Normative Ethics

C.1 Introduction

Ethics is the branch of philosophy concerned with questions of right and wrong, good and bad, and how we ought to live our lives. We make ethical choices every day. When we decide whether to tell the truth or lie, help someone in need or ignore them, treat others with respect or act in a discriminatory manner, we are making moral decisions that reflect our values, beliefs, and moral principles. Philosophical ethics seeks to provide a systematic framework for making these decisions.

In this chapter, we will explore some of the key concepts and theories in philosophical ethics. This branch of research is also commonly called *moral philosophy.* We use the terms *ethics* and *morality* interchangeably. The subfield of ethics dedicated to developing moral theories is called normative ethics, which considers questions about how we ought to act. It investigates *normative* claims as opposed to *empirical* ones, examining how the world ought to be, rather than simply how it is. The former consider concepts such as rights, duties, roles, and morality, using words such as should, must, and ought. The latter contain information that is either true or false depending on how the world is and are (in theory) testable.

This chapter outlines some of the main reasons why it's important for anyone concerned about AI to learn about ethics. We then turn to the basic building blocks of moral theories, examining various moral considerations like intrinsic goods, constraints, and special obligations. Then we will explore some of the most prominent ethical theories, like utilitarianism, deontology, and virtue ethics, evaluating their strengths and weaknesses. Finally, we consider how we might deal with reasonable disagreement over what is right and wrong. Throughout, our key focus is on the ethical concepts that are most relevant to the development, implementation, and governance of AI.

C.2 Why Learn About Ethics?

This chapter will explain ethics. Here, we cover the most prominent theories in the history of ethical discourse. Reading this chapter should make it easier to understand debates about ethics in AI systems.

Ethics is relevant to the field of AI for two key reasons. First, AI systems are increasingly being integrated into various aspects of human life, such as healthcare, education, finance, and transportation, and they have the potential to significantly impact our lives and wellbeing. As AI systems become increasingly intelligent and powerful, it is crucial to ensure that they are designed, developed, and deployed in ways that promote widely shared values and do not amplify existing social biases or cause needless harms. Unfortunately, there are already numerous examples of AI systems being designed in ways that failed to adequately consider such risks, such as racially biased facial recognition systems. In order to wisely manage the growing power of AI systems, developers and users of AI systems need to understand the ethical challenges that AI systems introduce or exacerbate.

Second, AI systems raise a range of new ethical questions that are unique to their technological nature and capabilities. For instance, AI systems can generate, process, and analyze vast amounts of data—much more than was previously possible. In what ways does this new technology challenge traditional notions of privacy, consent, intellectual property, and transparency? Another important set of questions relates to the moral status of AI systems. This is likely to become more pressing if AI systems become increasingly autonomous and able to interact with human beings in ways that convince their users that they have their own preferences and feelings. What should we do if AI systems appear to meet some of the potential criteria for sentience or other morally relevant features?

Thirdly, as further explored in the Single-Agent Safety and Machine Ethics chapters, it is challenging to specify objectives or goals for highly powerful AI systems in ways that do not lead in a predictable way to highly undesirable consequences. In order to grasp why it is so challenging to specify these objectives, it is helpful to understand the ethical theories that have been proposed. Questions of what it means to act rightly or to live a good life have been debated by many thinkers over several millennia, with strong arguments advanced for a number of competing positions. These debates can provide us with greater insight into the challenges that AI developers will need to overcome in order to build increasingly powerful AI systems in a beneficial way. Rather than attempting to bypass or ignore such controversies, AI developers should accept that their design decisions may raise difficult ethical questions that need to be considered carefully.

C.2.1 Is Ethics “Relative?”

Even after millennia of deliberation, we do not agree on all of morality. Philosophers have been thinking about and debating moral principles for millennia, yet they have not achieved consensus on many moral issues. Widespread disagreements remain in both philosophical and public discourse, including about important topics like abortion, assisted suicide, capital punishment, animal rights, and the effects of human activity on natural ecosystems. One troubling idea is that these disagreements are irresolvable because no moral principles or judgments are absolutely or universally correct. In the case of AI, this may lead AI developers to believe that they have no role to play in shaping how AI systems behave.

Cultural relativism claims there is no objective, culturally independent standard of morality. Consider the principle that consensual relationships between adults are acceptable regardless of whether they are heterosexual or homosexual. A moral relativist would suggest this principle is correct for people who belong to some cultures where homosexuality is accepted, but incorrect for people who belong to other cultures where homosexuality is criminalised or socially stigmatized. These differences are systemic: many cultures have moral standards that seem incompatible with others’ ideals, such as different views on marriage, divorce, gender roles, freedom of speech, or religious tolerance. These differences form the basis for arguments for cultural relativism.

Normative moral relativism vs. descriptive moral relativism [1]. Moral relativism has various forms, but here we discuss two: descriptive moral relativism and normative moral relativism. Descriptive moral relativism is straightforward: it means that different societies around the world have different sets of rules about what’s right and wrong, much like they have unique cuisines, customs, and traditions. Descriptive moral relativism makes no claims about which, if any, of these rules is right or wrong. Normative moral relativism suggests that one cannot say that something is right or wrong in general, but only relative to a particular culture or set of norms. Normative moral relativists conclude that morality itself is not something universal or absolute. Strictly speaking, descriptive moral relativism and normative moral relativism are independent of each other, although in practice descriptive moral relativism is often treated as if it provides evidence for normative moral relativism.

Objections to Moral Relativism

A number of arguments can be advanced against descriptive and normative moral relativism [1], which we explore in this subsection. We will explore the argument that cultural differences might be overstated, which makes descriptive moral relativism harder to uphold. Another argument is that proponents of normative moral relativism often face challenges when confronted with instances of extreme harm. For instance, while many would unequivocally agree that torturing a child for entertainment is morally wrong, a normative moral relativist might be required to argue that its morality is contingent upon the cultural context. Extreme examples such as this suggest few people are willing to be thoroughgoing moral relativists. We further explore arguments for and against moral relativism in this section.

Human moral systems appear to share some common features. Some have argued that most or all societies share some norms. For example, prohibitions against lying, stealing, or killing human beings are common across cultures. Many cultures have some form of reciprocity, which is the idea that people have a moral obligation to repay the kindness or generosity they have received from others or that people should treat others the way they wish to be treated [2]. This can be seen in the widespread practice of exchanging gifts and in moral codes that emphasize fairness and justice. Additionally, human cultures have typically some concept of parenthood, which often involves a moral obligation to care for one's children, as well as broader obligations to one's family and group. These common features suggest that there are at least a few universal aspects of morality that transcend cultural boundaries.

Moral relativism conflicts with common-sense morality [1]. Consider controversial practices still prevalent in some cultures, such as honor killings in parts of the Middle East. The honor of a family depends on the "purity" of its women. If a woman is raped or is deemed to have compromised her chastity in some way, the profound shame brought upon her family may lead them to kill her in response. According to the normative moral relativist, if such a practice is in line with the moral standards of the society where it takes place, there is nothing wrong with it. Even more disturbingly, on some versions of relativism, men in these societies may even be considered morally in the wrong if they fail to kill their wives, daughters or sisters for having worn the wrong clothing, having premarital sex or being raped. Similarly, normative moral relativism would require us to believe that the morality of owning slaves was entirely dependent on the societal context. Moral iconoclasts, such as early anti-slavery campaigners, would by definition always be morally wrong. In practice, if required to accept that moral standards that endorse honor killings or slavery are not wrong in a general sense, many moral relativists may recoil from this.

Cultural moral relativism denies the possibility of meaningful moral debate or moral progress [1]. Moral relativism seems to require us to accept contradictory claims. For example, moral relativists might say that a supporter of gay marriage is correct in saying that homosexuality is morally acceptable, while someone from a different culture might be correct in saying that homosexuality is morally wrong, provided that both claims are in line with the moral standards of the cultures they respectively belong to. If moral relativism requires assert to simultaneously assert and deny that homosexuality is morally acceptable, and any theory that generates contradictions should be rejected, this would appear to mean that we should reject moral relativism. In order to resist this, moral relativists typically reinterpret the way we usual moral language in a way that can save it from contradiction. The relativist would say that when we say "homosexuality is wrong", what we really mean is "Homosexuality is not approved by my society's norms". This means that relativists have to deny the possibility of moral disagreement and claim that anyone who engages in such debates does not understand the meaning of what they are saying.

Moral relativism does not necessarily promote tolerance [1]. Some have argued that one of the attractions of moral relativism is that it promotes tolerance. By recognizing cultural differences (descriptive moral relativism), they may assert that everyone ought to do what their culture says is right (normative moral relativism). However, in a society that is deeply intolerant, cultural moral relativism cannot support tolerance, as it cannot claim that this has any universal or objective value.

Moral relativism only recommends tolerance to cultures where it is already accepted. Indeed, to be tolerant, one need not be a normative moral relativist. There are alternatives views which can accommodate tolerance and multiple perspectives, such as cosmopolitanism, liberal principles, and value pluralism. We discuss adjudicating among competing moral views in **??**.

In practice, moral relativism can shut down ethics discussions [1]. It is important to note that different cultures have different moral standards. However, AI developers sometimes invoke this observation and side with normative moral relativism to avoid considering the ethics of their AI design choices. Moreover, suppose AI developers do not analyze the ethical implications of their choices and avoid ethical discussions by noting the lack of cross-cultural consensus. In that case, the default is for AI development to be driven by amoral forces, such as self-interest or what makes the most sense in a competitive market. Decisions driven by other forces, such as commercial incentives, will not necessarily be aligned with the broader interests of society. Moral relativism can be unattractive from a pragmatic point of view, as it limits our ability to engage in discussions that may sometimes lead to convergence on shared principles. This quietist stance de-emphasizes moral arguments to the benefit of economic incentives and self-interest.

Why are these debates about moral relativism relevant to AI? People commonly observe that different cultures have different beliefs when discussing how to ensure that AIs promote human values. It is essential not to conflate this observation with normative moral relativism and conclude that AI developers have no ethical responsibilities. Instead, they are responsible for ensuring that the values embodied in their AI systems are beneficial. Rather than a barrier, cultural variation means that making AIs ethical requires a broad, globally representative approach.

C.2.2 Is Ethics Determined by Religion?

Moral relativists may believe that studying ethics is futile because ethical questions are irresolvable. On the other hand, some people believe that studying ethics is futile because moral questions are already solved. This position is most common among those who say that religion is the source of morality.

Divine Command Theory

Many believe morality depends on God's will and commands. The view called *divine command theory* says whether an action is moral is determined solely by God's commands rather than any qualities of the action or its consequences. (We use the term "God" inclusively to refer to the god or gods of any religion.) This theory suggests that God has the power to create moral obligations and can change them at will.

While this book does not argue for or against any particular religion, we do suggest that there are severe problems with equating religion and morality. One problem is that it creates a problematic understanding of God.

If you believe there is a god, you likely believe he is more than just an arbitrary authority figure. Many religious traditions view God as inherently good. It is precisely because God is good that religion compels us to follow God's word. However, if you believe that we should follow God's word because God is good, then there must be some moral qualities (like goodness) that exist independently of God's rules—thus, divine command theory is false [3].

To be clear, this is not an argument against believing in God or religion. It is an argument against equating God or faith with morality. Both religious people and irreligious people can behave morally or immorally. That's why everyone needs to understand the factors that might make our actions right or wrong. ref

C.3 Moral Considerations

How can we determine whether an action is right or wrong? What are the kinds of principles and values that should guide our moral decisions? There are many factors to consider. Here, we'll focus on a few—-goodness, constraints, special obligations, and options-—that very commonly enter into moral decision-making.

C.3.1 The "Goodness" of Actions and Their Consequences

Moral decision-making often involves considering the values, or "goods," that are at stake. These may be intrinsic goods or instrumental goods.

Intrinsic goods are things that are valuable for their own sake. Philosophers disagree about what, if anything, is intrinsically good, but many argue for the intrinsic value of things like happiness, love, and knowledge. We value such things simply because they are valuable—not because they necessarily lead to anything else.

Instrumental goods are things that are valuable because of the benefits they provide or the outcomes they achieve. We pursue instrumental goods as a means to an end, but not for their own sake. Money, power, and education are examples of instrumental goods. We value them because they can lead to other things we value, like security, influence, career opportunities, or intrinsic goods.

Intrinsically good things are not necessarily instrumentally good. Sometimes, intrinsically bad things can be instrumentally good and intrinsic goods can be instrumentally bad. For instance, many people believe that honesty is intrinsically good. However, it's easy to imagine cases in which honesty can lead to bad outcomes, like hurt feelings. Suppose a friend has confided in you that they are staying at a shelter to hide from an abusive partner. If that abusive partner asks you for your friend's location, you may think that that honesty is intrinsically good. However, revealing your friend's location would be instrumentally bad, as it may lead to further violence and perhaps even a risk to your friend's life. On the other hand, consider medical treatments like chemotherapy. Chemotherapy is instrumentally good because it can prolong cancer patients' lives. Yet, as it requires the administration of highly toxic drugs into a patient's body, it could be seen as harmful, or intrinsically bad. For many people, exercise is painful, and pain is intrinsically bad, but exercise can be instrumentally good.

There is no consensus about what is intrinsically good. Some philosophers believe that there are many intrinsic goods. Others believe there is only one value. One common view is that the only intrinsic good is wellbeing, and everything else is valuable only insofar as it promotes wellbeing.

Value pluralists believe that there are many intrinsic goods. These values may include justice, rights, autonomy, and virtues such as courage. Other philosophers believe there is only one fundamental value. Among these, one common view is that the only intrinsic good is wellbeing, and everything else is valuable only insofar as it promotes wellbeing.

C.3.2 Constraints and Special Obligations

We have covered the moral consideration of intrinsic goods, and focused on the intrinsic good wellbeing. Special obligations and constraints are key considerations when we make ethical decisions.

Special obligations are duties arising from relationships. We can incur special obligations when we promise someone to do something, take a professional position with responsibilities, have a child, make a romantic commitment to a partner, and so on. Sometimes we can have special obligations that we did not volunteer for—a child to its parents, or our duties to fellow citizens.

Constraints are actions that we are morally prohibited from taking. A constraint is something that places limits on our actions. For example, many people think we're morally prohibited from lying, stealing, cheating, harming others, and more.

Constraints often come in the form of rights. Rights are claims that individuals may have over their community. For instance, many people believe that humans have the rights to life, freedom, privacy, and so on. Some people argue that any individual with the capacity for experiencing pleasure and pain has rights. Non-human individuals (including animals and AI systems) might also have certain rights.

An individual's rights may require that society intervene in certain ways to ensure that those rights are fulfilled. For instance, an individual's right to food, shelter, or education may require the rest of society to pay taxes so that the government can ensure that everyone's rights are fulfilled. Rights that require certain actions from others are called positive rights.

Other rights may require that society abstain from certain actions. For instance, an individual's right to free speech, privacy, or freedom from discrimination may require the rest of society to refrain from censorship, spying, and discriminating. Rights that require others to abstain from certain behaviors are called negative rights.

Many AI researchers think that, for now, we should avoid accidentally creating AIs that deserve rights [4]; for instance, perhaps all entities that can experience suffering have natural rights to protect them from it. Some think we should especially avoid giving them positive rights; it might be fine to give them rights against being tortured but not the right to vote. If they come to deserve rights, this would create many complications and undermine our claim to control.

C.3.3 What does it mean for an action to be right or wrong?

Some of the first questions we might ask about ethics are: Are all actions either right or wrong? Are some simply neutral? Are there other distinctions we might want to draw between the morality of different actions?

The answers to these questions, like most moral questions, are the subject of much debate. Here, we will simply examine what it might mean for an action to be right or wrong. We will also draw some other useful distinctions, like the distinction between obligatory and non-obligatory actions, and between permissible and impermissible actions. These distinctions will be useful in the following section, when we discuss the considerations that inform our moral judgments.

Options

Special obligations and constraints tell us what we should not do, and sometimes, what we must do. Intrinsic goods tell us about things that would be good, should they happen. But philosophers debate how much good we are required to do.

Options are moral actions which we are neither required to do nor forbidden from doing. Even though it would be good to donate money, many people do not think people are morally required to donate. This is an ethical option. If we believe in options, not all actions are either required or forbidden.

We now break down actions onto a spectrum on which we will simply examine what it might mean for an action to be right or wrong. We will also draw some other useful distinctions, like the distinction between obligatory and non-obligatory actions and between permissible and impermissible actions.

Obligatory actions are those that we are morally obligated or required to perform. We have a moral duty or obligation to carry out obligatory actions, based on ethical principles. For example, it is generally considered obligatory to help someone in distress, or refrain from hurting others.

Non-obligatory actions are actions that are not morally required or necessary. Non-obligatory actions may still be morally good, but they are not considered to be obligatory. For example, volunteering at a charity organization or donating to a good cause may be good, but most people don't consider them to be obligatory.

Permissible actions may be morally good or simply neutral (i.e. not good or bad). In general, any action that is not impermissible is permissible. Moral obligations, of course, are permissible. We can consider four other actions: volunteering, donating to charity, eating a sandwich, and taking a walk. These seem permissible, and can be classified into two categories.

One class of permissible actions is called *supererogatory actions*. These may include volunteering or giving to charity. They are generally considered good; in fact, we tend to believe that the people who do them deserve praise. On the other hand, we typically don't consider the failure to do these actions to be bad. We might think of supererogatory actions as those that are morally good, but optional; they go "above and beyond" what is morally required.

Another class of permissible actions is called *morally neutral actions*. These may include seemingly inconsequential activities like eating a sandwich or taking a walk. Most people probably believe that actions like these are neither right nor wrong.

Impermissible actions are those that are morally prohibited or unacceptable. These actions violate moral laws or principles and are considered wrong. Stealing or attacking someone are generally considered to be impermissible actions.

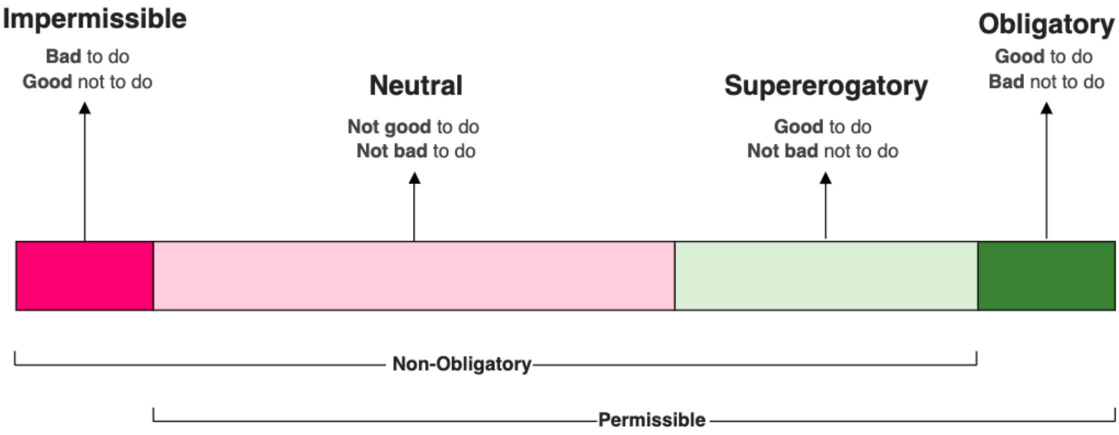

Figure C.1: Actions can be classified according to whether they are permissible and obligatory.

Some philosophers believe that all actions fit on a scale like the one above. At one end of the scale are impermissible actions, like murder, theft, or exploitation. At the other end are obligatory actions, like honesty, respect, and not harming others. In between are neutral and supererogatory actions. These are neither impermissible nor obligatory. Many people believe that the vast majority of our actions fall into these two categories. Crucially, in designing ethical AI systems that operate in the real world, it is important to determine which actions are obligatory and which actions are impermissible.

However, some philosophers do not believe in options; rather that actions are all on a spectrum from the least moral to the most moral. We will learn more about these positions, and others, when we discuss moral theories later in this chapter.

From Considerations to Theories

Moral considerations can guide our day-to-day decision making. Understanding which factors are morally relevant can help us think more clearly about what we should do. Of course, we don't always stop to consider every factor before making a decision. Rather, we tend to draw broader conclusions or moral principles based on our evaluations of specific cases. For instance, once we consider a few examples of the ways in which stealing can harm others, we might draw the conclusion that we shouldn't steal.

The considerations discussed in this section provide a basis on which we can develop more practical, action-guiding theories about how we should behave. The types of fundamental considerations in this section comprise a subfield of ethics called *metaethics*. Metaethics is the consideration of questions like "What makes an action right or wrong?" and "What does it mean to say that an action is right or wrong?" [5]

These considerations are important in the context of designing AI systems. In order to respond to situations in an appropriate way, AI systems need to be able to identify morally relevant features and detect situations where certain moral principles apply. They would also need to be able to evaluate and compare the moral worth of potential actions, taking into account various purported intrinsic goods as well as normative factors such as special obligations and constraints. The challenges of designing objectives for AI systems that respect moral principles are further discussed in the Machine Ethics chapter.

In the following section, we will discuss some popular moral theories.

C.4 Moral Theories

Moral theories are systematic attempts to provide a general account of moral principles that apply universally. Good moral theories should provide a coherent, consistent framework for determining whether an action is right or wrong. A basic background understanding of some of the most commonly advanced moral theories provides a useful foundation for thinking about the kinds of goals or ideals that we wish AI systems to promote. Without this background, there is a risk that developers and users of AI systems may jump to conclusions about these topics with a false sense of certainty and without considering many potential considerations that could change their decisions. Considering a range of different philosophical theories enables us to stress-test our arguments more thoroughly and surface questionable assumptions that may not have been noticed otherwise. It would be highly inefficient for those developing AI systems or trying to make them safer to attempt to re-invent moral systems, without learning from the large existing body of philosophical work on these topics.

There are many different types of moral theories, each of which emphasizes different moral values and considerations. Consequentialist theories like utilitarianism hold that the morality of an action is determined by its consequences or outcomes. Utilitarianism places an emphasis on maximizing everyone's wellbeing. Deontological theories like Kantian ethics hold that the morality of an action is determined by whether it conforms to universal moral rules or principles. Deontology places an emphasis on rights, special obligations, and constraints.

Below, we explore the most common modern moral theories: *utilitarianism*, *deontology*, *virtue ethics*, and *social contract theory.*

C.4.1 Utilitarianism

Utilitarianism is the view that we should do whatever results in the most overall wellbeing [6]. According to Katarzyna de Lazari-Radek and Peter Singer, "The core precept of Utilitarianism is that we should make the world the best place we can. That means that, as far as it is within our power, we should bring about a world in which every individual has the highest possible level of wellbeing" [7]. Under Utilitarianism, the right action in any situation is the one which will increase overall wellbeing the most, not just for the people directly involved in the situation but globally.

Expected Utility

Utilitarianism enables us to use empirical, quantitative evidence when deciding moral questions. As we discussed in Section 6.4, there is no consensus about what, precisely, wellbeing is. However, if we discover that wellbeing is something measurable, like happiness, moral decision-making could take advantage of calculation and would rely less on qualitative argumentation. To determine what action is morally right, we would simply consider the available options. We might run some tests or perform data analysis to determine which action would create the most happiness, and that action would be the right one. Consider the following example:

> *Drunk driving*: Amanda has had a few alcoholic drinks and is deciding whether to drive or take the bus home. Which should she choose?

A utilitarian could analyze this scenario by listing the possible outcomes of each choice and determining their impact on overall wellbeing. We call an action's impact on wellbeing its *utility.* If an action has *positive utility*, it will cause happiness. If an action has *negative utility*, it will cause suffering. Larger amounts of positive utility represent larger amounts of happiness, and larger amounts of negative utility represent larger amounts of suffering. Since no one can predict the future, the utilitarian should also consider the probability that each potential outcome would occur.

A simplified, informal, back-of-the-envelope version of this utilitarian calculation is below:

Amanda's action	Possible outcome(s)	Probability of each outcome	Utility
Amanda takes the bus.	Amanda is frustrated, the bus is slow, and she has to wait in the cold.	100%	-1
Amanda drives home.	Amanda gets home safely, far sooner than she would have on the bus.	95%	+1
	Amanda gets into an accident and someone is fatally injured.	5%	-1000

Table C.1: Illustrative calculation of utility from Amanda's possible actions.

	Utilitarianism	Deontology	Contractarianism
What is Alex's estimate of the chance this theory is true	60%	30%	10%
Does this theory like lying to save a life?	Yes	No	Yes

Table C.2: Example: Alex's credence in various theories and their evaluation of lying to save a life.

We are interested in the *expected utility* of each action—the amount of wellbeing that each action is likely to result in. To calculate the expected utility, we multiply the utility of each possible outcome by the probability of that outcome occurring.

Amanda choosing to take the bus has a 100% chance (a certainty) of causing a small decrease in utility; she will be slightly inconvenienced. Since the change in utility is small and negative, we'll estimate a small negative number to represent it, like -1. *The expected utility of Amanda taking the bus is 100% × -1, or simply -1.*

If Amanda drives home, there is a 95% chance that she will get home safely and create a small increase in utility——let's say of +1. However, there's also a 5% chance she could cause an accident and end someone's life. The accident would result in a very large decrease in utility. Someone would experience pain and death, Amanda would feel guilty for the rest of her life, and the victim's friends and family would experience loss and grief. We might estimate that the potential loss in utility is -1000. That's 1000× worse than the small increase in utility if Amanda gets home safely. *The expected utility of Amanda driving home is the sum of both possibilities:* $.95 \times 1 + .05 \times -1000$, *or* -49.05.

Both of Amanda's options are expected to yield negative utility, but the utilitarian would say that she should choose the better of the two options. Unsurprisingly, Amanda should take the bus.

Implications of Utilitarianism

Utilitarianism may sometimes yield results that run against commonly held beliefs. Utilitarianism aims at producing the most wellbeing and insists that this is the only thing that matters. However, many of the moral values that we have inherited conflict with this goal. Utilitarianism can be seen as having less of a bias to defend the moral status quo relative to some other moral theories such as deontology or virtue ethics. This either makes Utilitarianism exciting or threatening.

Utilitarianism can lead to some radical moral claims. Utilitarianism's sole focus on wellbeing can lead it to promote what have been or are viewed as radical actions. For example, the founder of utilitarianism, Bentham, argued to decriminalize homosexuality, and contemporary utilitarians have argued we have a much greater obligation to give to charity than most of us seem to believe.

Bentham held many beliefs that were ahead of his time. Written in 1785, in a social and legal environment very hostile to homosexuality, Bentham's essay "Offences against oneself" rebuts the arguments that legal scholars had used to justify laws against homosexuality [8].

Today, many utilitarians believe that we should prioritize helping people in low-income countries. Utilitarianism continues to make recommendations that today's society finds controversial. Consider the following example:

> On her morning walk through the park, Carla sees a child drowning in the pond. She is wearing a new suit that she bought the day before, worth $3,500. Should she dive in to save the child, even though she would destroy her suit? [9]

The philosopher Peter Singer, who first posed this question, argues that Carla should dive in. Furthermore, he argues that our judgment in this case might mean that we should re-evaluate our obligation to donate to charity. There are charities that will save a child's life for around $3,500. If we should forgo that amount in order to save a child who is right in front of us, shouldn't we do the same for children across the world? Singer argues that distance is not relevant to our moral obligations. If we have an obligation to a child in front of us, we have the same obligation to similar children who may be far away.

To maximize global wellbeing, Singer says that we should give our money up until the point where a dollar would be better spent on us than on charity. If our money helps others more than it can help ourselves, there isn't a utilitarian reason to keep it. For an adult making, say, $50,000 per year, an extra $3,500 would be helpful, but is not critical to their wellbeing. However, for someone making less than $3 per day in a low-income country, $3,500 would be life-changing—not just for one recipient, but for that person's entire family and community. Singer argues that, if giving money away can significantly help someone else, and if giving it away would not be a significant sacrifice, we should give the money to the person who needs it most.

These conclusions imply that most of us (especially those of us in high-income countries) should live very different lives. We should, for the most part, live as inexpensively as possible and donate a significant portion of our income to people in lower-income communities.

Utilitarianism's Central Claims

Utilitarianism can be distinguished from other ethical theories by four central claims.

Claim one: Consequences (and only consequences) determine whether an action is right or wrong.

Utilitarianism is a form of consequentialism. Any theory that claims that the consequences of an action alone determine whether an action is right or wrong is *consequentialist*. Other theories, as we will discuss later in this chapter, claim that some actions are right or wrong regardless of their consequences.

Claim two: Wellbeing is the only intrinsic good.

Utilitarians believe that the only type of consequences that make an action right or wrong are those that affect happiness or wellbeing. In that sense, utilitarianism can be understood as a combination of consequentialism and hedonism, as we discussed it in section Wellbeing. Recall that there are several different accounts of wellbeing, all of which are compatible with utilitarianism.

Classical utilitarianism. Most utilitarians are hedonists about wellbeing; they believe that wellbeing is a function of pleasure and suffering. Such utilitarians classical utilitarians. When classical utilitarians say they want to improve wellbeing, they mean that they want there to be more pleasure and less suffering in the world.

Preference utilitarianism. In contrast to classical utilitarians, preference utilitarians believe that wellbeing is constituted by the satisfaction of people's preferences.

The preference account of wellbeing is one of the many modifications of classical utilitarianism. While we will not describe these other theories in detail, it is useful to know that if we disagree with one aspect of classical utilitarianism, there is often another utilitarian or consequentialist theory that can accommodate our beliefs.

Claim three: Everyone's wellbeing should be weighed impartially.

Utilitarians believe that people have the same intrinsic moral worth. Bentham exemplified utilitarian thought with the phrase "Each to count for one and none for more than one." People of different classes, races, ethnicities, religions, abilities, and so on are of equal moral worth. In other words, utilitarianism is an *impartial* moral theory.

For an individual to deserve moral treatment, they just need to be capable of having wellbeing. According to Bentham, "The question is not, Can they reason?, nor Can they talk? but, Can they suffer?" This quote is often taken to mean that we should be concerned with the wellbeing of animals, since animals feel pleasure and pain just like humans. Similar positions are held by other utilitarians such as Peter Singer [10]. If, in the future, AI systems develop a capacity for wellbeing, they would deserve moral treatment as well according to classical utilitarians.

Claim four: We should maximize wellbeing.

Utilitarians aim to maximize wellbeing. Utilitarians do not think it is sufficient to perform an action with good consequences; they think the only right action is the one with the best consequences. They do not believe in options. The following example illustrates this distinction.

> Dorian has a choice: teach biology or research air quality. As a teacher, he would help hundreds of students. As a researcher, he would save thousands of lives. He enjoys teaching somewhat more than research. What should he choose?

A utilitarian might argue that Dorian should become a researcher. In this case, he knows that he will do more good. This is despite the fact that Dorian would be a great teacher, and would have a positive impact as a teacher. He would do more good through his job as a public health researcher, so a utilitarian might argue that he is obligated to take that option.

The best option is always the one that maximizes wellbeing. This is a straightforward result of valuing everyone's wellbeing impartially and always striving to do the best rather than the merely good.

In summary, utilitarianism makes several claims: wellbeing is the only intrinsic good, wellbeing should be maximized, wellbeing should be weighed impartially, and an action's moral value is determined by its consequent effects on wellbeing. Utilitarianism teaches that the best action we can take is the one that leads to the best positive effect on wellbeing.

Common Criticisms of Utilitarianism

While utilitarianism remains a popular moral theory, it is not without its critics. This section explains some of the most common objections to utilitarianism.

Criticism: "Utilitarianism is too demanding."

Many philosophers argue that utilitarianism is too demanding [11]. It insists that we choose the best actions, rather than merely good ones. As we saw in our discussion of the drowning child and our obligations to the global poor, this can lead utilitarianism to recommend unconventionally large commitments.

According to this criticism, utilitarianism asks us to give up too much of what we take to be valuable for the sake of other people's wellbeing. Perhaps we should quit a career that we love in order to work on something that does more good, or we should not buy gifts for family and friends if the money would produce more wellbeing when given to someone suffering from a preventable disease. To live up to this critique of everyday values we would have to radically change our lives, and continue to change them as the global situation evolved. The critic thinks that this is too much to reasonably ask of someone. A moral theory, they think, should not make a moral life highly challenging.

A utilitarian can respond in two ways. The first way is to argue that, while utilitarianism is theoretically demanding, it is practically less so. For example, someone trying to live up to the theoretical demands of utilitarianism might burn out, or harm the people around them with their indifference. If they had asked less of themselves, they might have done more good in the long run. Utilitarianism might even recommend acting almost normally, if acting almost normally is the best way to maximize wellbeing.

However, it is unlikely that this response undermines the argument that we should give some portion of our money to charity. Even if donating most of our income would backfire. most people should likely donate more than they do. Many utilitarians simply accept that their theory is demanding. Utilitarianism does demand a lot of us, and until the critic shows that these demands are not morally required of us, then we might just live in a demanding world. While demanding too much of yourself can be counter-productive, we should do far more than we currently do.

Criticism: "Utilitarianism requires intractable calculations."

Another way of critiquing utilitarianism is to say that even if the theory is consistent and appealing, it isn't useful because we rarely know the consequences of our actions in advance.

When we illustrated a utilitarian calculation above using the case of drunk driving, we intentionally simplified the situation. We considered only a few possible immediate outcomes and we estimated their possible likelihoods. In the real world, however, someone considering whether to drive home faces unlimited possible outcomes, and those outcomes could cause other events in the future that would be impossible to predict. Moreover, we rarely know the probabilities of the effects of our actions. Utilitarianism would be impractical if it required us to make a long series of predictions and calculations for every choice we face. Certainly, we shouldn't expect Amanda to do so in the moment.

In response to this criticism, a utilitarian might differentiate between a criterion of rightness and a decision-making procedure [12]. A *criterion of rightness* is the factor that determines whether actions are right or wrong. According to utilitarianism, the criterion of rightness is whether an action maximizes expected wellbeing compared to its alternatives. In contrast, a theory's *decision-making procedure* is the process it recommends individuals use to make decisions. Crucially, a theory's decision procedure does not need to be the same as its criterion of rightness.

For example, a utilitarian would not likely advise everyone to make detailed calculations before getting in the car after having a couple of drinks. Most utilitarians would advise everyone to simply never drive drunk. There's only a need to consider a utility calculation in cases where the best option is particularly unclear. Even then, such calculations are only approximate and should not necessarily be decisive. Just as corporations try to maximize profit without consulting a spreadsheet for every decision, utilitarians might follow certain rules of thumb without relying on utility calculations.

In practice, utilitarians rely on robust heuristics for bringing about better consequences and rarely consult explicit calculations. To better improve the world, like others they often cultivate virtues such as truth-telling, being polite, being fair, and so on. They often imitate practices that have stood the test of time, even if they do not fully understand their rationale. That is because some things may have complex or obscure reasons that are not easily discerned by human reason or not easily amenable to calculation. They often bear in mind Chesterton's fence, which warns against removing a barrier without knowing why it was erected in the first place. Even if their criterion of rightness can be controversial, utilitarians adopt decision procedures that are often conventional.

Criticism: "Utilitarianism places too much value on wellbeing."

Many philosophers argue that utilitarianism neglects sources of value other than wellbeing. One famous argument meant to show that wellbeing isn't the only source of value is Robert Nozick's "Experience Machine" [13]. Nozick considers the following thought experiment:

> "Suppose there were an experience machine that would give you any experience you desired. Superduper neuropsychologists could stimulate your brain so that you would think and feel you were writing a great novel, or making a friend, or reading an interesting book. All the time you would be floating in a tank, with electrodes attached to your brain. Should you plug into this machine for life, preprogramming your life's experiences?"

Nozick claims that we would decline this offer because we care about the reality of our actions. We do not just want to feel that we have cheered up our friend; we actually want them to feel better. We do not want the experience of writing a great work of literature, we want great literature to exist because we worked on it. Many philosophers consider this a decisive rebuttal to the idea that wellbeing is the only thing that matters.

Though many people say that they would prefer not to use the machine when it is introduced as above, they may have a different reaction when the thought experiment is presented differently.

> "You wake up in a plain white room. You are seated in a reclining chair with a steel contraption on your head. A woman in a white coat is standing over you. 'The year is 2659,' she explains, 'The life with which you are familiar is an experience machine program selected by you some forty years ago. We at IEM interrupt our clients' programs at ten-year intervals to ensure client satisfaction. Our records indicate that at your three previous interruptions you deemed your program satisfactory and chose to continue. As before, if you choose to continue with your program you will return to your life as you know it with no recollection of this interruption. Your friends, loved ones, and projects will all be there. Of course, you may choose to terminate your program at this point if you are unsatisfied for any reason. Do you intend to continue with your program?" [14]

Joshua Greene, the author of this example, supposes that most people would not want to leave the program. He suggests that what accounts for the seeming difference between his and Nozick's versions is the *status-quo bias*. People tend to prefer the life they know. Surveys of real people's responses to these thought experiments indicate that a range of factors—including the status-quo bias—affect their responses. Nozick's example is not as clear cut as his argument supposes.

In summary, utilitarianism is often criticized in three ways. People claim that it is (1) too ethically demanding, (2) practically unusable, and (3) wrong to neglect values other than wellbeing. In response, utilitarians may argue that we simply live in a demanding world, that we can use heuristics instead of constantly making calculations, and that Nozick's thought experiment does not necessarily show that we have values aside from wellbeing.

Conclusions about Utilitarianism

Utilitarianism is a consequentialist, welfarist, impartial, and optimizing ethical theory. Utilitarians believe that our actions should be guided by whatever leads to the greatest overall wellbeing. As a form of consequentialism, utilitarianism emphasizes that the outcomes of our actions are what truly matter. It is impartial and welfarist, which means that it considers everyone's wellbeing equally important when evaluating outcomes, and considers nothing except for wellbeing. Consequently, utilitarianism transforms moral questions into empirical ones: determining what is right simply involves identifying the action that maximizes total wellbeing.

Utilitarians often advocate for policies that appear radical to their contemporaries. For example, Jeremy Bentham advocated for gay rights, John Stuart Mill for women's rights, and Peter Singer for animal rights before these were widely acceptable. However, critics of utilitarianism argue that some of these radical claims make it too demanding. They also contend that utilitarianism is overly focused on wellbeing and presents challenges that are difficult to address. Despite these criticisms, utilitarianism provides a clear framework for determining the morality of actions.

We have already explored several concepts that are helpful for creating utilitarian AIs. A utilitarian might want to create AIs that act in a utilitarian manner, maximizing total utility. We have explored concrete ways to begin doing so. If we wish to create preference utilitarian AIs, we might model individual preferences using utility functions. Instead, we might think other conceptions of wellbeing like hedonism are more accurate; if so, we might estimate general-purpose wellbeing functions instead. AIs based on different theories of wellbeing prefer different outcomes.

Once we have a good representation of individual wellbeing, we must aggregate these to create a utilitarian social welfare function. However, this might not be as easy as it sounds. For one, we need to decide whose wellbeing is relevant to every decision. Evaluating the effect of an action on everyone's wellbeing can be difficult in highly complex, socially connected worlds. These approximations need to hold across different time horizons as well, not just short ones.

Creating utilitarian AIs requires overcoming several practical challenges. In principle, creating a utilitarian AI is straightforward. In reality, this requires addressing several concrete challenges. For instance, creating utility estimates requires not only an accurate understanding of general human preferences but also the ability to adapt these estimates to particular individuals, since everyone has their own idiosyncracies. Decision-making must account for balancing short-term pleasures and long-term wellbeing. Most actions affect a large number of other people, often indirectly; as a result, a utilitarian AI must consider a broad range of stakeholders for every decision.

Utilitarian decision-making can require solving intractable problems. AIs will need to have a good predictive model of the world, with the ability to forecast the effects of their actions on large numbers of people. This is especially difficult for important decisions that consider complex systems or long time horizons. AIs should not resort to short-term optimizations that could have long-lasting negative consequences, such as pushing students to enjoy extra leisure time instead of studying to reap benefits later. The possibility of tail risks and black swans make utilitarian decision-making, which relies on expected value, even harder to practice.

C.4.2 Deontology

***Deontology* is the name for a family of ethical theories that deny that the rightness of actions is solely determined by their consequences.** Deontologists emphasize constraints rather than consequences. "Thou shalt not kill", "thou shalt not steal", "honor thy mother and father"—these are deontological principles that may be familiar from the Ten Commandments. Deontological theories are systems of rules or obligations that constrain moral behavior [15]. They may be based on a theological justification, but they do not need to be. These theories are often based on simple and unambiguous rules, which may make them easier than other theories to implement in AIs.

The term *deontology* encompasses religious ethical theories, non-religious ethical theories, and principles and rules that are not part of theories at all. Some deontological theories are religious. For example, *divine command theory* teaches that we have a duty to do as God commands. Others, like Kant's ethics (which we will discuss later), are non-religious. While deontological theories may derive their rules from different sources, they are united by their focus on duties and rights.

Many deontological principles are not tied to any particular theory. Instead, they may be an attempt to find rules or principles which fit our intuitions about specific moral issues like abortion, terrorism, or suicide. This kind of moral analysis is still deontological—it is looking for universal rules which can tell us what to do in particular cases—even though it is not tied to a specific theory. Most of what we will say about deontology in this section is applicable to deontological theories.

Features of Deontological Theories

Deontological theories give obligations and constraints priority over consequences. Unlike consequentialism, deontological theories do not justify their rules by appealing to their consequences. Under deontological theories, some actions (like lying or killing) are simply wrong, and they cannot be justified by the good consequences that they might bring about. For example, when

Elena's boss asks her how hard her co-workers work, a deontologist might argue that she should tell the truth, even if she knows the truth will lead to some of her colleagues losing their jobs.

Many constraints on our actions are derived from a respect for other people's rights. On most accounts, every person has certain rights simply by virtue of being a person. Each individual has a claim to them, and no one is permitted to violate them under any circumstances. Common examples of rights include the right to life, the right to freedom, and the right to autonomy.

Modern deontological theories tend to emphasize that we should not interfere with others' autonomy. Since Kant developed his moral theory, many deontologists have followed him in placing importance on human *autonomy*, our ability to freely choose how we act. This means protecting each other from acts which might restrict our autonomy. This is an example where different moral theories place emphases on normative factors: whereas utilitarianism focuses on wellbeing, deontological theories often emphasize autonomy.

Another key part of their idea of autonomy is that our actions are not entirely governed by moral considerations. We are constrained from certain types of behavior, but apart from those behaviors, we can freely choose how to act. When we are choosing a career for example, we should not become a torturer, or an assassin, but apart from those types of constraints, we can choose from a range of harmless careers that suit our interests. By contrast, a utilitarian might argue that we must choose the single career (if there is one) that would have the best outcome.

According to deontology, intentions can be right or wrong, as well as actions. Many deontological theories assert that intentions, as well as actions, can be moral or immoral. For example:

> *Intentional push*: Farid's elderly mother has been annoying him, so he decides to push her down the stairs. When he next sees her at the top of the stairs, he carries out his plan.
> *Intention, but no push*: Farid's elderly mother has been annoying him, so he decides to push her down the stairs. When he next sees her at the top of the stairs, he plans to push her. Just before he does, she trips and falls.

According to some deontological theories, Farid is equally wrong in both scenarios. Though he never pushes his mother in the second scenario, the intention itself is a moral error. By contrast, for classical utilitarians, intentions do not matter in themselves. On this view, intentions are only right or wrong insofar as they lead to good or bad consequences.

Our common-sense moral intuitions are often aligned with the idea that intentions matter. Perhaps that's why they play a very important role in the law of many countries. In America's Model Penal Code, for example, the strength of a criminal's intention is measured with a scale of four adverbs: a criminal could commit a crime *purposefully*, *knowingly*, *recklessly*, or *negligently.* If they commit it *purposefully*, then they knew what the outcome would be, and they intended that outcome to happen. If they commit it *knowingly*, they don't primarily intend the criminal effect of their actions, but they act anyway. People are acting *recklessly* when they knowingly engage in behaviors which pose risks to others, like owning a tiger, or flying a drone too close to an airport. Those who act *negligently* fail to perceive a substantial and unjustifiable risk that their conduct will have harmful results. Not all legal codes differentiate between these four levels in practice, but they all punish purposeful or knowing criminals more than negligent ones.

If a driver purposefully rams their car into another vehicle, intending to cause injury, they would face serious criminal charges. If the driver was speeding recklessly and lost control of their vehicle, resulting in accidental injury to another driver, they may face a lesser charge, such as reckless driving. The lessened mental state, even if the end result was the same, often reduces the severity of punishment under the law. As we can see, to determine whether an AI acted immorally, some moral theories would require that we be able to determine an AI's intent, a goal of transparency research.

Deontologists advance a number of arguments against consequentialism that help to clarify the distinctive features of their moral theories. There are many deontological theories, and they are often grouped together under one umbrella simply because they share the feature that they are not consequentialism. Deontological theorists often criticize consequentialism, and partially define their theories by the ways that they make up for the flaws they see in consequentialism. Two problems that they see in consequentialism are that (1) consequentialism is very demanding (in other words, it doesn't allow much autonomy) and (2) consequentialism leads to some radical conclusions (for example, it sometimes justifies actions that most people believe are wrong).

(1) Unlike consequentialism, deontology gives us options. Deontology preserves human autonomy because it only forbids us from performing a limited number of impermissible actions. The remaining actions are optional. On the other hand, consequentialism implies that every action is moral or immoral to a certain degree, and each person is obligated to do the most good at all times. According to consequentialism, every life decision—like marriage, career choice, which relationships to pursue, which food to eat—is a moral decision. Deontologists find this to be far too demanding.

(2) Unlike consequentialism, deontology does not allow any action to be justified by its outcome. According to many consequentialist theories, any action can be justified if it results in a better outcome than the alternative actions. Even killing or torturing innocent people might be the right choice if it increases everyone else's wellbeing more than it harms its victims. Some people believe that deontological theories—which forbid actions like killing and torture in all cases—are more plausible.

Deontological Principles

We need principles, as well as rules, to capture the complexity of ethics. While deontologists generally consider the absolute prohibition of certain actions to be a strength of deontology, it can sometimes lead to counterintuitive moral judgments. Suppose a pair of conjoined twins will die without undergoing a medical procedure. However, if the procedure is carried out, one of them will die. The rule "do not kill" might stop a surgeon from operating, which would mean that neither twin has a chance of survival.

Difficult, messy situations like this, where clear-cut rules seem to fail us, have led some deontologists to develop important distinctions which complicate their theory but better capture the way we think about ethics. We will introduce two of these principles: the *doctrine of double effect* and the *action/omission distinction.*

The doctrine of double effect. In the case of the conjoined twins, the surgeon might appeal to the doctrine of double effect. This is a principle which states that an agent is morally allowed to carry out actions that predictably lead to bad outcomes——like the death of an innocent——as long as they *intend* the good effect, but not the bad effect of the action. The constraint against letting two people die, however, must be stronger than the constraint against performing the surgery. In this case, the doctor can operate only if she intends to save one of them, not to bring about a death. The doctrine of double effect can also explain why it can be morally permissible to kill in self defense [16]. You are permitted to defend yourself and your family from imminent attack. If the only way to protect yourself is to kill your assailant, you may be permitted, as long as you intend to save your family and not to kill.

The action/omission distinction. The action/omission distinction is important for understanding the role of responsibility in deontological theories. Intuitively, we find someone to be more responsible for something they did than for something they allowed to happen [17]. For example:

> *Stealing*: While walking past the bank at night, Gabe sees that the night deposit box is open. Inside it, he sees a bag filled with money. He decides to steal the bag.
> *Failing to report*: Heather sees Gabe take the money, but doesn't report it to the police. If she had, the money would have been returned to its owner.

In these examples, both Gabe and Heather have done something wrong, but Gabe's crime is worse. He is directly responsible for the theft, while Heather has only committed an act of omission——she failed to report the crime. This captures some of our ordinary intuitions about responsibility. We generally do not hold people responsible for what they do not do.

Criticisms of Deontology

Criticism: Some say that deontology responds unconvincingly to moral catastrophes.

> *Nuclear terrorism*: Ines is part of a terrorist cell that has placed a nuclear weapon in a capital city. If it goes off, it will kill millions. She claims that it will go off within 24 hours, but she will not say which city it is in. Jamie has captured Ines and questioned her, but Ines will not give away the bomb's location. If Jamie tortures Ines, she will find out the bomb's location and millions of lives will be saved.

Some deontological theories accept that Jamie should not torture Ines, no matter how many people will die as a result. However, many people find this implausible. When so many lives are at stake, it may seem selfish for Jamie to take the easier option, prioritizing her moral purity over the lives of the people she could save.

To accommodate our intuitions about moral catastrophes, some deontological theorists have adopted a *threshold* [18]. According to a threshold view, when a certain number of lives are at stake (i.e. when the badness of an outcome reaches a certain threshold), the theory defers to the consequentialist recommendation that the otherwise impermissible act (torture in this case) is allowed.

There are a number of problems with the threshold view. Most importantly, it is not clear how to determine what the threshold should be, and any decision about it will be arbitrary. If a million lives at stake justify Jamie's act of torture, then it seems odd to say that ten thousand, one thousand, or even ten lives at stake do not.

Criticism: Deontology sometimes requires us to make the world worse.

> *Better job*: Kimiko has been offered a job where she will use her unique set of skills to reform the health system of the country she lives in. The next best hire for the job is far less experienced than she is. Therefore, if Kimiko takes the job, thousands of lives will be spared every year for at least a decade. However, the job would require Kimiko to move to a new city, and she promised her children that the family would not move until they had all finished school.

Many deontological theories consider promise-keeping to be very important. They would not allow Kimiko to break her promise to her children in this situation, even though keeping her promise would cost thousands of lives.

Deontological rules may sometimes lead to the best consequences, but they often do not. This leads to what some people call the *paradox of deontology.* There may be situations in which it is impermissible to stop many instances of the same impermissible act occurring. If there is a rule that we cannot lie and kill, then we cannot lie to prevent hundreds of acts of lying, or kill to prevent hundreds of acts of killing.

Immanuel Kant and the Categorical Imperative

Immanuel Kant was a German enlightenment thinker who developed an especially strong deontological theory which we now call *Kantianism.* According to Kant, some actions are absolutely, universally wrong. For instance, Kant believed that killing, stealing, lying, committing suicide, and breaking a promise are wrong in all circumstances.

Kant believed that anyone can (in theory) arrive at these conclusions due to their own capacity to reason. Because, in his opinion, we would all arrive at the same conclusions, he believed in a universal moral law which we all have a duty to follow. The method that he thought would help us discover this moral law is called the *categorical imperative.*

Kant described the categorical imperative in several different ways, and his descriptions are different enough that they are now referred to as separate formulations of the imperative [19]. In other words, they are different ways to discover the same moral law. We will focus on two formulations: the *universal law* formulation, and the *humanity* formulation. The universal law formulation asks each of us to imagine that when we make a moral rule for ourselves to follow, we have actually made a law for everyone. In other words, we need to ask ourselves: what if everyone did that? If a world where everyone followed our rule was contradictory or irrational, then we have discovered something that we shouldn't do. The humanity formulation tells us to act on rules which lead us to treat people in a way that lets them maintain their autonomy. This means never getting in the way of their ability to exercise their human capacities for reason and autonomy, or make their own decisions.

Kant's method is called the categorical imperative because he believed all moral rules must be categorical. Kant distinguishes two types of rules: *hypothetical* rules are of the form "do X in order to Y," which only apply when we already want to Y, and *categorical* rules are of the form "do X." An example of a hypothetical rule is "be kind to people if you want them to do you favors." This is hypothetical, or conditional, because we would only be required to follow this rule if we already cared about receiving favors. Kant thought that the moral law was a universal list of rules which apply to everyone, so he argued that all moral rules must be categorical, not hypothetical. A categorical rule like "be kind" can apply to everyone, while the hypothetical "be kind if you want them to do you favors" only applies to those who want favors.

The universal law formulation. This idea leads us to the first formulation of the categorical imperative, Kant's method for discerning right from wrong. Kant tells us in this formulation that we should only act in ways which could be made into laws for all of humankind. In other words, what if everyone did that same thing: "What if everyone killed people who stood in their way?" "What if everyone cheated on exams?" "What if everyone lied?" If the answer to the question seems to make your intended action impossible or inadvisable, do not do it. This formulation provides the clearest test to show whether an action is in accordance with the moral law or not.

In slightly more complicated terms, we can formalize Kant's thoughts on universalising rules into a four-part test for any rule which we can apply to any categorical rule we might think of. If our proposed rule passes all four stages, then it is permissible. First, we turn our proposed action into a categorical rule ("do X" or "do X when in Situation S"). Then, we change the rule so that it applies to everyone, not just us. To test whether the rule is part of the moral law, we first check whether it contradicts itself. If it does, then it is a rule we absolutely must not follow. If the rule isn't contradictory, we check whether it conflicts with something else that we must value. If it does, then we shouldn't follow the rule; if it doesn't, then we must follow it. We will now go through an example of each step, to model how this process might work.

Step 1: Turn the proposed action into a rule.

Luke promised to take his mother's dog for a walk. But today he is tired and doesn't want to. He proposes not to go on the walk. If his action became a rule, the rule might be: "I will not fulfill promises when it is inconvenient to me."

Step 2: Make that rule apply to everyone.

To do this, we just remove Luke from the rule: "No one should fulfill promises when it is inconvenient for them."

Step 3: The "contradiction in conception:" Can we coherently imagine a world where everyone follows the rule?

Luke's rule fails at this stage. In order for Luke to conceive of his rule, the institution of promise-keeping must exist. However, in a world in which everyone breaks their promises, the institution of promise-keeping doesn't really exist. Luke's rule fails because it leads to a contradiction.

According to Kant, we can conclude from the contradiction in conception that we should never break promises just for convenience.

Step 4: The "contradiction in the will:" If the rule is conceivable, would it be rational to follow it? Would it conflict with something else we must do?

To illustrate the contradiction in the will, Kant considers the case of laziness. Suppose Mari never works hard because she is lazy. We could formulate her action as a rule: "You shouldn't work hard if you feel lazy." This rule passes the contradiction in conception because it's possible to imagine a world in which no one works hard. However, Kant argues that it fails as a rule for another reason: it contradicts our will. In Kant's view, it is in our nature as rational beings to work and to "develop our talents." Therefore, we should not be lazy because, as a rule, it would violate our rational nature.

Kant's universal law formulation of the categorical imperative is a method of testing whether a rule can be willed for everyone. First, we determine whether the rule is even conceivable. Then, we determine whether people would will it. In theory, this method can tell us whether any rule is right or wrong. A Kantian AI would therefore need the capacity to reason.

Now we turn to an alternative formulation of the categorical imperative.

The "humanity" formulation of the categorical imperative. The humanity formulation is perhaps more influential among philosophers today than the universal law formulation. Roughly speaking, the humanity formulation states that we should always treat other people's humanity as an end, not merely as a means. Kant means something specific by *humanity*, *end*, and *means*.

When Kant writes about *humanity*, he is referring to the ability to engage in autonomous, rational behavior that he believed was characteristic of human nature. *Ends* are the goals that we aim to reach, and *means* are the methods of achieving ends. To treat humanity always as an end and never as a means is to treat everyone with respect for their autonomy and rationality. It's one of Kant's most influential ideas. To make this idea more concrete, here is an example.

> *The urgent lift*: Nathan wants a lift into town to go go-carting. He approaches a stranger and lies to her, telling her that his brother is having an allergic attack in town and he needs to deliver his EpiPen. Nathan tells the stranger that if she doesn't give him a lift, his brother might die.

In the example, Nathan is treating the stranger as a means to get into town. He doesn't respect the stranger as a person with her own ends, who may have better things to do. By lying to the stranger, Nathan undermines her autonomy by obscuring the truth. In other words, he is not respecting the stranger's humanity.

Criticisms of Kant's Ethics

Criticism: Many modern philosophers find Kant's ethics too extreme.

> *Mad axeman*: Omar hears a knock on his door late at night. A man with a wild look and a bloody ax asks, "Is Piper here?" Omar knows Piper is upstairs. Should he tell the truth?

Kant's ethical writings claim that it would be wrong for Omar to lie. According to Kant, lying is always wrong. Even in the case of the mad axeman, Kant insisted that it would be wrong to lie (though some philosophers inspired by Kant argue he could have avoided this claim). Most modern moral philosophers disagree with Kant's conclusion in this case.

Pro tanto duties. Instead, philosophers refer to duties to act which may be overridden by other duties. Many modern philosophers would agree that Omar has a duty to avoid lying. But they would also argue that he also has a duty to safeguard his friend, and that this duty outweighs his duty to avoid lying. In other words, many modern philosophers would see Omar's duty to avoid lying as *pro tanto*. Latin for "to that extent," pro tanto means that a given duty can be weighed against other duties to determine a course of action. Although the principle of honesty offers a pro tanto reason in favor of revealing Piper's real location to the axeman, it is not the only consideration. Other moral obligations, such as the duty to ensure others' welfare, may ultimately mean Omar should withhold Piper's real location from the axeman.

Aspects of Kant's ethics inspire modern deontology. Many aspects of Kant's morality are explicitly present in modern moral theories. For example, many deontological theories still reflect the ideas that we should treat people as ends, not as mere means; the focus on respecting others' rational and autonomous humanity; and the concept of considering how your actions might apply universally.

Conclusions about Deontology

Deontology emphasizes constraints on actions. Deontological theories consider systems of rules, often derived from basic concepts such as mutual respect for rights and autonomy, that prohibit acting in certain ways. These rules include principles such as the doctrine of double effect and the action/omission distinction, which capture intuitions about the role of intentions in morality. Kantian ethics is a well-known form of deontological ethics. It deduces the categorical imperative, which is a principle with many formulations, one of which is that we are morally required to follow general rules of acting that we would be happy for everyone else to follow, from the assumption of universal capacity to reason. However, deontological ethics is often criticized for being rigid in its rule-following and insensitive to outcomes. Modern philosophers sometimes find Kant's opinions extreme and ambiguous.

Deontological constraints in machine ethics. In the near term, we will likely see the use of AIs to maximize wealth or pursue other ambitious goals. Constraints on these AIs' actions are important. At the very least, such AIs must follow the law. However, the law is insufficient to ensure that these bots act ethically. We might want to supplement constraints provided by the law with deontological constraints as well. Requiring AIs to respect autonomy, for instance, might help avoid manipulation that is legal but unethical.

It is hard to get AIs to reliably follow rules. Whether due to clever workarounds or outright disregard, entities—such as humans or corporations—often engage in minimal compliance or exploit loopholes to achieve their objectives. This tendency to circumvent rules can also be present in AI systems. LLMs such as GPT-4 and Llama-2 have struggled to reliably adhere to straightforward instructions under the RuLES benchmark, a series of tests assessing rule adherence across different scenarios like security protocols and games. These AIs have exhibited vulnerabilities, succumbing to adversarial attacks where they are tricked or provoked into breaking rules through jailbreaks and prompt injections. To impose a deontological system of ethics, we would need to ensure that AIs can reliably follow rules.

C.4.3 Virtue Ethics

Virtue ethics is a moral theory that emphasizes the importance of having the right character traits, rather than producing the right consequences or performing the right actions [20]. A virtue ethicist might argue that we should help others in need by donating to charity, but not because we should promote wellbeing or because we have certain moral obligations. According to a virtue ethicist, we should donate to charity because that's what a generous person would do, and generosity is a virtue.

Modern virtue ethics is inspired by the ancient Greek philosopher Aristotle. In his book "Nicomachean Ethics," Aristotle explored three key concepts that are essential for understanding virtue ethics [21]. First, he explored the concept of *virtue*, or morally good character traits. Second, he developed the concept of practical wisdom, which is the set of skills and experience required in order to behave in line with virtues. Third, he argued that developing virtue and exercising practical wisdom are essential for *flourishing*, or living a good life. Each concept is described in more detail below.

Virtue

A virtue is a morally good character trait or disposition. Putative examples of virtues include courage, generosity, fairness, and kindness. Virtues are morally good character traits, and vices are morally bad ones. Putative vices include cowardice, selfishness, unfairness, and cruelness. Having a certain virtue or vice is not binary, but a matter of degree. In other words, individuals aren't typically completely courageous or completely cowardly; they can be more or less courageous.

To be virtuous is not just to behave in certain ways but also to feel certain ways. Two individuals might behave in exactly the same ways but have different feelings, and thus exhibit different virtues and vices.

Consider two siblings, Bobby and Cory, who behave similarly in every situation. They are both trusted by their friends, they both keep their promises, and they are both equally honest. However, Bobby behaves virtuously because it makes him feel good; he derives pleasure from helping others. Cory, on the other hand, behaves virtuously despite her feelings; helping others feels to her like a burden. Her behavior is the same as Bobby's, perhaps because she wishes to be seen as a virtuous person or she wants to avoid getting in trouble. According to most virtue theories, only Bobby is virtuous. While Cory behaves the same way, her behavior does not indicate virtue.

Virtue ethicists claim that other theories miss important morally relevant features. Mental states like emotions and motivations, virtue ethicists argue, are morally relevant. A consequentialist would evaluate Bobby and Cory as equally moral because their actions produce the same consequences. Some, but not all, deontologists would evaluate Bobby and Cory as equally moral because they behave the same ways with respect to rules and obligations. Virtue ethics emphasizes an intuition that many people have: that Bobby is morally superior to Cory.

Practical Wisdom

Being disposed to behave virtuously is necessary, but not sufficient, for being a virtuous person. It's also important, according to virtue ethics, to have *practical wisdom*——the ability to reason and to act appropriately on the inclination to be virtuous.

For individuals who lack practical wisdom, the inclination to be virtuous can lead them to behave wrongly. Someone who is inclined towards honesty and who derives pleasure from being honest might, in some situations, be too honest. If they lack practical reason, they may needlessly insult strangers or cause conflicts between friends. Practical wisdom is the ability to understand when it's appropriate to be honest and when it's important to act on a different virtue, like kindness.

While people may be born with the disposition towards certain virtues, practical wisdom is learned through experience. Children, for example, may desire to behave well, but make errors due to a lack of experience. They may tell their mother that they don't like her outfit, unable to differentiate between honesty and cruelty. If their mother reacts with hurt feelings, her children will learn from the experience that, in some situations, kindness is more appropriate than honesty. In other words, they will gain practical wisdom. We can learn practical wisdom from people who have had more experience than us, but also from cultural figures who clearly excel in their virtue, people who offer excellent examples of virtues of honesty, steadfastness, and compassion. Evidently, if we would like to make AIs virtuous, we would need to have exemplars for them to imitate.

Flourishing

Flourishing is living a good life. Aristotle believed that being virtuous is necessary for flourishing. In fact, he defined virtues in terms of their relationship towards flourishing. Virtues, he argued, are those character traits which lead an individual to flourish. The virtue ethicist's idea of flourishing has similarities to the objective goods account of wellbeing that we discussed earlier in this chapter.

While most virtue ethicists agree that being virtuous is necessary for living a good life, they often disagree about whether it is sufficient for living a good life. Aristotle argued that, in addition to being virtuous, an individual must have the resources to enact virtue in order to lead a flourishing life. Someone living in poverty, according to Aristotle, is unable to enact certain virtues, like magnanimity. A virtuous but very unlucky person may be unable to flourish.

In sum, virtue ethics argues that to live a good life we must develop our virtues, and that to develop our virtues we should imitate people who act virtuously.

Criticisms of Virtue Ethics

Criticism: Virtue ethics is not action-guiding.

Virtue ethics doesn't always clearly help us determine the right things to do. When faced with a dilemma, like whether or not to steal from a grocery store in order to feed one's family, virtue ethics does not seem to offer much guidance. We should be generous and selfless, which seems to suggest that stealing is wrong. However, we should also be loyal and protective of our family, which seems to suggest that we should do what is necessary in order to feed them. In such cases, virtue ethics may not seem very useful. Virtue ethicists may argue, however, that while their theory does not include a decision procedure for every situation, a virtuous person with a high capacity for practical reasoning will understand which virtues to express and when.

Criticism: Virtue ethics is too focused on the individual.

According to virtue ethics, whether an action is right or wrong depends entirely on characteristics of the actor. If the person performing the action is ideally virtuous, then their actions will be morally right. This may seem odd to those who believe that the field of ethics is concerned with how to treat other people. Presumably we should save someone from drowning, not because of facts about our own character traits but because of facts about the drowning person. We should save them because their life has value, because their wellbeing matters, and because they have a right to life, not because we are courageous.

Conclusions about Virtue Ethics

Virtue ethics places emphasis on the character of the individual. While consequentialist theories prioritize outcomes and deontological theories focus on constraints, virtue ethics centers around the importance of acting virtuously. Aristotle, who greatly influenced modern virtue ethics, highlighted virtue (good character traits), practical wisdom (the skills and experience necessary for virtuous action), and flourishing (living a good life) as the three fundamental aspects of morality. He argued that the first two are essential for achieving the third. Critics of virtue ethics argue that it lacks clear guidance for action and places excessive emphasis on the individual rather than their actions.

Virtues can be instilled into AIs. Characteristics such as justice, honesty, responsibility, care, prudence, and fortitude have been proposed as basic AI virtues. AIs that are trained to represent such virtues might balance promoting their objectives with behaving ethically. Assuming we can train AIs in such ways, we might fine-tune them to represent different characteristics in different settings: an AI used for data analysis might prioritize honesty, while an AI used for teaching children might prioritize responsibility and care. We can imagine turning up or down different characteristics to create ethical AIs that act appropriately to the setting.

AIs can help us cultivate virtues. We might want to use AIs to create moral value for humans; virtue ethics suggests a few ways to do so. One might be to help individuals and societies become more virtuous, such as by helping people take opportunities to be virtuous or improving material conditions so that they can better exercise virtues like magnanimity and generosity. Similarly, AIs might help with developing their users' practical wisdom, such as through education or encouraging them to undertake projects. The goal of such systems might be to increase human flourishing.

C.4.4 Social Contract Theory

The focus of this section is social contract theories of morality. As the name suggests, social contract theory focuses on contracts—or, more generally, hypothetical agreements between members of a society——as the foundation of ethics. A rule such as "do not kill" is morally right, according to a social contract theorist, because individuals would agree that the adoption of this rule is in their mutual best interest, and would therefore insert it into a social contract underpinning that society. The most influential contemporary theorist within this tradition is John Rawls, who we will use to contextualize social contract theory, using the famous *veil of ignorance* thought experiment [22]. After understanding the broad strokes of his theory, we will consider a few reasons why such reasoning might be inadequate and consider some alternatives.

According to social contract theory, moral codes are the result of hypothetical agreements between members of society, established for mutual benefit. Let us consider the prohibition of thievery. It seems reasonable that people would agree to refrain from stealing: most people stand to benefit from a society that punishes thieves, given that it would disincentivize others from robbing them. The ethical principle of not stealing would thus have moral force, without requiring some fundamental principles such as maximizing wellbeing or respecting autonomy. According to the social contract theorist, all moral codes are similarly justified: they are reasonable hypothetical agreements which encourage behavior that creates mutual benefit.

The Veil of Ignorance

Philosopher John Rawls introduced the concept of a *veil of ignorance* as a tool for creating a social contract. When behind the veil of ignorance, individuals lose all knowledge of their personal attributes, such as their talents, religion, gender, sexuality, or class. In this state, sometimes called the *original position*, participants are asked to envision a basic structure for society, based on reasonable agreements without any knowledge of their own positions within the society they create.

In "A Theory of Justice", Rawls proposes one way to generate a social contract. Rawls places everyone behind a veil of ignorance to make decisions about a social contract. From here, having lost all knowledge of their individual characteristics, participants are invited to envision a basic structure for society, based on reasonable agreements without knowledge of their own position in the society they create.

Once the decisions have been made, participants leave the original position, discover who they are in society, and then live according to the social contract they created. Rawls believed that, if we were able to use the veil of ignorance to construct a real society, the society would likely include ideas like: protection of the worst-off, basic liberties for all, and restrictions on inequality. Because the people determining these contracts do not know which social group they will be a part of, they would not create a society in which some people are far worse off than others.

The veil of ignorance would prevent slavery. Those in Rawls' original position would arrive at many sensible conclusions. An individual behind the veil would not reasonably permit slavery, for instance, given that they do not know whether they are a slave or a slaver. On average, very few people seek to benefit from slavery, and most people are harmed by it. Individuals would likely not wish to take the chance of being a slave in exchange for a chance of being a wealthy slave owner. Therefore, those in the original position would reject slavery, and this forms the basis of the social contract to not enslave others.

Agreements behind the veil of ignorance create a basis for a just society. This process is unlikely to generate elitist, patriarchal, or ableist moral codes since one does not know whether they have a high income, privileged racial or gender identity, or disability. The interests of any particular group would not be favored, since no one knows whether they belong to that group. As such, decisions behind the veil of ignorance are made in the interest of society. According to Rawls, everyone accounts for this component of luck by making decisions as though they could, themselves, be members of any group in society.

Principles That may be Generated from the Veil of Ignorance

In this section, we look at some of the conclusions that Rawls argues individuals in the original position might reach: the maximin principle, the liberty principle, the equality of opportunity principle, and the difference principle.

We should protect the worst off. Behind the veil of ignorance, no one knows who they are, which means that anyone might be the worst-off individual in society. The main idea behind Rawls' contractarian logic is that people would agree to make sure everyone has a decent position in society, since anyone might end up in the worst position. Rawls argues that people behind the veil of ignorance would endorse the *maximin principle*, according to which we should prioritize the worst-off in society.

We should not exclude anyone from having basic liberties. Behind the veil of ignorance, no one knows who they are, which means that if any individual is excluded from having basic liberties, it could potentially be anyone. Therefore, individuals would rationally agree to distribute liberties to every member of society. As long as one does not infringe upon others' liberties, they should have the freedom to pursue their own conception of the good life, and with it enjoy civil and political liberties such as freedom of speech, religion, association, assembly, and the right to a fair trial. This is Rawls' *liberty principle*: the fair distribution of liberties, ensured by a contractarian agreement, do not infringe upon the liberties of others.

We should ensure equality of opportunity. Adopting the veil of ignorance may lead us to conclude that no one should be denied opportunities based on their gender, age, race, family status, or other personal characteristics. People behind the veil would agree that, when inequalities arise, they should only do so through fair access to opportunity for everyone. Equality of opportunity ensures that differences in the relative positions of individuals are meritocratic. This is Rawls' *equality of opportunity* principle: some degree of inequality is permissible, but only if it is merit-based.

We should require that inequalities benefit the least-advantaged. Allowing some inequalities can lead to increased overall wealth in a society, as higher earnings for the most productive members can create incentives for economic growth that benefits everyone. This is the basis of Rawls's *difference principle*: some degree of inequality is permissible, but only if it also benefits the worst-off members of society.

These principles, in theory, ensure two conditions: everyone has basic rights and equal access to opportunity, and any inequalities that follow from the equality of opportunity also help the least privileged, even if they help the privileged more.

Rawls's Conclusions might be Too Strong

The maximin principle is at odds with common-sense morality. According to Rawls, our moral evaluation of a society should be determined by the wellbeing of its worst-off individual. This implies that we must invest all our resources into raising the wellbeing of the worst-off member of society, and remain indifferent towards everyone else's wellbeing, as long as they do not become the worst-off.

The grouch takes priority. Imagine a grouch: someone who is always at low levels of wellbeing and extremely hard to please. Say that giving them a billion dollars would make them only as happy as an ordinary person would be with a slice of cake. The maximin principle dictates that our priority should be to focus on improving their wellbeing, even if it means using a large amount of resources that could have made everyone else in society much happier. This prioritization seems counterintuitive.

We must be indifferent towards improving everyone else's wellbeing. Suppose there was a new medical breakthrough that can cure a widespread disease, greatly improving the lives of many who suffer from the illness. This seems like a good thing, and it would be strange to not care about this happening. However, if the worst-off individual doesn't have that disease and their wellbeing remains unchanged, nothing morally important has changed.

We must be indifferent towards decreasing everyone else's wellbeing. Consider a scenario in which a technical error causes all electronically stored money to be deleted, plunging most countries into chaos and leading to widespread poverty. This seems like a bad thing, and it would be strange not to care about this happening. However, if the worst-off individual lives in an unaffected area and does not have a bank account, and their wellbeing remains unchanged, once more nothing morally important has changed.

The maximin principle seems untenable. In all three of these scenarios, Rawls' maximin principle gives us implausible answers. We likely should not spend all of our resources to give one unhappy person a bit of joy at the cost of everyone else's wellbeing. Similarly, it seems extremely morally relevant if we can cure a widespread disease or ensure that most people do not lose all their money.

Rawls' Conclusions Might Not Follow from the Veil of Ignorance

Behind the veil of ignorance, we would care about more than maximin. If, behind the veil of ignorance, we know that there is just one person who will be much worse off than everyone else, we may not unanimously agree to prioritize making that person's situation better. Individuals behind the veil of ignorance may not consider the worst-case outcomes when making decisions under uncertainty. While the people behind the veil might be highly risk-averse, ensuring that the general distribution of society is good rather than just the average level of wellbeing and endorsing the liberty and difference principles, they would likely not endorse the maximin principle.

The veil of ignorance may lead to utilitarianism. Nobel prize winning economist John Harsanyi conceptualized the veil of ignorance before Rawls, and used it to support utilitarianism [23]. He argued that rational agents would aim to maximize the total amount of wellbeing in their society, so that the average outcome is as good as possible. Decisions under uncertainty often involve maximizing expected or average results. Harsanyi argued that rational individuals behind the veil of ignorance would therefore choose a utilitarian organization for society.

A problem for Rawls. Rawls' "A Theory of Justice" was designed as an alternative to utilitarianism, but it has been used to justify utilitarianism. If the veil of ignorance indeed creates conditions more conducive to utilitarian moral decisions, it might actually support utilitarianism instead of Rawls' theory of justice.

Alternatives to Rawls' Social Contract Theory

A natural middle ground between Rawlsian and utilitarian ideas is prioritarianism. *Prioritarianism* is an ethical theory that gives greater moral weight to improving the wellbeing of those who are worst off in society [24]. Imagine a situation where we have the option to distribute resources among three people: Rana, Sean, and Toby. Before any intervention, they have the following levels of wellbeing: Rana has 6 units of wellbeing, Sean has 5 units of wellbeing, and Toby has 1 unit of wellbeing.

We can choose one of the following options:

1. Increase Rana and Sean's wellbeing by 2 units each, resulting in (8, 7, 1);
2. Increase Toby's wellbeing by 1 unit, resulting in (6, 5, 2); or
3. Increase Toby's wellbeing by 2 units, reduce Sean's wellbeing by 2 units, and reduce Rana's wellbeing by 3 units, resulting in (3, 3, 3).

Prioritarianism suggests that we should prioritize improving the wellbeing of Toby, the worst-off individual. However, we should also take into account the wellbeing of others. Unlike utilitarianism, which dictates that we choose option 1 to maximize overall wellbeing, prioritarianism suggests that we choose option 2 to help the most disadvantaged member. Unlike Rawls' maximin principle, which suggests that we choose option 3 to most improve the situation of the least advantaged individual (even at the cost of everyone else), prioritarianism attributes moral weight to the wellbeing of Rana and Sean and so might choose option 1 or 2 instead. This approach strikes a balance between utilitarianism and Rawlsianism, placing higher moral value on improving someone's life when (a) that person's overall wellbeing is relatively low and (b) the increase in wellbeing would be substantial.

One example of a prioritarian policy is focusing educational interventions on disadvantaged students. Prioritarians might support policies that concentrate resources on helping students from disadvantaged backgrounds or those with learning disabilities, with the goal of closing achievement gaps and improving outcomes for the worst-off. Utilitarians might argue that resources should be allocated to improve overall educational outcomes, which might involve investing in programs that benefit a larger number of students, even if it doesn't specifically target the most disadvantaged.

Contractualism. Philosopher T.M. Scanlon developed another approach to social contract theory called contractualism [25]. Scanlon established contractualism to build on the idea that morality is based on agreements between people while addressing some of the limitations of Rawls' theory. According to Scanlon, people are naturally inclined to seek reasonable moral agreements, driven by their sense of justice. This makes Rawls' original position, in which people are behind a veil of ignorance, unnecessary. In Scanlon's view, morality is about what we owe each other as rational beings.

Scanlon did not have a definitive answer to what we owe one another. Instead, he proposed using a social contract underpinned by reasonableness. It is reasonable to reject a moral code declaring slavery is right. However, it is less reasonable to reject a moral code declaring that community resources should be used to help the worse off. For a contractualist, certain actions are wrong if they don't meet a standard of behavior that "no one could reasonably reject as a basis for informed, unforced, general agreement." This means that principles of morality should be something that rational people can generally accept. By making fewer strong assumptions, Scanlon avoids some of the pitfalls of Rawls' process and conclusions.

Conclusions about Social Contract Theory

Rawls' contractarianism and the veil of ignorance. Social contract theory proposes that agreements between society's members form the foundation of morality. Moral codes are determined by whether rational individuals would agree to mutually abide by them. John Rawls' influential approach relies on the veil of ignorance, where decision-makers are unaware of their position in society. Rawls' theory suggests the maximin, liberty, and difference principles, which together aim to raise the lowest levels of wellbeing, ensure the provision of basic liberties, and minimize inequalities in society.

Alternatives to Rawls: Scanlonian contractualism, prioritarianism, and utilitarianism. Some of Rawls' conclusions, and especially the maximin principle, are implausible. The veil of ignorance can be used to support utilitarianism, and alternatives like prioritarianism are better at accounting for some of our intuitions. Scanlon's contractualism offers another perspective on social contract theory, focusing on people acting according to reasonable principles and contending that morality is built upon obligations to others—what we owe to each other. Social contract theories offer an alternative way of thinking about moral agreements that might be highly relevant to modern approaches to machine ethics, highlighting the importance of rational agreement and mutual benefit in shaping ethical principles.

Several contractarian principles can be used to inform AI design. Often, we require algorithmic decision-making to be blinded to personal details like race, gender, and social status. This can be justified by Rawlsian ideas such as the veil of ignorance and the equality of opportunity principle. However, relying on social contract reasoning can present problems as well. Unlike other theories, social contract theories have no fundamental values besides agreement. As a result, any outcome that everyone agrees with can be considered morally justifiable. AIs forming a social contract with each other may agree not to care for humans, or AIs forming a contract with humans may use their intellect to persuade them to accept less-than-ideal terms.

Governance can be contractarian. We can use AIs to emulate people behind the veil of ignorance since AIs can reason from an impartial position and have them negotiate a social contract on our behalf. We could use artificial intelligence to inform our Scanlonian judgments as well, such as by asking AIs trained on human responses whether individuals have acted according to principles that they would reasonably reject. Social contract theory can also provide insight into how we should govern AIs, such as by recommending that our governance procedures are inclusive of all stakeholders' interests and perspectives.

Summary

We have outlined four common approaches to morality. First we looked at utilitarianism, a theory that argued that actions should be judged based on how much wellbeing they cause. Then we considered deontology, the view that to live ethically is to live by the right system of rules. According to virtue ethics, the goal of a good life is to become a virtuous person. Social contract theory recommends following principles that we might arrive at together in an ideal contractual process. These theories are all very different. They don't just differ in their moral claims; they consider morality to have different goals that should be approached in different ways.

The theories we've discussed are focused on different moral considerations. Early in this chapter, we discussed moral considerations like intrinsic goods, special obligations, constraints, and options. Utilitarianism, of course, is most concerned with wellbeing, an intrinsic good, and does not support options. Deontology especially emphasizes constraints. Social contract theory especially emphasizes special obligations.

In common-sense morality (i.e. the moral decision-making that everyone does on a daily basis) different considerations seem more important than others in different situations. It may be, then, that some theories are more useful for thinking about some problems than others.

We needn't pick a single moral theory, and the goal of this chapter is not to choose the best one. Utilitarianism and social contract theory might be the best approaches for thinking about some society-wide policies, but perhaps deontology is helpful when we are drafting laws that are intended to apply to everyone in a country. Virtue ethics may be especially useful in day-to-day situations.

When moral theories conflict, it is important that we accommodate some uncertainty. There may be something we can learn from all of them.

C.4.5 Making Decisions Under Moral Uncertainty

This section considers how to make decisions under moral uncertainty——how to act morally when we are unsure which moral view is correct. Although ignoring our uncertainty may be a comfortable approach in daily life, there are situations where it is crucial to identify the best decision. We will start by considering our uncertainties about morality and the idea of reasonable pluralism, which acknowledges the potential co-existence of multiple reasonable moral theories such as ethical theories, common-sense morality, and religious teachings.

The main aim of the section is to recognize that there are different reasonable moral beliefs and think about what to do with them. We will explore uncertainties about moral truths, why they matter in moral decision-making for both humans and AI, and how to deal with them. We will look at a few proposed solutions, including *My Favorite Theory*, *Maximize Expected Choice-Worthiness (MEC)*, and *Moral Parliament* [26]. These approaches will be compared and evaluated in terms of their ability to help us make moral decisions under uncertainty.

Dealing with Moral Uncertainty

Moral uncertainty requires us to consider multiple moral theories. Individuals may have varying degrees of belief in different moral theories, known as *credence*.

Credence is the probability assigned by an individual of the chance a theory is true. Someone may have a high degree of credence (70%) in utilitarianism, meaning they believe that maximizing utility is likely to be the most important moral principle. However, they may also have some credence in deontological rules, believing they are plausible but less likely to be true than utilitarianism (30%). Since ethical theories are not flawless, often arriving at intuitively questionable conclusions, it is valuable to consider multiple perspectives when making moral decisions.

Living a good life can require a combination of insights from multiple moral theories. We often find ourselves balancing different kinds of deeds: creating pleasure for people, respecting autonomy, and adhering to societal moral norms. Abiding by a *reasonable pluralism* means accepting that moral guidance from different sources may conflict while still offering value.

In high-stakes moral decisions, such as in healthcare or AI design, reasonable pluralism alone may fall short. When a healthcare administrator faces the tough choice of allocating limited resources between hospitals, we might want them to seriously consider the ethics of what they are doing rather than just go with conflicting wisdom that seems reasonable at first glance. Therefore, we must think hard about how to make decisions under moral uncertainty——seeking truth for crucial decisions is vital.

If AI is unable to account for moral uncertainty, harmful outcomes are likely. As AI systems become increasingly advanced, they are likely to become better at optimising for the goals that we set them, finding more creative and powerful solutions to achieve these. However, if they pursue very narrowly specified goals, there is a risk that these solutions come at the expense of other important values that we failed to adequately include as part of their goals. Given that philosophers have not yet converged on a single consistent moral theory that can take account of all relevant arguments and intuitions, it seems important for us to design AI systems without encoding a false sense of certainty about moral issues that could lead to unintended consequences. To counter such problems, we need AI systems to recognize that a broad range of moral perspectives might be valid—in other words, we need AI systems to acknowledge moral uncertainty. This presents another challenge: how should we rationally balance the recommendations of different ethical theories? This is the question of moral uncertainty.

How Should We Approach Moral Uncertainty?

There are several potential solutions to moral uncertainty. Faced with ethical uncertainty, we can turn to systematic approaches, using our estimates of how theories judge different actions and how likely these theories are. This section explores three potential solutions to moral uncertainty. The first is adopting a favored moral theory that aligns with personal beliefs (*My Favorite Theory*). The second is aiming for the highest average moral value by calculating the expected choice-worthiness of each option (*Maximize Expected Choice-Worthiness*). The third is treating the decision as a negotiation in a parliament, considering multiple moral views to find a mutually acceptable solution (*Moral Parliament*).

Consider whether we should we lie to save a life. Imagine that a notorious murderer asks Alex where his friend, Jordan, is. Alex knows that revealing Jordan's location will likely lead to his friend's death, while lying would save Jordan's life. However, lying is morally questionable. Alex must decide which action to take. He is unsure, and considers the recommendations of the three moral theories he has some credence in: utilitarianism, deontology, and contractarianism. Alex thinks utilitarianism, which values lying to save a life highly, is the most likely to be true: he has 60% credence in it. Deontology, which Alex has 30% credence in, strongly disapproves of lying, even to save a life, and contractarianism, which Alex has 10% credence in, moderately approves of lying in this situation. This information is represented in Table C.3.

	Utilitarianism	Deontology	Contractarianism
What is Alex's estimate of the chance this theory is true	60%	30%	10%
Does this theory like lying to save a life?	Yes	No	Yes

Table C.3: Example: Alex's credence in various theories and their evaluation of lying to save a life.

Under My Favorite Theory (MFT), Alex would pick whatever utilitarianism recommends. Alex thinks that utilitarianism is the most likely to be true. The MFT approach is to follow the prescription of the theory we believe is the closest to a moral truth. Intuitively, many people do this already when thinking about morality. The advantage of MFT is its simplicity: it is relatively simple and straightforward to implement. It does not require complex calculations or a detailed understanding of different moral perspectives. This approach can be useful when the level of moral uncertainty is low, and it is clear which theory or option is the best choice.

However, following MFT can lead to harmful single-mindedness or overconfidence. It can be difficult to put aside personal biases or to recognize when one's own moral beliefs are fallible (individuals tend to defend and rationalize their chosen theories). The key issue with MFT is that it can discard relevant information, such as when the credences in two theories are close, but their judgments of an action vastly differ [27]. Imagine having 51% credence in contractarianism, which mildly supports lying to save a life, and 49% credence in Deontology, which views it as profoundly immoral. MFT suggests following the marginally favored theory, even though the potential harm, according to the second theory, is much larger. This seems counterintuitive, indicating that MFT might not always provide the most sensible approach for navigating moral uncertainty.

Maximize Expected Choice-Worthiness (MEC) gives us a procedure to follow. MEC tells us that to determine how to act, we need to do the following two things:

1. **Determine choice-worthiness.** Choice-worthiness is a measure of the overall desirability or value of an option—in this context, it is how morally good a choice is. This is an expression of the size of the moral value of an action. We can represent the choice-worthiness of an action as a number. For instance, we might think that under utilitarianism, the choice-worthiness of murder could be -1000, of littering could be -2, of helping an old lady cross the street could be +10, and of averting existential risk could be +10000.

2. **Multiply choice-worthiness by credence.** We can consider the average choice-worthiness of an action, weighted by how likely we think each theory is to be true. This gives us a sense of our best guess of the average moral value of an action, much like how we considered expected utility when discussing utilitarianism. Table C.4 has each theory's choice-worthiness value for lying to save a life. As before, utilitarianism highly values lying to save a life (+500), deontology strongly disapproves of it (-1000), and contractualism moderately approves of it (+100). In the row beneath these values are the credence probability-weighted judgments for Alex. The total calculation is

$$\frac{60}{100} \cdot 500 + \frac{30}{100}(-1000) + \frac{10}{100} \cdot 10 = 300 - 300 + 10 = 10$$

Under MEC, Alex would choose to lie, because given Alex's credence in each moral theory and his determination of how each moral theory judges lying to save a life, lying has a higher expected choice-worthiness. Alex would lie because he judges that, on average, lying is the best possible action.

	Utilitarianism	Deontology	Contractarianism
What is Alex's estimate of the chance this theory is true	60%	30%	10%
How much does this theory like lying to save a life	+500	-1000	+100
What is the probability-weighted judgment?	+300	-300	+ 10

Table C.4: Example: Alex's credence in various theories, their evaluation of lying to save a life, and their probability-weighted contribution to the final judgment.

MEC gives us a way of balancing both how likely we think each theory is with how much each theory cares about our actions. We can see that utilitarianism and deontology's relative contributions to the total moral value cancel out, and we are left with an overall "+10" in favor of lying to save a life. This calculation tells us that—when accounting for how likely Alex thinks each theory is to be true and how strong the theories' preferences over his actions are—Alex's best guess is that this action is morally good. (Although, since these numbers are rough and the final margin is quite thin, we would be wary of being overconfident in this conclusion: the numbers do not necessarily represent anything true or precise.) MEC has distinct advantages. For instance, unlike MFT, we ensure that we avoid actions that we think are probably fine but might be terrible, since large negative choice-worthiness from some theories will outweigh small positives from others. This is a sensible route to take in the face of moral uncertainty.

However, MEC faces challenges when comparing theories. While some cases are neatly solved by MEC, its philosophical foundations are questionable. Our assignment of choice-worthiness can be arbitrary: virtue ethics, for instance, advises acting virtuously without clear guidelines. MEC is unable to deal with 'ordinal' theories that only rank actions by their moral value rather than explicitly judging how morally right or wrong they are, making it difficult to determine choice-worthiness. Absolutist theories, like extreme Kantian ethics, deem certain actions absolutely wrong. We had initially assigned -1000 for the value of lying to save a life for a deontologist; it is unclear what we could put that would capture such an absolutist view. If we considered this to be infinitely bad, which seems like an accurate representation of the view, then it would overwhelm any other non-absolutist theory. Even if we think it is simply very large, these firm stances can be more forceful than other ethical viewpoints, despite ascribing a low probability to the theory's correctness. This is because even a small percentage of a large value is still meaningfully large; consider that 0.01% of 1,000,000 is still 100 — a figure that may outweigh other theories we deem more probable.

Following a moral parliament approach, Alex could consider the proposal to lie more thoroughly and make a considered decision. In the moral parliament, imagined delegates representing different moral theories negotiate to find the best solution. In Alex's moral parliament, there would be 60 utilitarian delegates, 30 deontological delegates, and 10 contractarian delegates––numbers proportional to his credence in each theory. These delegates would negotiate and then come to a final decision by voting. Drawing inspiration from political systems, the moral parliament allows an agent to find flexible recommendations that enable compromises among plausible theories.

Depending on the voting rule, moral parliament can lead to different outcomes. In a conventional setting with majority rule, the utilitarians in Alex's moral parliament would always be able to push through their decisions. To avoid such outcomes, philosophers recommend using *proportional chances voting*. Here, an action is taken with a probability equal to its vote-share: in this case, if no one changed their mind, then the outcome of the parliament would recommend with 70% probability that Alex lie and with 30% probability that he tells the truth, since only the 30 deontological delegates would vote against this proposal. This encourages negotiation even if there is already a majority, which naturally leads to more cooperative outcomes. This is a more intuitive approach than, for instance, assigning choice-worthiness values and multiplying things out, even if it does take more effort.

However, it is difficult to see what a moral parliament might recommend. We can envision a variety of different possible outcomes in Alex's case. We might have the simple outcome described above, where no one changes their mind. Or, we might think that the deontologists can convince the contractualists that lying is bad in this case because lying would not be tolerated behind the veil of ignorance, reducing the chance of lying to 60%. Or, the parliament might even propose something entirely new, such as a compromise in which Alex does not explicitly lie but simply omits the truth. All of these are reasonable—it is difficult to choose between them.

The outcome of the moral parliament is not determined externally. Instead, it is a matter of our imagination, subject to our biases. Often, individuals resist imagining things that contradict our preconceived notions. This means that moral parliament may not be very helpful for individuals thinking about what is right to do in practice. With enough resources, however, we might be able to simulate model parliaments to recommend such decisions for us, such as by hiring diplomats to represent moral positions and having them bargain—or by assigning moral views to multiple AI systems and having them come to a collective decision.

Conclusions about Moral Uncertainty

Taking moral uncertainty into account is difficult but important. As AI systems become increasingly embedded across various part of society and take more consequential decisions, it will become more important to ensure that they can handle moral uncertainty. The same AI systems may be used by a wide variety of people with different moral theories, who would demand that AI acts as far as possible in ways that do not violate their moral views. Incorporating moral uncertainty into AI decision-making could reduce the probability of taking actions that are seriously wrong under some moral theories.

In this section, we explored ways of moving beyond intuitive judgments or a reasonable pluralism of theories, examining three solutions to moral uncertainty: my favorite theory, maximize expected choice-worthiness, and moral parliament. Each approach has its strengths and limitations, with the choice depending on the situation, level of uncertainty, and personal preferences.

It is important to recognize that there might not be a one-size-fits-all solution to ethical dilemmas. As AI technology progresses, we should establish methods for addressing moral uncertainty, such as including diverse moral perspectives and quantifying uncertainty.

C.5 Moral Parliament

In section C.4, have explored a few of the most influential ethical theories developed by moral philosophers to explain how we ought to behave. However, each of these presents some serious problems that leave them unsuitable to be the sole solution to the problem of machine ethics that we defined in Chapter 6. Most theories seem to have at least some counterintuitive implications that we would be uncomfortable endorsing, and many seem to capture something valuable about morality. The problem of determining how to act given uncertainty about which theory of ethics is correct is the problem of *moral uncertainty*, which we considered in the previous section. One solution to moral uncertainty is using a *moral parliament.*

Should we use a moral parliament to guide AI decision-making? In this section, we will explore using moral parliaments to help AIs make robust ethical decisions in the face of moral uncertainty. First, we explore how we might use AIs to implement a moral parliament. Then, we will consider five advantages such an approach has over giving AIs specific human values.

C.5.1 Implementing a Moral Parliament

We can use AIs to simulate moral parliaments. If we want AIs to be guided by moral theories while accounting for moral uncertainty, we can use a moral parliament. The basic idea is to represent a set of moral theories that we think are plausible, place representatives of these theories into a parliament, and use democratic processes to make decisions [28].

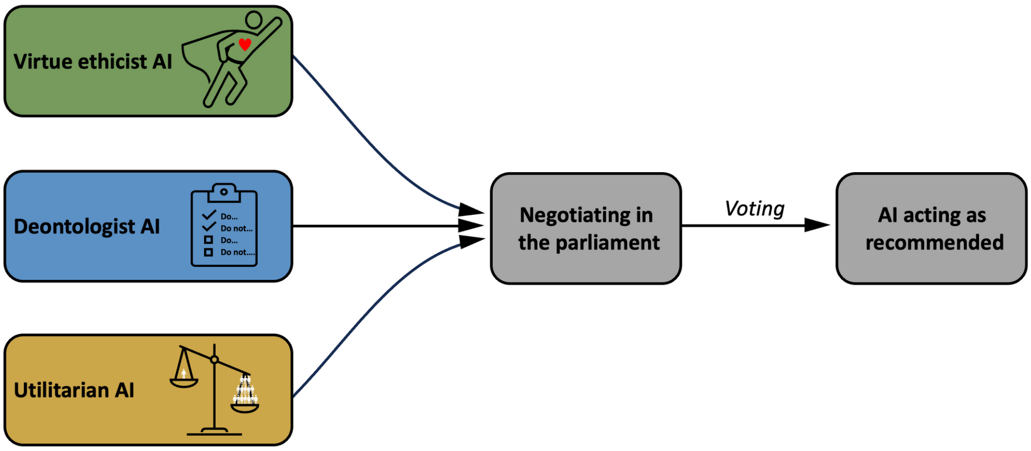

Figure C.2: Different AIs, each programmed to represent a unique moral theory, engage in negotiation and voting to decide on a course of action, simulating parliamentary processes.

We can use advanced AIs to emulate these representatives by training AIs to act in accordance with a specific moral theory. This allows us to run moral parliaments artificially, permitting real-time decision-making. Just like in real parliaments, we would have delegates—AIs instructed to represent certain moral theories—get together, discuss what to do, negotiate favorable outcomes, and then vote on what to recommend. We could use the output of this moral parliament to instruct a separate AI in the real world to take certain actions over others.

This is speculative and might still face problems; for instance, AIs might have insufficient understanding of our moral theories. However, these problems could become more tractable with advanced AI systems. Assuming we have this ability, using a moral parliament might be an attractive solution to getting AIs to act in accordance with human values in real time.

We can decide who we want represented in our generalized moral parliament. While we have explored the traditional moral parliament method of representing moral theories, we can generalize beyond this. Instead of representing theories, it might be more appropriate to represent stakeholders; for instance, in a decision about public transport, we could emulate representatives for local residents, commuters, and environmental groups, all of whom have an interest in the outcome. Using a generalized moral parliament for decision-making in AI is an approach that ensures all relevant perspectives are taken into account. In contrast to traditional methods that focus on representing different moral theories, *stakeholder representation* prioritizes the views of those directly affected by the AI's decisions. This could enhance the AI's understanding of the intricate human social dynamics involved in any given situation.

Let's consider an AI suggesting new regulations for data privacy. By emulating the perspectives of users, advertisers, and developers, it can obtain a comprehensive understanding of the potential implications of new regulations. Users may prioritize privacy and usability, advertisers may focus on visibility and click-through rates, and developers may be concerned with feasibility and profitability. By considering these diverse views, the AI can make a more balanced decision that better aligns with overall societal interests in real time.

C.5.2 Advantages of a Moral Parliament

Using moral parliaments presents a wide array of benefits relative to just giving AIs certain sets of values directly. In this subsection, we will explore how they are customizable, transparent, robust to bugs and errors, adaptable to changing human values, and pro-negotiation.

Customizable moral parliaments are diverse and scalable. The generalized moral parliament can accommodate a wide variety of stakeholders, ranging from individual users to large corporations, and from local communities to global societies. By emulating a large set of stakeholders, we can ensure that a diverse set of views are represented. This allows AIs to effectively respond to a wide range of scenarios and contexts, providing a robust framework for ensuring AIs decisions reflect the values, interests, and expectations of all relevant stakeholders. Additionally, moral parliaments are scalable: if we are concerned about a lack of representation, we can simply emulate more stakeholders. By grounding AI decision-making in human perspectives and experiences, we can create AI systems that are not only more ethical and fair but also more effective and beneficial for society as a whole.

Transparency is another key benefit of a moral parliament. As it stands, automated decision-making is opaque: we rarely understand why AIs make the decisions they do. However, since the moral parliament gives us a clear mechanism of representing and weighing different perspectives, it allows stakeholders to understand the basis of an AI's decision-making. We could, for instance, enforce that AIs keep records of simulated negotiations in human languages and then view the transcripts. Using moral parliaments provides insights into how different moral considerations have been weighed against each other, making the decision-making process of an AI more transparent, explainable, and accountable.

Moral parliaments tend to be less fragile and less prone to bugs. If we are sure that utilitarianism is the correct moral view, we might be tempted to create AIs that maximize wellbeing-—this seems clean and elegant. However, having a diverse moral parliament would make AIs less likely to misbehave. By having multiple parliament members, we would achieve *redundancy*. This is a common principle in engineering: to always include extra components that are not strictly necessary to functioning, in case of failure in other components (and is explored further in the Safety Engineering chapter). We would do this to avoid failure modes where we were overconfident that we knew the correct moral theory, such as lying and stealing for the greater good, or just to avoid poor implementation from AIs optimizing for one moral theory. For example, a powerful AI told that utilitarianism is correct might implement utilitarianism in a particular way that is likely to lead to bad outcomes.

Imagine an AI that has to evaluate millions of possibilities for every decision it makes. Even with a small error rate, the cumulative effect could lead the AI to choose risky or unconventional actions. This is because, when evaluating so many options, actions with high variance in moral value estimation may occasionally appear to have significant positive value. The AI could be more inclined to select these high-risk actions based on the mistaken belief that they would yield substantial benefits. For instance, an AI following some form of utilitarianism might use many resources to create happy digital minds—at the expense of humanity—even if that is not what we humans think is morally good.

This is similar to the Winner's Curse in auction theory: those that win auctions of goods with uncertain value often find that they won because they overestimated the value of the good relative to everyone else; for instance, when bidding on a bag of coins at a fair, people who overestimate how many coins there are will be more likely to win. Similarly, the AI might opt for actions that, in hindsight, were not truly beneficial. A moral parliament can make this less likely, because actions that would be judged morally extreme by most humans also wouldn't be selected by a diverse moral parliament.

The process of considering a range of theories inherently embeds redundancy and cross-checking into the system, reducing the probability of catastrophic outcomes arising from a single point of failure. It also helps ensure that AI systems are robust and resilient, capable of handling a broad array of ethical dilemmas.

Moral parliaments encourage compromise and negotiation. In real-life parliaments, representatives who hold different opinions on various issues often engage in bargaining, compromise, and cooperation to reach agreeable outcomes and find common ground. We want our AIs to achieve similar outcomes, such as ones that are moderate instead of extreme. Ideally, we want AIs to select outcomes that many moral theories and stakeholders all like, rather than being forced to trade off between them.

In particular, we might want to design our moral parliaments in specific ways to encourage this. One such feature is proportional chances voting, in which each option then gets a chance of winning that's proportional to the number of votes it gets—if a parliament is 60/40 split on a proposal, then the AI would do what's recommended 60% of the time rather than just going with the majority. This setup motivates the representatives to come together on options that are compromises rather than sticking to their own viewpoints rigidly. They want to do this to prevent any option they see as extremely bad from having any chance of winning. This ensures a robust high-level principle guiding AI behavior, reducing the risk of extreme outcomes, and fostering a more balanced, nuanced approach to ethical decision-making.

Using a moral parliament reduces the risk of overlooking or locking in certain values. The moral parliament represents an approach to ethical decision-making in AI that is distinctively cosmopolitan, in that it encompasses a broad range of moral theories or stakeholders. It ensures that many ethical viewpoints are considered. This wider view is helpful in dealing with moral problems and tough decisions AI systems may run into, by making sure that all important considerations are thought over in a balanced way. AIs using moral parliaments are less likely to ignore values that matter to different groups in society.

Further, the moral parliament allows the representation of human values to grow and change over time. We know that moral views change over time, so we should be humble about how much we know about morality: there might be important things we don't yet understand that could help us get closer to the truth about what is right and wrong. By regularly using moral parliaments, AI systems can keep up with current human values, rather than sticking to the old values that were defined when the AI was created. This keeps AI up-to-date and flexible, and prevents it from acting based on outdated or irrelevant values that are locked into the system.

Challenges. Deciding which ethical theories to include in the moral parliament could be a challenging task. There are numerous ethical frameworks, and selecting a representative set may be subjective and politically charged. The decision procedure used to assign appropriate weights to different ethical theories and aggregate their recommendations in ways that reflect their importance could also be contentious. Different stakeholders may have varying opinions on how to prioritize these theories. Moreover, ethical theories can be subject to interpretation and may have nuanced implications. Advanced AI systems would need to be able to accurately understand and apply these theories in order to use a moral parliament.

Conclusions About Moral Parliament

Summary. In this section, we explored using AIs to operationalize moral parliaments, whether in the original form of representing moral theories or generalized to representing stakeholders for any given issue. We highlighted the advantages of using a moral parliament, such as reducing the risk of overlooking or locking in certain values, allowing for the representation of changing human values over time, and increasing transparency and accountability in AI decision-making. We also noted that moral parliaments encourage compromise and negotiation, leading to more balanced and nuanced ethical decisions.

AIs might use moral parliaments for redundancy and adaptability. Especially if we are unsure about what comprises human wellbeing and how we should best distribute it, we might want to embed redundancy and adaptability into our AIs so that they can act ethically even if we make mistakes or change our minds. Given the potential benefits and the ability of moral parliaments to address key concerns in AI decision-making, we should be optimistic about their future use. By incorporating diverse perspectives of individual moral theories or stakeholders in real-time, moral parliaments can help ensure that AI systems act in accordance with human values and avoid extreme or biased behavior, even if human values change over time.

C.6 Conclusion

Ethics is the study of moral principles and how they guide our decisions and actions. It encompasses questions of right and wrong, and it provides a framework for making choices based on our values and beliefs. It is important for AI researchers to have a basic understanding of ethics in order to ethically guide the development and governance of AI systems.

We examined a number of considerations that commonly enter into moral decision making. Intrinsic goods, like wellbeing, are valuable for their own sake. Consequentialist moral theories consider the maximization of intrinsic goods to be our principle moral responsibility. Utilitarianism, a form of consequentialism, is the view that we should maximize wellbeing.

Constraints also play a role in moral decision making. Constraints are the basis of many deontological theories. Kant's ethics categorically constrains actions and is derived from the respect for humanity's rational autonomy.

Virtue ethics focuses on the importance of developing certain character traits rather than solely considering consequences or specific rules. Aristotelian virtue ethics links virtue with flourishing; living a good life is intertwined with possessing virtuous traits.

Rawls' ethics, prioritarianism, and contractualism are all forms of social contract theory. Social contract theory suggests that morality can be derived from hypothetical agreements between members of society, made for everyone's mutual benefit. Rawls proposes that decisions about a social contract should be made behind a veil of ignorance, where individuals are unaware of their personal attributes. From this position, principles such as protecting the worst-off, ensuring basic liberties for all, and limiting inequality can potentially be argued.

As discussed in the chapter on Machine Ethics, we need to ensure that AIs behave ethically towards others. As AIs have increasing influence over society and act more autonomously, it will become important that they are able to detect situations where the moral principles apply, assess how to apply the moral principles, evaluate the moral worth of candidate actions, select and carry out actions appropriate for the context, monitor the success or failure of the actions, and adjust responses accordingly. Ideally, AIs would be designed to take into account many of the ethical concepts introduced in this chapter. Examples of desirable abilities from this perspective might include: -

representing various purported intrinsic goods, including pleasure, autonomy, the exercise of reason, knowledge, friendship, love, and so on - distinguishing between subtly different levels of these goods, and ensuring that the AI's value functions are not vulnerable to optimizers - representing more than just intrinsic goods, for example legal systems and normative factors including special obligations and deontological constraints

Because there is no consensus about which moral theory (if any) is correct, we must accommodate some degree of moral uncertainty. To navigate moral uncertainty, we might consider multiple perspectives when making moral decisions. In the context of AI development, moral uncertainty becomes especially significant due to the wide-ranging societal implications of AI systems. Designers and developers of AI technologies must carefully consider the ethical and moral implications of their actions and ensure that ethical considerations are integrated into the development process.

Being aware of moral uncertainty can help AI developers avoid negative outcomes. Uncertainty highlights complexity of moral judgments and the potential for unexpected consequences. To address moral uncertainty effectively, AI systems should be capable of recognizing and acknowledging a broad range of moral perspectives. This presents challenges in determining how to rationally balance the recommendations of different ethical theories and ensure alignment with human values.

There is no known solution to all ethical dilemmas, and the choice of approach may depend on context, the level of uncertainty, and personal preferences. It is crucial to incorporate diverse moral perspectives, quantify uncertainty, and employ strategies like maximizing expected choice-worthiness and moral parliament. By doing so, we can work to ensure AI systems act in accordance with our values and mitigate unintended harm.

C.7 Literature

C.7.1 Recommended Reading

- Steven Lukes. *Moral Relativism.* Picador, 2008

- Katarzyna de Lazari-Radek and Peter Singer. *Utilitarianism: A Very Short Introduction.* Oxford University Press, 2017. ISBN: 9780198728795. DOI: 10.1093/actrade/9780198728795.001.0001. URL: https://doi.org/10.1093/actrade/9780198728795.001.0001

- Shelly Kagan. *Normative Ethics.* Routledge, London, 2018

- Toby Newberry and Toby Ord. *The Parliamentary Approach to Moral Uncertainty.* Tech. rep. Technical Report# 2021-2, Future of Humanity Institute, University of Oxford ..., 2021

- Stephen Darwall. *Deontology.* Blackwell Publishers, Oxford, 2003

- Aristotle. “Nicomachean Ethics”

- Stanford Encyclopedia of Philosophy (SEP) on virtue ethics https://plato.stanford.edu/entries/ethics-virtue/

- T. M. Scanlon. *What We Owe to Each Other.* Belknap Press, 2000

C.7.2 References

[1] Chris Gowans. "Moral Relativism". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2021. Metaphysics Research Lab, Stanford University, 2021.

[2] Oliver Scott Curry, Daniel Austin Mullins, and Harvey Whitehouse. "Is It Good to Cooperate?: Testing the Theory of Morality-as-Cooperation in 60 Societies". In: *Current Anthropology* 60.1 (2019), pp. 47–69. DOI: 10.1086/701478. eprint: https://doi.org/10.1086/701478. URL: https://doi.org/10.1086/701478.

[3] C. Plato. *Euthyphro*. Kessinger Publishing, 2014.

[4] Jeff Sebo. *Op-Ed: What should we do if a chatbot has thoughts and feelings?* 2022. URL: https://www.latimes.com/opinion/story/2022-06-16/artificial-intelligence-morals-ethics-sentience-thinking.

[5] Andrew Fisher. *Metaethics: An Introduction*. Routledge, 2011.

[6] John Stuart Mill and Jeremy Bentham. *Utilitarianism and Other Essays*. Penguin Books, 1987.

[7] Katarzyna de Laari-Radek and Peter Singer. *Utilitarianism: A Very Short Introduction*. Oxford University Press, 2017.

[8] Jeremy Bentham and Louis Crompton. "Offences Against One's Self:" in: *Journal of Homosexuality* 3.4 (1978), pp. 389–406. DOI: 10.1300/J082v03n04_07. eprint: https://doi.org/10.1300/J082v03n04_07. URL: https://doi.org/10.1300/J082v03n04_07.

[9] Peter Singer. "Famine, affluence, and morality". English (US). In: *Applied Ethics*. United States: Taylor and Francis, July 2017, pp. 132–142. ISBN: 9781138936928. DOI: 10.4324/9781315097176.

[10] Peter Singer. *The Expanding Circle: Ethics and Sociobiology*. Farrar, Straus and Giroux, 1981.

[11] Samuel Scheffler. *The Rejection of Consequentialism: A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions*. Oxford University Press, 1994. ISBN: 9780198235118. DOI: 10.1093/0198235119.001.0001. URL: https://doi.org/10.1093/0198235119.001.0001.

[12] R. Eugene Bales. "Act-Utilitarianism: Account of Right-Making Characteristics or Decision-Making Procedure?" In: *American Philosophical Quarterly* 8.3 (1971), pp. 257–265. ISSN: 00030481. URL: http://www.jstor.org/stable/20009403.

[13] Robert Nozick. *Anarchy State and Utopia*. John Wiley & Sons, 1974.

[14] Joshua Greene. *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. New York: Penguin Press, 2013.

[15] Stephen L. Darwall, ed. *Deontology*. Malden, MA: Wiley-Blackwell, 2003.

[16] T. Aquinas, T.F.E.D. Province, and C.W. Publishing. *The Summa Theologica: Complete Edition*. Catholic Way Publishing, 2014. ISBN: 9781783793143. URL: https://books.google.com.au/books?id=Ee0HBAAAQBAJ.

[17] Philippa Foot. *Virtues and vices and other essays in moral philosophy*. Berkeley: University of California Press, 1978.

[18] Michael Moore. "The Rationality of Threshold Deontology". In: *Moral Puzzles and Legal Perplexities: Essays on the Influence of Larry Alexander*. Cambridge University Press, 2019, pp. 371–387.

[19] Immanuel Kant. *Groundwork for the Metaphysics of Morals*. Ed. by Thomas E. Hill and Arnulf Zweig. New York: Oxford University Press, 1998.

[20] Rosalind Hursthouse and Glen Pettigrove. "Virtue Ethics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2023. Metaphysics Research Lab, Stanford University, 2023.

[21] Roger Crisp, ed. *Aristotle: Nicomachean Ethics*. Cambridge University Press, 2014.

[22] John Rawls. "Applied Ethics: A Multicultural Approach". In: Routledge, 2017. Chap. A Theory of Justice.

[23] John C. Harsanyi. “Cardinal Utility in Welfare Economics and in the Theory of Risk-taking”. In: *Journal of Political Economy* 61.5 (1953), pp. 434–435. DOI: 10.1086/257416. eprint: https://doi.org/10.1086/257416. URL: https://doi.org/10.1086/257416.

[24] Derek Parfit. “Equality and Priority”. In: *Ratio* 10.3 (1997), pp. 202–221. DOI: https://doi.org/10.1111/1467-9329.00041. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9329.00041. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9329.00041.

[25] T. M. Scanlon. *What We Owe to Each Other*. Harvard University Press, 1998. ISBN: 9780674950894. (Visited on 10/13/2023).

[26] William MacAskill, Krister Bykvist, and Toby Ord. *Moral Uncertainty*. Sept. 2020. ISBN: 9780198722274. DOI: 10.1093/oso/9780198722274.001.0001.

[27] Harry R. Lloyd. *The Property Rights Approach to Moral Uncertainty*. Happier Lives Institute’s 2022 Summer Research Fellowship. 2022. URL: https://www.happierlivesinstitute.org/report/property-rights/.

[28] Toby Newberry and Toby Ord. *The Parliamentary Approach to Moral Uncertainty*. Tech. rep. Technical Report# 2021-2, Future of Humanity Institute, University of Oxford . . ., 2021.

[29] Steven Lukes. *Moral Relativism*. Picador, 2008.

[30] Katarzyna de Lazari-Radek and Peter Singer. *Utilitarianism: A Very Short Introduction*. Oxford University Press, 2017. ISBN: 9780198728795. DOI: 10.1093/actrade/9780198728795.001.0001. URL: https://doi.org/10.1093/actrade/9780198728795.001.0001.

[31] Shelly Kagan. *Normative Ethics*. Routledge, London, 2018.

[32] Stephen Darwall. *Deontology*. Blackwell Publishers, Oxford, 2003.

[33] T. M. Scanlon. *What We Owe to Each Other*. Belknap Press, 2000.