are not ideally rational agents. We'd like to model how people, like those leading the development and regulation of AI, will behave, and they will often behave in ways that perfect expected utility maximizers would not. If AIs learn from and interact with humans, they too might exhibit some aspects of human irrationality. Therefore, we like to model human behavior more accurately.

Economists have proposed alternative theories, such as Daniel Kahneman and Amos Tversky's Prospect theory, to explain why humans deviate from rationality, as defined under the von Neumann-Morgenstern axioms. In this section, we examine some major ways that humans break rationality, as defined under the von Neumann-Morgenstern model, and how non-expected utility theories help us

better understand human choices. Having a stronger model of human behavior will ultimately help us design AIs that behave in ways more aligned with humans.

D.5.1 Humans and Rationality

Humans are not ideally rational agents. As we discussed before, rational agents have a thorough understanding of their own preferences, have complete and stable preferences, and are able to make which decisions will help them maximally satisfy these preferences. Human decision-making deviates from ideal rationality. For example, we prioritize fairness, may have incommensurable values, and value the desires of others even when these mean we must compromise self-interest. Human preferences are also unstable, susceptible to persuasion, and can change over the course of our lives or in light of new information.

Humans often satisfice rather than maximize. According to the theory of *bounded rationality*, humans often make choices that result in outcomes that are "good enough" rather than the most ideal. Human rationality is limited by cognitive abilities, time, and available information, meaning we must frequently make decisions without considering all possible scenarios and outcomes. Take an everyday example: suppose a group of friends is deciding what restaurant to dine at. They will likely choose the first satisfactory option they come across, rather than methodically consider all possible places in the city. Thus, humans are said to *satisfice*—choose the first option that is satisfactory—rather than exhaustively maximize.

AIs that are not ideally rational can have varying safety. It is plausibly safer for AI systems to be satisficers, since maximizers may behave in undesirable ways while trying to relentlessly maximize some metric [62]. As discussed in Proxies, maximizers may optimize proxies in ways that differ from the idealized result. However, when AI agents are trained on human data and trained to interact with humans, they may pick up some of our biases and thereby not be ideally rational. AIs with many irrational behaviors could be harder to predict and therefore be harder to control. We will now proceed to formalize our understanding of why expected utility theory fails to capture human behavior.

D.5.2 Evidence Against Expected Utility Theory

Overview. Humans violate the von Neumann-Morgenstern axioms in many different ways. We consider three examples in this section: the *Allais Paradox*, *fairness*, and the *problem of framing*, in which humans make inconsistent choices over lotteries that agents following von Neumann-Morgenstern would be indifferent between. These problems show violations of von Neumann-Morgenstern rationality: specifically, the independence and decomposability axioms.

Allais Paradox

Humans are not perfect expected utility maximizers. The Allais Paradox, described in 1953 to highlight an inconsistency between utility theory and human behavior, presents two scenarios where players must make a choice between two gambles [63].

Choice:	Gamble 1	Gamble 2
Scenario A	100% chance of \$1 million	10% chance of \$5 million 89% chance of \$1 million 1% chance of \$0

Table D.3: Allais Paradox: Scenario A

Choice:	Gamble 1	Gamble 2
Scenario B	11% chance of $1 million 89% chance of $0	10% chance of $5 million 90% chance of $0

Table D.4: Allais Paradox: Scenario B

In reality, most players favor Gamble 1 in Scenario A, due to the certainty of a large payout of $1 million. Simultaneously, they favor Gamble 2 in Scenario B, since the larger potential payout of $5 million outweighs its slightly lower likelihood. Both these preferences are individually sensible and unproblematic. However, holding both preferences simultaneously is in violation of von Neumann-Morgenstern's independence axiom, and thus inconsistent with expected utility theory.

Humans violate the independence axiom. We explained above that the independence axiom holds that preferences between two lotteries are not impacted by the addition of equal probabilities of a third, independent lottery to each lottery. It follows that we can subtract the same lottery from two equivalent lotteries and preserve the original preference.

In Scenario A, an 89% chance of winning $1 million is common to both choices. Therefore, the decision between Gamble 1 and Gamble 2 in Scenario A is equivalent to the reduced game described, in which the 100% chance of winning $1 million has simply been divided into an 89% chance and an 11% chance of winning the same $1 million.

	Gamble 1	Gamble 2
Scenario A	89% chance of $1 million 11% chance of $1 million	10% chance of $5 million 89% chance of $ 1 million 1% chance of $0
Scenario A reduced	11% chance of 1 million	10% chance of $ 5 million 1% chance of $0

Table D.5: Allais Paradox: Scenario A reduced

Similarly, there is a common lottery between Gamble 1 and Gamble 2 in Scenario B. We can ignore an 89% chance of winning $0 from both choices and we will be left with the reduced game described, in which the 89% chance of winning $0 has simply been ignored.

	Gamble 1	Gamble 2
Scenario B	11% chance of $1 million 89% chance of $0	10% chance of $5 million 90% chance of $0
Scenario B reduced	11% chance of $1 million	10% chance of $5 million 1% chance of $0

Table D.6: Allais Paradox: Scenario B reduced

Scenario A Reduced and Scenario B Reduced are exactly the same! Therefore, a rational agent should be consistent and select the same gamble in either simplified game. Since the simplified scenarios are equivalent to the original scenarios via the independence axiom—adding third lotteries to both options should make no difference to the choice in either scenario—we can conclude that a rational agent should also select the same gamble in Scenario A and Scenario B. In reality, people overwhelmingly favor Gamble 1 in Scenario A and Gamble 2 in Scenario B. This behavior is inconsistent with rational

behavior under the independence axiom.

We can interpret this discrepancy in two ways. We could assume that humans are making an error of judgment. Or, we could conclude that human choice isn't unreasonable, but that the von Neumann-Morgenstern axioms are a flawed characterization of rationality. Indeed, Kahneman and Tversky concluded that the expected utility model fails to capture important nuances in human decision making [64].

Fairness

It is sometimes thought that the independence axiom is at odds with concepts of fairness. The independence axiom tells us that adding equal probabilities of a third lottery makes no difference to an agent's preference between two other lotteries. However, in some cases, we do care what else could happen: we make all-things-considered judgments of how we value outcomes. In particular, we care about fairness [65].

If an agent cares about fairness, they may be forced to abandon independence. Suppose Rachel, on her deathbed, has to leave everything she owns to one person. She is indifferent between everything going to her son and everything going to her daughter, but would like to treat them equally. There is a 50% chance that the law will change such that everything goes to her son, no matter what her will says. She is now considering two options, which we can write down as lotteries.

a. She leaves everything to her daughter. Let this be the "fair lottery" L_F, wherein her daughter receives everything with probability 0.5 (once her will is executed), and her son receives everything with probability 0.5 (once the laws change).

b. She leaves everything to her son. Let this be the "unfair lottery" L_U, wherein her son receives everything with probability 1, since he receives it once her will is executed or once the laws change.

According to the independence axiom, if she is indifferent between her son and daughter receiving her possessions for sure, then she should also be indifferent between the above two lotteries because we can obtain them by adding equal probabilities of a different outcome—a 50% chance her son gets everything—to the original choice. However, if she cares at all about fairness, then she should prefer L_F over L_U—a preference that is incompatible with the independence axiom!

Again, we see that the von Neumann-Morgenstern axioms lack descriptive power. While, in principle, adding alternatives should be irrelevant, they are not always so. The Allais paradox demonstrated that adding alternatives changes how we think about monetary lotteries—this is sometimes attributed to factors like avoiding regret. In this case, we are concerned with fairness. Next, we will consider how humans systematically violate rationality based on the description of options.

Framing

Human decision making is swayed by the presentation of choices. Kahneman and Tversky noticed that humans will make different decisions depending on how options are presented, even when the underlying lotteries remain unchanged [66]. While a vNM-rational decision maker would ignore the presentation of options and focus only on its probabilities and outcomes, humans tend to be unaware of the large degree of influence that *framing* has on their decision making. Furthermore, human decision making usually uses vague feelings of subjective probability rather than well-considered

mathematically defined lotteries. Consider the following example:

An illness is expected to kill 600 people if left untreated. Participants playing the role of health officials must choose between two policy options in two different scenarios. These options are presented in the table below.

Choice:	Policy A1	Policy A2
Scenario A	Save 200 people with certainty.	1/3 chance save 600 people. 2/3 chance saves no one.

Table D.7: Framing effects: Scenario A

Choice:	Policy B1	Policy B2
Scenario B	400 people die with certainty.	1/3 chance save of no deaths. 2/3 chance of 600 deaths.

Table D.8: Framing effects: Scenario B

Framing effects violate von Neumann-Morgenstern rationality. Participants tend to choose Policy A1 in Scenario A, because of the certainty of saving lives. Participants also tend to choose Policy B2 in Scenario B, because of the possibility of averting more death and not condemning people to certain death. However, the only difference between Scenario A and B is the manner of presentation, or framing. Scenario A is presented in terms of lives saved, and Scenario B in terms of deaths caused. The equivalence is shown in this table.

Choice:	Policy A1/B1	Policy A2 B2
Scenario A	Save 200 people with certainty (and so causes 400 deaths for sure).	1/3 chance save 600 people (or no deaths). 2/3 chance saves no one (or 600 deaths).
Scenario B	Causes 400 deaths with certainty (and so saves 200 people for sure)	1/3 chance save of no deaths (or saving 600 people). 2/3 chance of 600 deaths (or saving no one at all).

Table D.9: Framing effects: Scenarios A and B compared

Together, the choice of Policy A1 in Scenario A and Policy B2 in Scenario B violate a requirement of von Neumann-Morgenstern rationality: that an agent be indifferent between two lotteries with the same probabilities and outcomes, even if they are presented differently. In reality, an individual's habits, environment, and cognitive biases can cause different responses to framings of the same underlying lottery. These are ways individuals can deviate from expected utility theory.

D.5.3 Prospect Theory

Prospect theory is a prominent non-expected utility model of human behavior. Prospect theory seeks to accurately describe how people make choices in risky situations by providing a psychological model of decision making. The model's features are designed to more accurately describe human behavior in situations when facing risk, rather than provide a description of how

humans "should" behave, like in the vNM utility model. Thus, prospect theory explains common behavioral patterns that are considered irrational in expected utility theory [67].

Prospect theory is a multi-stage decision-making model. Prospect theory views decision making over two stages: editing and evaluation. In the editing stage, outcomes are re-framed as either gains or losses instead of just final wealth. This is important because people can perceive the same outcome differently based on how the problem is presented and their own biases. This stage also accounts for framing effects. In the evaluation stage, gains and losses are multiplied by weighted probabilities to determine the preferred outcome. This stage takes into account two factors: a *value function*, how people value different outcomes, and a *weighting function*, how people weigh the probabilities of each outcome. Next, we will explore how these two functions help to explain how people make decisions in uncertain situations.

Prospect theory's value function approximates how humans think about wealth. A value function assigns value to an outcome, much like a Bernoulli utility function assigns utility to an outcome. In prospect theory, the value function has a few key characteristics. The input is defined in gains and losses, modeling the idea that people are sensitive to changes in wealth, not only to their level of wealth. The curve is sigmoid to reflect that people are risk seeking towards losses and risk averse towards gains. That is, people like a chance at avoiding losses but dislike a chance of losing gains. The curve is steeper in losses since people are modeled more sensitive to a loss compared to a gain of equal amount.

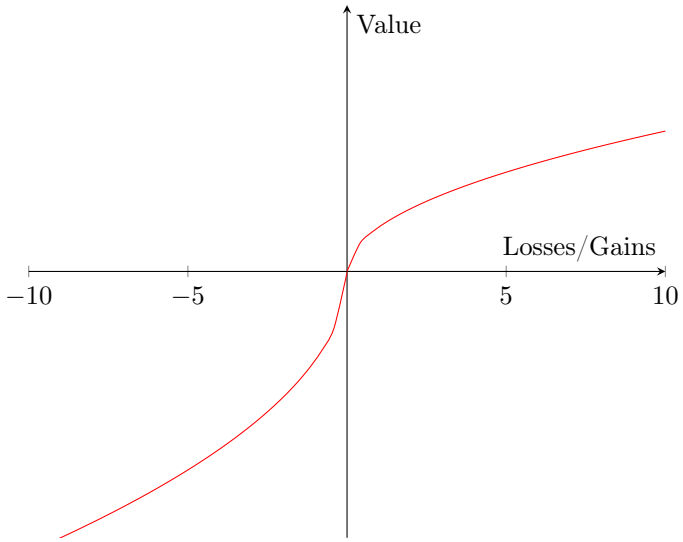

Figure D.7: Prospect theory's value function is steeper in losses than in gains.

Prospect theory's decision weights describe how humans think about probabilities. A decision weight is the scaling factor applied to an outcome that quantifies how much that outcome contributes to the decision. In expected utility theory, decision weights are just probabilities of outcomes. Instead of assuming people accurately assess the likelihood of outcomes, however, prospect theory accounts for the way humans actually process probabilities. People often overestimate the risk of unlikely events with extreme consequences. Humans significantly overestimate the risk of dying in a shark attack, possibly because of the graphic nature of the attacks and their over-representation in pop media, when in reality the true probability of death by shark attack is quite low. Humans also

place a greater weight on relative certainty: people are usually willing to pay much more to improve their odds from 0% to 1% than from 1% to 2%, since certainty of failure is removed. Prospect theory's proposed weighting for probabilities is shown in the next figure. This curve illustrates how people tend to persistently overestimate relatively small probabilities while persistently underestimating relatively large probabilities.

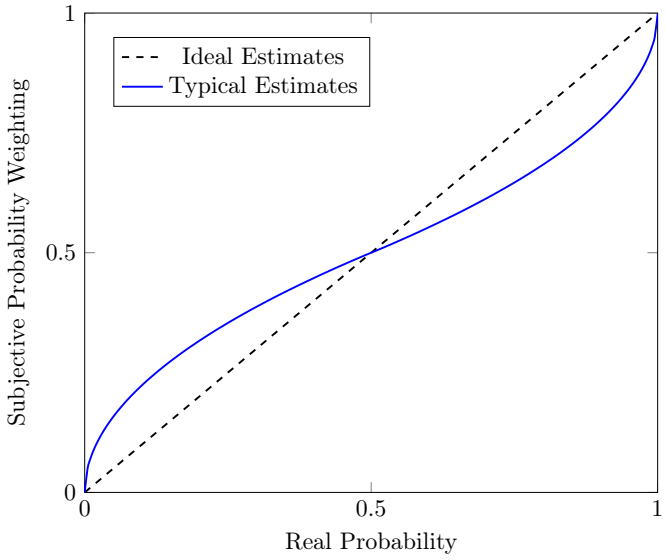

Figure D.8: Humans often perceive probabilities incorrectly, overestimating the likelihood of low-probability events and underestimating the likelihood of high-probability events.

D.5.4 Models of Decision Making

Generalized decision-making models. We can put the above together to describe many different types of decision-making models more than expected utility theory. Recall that expected utility is of the form:

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

for a lottery L with outcomes o_i and associated probabilities p, over a utility function u. We can incorporate value functions and weighting functions to construct a theory of decision making that is more general than expected utility theory while following the same structure of adding together the products of decision weights and values:

$$V(L) = w(p_1) \cdot v(o_1) + w(p_2) \cdot v(o_2) + \cdots + w(p_n) \cdot v(o_n).$$

Here, V is the value assigned to the lottery. We have also added w_i, a weighting function that transforms each probability into a corresponding decision weight, and v_i, a value function which—much like a utility function—values each outcome.

Prospect theory, formalized. Prospect theory uses this structure while using a slightly different value function which evaluates gains and losses in wealth, rather than total wealth. The S-shaped prospect theory value function is represented as $v_p(o_i - o_0)$, with o_i representing the outcome being evaluated, and o_0 representing the agent's initial state:

$$V_p = w_p(p_1) \cdot v_p(o_1 - o_0) + \cdots + w_p(p_n) \cdot v_p(o_n - o_0).$$

In monetary lotteries, this would contain $v_p(w_i - w_0)$, which tells us that the value function considers the loss or gain in wealth in each possible outcome. Prospect theory incorporates the specific weighting and value functions determined by Kahneman and Tversky's research on human behavior. We can modify the model by substituting each function with functions of our own choosing.

Summary. In this section, we considered the Allais Paradox, a thought experiment on fairness, and two instances where the von Neumann-Morgenstern axioms fail to capture human preferences. We also examined a study conducted by Kahneman and Tversky on framing effects, where humans behave in a clearly irrational manner. We then examined the value function and the weighting function, which are the two main innovations that comprise prospect theory and other non-expected utility theories that seek to capture insights into human behavior where conventional theory fails.

Understanding the limitations of expected utility theory, and having more descriptive models of human behavior helps us understand how humans, human-led organizations, and AIs imitating humans will behave. In this chapter, and in the Single-Agent Safety chapter, we present reasons that expected-utility maximizers can be dangerous. Non-expected utility theories provide some concepts to consider when designing agents that are not expected-utility maximizers.

D.6 Conclusion

In this chapter, we studied the properties of utility functions and how agents use utility functions to make decisions. Utility functions have been part of a significant paradigm within decision theory in economics, psychology, and other fields, and are increasingly relevant to understanding and designing artificial intelligence. Artificial agents in many cases are expressly designed to optimize objects (such as reward functions) that strongly shape their utility functions.

We outlined the properties of Bernoulli utility functions, which allow us to express preferences over goods and situations with precise numbers, and von-Neumann-Morgenstern utility functions, which extend utility functions over probabilistic situations. From von Neumann-Morgenstern utility functions, we derive the idea of expected utility theory: the idea that rational agents do and should make choices that maximize the expectation of their utility function. This simple-sounding idea helps us understand decision making, but also often fails to perfectly describe human behavior.

We applied utility functions to the problem of AI corrigibility—whether AI systems are receptive to corrections. AIs with complete and transitive preferences will establish preferences about ceasing to pursue their current objective, and consequently may attempt to thwart corrective measures. Non-corrigible AI systems are a significant concern, since they create difficulties in making them safe.

We worked through examples of when it may be advisable to behave in risk-averse, risk-neutral, and risk-seeking manners, which correspond to concave, linear, and convex utility functions respectively. Risk aversion is a natural instinct for animals and humans, and helps maximize median value in the long run. Risk neutrality maximizes expected value, but faces risk of ruin. Risk-seeking behavior is often applied in situations where an agent has little to lose and a lot to gain. People and organizations adopt different risk attitudes depending on the context and situation of the decision.

However, expected utility theory is a flawed theory—human behavior that we consider to be reasonable often violates the strict rationality outlined by the von Neumann-Morgenstern axioms. Paradigms outside expected utility theories, such as prospect theory, attempt to more accurately describe human decision-making processes by incorporating additional functions that describe how humans think

about wealth and subjectively weigh perceived probabilities.

An essential concern in designing artificial agents is that they must reflect human values. The broader study of utility functions, and how humans and other agents do and should make decisions, is essential context for ensuring that artificial agents avoid catastrophic risks and behave in accordance with human values.

D.7 Literature

D.7.1 Recommended Reading

While this chapter is mostly self-contained, most college level microeconomics textbooks can serve as a **primary supplementary reading**. A few examples, in increasing order of difficulty, are:

1. CORE. The Economy. (Section 3)
2. Hugh Gravelle and Ray Rees. *Microeconomics*. (Chapter 2)
3. Andreu Mas-Colell, Michael D. Whinston, Jerry R. Green. *Microeconomic Theory*. (Chapter 2)

By default, consult (2). In addition, you can look at further readings:

- Stanford Encyclopedia of Philosophy entry on the St. Petersburg Paradox:
- Adam Bales. "Will AI avoid exploitation? Artificial general intelligence and expected utility theory". In: *Philosophical Studies* (2023). URL: https://doi.org/10.1007/s11098-023-02023-4
- The Shutdown Problem: Two Theorems, Incomplete Preferences as a Solution, Thornley (forthcoming)
- D. Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, 2011
- Lara Buchak. *Risk and Rationality*. Oxford University Press, 2013. ISBN: 9780199672165. DOI: 10.1093/acprof:oso/9780199672165.001.0001. URL: https://doi.org/10.1093/acprof:oso/9780199672165.001.0001

D.7.2 References

[1] P. Cilliers and David Spurrett. "Complexity and post-modernism: Understanding complex systems". In: *South African Journal of Philosophy* 18 (Sept. 2014), pp. 258–274. DOI: 10.1080/02580136.1999.10878187.

[2] Alethea Power et al. *Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets*. 2022. arXiv: 2201.02177 [cs.LG].

[3] Jordan Hoffmann et al. *Training Compute-Optimal Large Language Models*. 2022. arXiv: 2203.15556 [cs.CL].

[4] Hang Zhao et al. "Response mechanisms to heat stress in bees". In: *Apidologie* 52 (Jan. 2021). DOI: 10.1007/s13592-020-00830-w.

[5] Trevor Kletz. *An Engineer's View of Human Error*. CRC Press, 2018.

[6] Lucas Davis. "The Effect of Driving Restrictions on Air Quality in Mexico City". In: *Journal of Political Economy* 116 (Feb. 2008), pp. 38–81. DOI: 10.1086/529398.