CHAPTER 7

Collective Action Problems

7.1 MOTIVATION

Introduction

In the chapters "Complex Systems" and "Safety Engineering," we considered AI risks that arise not only from the technologies themselves but from the broader social contexts in which they are embedded. In this chapter, we extend our exploration of these systemic risks by exploring how the collective behavior of a *multi-agent system* may not reflect the interests of the agents that comprise it. The agents may produce conditions that none of them wants, even when every one of them has the same goals and priorities. In the words of economics Nobel laureate Thomas Schelling, "Micromotives do not equal macrobehavior" [364]. Let us explore this idea using some examples.

Example: traffic jams. Consider a traffic jam, where the only obstacle to each motorist is the car in front. Everyone has the same goal, which is to reach their destination quickly. Since nobody wants to be stuck waiting, the solution might appear obvious to someone unfamiliar with traffic: everyone should simply drive forward, starting at the same time and accelerating at the same rate. And yet, without external synchronization, achieving this preferable state is impossible. All anyone can do is start and stop in response to each others' starting and stopping, inching towards their destination slowly and haltingly.

Example: tall forests [365]. In the Rockefeller forest of Northern California, the trees are more than 350 feet tall, on average. We can model these trees as agents competing for sunlight access. The taller a tree is, the more sunlight it can access, as its leaves are above its neighbors'. However, there is no benefit to being tall other than avoiding being overshadowed by other trees. In fact, growing so tall costs each tree valuable resources and risks their structural integrity failing. If all the trees were 200 feet shorter, each tree would occupy the same position in the competition as they do and each would get the same amount of sunlight as they do, but with greatly reduced growing costs. All the trees would profit from this arrangement. However, as there is

no way to impose such an agreement between the trees, each races its neighbor ever higher, and all pay the large costs of growing so tall.

Example: excessive working hours [366]. People often work far longer hours than they might ideally like to, rather than taking time off for their other interests, in order to be competitive in their field. For instance, they might be competing for limited prestigious positions within their field. In theory, if everyone in a given field were to reduce their work hours by the same amount, they could all free up time and increase their quality of life while maintaining their relative position in the competition. Each person would get the work outcome they would have otherwise, and everyone would benefit from this freed-up time. Yet no one does this, because if they alone were to decrease their work efforts, they would be out-competed by others who did not.

Example: military arms races. Like tree height, the major benefit of military power is not intrinsic, but relative: being less militarily capable than their neighbors makes a nation vulnerable to invasion. This competitive pressure drives nations to expend vast sums of money on their military budgets each year, reducing each nation's budgets for other areas, such as healthcare and education. Some forms of military investment, such as nuclear weaponry and military AI applications, also exacerbate the risks of large-scale catastrophes. If every nation were to decrease its military investment by the same amount, everyone would benefit from the reduced expenses and risks without anyone losing their relative power. However, this arrangement is not stable, since each nation could improve its security by ensuring its military power exceeds that of its competitors, and each risks becoming vulnerable if it alone fails to do this. Military expenditure therefore remains high in spite of these seemingly avoidable costs.

Competitive and evolutionary pressures. The same basic structure underlies most of these examples [367]. A group of agents is engaged in a competition over a valuable and limited item (sunlight access, housing quality, military security). One way an agent can gain more of this valuable item is by sacrificing some of their other values (energy for growth, social life, an education budget). Agents who do not make these sacrifices are outcompeted by those who do. Natural selection weeds out those who do not sacrifice their other values sufficiently, replacing them with agents who sacrifice more, until the competition is dominated by those agents who sacrificed the most. These agents gain no more of the valued item they are competing for than did the original group, yet are worse off for the losses of their other values.

Steering each agent ≠ steering the system. These phenomena hint at the distinct challenges of ensuring safety in multi-agent systems. The danger posed by a collective of agents is greater than the sum of its parts. AI risk cannot be eradicated by merely ensuring that each individual AI agent is loyal and each individual human operator is well-intentioned. Even if all agents, both human and AI, share a common set of goals, this does not guarantee macrobehavior in line with these goals. The agents' *interactions* can produce undesirable outcomes.

Chapter focus. In this chapter, we use abstract models to understand how intelligent agents can, despite acting rationally and in accordance with their own self-interest, collectively produce outcomes that none of them wants, even when they could seemingly have achieved preferable alternative outcomes. We can characterize these risks by crudely differentiating them into the following two sets:

- **Multi-human dynamics.** These risks are generated by interactions between the human agencies involved in AI development and adoption, particularly corporations and nations. The central concern here is that competitive and evolutionary pressures could drive humanity to hand over increasing amounts of power to AIs, thereby becoming a "second-class species." The frameworks we explore in this chapter are highly abstract and can be useful in thinking more generally about the current AI landscape.
  Of particular importance are *racing dynamics*. We see these in the corporate world, where AI developers may cut corners on safety in order to avoid being outcompeted by one another. We also see these in international relations, where nations are racing each other to adopt hazardous military AI applications. By observing AI races, we can anticipate that merely persuading these parties that their actions are high-risk may not be sufficient for ensuring that they act more cautiously, because they may be willing to tolerate high risk levels in order to "stay in the race." For example, nations may choose to continue investing in military AI technologies that could fail in catastrophic ways, if abstaining from doing so risks losing international conflict.
- **Multi-AI dynamics.** These risks are generated by interactions with and between AI agents. In the future, we expect that AIs will increasingly be granted autonomy in their behavior, and will therefore interact with others under progressively less human oversight. This poses risks in at least three ways. First, evolutionary pressures may promote selfish behavior and generate various forms of intrasystem conflict that could subvert our goals. Second, many of the mechanisms by which AI agents may cooperate with one another could promote undesirable behaviors, such as nepotism, outgroup hostility, and the development of ruthless reputations. Third, AIs may engage in conflict, using threats of extreme scale in order to extort others, or even promoting all-out warfare, with devastating consequences.

We explore both of the above sets of multi-agent risks using generalizable frameworks from game theory, bargaining theory, and evolutionary theory. These frameworks help us understand the collective dynamics that can lead to outcomes that were not intended or desired by anyone individually. Even if AI systems are fully under human control and leading actors such as corporations and states are well-intentioned, humanity could still end up eroding away our power gradually until it cannot be recovered.

Game Theory

Rational agents will not necessarily secure good outcomes. Behavior that is individually rational and self-interested can produce collective outcomes that

are suboptimal, or even catastrophic, for all involved. This section first examines the Prisoner's Dilemma, a canonical game theoretic example that illustrates this theme—though cooperation would produce an outcome that is better for both agents, for either one to cooperate would be irrational.

We then build on this by introducing two additional levels of sophistication. The first addition is time. We explore how cooperation is possible, though not assured, when agents interact repeatedly over *time*. The second addition is the introduction of *more than two agents*. We explore how collective action problems can generate and maintain undesirable states. Ultimately, we see how these natural dynamics can produce catastrophically bad outcomes. They perpetuate military arms races and corporate AI races, increasing the risks from both. They may also promote dangerous AI behaviors, such as extortion.

Cooperation

Cooperation is necessary, but not sufficient, for multi-AI agent safety. In this section, we turn to assessing how cooperation can help with addressing the challenges outlined above. However, we also consider what problems cooperation may pose itself, in the context of AI. We explore five mechanisms that can promote or maintain cooperation:

- *Direct reciprocity*: the chance of a future meeting incentivizes cooperative behavior in the present.
- *Indirect reciprocity*: cooperative reputations are rewarded.
- *Group selection*: inter-group competition promotes intra-group unity.
- *Kin selection*: indirect benefits of cooperation outweigh direct costs, motivating altruism towards genetic kin.
- *Institutions*: large-scale external forces motivate cooperation through enforcement.

Conflict

Rational agents may sometimes choose destructive conflict instead of peaceful bargaining. This section explores the nature of conflict between agents. We start with an overview of bargaining theory, which provides a lens for understanding why rational agents sometimes choose mutually-costly conflict over peaceful alternatives. We next explore several specific factors that drive conflict.

1. *Power shifts*: a shift in political power triggers preventative conflict.
2. *First-strike advantage*: time-sensitive offensive advantages motivate a party to initiate conflict preemptively.
3. *Issue indivisibility*: wherever the entity over which parties are contesting is indivisible, it is harder to avoid resorting to conflict.
4. *Information problems*: mis- and dis-information kindle defensive or offensive action over cooperation.

5. *Inequality*: inequality may increase the probability of conflict, due to factors such as relative deprivation and social envy.

Evolutionary Pressure

Natural selection will promote AIs that behave selfishly. In this final section, we use evolutionary theory to study what happens when a large number of agents interact many times over many generations. We start with generalized Darwinism: the idea that evolution by natural selection can take place outside of the realm of biology. We explore examples in linguistics, music, philosophy and sociology. We formalize generalized Darwinism using Lewontin's conditions for evolution by natural selection and the Price equation for evolutionary change. Using both, we show that AIs are likely to be subject to evolution by natural selection: they will vary in ways that produce differential fitness and so influence which traits persist through time and between "generations" of AIs.

Next, we explore two AI risks generated by evolutionary pressures. The first is that correctly-specified goals may be subverted or distorted by "intrasystem goal conflict." The interests of propagating information (such as genes, departments, or sub-agents) can sometimes clash with those of the larger entity that contains it (such as an organism, government, or AI system), undermining unity of purpose. The second risk we consider is that natural selection tends to favor selfish traits over altruistic ones. A future shaped by evolutionary pressures is, therefore, likely to be dominated by selfish behavior, both in the institutions that produce and use AI systems, and in the AIs themselves.

The conclusions of this section are simple. Natural selection will by default be a strong force in determining the state of the world. Its influence on AI development carries the risk of intrasystem goal conflict and the promotion of selfish behavior. Both risks could have catastrophic effects. Intrasystem goal conflict could prevent our goals from being carried out and generate unexpected actions. AI agents could develop selfish tendencies, increasing the risk that they might employ harmful strategies (including those covered earlier in the chapter, such as extortion).

7.2 GAME THEORY

7.2.1 Overview

This chapter explores the dynamics generated by the interactions of multiple agents, both human and AI. These interactions create risks distinct from those generated by any individual AI agent acting in isolation. One way we can study the strategic interdependence of agents is with the framework of *game theory*. Using game theory, we can examine formal models of how agents interact with each other under varying conditions and predict the outcomes of these interactions.

Here, we use game theory to present natural dynamics in biological and social systems that involve multiple agents. In particular, we explore what might cause agents to

come into conflict with one another, rather than cooperate. We show how these multi-agent dynamics can generate undesirable outcomes, sometimes for all the agents involved. We consider risks created by interactions within and between human and AI agents, from human-directed companies and militaries engaging in perilous races to autonomous AIs using threats for extortion.

We start with an overview of the fundamentals of game theory. We begin this section by setting out the characteristics of game theoretic agents. We also categorize the different kinds of games we are exploring.

We then focus on the Prisoner's Dilemma. The Prisoner's Dilemma is a simple example of how an interaction between two agents can generate an equilibrium state that is bad for both, even when each acts rationally and in their own self-interest. We explore how agents may arrive at the outcome where neither chooses to cooperate. We use this to model real-world phenomena, such as negative political campaigns. Finally, we examine ways we might foster rational cooperation between self-interested AI agents, such as by altering the values in the underlying payoff matrices. The key upshot is that intelligent and rational agents do not always achieve good outcomes.

We next add in the element of time by examining the Iterated Prisoner's Dilemma. AI agents are unlikely to interact with others only once. When agents engage with each other multiple times, this creates its own hazards. We begin by examining how iterating the Prisoner's Dilemma alters the agents' incentives—when an agent's behavior in the present can influence that of their partner in the future, this creates an opportunity for rational cooperation. We study the effects of altering some of the variables in this basic model: uncertainty about future engagement and the necessity to switch between multiple different partners. We look at why the cooperative strategy *tit-for-tat* is usually so successful, and in what circumstances it is less so. Finally, we explore some of the risks associated with iterated multi-agent social dynamics: corporate AI races, military AI arms races, and AI extortion. The key upshot is that cooperation cannot be ensured merely by iterating interactions through time.

We next move to consider group-level interactions. AI agents might not interact with others in a neat, pairwise fashion, as assumed by the models previously explored. In the real world, social behavior is rarely so straightforward. Interactions can take place between more than two agents at the same time. A group of agents creates an environmental structure that may alter the incentives directing individual behavior. Human societies are rife with dynamics generated by group-level interactions that result in undesirable outcomes. We begin by formalizing "collective action problems." We consider real-world examples such as anthropogenic climate change and fishery depletion. Multi-agent dynamics such as these generate AI risk in several ways. Races between human agents and agencies could trigger flash wars between AI agents or the automation of economies to the point of human enfeeblement. The key upshot is that achieving cooperation and ensuring collectively good outcomes is even more difficult in interactions involving more than two agents.

7.2.2 Game Theory Fundamentals

In this section, we briefly run through some of the fundamental principles of game theory. Game theory is the branch of mathematics concerned with agents' choices and strategies in multi-agent interactions. Game theory is so-called because we reduce complex situations to abstract games where agents maximize their payoffs. Using game theory, we can study how altering incentives influences the strategies that these agents use.

Agents in game theory. We usually assume that the agents in these games are self-interested and rational. Agents are "self-interested" if they make decisions in view of their own utility, regardless of the consequences to others. Agents are said to be "rational" if they act as though they are maximizing their utility.

Games can be "zero sum" or "non-zero sum." We can categorize the games we are studying in different ways. One distinction is between zero sum and non-zero sum games. A **zero sum** game is one where, in every outcome, the agents' payoffs all sum to zero. An example is "tug of war": any benefit to one party from their pull is necessarily a cost to the other. Therefore, the total value of these wins and losses cancel out. In other words, there is never any net change in total value. Poker is a zero sum game if the players' payoffs are the money they each finish with. The total amount of money at a poker game's beginning and end is the same — it has simply been redistributed between the players.

By contrast, many games are non-zero sum. In *non-zero* sum games, the total amount of value is not fixed and may be changed by playing the game. Thus, one agent's win does not necessarily require another's loss. For instance, in cooperation games such as those where players must meet at an undetermined location, players only get the payoff together if they manage to find each other. As we shall see, the Prisoner's dilemma is a non-zero sum game, as the sum of payoffs changes across different outcomes.

Non-zero sum games can have "positive sum" or "negative sum" outcomes. We can categorize the outcomes of non-zero sum games as *positive sum* and *negative sum*. In a positive sum outcome, the total gains and losses of the agents sum to greater than zero. Positive sum outcomes can arise when particular interactions result in an increase in value. This includes instances of mutually-beneficial cooperation. For example, if one agent has flour and another has water and heat, the two together can cooperate to make bread, which is more valuable than the raw materials. As a real-world example, many view the stock market as positive sum because the overall value of the stock market tends to increase over time. Though gains are unevenly distributed, and some investors lose money, the average investor becomes richer. This demonstrates an important point: positive sum outcomes are not necessarily "win-win." Cooperating does not guarantee a benefit to all involved. Even if extra total value is created, its distribution between the agents involved in its creation can take any shape, including one where some agents have negative payoffs.

In a negative sum outcome, some amount of value is lost by playing the game. Many competitive interactions in the real world are negative sum. For instance, consider "oil wars"—wars fought over a valuable hydrocarbon resource. Oil wars are zero-sum with regards to oil since only the distribution (not the amount) of oil changes. However, the process of conflict itself incurs costs to both sides, such as loss of life and infrastructure damage. This reduces the total amount of value. If AI development has the potential to result in catastrophic outcomes for humanity, then accelerating development to gain short-term profits in exchange for long-term losses to everyone involved would be a negative sum outcome.

7.2.3 The Prisoner's Dilemma

Our aim in this section is to investigate how interactions between rational agents, both human and AI, may negatively impact everyone involved. To this end, we focus on a simple game: the Prisoner's Dilemma. We first explore how the game works, and its different possible outcomes. We then examine why agents may choose not to cooperate even if they know this will lead to a collectively suboptimal outcome. We run through several real-world phenomena which we can model using the Prisoner's Dilemma, before exploring ways in which cooperation can be promoted in these kinds of interactions. We end by briefly discussing the risk of AI agents tending towards undesirable equilibrium states.

The Game Fundamentals

In the Prisoner's Dilemma, two agents must each decide whether or not to cooperate. The costs and benefits are structured such that for each agent, defection is the best strategy regardless of what their partner chooses to do. This motivates both agents to defect.

The Prisoner's Dilemma. In game theory, the *Prisoner's Dilemma* is a classic example of the decisions of rational agents leading to suboptimal outcomes. The basic setup is as follows. The police have arrested two would-be thieves. We will call them Alice and Bob. The suspects were caught breaking into a house. The police are now detaining them in separate holding cells, so they cannot communicate with each other. The police suspect that the pair were planning *burglary* (which carries a lengthy jail sentence). But they only have enough evidence to charge them with *trespassing* (which carries a shorter jail sentence). However, the testimony of either one of the suspects would be enough to charge the other with burglary, so the police offer each suspect the following deal. If only one of them rats out their partner by confessing that they had intended to commit burglary, the confessor will be released with *no jail time* and their partner will spend *eight years* in jail. However, if they each attempt to rat out the other by both confessing, they will both serve a medium prison sentence of *three years*. If neither suspect confesses, they will both serve a short jail sentence of only *one year*.

The four possible outcomes. We assume that Alice and Bob are both rational and self-interested: each only cares about minimizing their own jail time. We define the decision facing each as follows. They can either "cooperate" with their partner by remaining silent or "defect" on their partner by confessing to burglary. Each suspect faces four possible outcomes, which we can split into two possible scenarios. Let's term these "World 1" and "World 2"; see Figure 7.1. In World 1, their partner chooses to cooperate with them; in World 2, their partner chooses to defect. In both scenarios, the suspect decides whether to cooperate or defect themself. They do not know what their partner will decide to do.

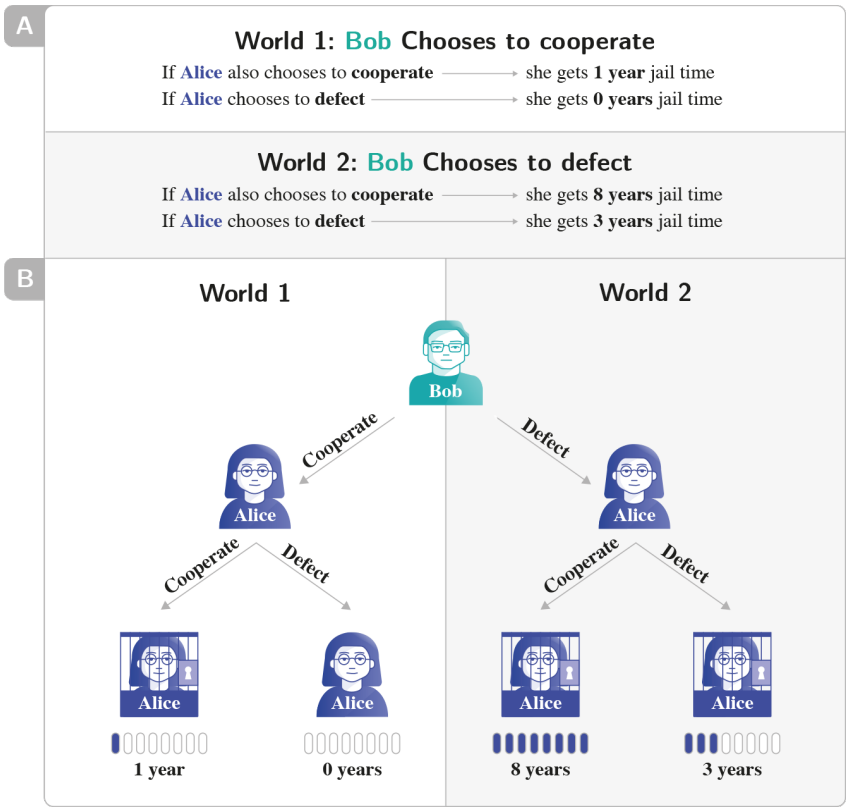

FIGURE 7.1. The possible outcomes for Alice in the Prisoner's Dilemma.

Defection is the dominant strategy. Alice does not know whether Bob will choose to cooperate or defect. She does not know whether she will find herself in World 1 or World 2; see Figure 7.1. She can only decide whether to cooperate or defect herself. This means she is making one of two possible decisions. If she defects, she is...

> ...in World 1: Bob cooperates and she goes free instead of spending a year in jail.
> ...in World 2: Bob defects and she gets a 3-year sentence instead of an 8-year one.

Alice only cares about minimizing her own jail time, so she can save herself jail time in either scenario by choosing to defect. She saves herself one year if her partner cooperates or five years if her partner defects. A rational agent under these circumstances will do best if they decide to defect, regardless of what they expect their partner to do. We call this the *dominant strategy*: a rational agent playing the Prisoner's Dilemma should choose to defect *no matter what their partner does*.

One way to think about strategic dominance is through the following thought experiment. Someone in the Arctic during winter is choosing what to wear for that day's excursion. They have only two options: a coat or a t-shirt. The coat is thick and waterproof; the t-shirt is thin and absorbent. Though this person cannot control or predict the weather, they know there are only two possibilities: either rain or cold. If it rains, the coat will keep them drier than the t-shirt. If it is cold, the coat will keep them warmer than the t-shirt. Either way, the coat is the better option, so "wearing the coat" is their dominant strategy.

Defection is the dominant strategy for both agents. Importantly, both the suspects face this decision in a symmetric fashion. Each is deciding between identical outcomes, and each wishes to minimize their own jail time. Let's consider the four possible outcomes now in terms of both the suspects' jail sentences. We can display this information in a *payoff matrix*, as shown in Table 7.1. Payoff matrices are commonly used to visualize games. They show all the possible outcomes of a game in terms of the value of that outcome for each of the agents involved. In the Prisoner's Dilemma, we show the decision outcomes as the payoffs to each suspect: note that since more jail time is worse than less, these payoffs are negative. Each cell of the matrix shows the outcome of the two suspects' decisions as the payoff to each suspect.

TABLE 7.1. Each cell in this payoff matrix represents a payoff. If Alice cooperates and Bob defects, the top right cell tells us that Alice gets 8 years in jail while Bob goes free.

	Bob cooperates	Bob defects
Alice cooperates	-1, -1	-8, 0
Alice defects	0, -8	-3, -3

Nash Equilibria and Pareto Efficiency

The stable equilibrium state in the Prisoner's Dilemma is for both agents to defect. Neither agent would choose to go back in time and change their decision (to switch to cooperating) if they could not also alter their partner's behavior by doing so. This is often considered counterintuitive, as the agents would benefit if they were both to switch to cooperating.

Nash Equilibrium: both agents will choose to defect. Defection is the best strategy for Alice, regardless of what Bob opts to do. The same is true for Bob. Therefore, if both are behaving in a rational and self-interested fashion, they will both defect. This will secure the outcome of 3 years of jail time each (the bottom-right

outcome of the payoff matrix above). Neither would wish to change their decision, even if their partner were to change theirs. This is known as the *Nash equilibrium*: the strategy choices from which no agent can benefit by unilaterally choosing a different strategy. When interacting with one another, rational agents will tend towards picking strategies that are part of Nash equilibria.

Pareto improvement: both agents would do better if they cooperated. As we can see in the payoff matrix, there is a possible outcome that is better for both suspects. If both choose the cooperate strategy, they will secure the top-left outcome of the payoff matrix. Each would serve 2 years less jail time at no cost to the other. Yet, as we have seen, selecting this strategy is irrational; the *defect* strategy is dominant and so Alice and Bob each want to defect instead. We call this outcome *Pareto inefficient*, meaning that it could be altered to make some of those involved better off without making anyone else worse off. In the Prisoner's Dilemma, the *both defect* outcome is Pareto inefficient because it is suboptimal for both Alice and Bob, who would both be better off if they both cooperated instead. Where there is an outcome that is better for some or all agents involved, and not worse for any, we call the switch to this more efficient outcome a *Pareto improvement*. In the Prisoner's Dilemma, the *both cooperate* outcome is better for both agents than the Nash equilibrium of *both defect*; see Figure 7.2. The only Pareto improvement possible in this game is the move from the *both defect* to the *both cooperate* outcome; see Figure 7.3.

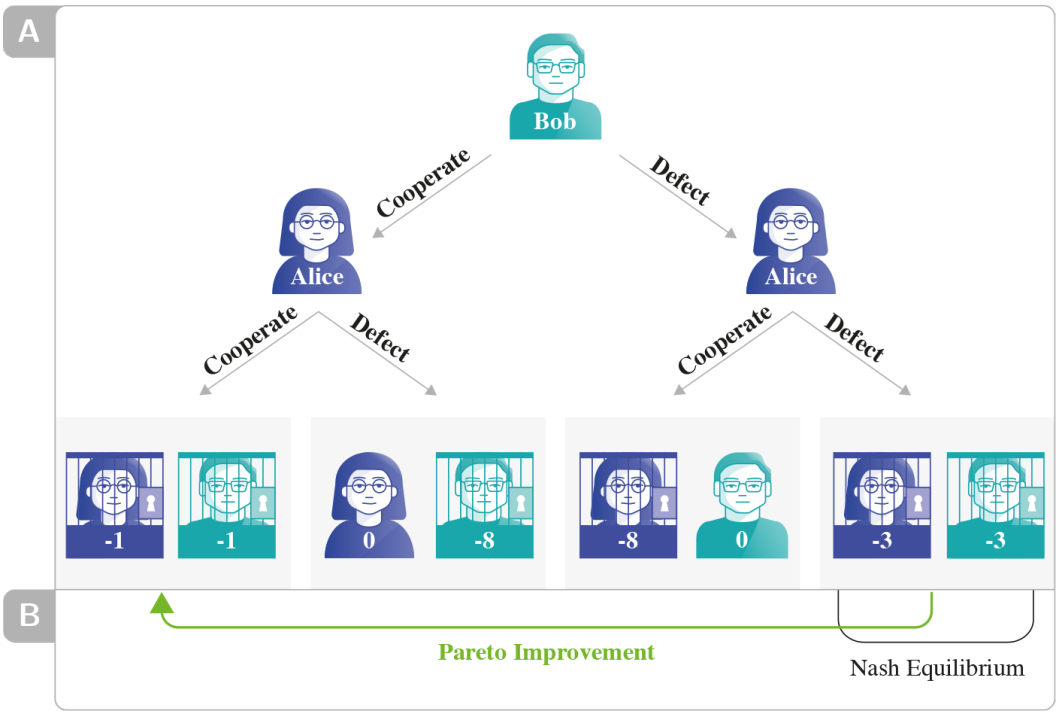

FIGURE 7.2. Looking at the possible outcomes for both suspects in the Prisoner's Dilemma, we can see that there is a possible Pareto improvement from the Nash equilibrium. The numbers represent their payoffs (rather than the length of their jail sentence).

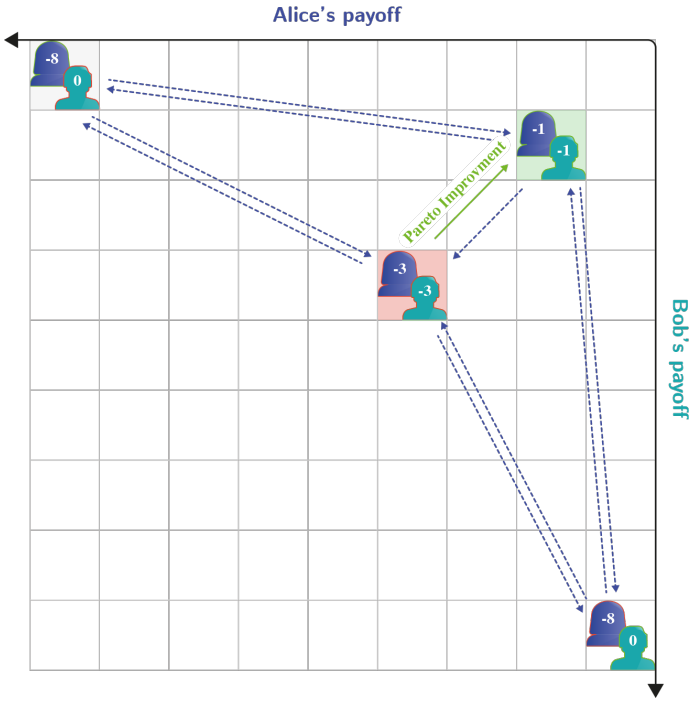

FIGURE 7.3. Both suspects' payoffs, in each of the four decision outcomes. Moving right increases Alice's payoff, and moving up improves Bob's payoff. A Pareto improvement requires moving right and up, as shown by the green arrow [368].

Real-World Examples of the Prisoner's Dilemma

The Prisoner's Dilemma has many simplifying assumptions. Nevertheless, it can be a helpful lens through which to understand social dynamics in the real world. Rational and self-interested parties often produce states that are Pareto inefficient. There exist alternative states that would be better for all involved, but reaching these requires individually irrational action. To illustrate this, let's explore some real-world examples.

Mud-slinging. Consider the practice of mud-slinging. Competing political parties often use negative campaign tactics, producing significant reputational costs. By running negative ads to attack and undermine the public image of their opponents, all parties end up with tarnished reputations. If we assume that politicians value their reputation in an absolute sense, not merely in relation to their contemporary competitors, then mud-slinging is undesirable for all. A Pareto improvement to this situation would be switching to the outcome where they all cooperate. With no one engaging in mud-slinging, all the parties would have better reputations. The reason this does not happen is that mud-slinging is the dominant strategy. If a party's opponent *doesn't* use negative ads, the party will boost their reputation relative to their opponent's by using them. If their opponent *does* use negative ads, the party will reduce the difference between their reputations by using them too. Thus, both parties converge on the Nash equilibrium of mutual mud-slinging, at avoidable detriment to all.

Shopkeeper price cuts. Another example is price racing dynamics between different goods providers. Consider two rival shopkeepers selling similar produce at similar prices. They are competing for local customers. Each shopkeeper calculates that lowering their prices below that of their rival will attract more customers away from the other shop and result in a higher total profit for themselves. If their competitor drops their prices and they do not, then the competitor will gain extra customers, leaving the first shopkeeper with almost none. Thus, "dropping prices" is the dominant strategy for both. This leads to a Nash equilibrium in which both shops have low prices, but the local custom is divided much the same as it would be if they had both kept their prices high. If they were both to raise their prices, they would both benefit by increasing their profits: this would be a Pareto improvement. Note that, just as how the interests of the police do not count in the Prisoner's Dilemma, we are only considering the interests of the shopkeepers in this example. We are ignoring the interests of the customers and wider society.

Arms races. Nations' expenditure on military arms development is another example. It would be better for all these nations' governments if they were all simultaneously to reduce their military budgets. No nation would become more vulnerable if they were all to do this, and each could then redirect these resources to areas such as education and healthcare. Instead, we have widespread military arms races. We might prefer for all the nations to turn some military spending to their other budgets, but for any one nation to do so would be irrational. Here, the dominant strategy for each nation is to opt for high military expenditure. So we achieve a Nash equilibrium in which all nations must decrease spending in other valuable sectors. It would be more Pareto efficient for all to have lower military spending, freeing money and resources for different domains. We will consider races in the context of AI development in the following section.

Promoting Cooperation

So far we have focused on the sources of undesirable multi-agent dynamics in games like the Prisoner's Dilemma. Here, we turn to the mechanisms by which we can promote cooperation over defection.

Reasons to cooperate. There are many reasons why real-world agents might cooperate in situations which resemble the Prisoner's Dilemma [369], as shown in Figure 7.4. These can broadly be categorized by whether the agents have a choice, or whether defection is impossible. If the agents do have a choice, we can further divide the possibilities into those where they act in their own self-interest, and those where they do not (altruism). Finally, we can differentiate two reasons why self-interested agents may choose to cooperate: a tendency toward this, such as a conscience or guilt, and future reward/punishment. We will explore two possibilities in this section — payoff changes and altruistic dispositions — and then "future reward/punishment" in the next section. Note that we effectively discuss "Defection is impossible" in the Single-Agent Safety chapter, and "AI consciences" in the Beneficial AI and Machine Ethics chapter.

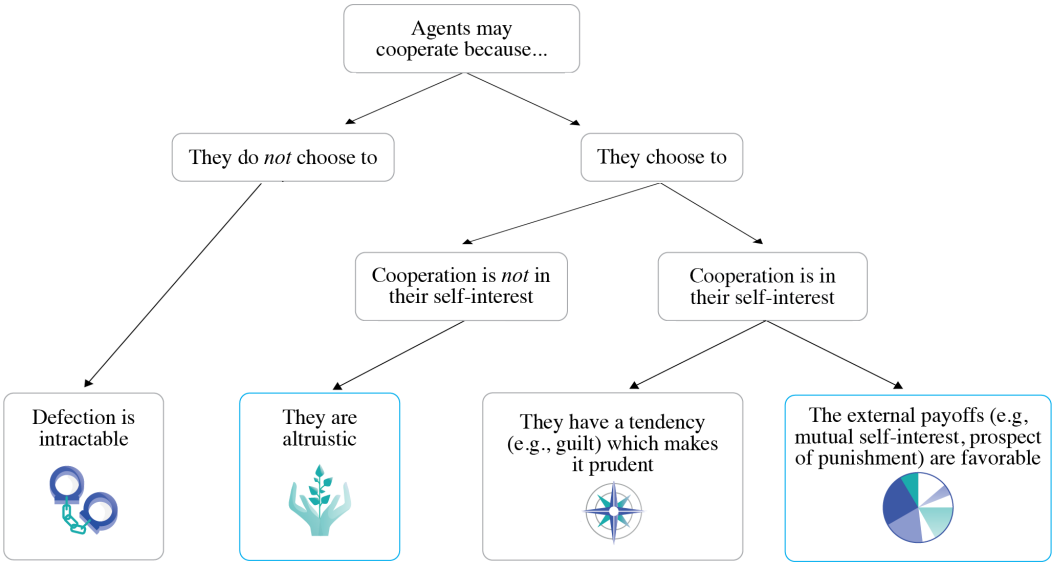

FIGURE 7.4. Four possible reasons why agents may cooperate in prisoner's Dilemma-like scenarios. This section explores two: changes to the payoff matrix and increased agent altruism [369].

External consideration: changing the payoffs to incentivize cooperation. By adjusting the values in the payoff matrix, we may more easily steer agents away from undesirable equilibria. As shown in Table 7.2, incentive structures are important. A Prisoner's Dilemma-like scenario may arise wherever an individual agent will do better to defect whether their partner cooperates ($c > a$) or defects ($d > b$). Avoiding this situation requires altering these constants where they underlie critical social interactions in the real world: changing the costs and benefits associated with different activities so as to encourage cooperative behavior.

TABLE 7.2. if $c > a$ and $d > b$, the highest payoff for either agent is to defect, regardless of what their opponent does: Defection is the dominant strategy. Fostering cooperation requires avoiding this structure.

	Agent B cooperates	Agent B defects
Agent A cooperates	a, a	b, c
Agent A defects	c, b	d, d

There are two ways to reduce the expected value of defection: lower the *probability* of defection success or lower the *benefit* of a successful defection. Consider a strategy commonly used by organized crime groups: threatening members with extreme punishment if they 'snitch' to the police. In the Prisoner's Dilemma game, we can model this by adding a punishment equivalent to three years of jail time for "snitching," leading to the altered payoff matrix as shown in Figure 7.5. The Pareto efficient outcome $(-1, -1)$ is now also a Nash Equilibrium because snitching when the other player cooperates is worse than mutually cooperating ($c < a$).

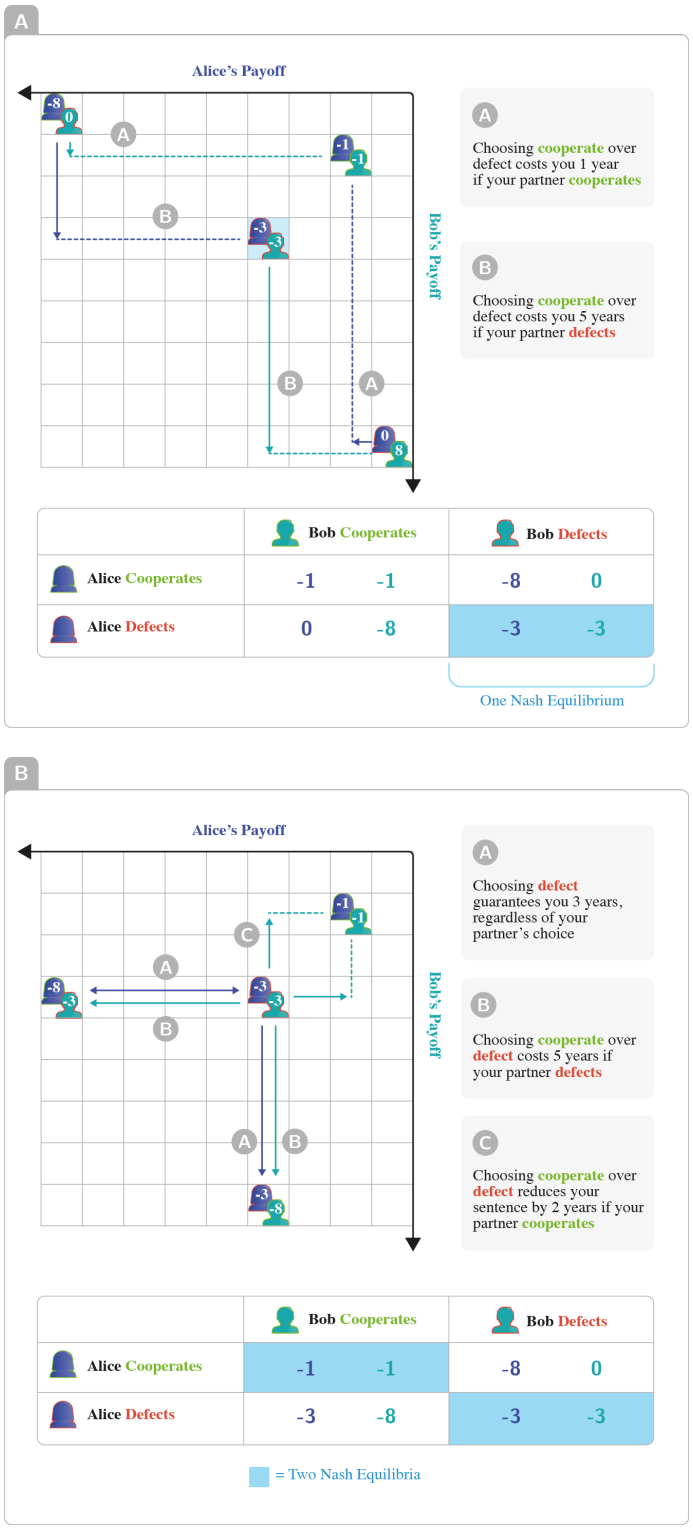

FIGURE 7.5. Altering the payoff matrix to punish snitches, we can move from a Prisoner's Dilemma (left) to a Stag Hunt (right), in which there is an additional Nash equilibrium.

Internal consideration: making agents more altruistic to promote cooperation. A second potential mechanism to foster cooperation is to make agents more altruistic. If each agent also values the outcome for their partner, this effectively changes the payoff matrix. Now, the length of their partner's jail sentence matters to each of them. In the Prisoner's Dilemma payoff matrix, the *both cooperate* outcome earns the lowest total jail time, so agents who valued their partners' payoffs equally to their own would converge on cooperation.

Parallels to AI safety. One possible example of such a strategy would be to target the values held by AI companies themselves. Improving corporate regulation effectively changes the company's expected payoffs from pursuing risky strategies. If successful, it could encourage the company building AI systems to behave in a less purely self-interested fashion. Rather than caring solely about maximizing their shareholder's financial interests, AI companies might cooperate more with each other to steer away from Pareto inefficient outcomes, and avoid corporate AI races. We explore this in more detail in *section 7.2.4 "AI races"* below.

Summary

Cooperation is not always rational, so intelligence alone may not ensure good outcomes. We have seen that rational and self-interested agents may not interact in such a way as to achieve good results, even for themselves. Under certain conditions, such as in the Prisoner's Dilemma, they will converge on a Nash equilibrium of both defecting. Both agents would be better off if they both cooperated. However, it is hard to secure this Pareto improvement because cooperation is not rational when defection is the dominant strategy.

Conflict with or between future AI agents may be extremely harmful. One source of concern regarding future AI systems is inter-agent conflict eroding the value of the future. Rational AI agents faced with a Prisoner's Dilemma-type scenario might end up in stable equilibrium states that are far from optimal, perhaps for all the parties involved. Possible avenues to reduce these risks include restructuring the payoff matrices for the interactions in which these agents may be engaged or altering the agents' dispositions.

7.2.4 The Iterated Prisoner's Dilemma

In our discussion of the Prisoner's Dilemma, we saw how rational agents may converge to equilibrium states that are bad for all involved. In the real world, however, agents rarely interact with one another only once. Our aim in this section is to understand how cooperative behavior can be promoted and maintained as multiple agents (both human and AI) interact with each other over time, when they expect repeated future interactions. We handle some common misconceptions in this section, such as the idea that simply getting agents to interact repeatedly is sufficient to foster cooperation, because "nice" and "forgiving" strategies always win out. As we shall see, things are

not so simple. We explore how iterated interactions can lead to progressively worse outcomes for all.

In the real world, we can observe this in "AI races", where businesses cut corners on safety due to competitive pressures, and militaries adopt and deploy potentially unsafe AI technologies, making the world less safe. These AI races could produce catastrophic consequences, including more frequent or destructive wars, economic enfeeblement, and the potential for catastrophic accidents from malfunctioning or misused AI weapons.

Introduction

Agents who engage with one another many times do not always coexist harmoniously. Iterating interactions is not sufficient to ensure cooperation. To see why, we explore what happens when rational, self-interested agents play the Prisoner' Dilemma game against each other repeatedly. In a single-round Prisoner's Dilemma, defection is always the rational move. But understanding the success of different strategies is more complicated when agents play multiple rounds.

In the Iterated Prisoner's Dilemma, agents play repeatedly. The dominant strategy for a rational agent in a one-off interaction such as the Prisoner's Dilemma is to defect. The seeming paradox is that both agents would prefer the cooperate-cooperate outcome to the defect-defect one. An agent cannot influence their partner's actions in a one-off interaction, but in an iterated scenario, one agent's behavior in one round may influence how their partner responds in the next. We call this the *Iterated Prisoner's Dilemma*; see Figure 7.6. This provides an opportunity for the agents to cooperate with each other.

Iterating the Prisoner's Dilemma opens the door to rational cooperation. In an Iterated Prisoner's Dilemma, both agents can achieve higher payoffs by fostering a cooperative relationship with each other than they would if both were to defect every round. There are two basic mechanisms by which iteration can promote cooperative behavior: punishing defection and rewarding cooperation. To see why, let us follow an example game of the Iterated Prisoner's Dilemma in sequence.

Punishment. Recall Alice and Bob from the previous section, the two would-be thieves caught by the police. Alice decides to defect in the first round of the Prisoner's Dilemma, while Bob opts to cooperate. This achieves a good outcome for Alice, and a poor one for Bob, who punishes this behavior by choosing to defect himself in the second round. What makes this a punishment is that Alice's score will now be lower than it would be if Bob had opted to cooperate instead, whether Alice chooses to cooperate or defect.

Reward. Alice, having been punished, decides to cooperate in the third round. Bob rewards this action by cooperating in turn in the fourth. What makes this a reward is that Alice's score will now be higher than if Bob had instead opted to defect, whether Alice chooses to cooperate or defect. Thus, the expectation that their defection will

be punished and their cooperation rewarded incentivizes both agents to cooperate with each other.

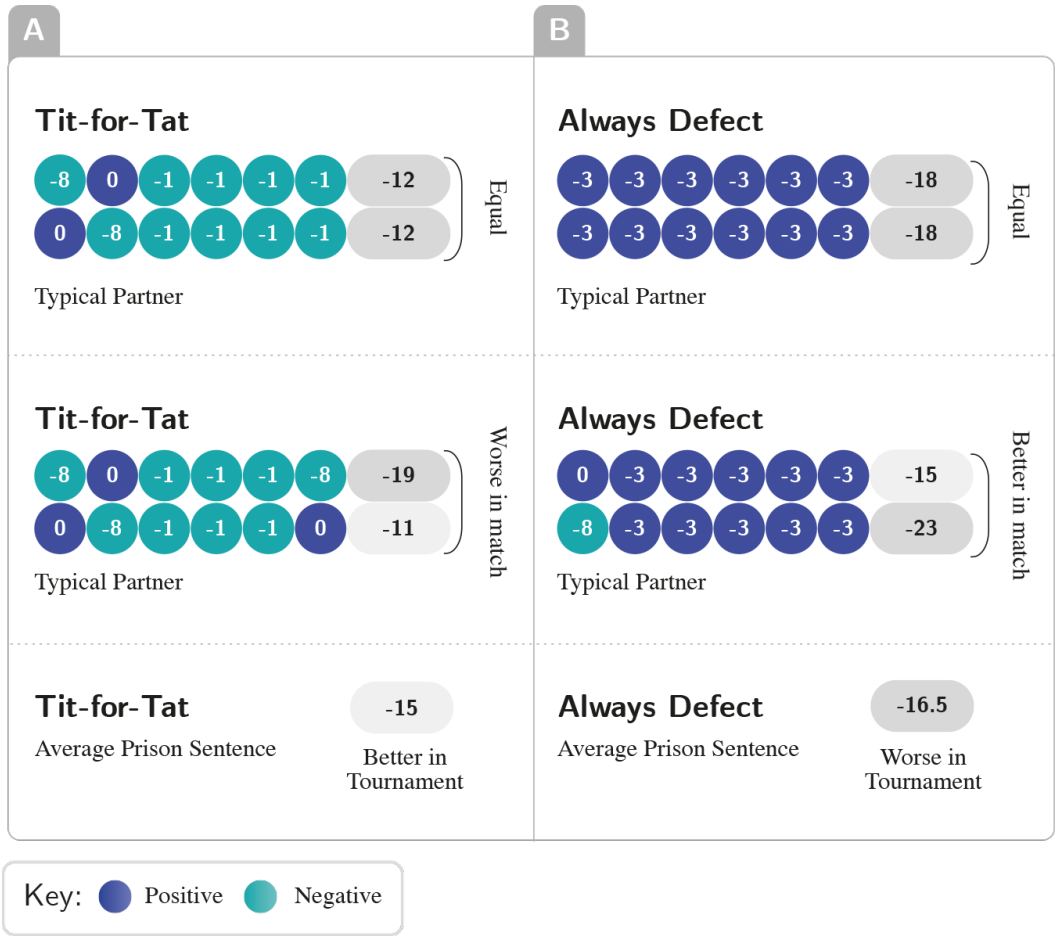

FIGURE 7.6. Across six rounds, both players gain better payoffs if they consistently cooperate. But defecting creates short-term gains.

In Figure 7.6, each panel shows a six-round Iterated Prisoner's Dilemma, with purple squares for defection and blue for cooperation. On the left is Tit-for-tat: An agent using this strategy tends to score the same as or worse than its partners in each match. On the right, always defect tends to score the same as or better than its partner in each match. The average payoff attained by using either strategy are shown at the bottom: Tit-for-tat attains a better payoff (lower jail sentence) on average—and so is more successful in a tournament—than always defect.

Defection is still the dominant strategy if agents know how many times they will interact. If the agents know when they are about to play the Prisoner's Dilemma with each other for the final time, both will choose to defect in that final round. This is because their defection is no longer punishable by their partner. If Alice defects in the last round of the Iterated Prisoner's Dilemma, Bob cannot punish her

by retaliating, as there are no future rounds in which to do so. The same is of course true for Bob. Thus, *defection is the dominant strategy for each agent in the final round*, just as it is in the single-round version of the dilemma.

Moreover, if each agent expects their partner to defect in the final round, *then there is no incentive for them to cooperate in the penultimate round either*. This is for the same reason: Defecting in the penultimate round will not influence their partner's behavior in the final round. Whatever an agent decides to do, they expect that their partner will choose to defect next round, so they might as well defect now. We can extend this argument by reasoning backwards through all the iterations. In each round, the certainty that their partner will defect in the next round regardless of their own behavior in the current round incentivizes each agent to defect. The reward for cooperation and punishment of defection have been removed. Ultimately, this removal pushes the agents to defect in every round of the Iterated Prisoner's Dilemma.

Uncertainty about future engagement enables rational cooperation. In the real world, an agent can rarely be sure that they will never again engage with a given partner. Wherever there is sufficient uncertainty about the future of their relationship, rational agents may be more cooperative. This is for the simple reason that uncooperative behavior may yield less valuable outcomes in the long term, because others may retaliate in kind in the future. This tells us that AIs interacting with each other repeatedly may cooperate, but only if they are sufficiently uncertain about whether their interactions are about to end. Other forms of uncertainty can also create opportunities for rational cooperation, such as uncertainty about what strategies others will use. These are most important where the Iterated Prisoner's Dilemma involves a population of more than two agents, in which each agent interacts sequentially with multiple partners. We turn to examining the dynamics of these more complicated games next.

Tournaments

So far, we have considered the Iterated Prisoner's Dilemma between only two agents: each plays repeatedly against a single partner. However, in the real world, we expect AIs will engage with multiple other agents. In this section, we consider interactions of this kind, where each agent not only interacts with their partner repeatedly, but also switches partners over time. Understanding the success of a strategy is more complicated in repeated rounds against many partners. Note that in this section, we define a "match" to mean repeated rounds of the Prisoner's Dilemma between the same two agents; see Figure 7.6. We define a "tournament" to mean a population of more than two agents engaged in a set of pairwise matches.

In Iterated Prisoner Dilemma tournaments, each agent interacts with multiple partners. In the 1970s, the political scientist Robert Axelrod held a series of tournaments to pit different agents against one another in the Iterated Prisoner's Dilemma. The tournament winner was whichever agent had the highest total payoff after completing all matches. Each agent in an Iterated Prisoner's Dilemma

tournament plays multiple rounds against multiple partners. These agents employed a range of different strategies. For example, an agent using the strategy named *random* would randomly determine whether to cooperate or defect in each round, entirely independently of previous interactions with a given partner. By contrast, an agent using the *grudger* strategy would start out cooperating, but switch to defecting for all future interactions if its partner defected even once. See Table 7.3 for examples of these strategies.

TABLE 7.3. Popular strategies' descriptions.

Strategy	**Characteristics**
Random	Randomly defect or cooperate, regardless of your partner's strategy
Always defect	Always choose to defect, regardless of your partner's strategy
Always cooperate	Always choose to defect, regardless of your partner's strategy
Grudger	Start by cooperating, but if your partner defects, defect in every subsequent round, regardless of your partner's subsequent behavior
Tit-for-tat	Start cooperating; then always do whatever your partner did last
Generous tit-for-tat	Same as *tit-for-tat*, but occasionally cooperate in response to your partner's defection

The strategy "Tit-for-tat" frequently won Axelrod's tournaments [370]. The most famous strategy used in Axelrod's tournaments was *Tit-for-tat*. This was the strategy of starting by cooperating, then repeating the partner's most recent move: if they cooperated, *Tit-for-tat* cooperated too; if they defected, *Tit-for-tat* did likewise. Despite its simplicity, this strategy was extremely successful, and very frequently won tournaments. An agent playing *Tit-for-tat* exemplified the two mechanisms for promoting cooperation, rewarding cooperation, yet also punishing defection. Importantly, *Tit-for-tat* did not hold a grudge—it forgave each defection after it retaliated by defecting in return, only once. This process of one defection for one defection is captured in the famous idiom "an eye for an eye." The *Tit-for-tat* strategy became emblematic as being one way to escape the muck of defection.

The success of Tit-for-tat is counterintuitive. In any given match, an agent playing *Tit-for-tat* will tend to score slightly worse than or the same as their partner; see Figure 7.6a. By contrast, an agent who employs an uncooperative strategy such as *always defect* usually scores the same as or better than its partner; see Figure 7.6b. In a match between a cooperative agent and an uncooperative one, the uncooperative agent tends to end up with the better score.

However, it is an agent's *average* score which dictates its success in a tournament, not its score in any particular match or with any particular partner. Two uncooperative partners will score worse on average than cooperative ones. Thus, the success of cooperative strategies such as 7.6 depends on the population strategy composition (the assortment of strategies used by the agents in the population). If there are enough

cooperative partners, cooperative agents may be more successful than uncooperative ones.

AI Races

Iterated interactions can generate "AI races." We discuss two kinds of races concerning AI development: corporate AI races and military AI arms races. Both kinds center around competing parties participating in races for individual, short-term gains at a collective, long-term detriment. Where individual incentives clash with collective interests, the outcome can be bad for all. As we discuss here, in the context of AI races, these outcomes could even be catastrophic.

AI races are the result of intense competitive pressures. During the Cold War, the US and the Soviet Union were involved in a costly nuclear arms race. The effects of their competition persist today, leaving the world in a state of heightened nuclear threat. Competitive races of this kind entail repeated back-and-forth actions that can result in progressively worse outcomes for all involved. We can liken this example to the Iterated Prisoner's Dilemma, where the nations must decide whether to increase (defect) or decrease (cooperate) their nuclear spending. Both the US and the Soviet Union often chose to increase spending. They would have created a safer and less expensive world for both nations (as well as others) if they had cooperated to reduce their nuclear stockpiles. We discuss this in more detail in 8.6.

Two kinds of AI races: corporate and military [273]. Competition between different parties—nations or corporations—is incentivizing each to develop, deploy, and adopt AIs rapidly, at the expense of other values and safety precautions. Corporate AI races consist of businesses prioritizing their own survival or power expansion over ensuring that AIs are developed and released safely. Military AI arms races consist of nations building and adopting powerful and dangerous military applications of AI technologies to gain military power, increasing the risks of more frequent or damaging wars, misuse, or catastrophic accidents. We can understand these two kinds of AI races using two game-theoretic models of iterated interactions. First, we use the *Attrition* model to understand why AI corporations are cutting corners on safety. Second, we'll use the *Security Dilemma* model to understand why militaries are escalating the use of—and reliance on—AI in warfare.

Corporate AI Races

Competition between AI research companies is promoting the creation and use of more appealing and profitable systems, often at the cost of safety measures. Consider the public release of large language model-based chatbots. Some AI companies delayed releasing their chatbots out of safety concerns, like avoiding the generation of harmful misinformation. We can view the companies that released their chatbots first as having switched from cooperating to defecting in an Iterated Prisoner's Dilemma. The defectors gained public attention and secured future investment. This competitive pressure caused other companies to rush their AI products to market, compromising safety measures in the process.

Corporate AI races arise because competitors sacrifice their values to gain an advantage, even if this harms others. As a race heats up, corporations might increasingly need to prioritize profits by cutting corners on safety, in order to survive in a world where their competitors are very likely to do the same. The worst outcome for an agent in the Prisoner's Dilemma is the one where only they cooperated while their partner defected. Competitive pressures motivate AI companies to avoid this outcome, even at the cost of exacerbating large-scale risks.

Ultimately, corporate AI races could produce societal-scale harms, such as mass unemployment and dangerous dependence on AI systems. We consider one such example in *7.2.5*. This risk is particularly vivid for emerging industries like AI which lack the better-established safeguards such as mature regulation and widespread awareness of the harm that unsafe products can cause found in other industries like pharmaceuticals.

Attrition model: a multi-player game of "Chicken." We can model this kind of corporate AI race using an "Attrition" model [371], which frames a race as a kind of auction in which competitors bid against one another for a valuable prize. Rather than bidding money, the competitors bid for the risk level they are willing to tolerate. This is similar to the game "Chicken," in which two competitors drive headlong at each other. Assuming one swerves out of the way, the winner is the one who does not (demonstrating that they can tolerate a higher level of risk than the loser). Similarly, in the Attrition model, each competitor bids the level of risk—the probability of bringing about a catastrophic outcome—they are willing to tolerate. Whichever competitor is willing to tolerate the most risk will win the entire prize, as long as the catastrophe they are risking does not actually happen. We can consider this to be an "all pay" auction: both competitors must pay what they bid, whether they win or not. This is because all of those involved must bear the risk they are leveraging, and once they have made their bid they cannot retract it.

The Attrition model shows why AI corporations may cut corners on safety. Let us assume that there are only two competitors and that both of them have the same understanding of the state of their competition. In this case, the Attrition model predicts that they will race each other up to a loss of one-third in expected value [372]. If the value of the prize to one competitor is "X", they will be willing to risk a 33% chance of bringing about an outcome equally disvaluable (of value "-X") in order to win their race [373].

As we have discussed previously, market pressures may motivate corporations to behave as though they value what they are competing for almost as highly as survival itself. According to this toy model, we might then expect AI stakeholders engaged in a corporate race to risk a 33% chance of existential catastrophe in order to "win the prize" of their continued existence. With multiple AI races, long time horizons, and ever-increasing risks, the repeated erosion of safety assurances down to only 66% generates a vast potential for catastrophe.

Real-world actors may mistakenly erode safety precautions even further. Moreover, real-world AI races could produce even worse outcomes than the one pre-

dicted by the Attrition model [373]. One reason for this is that competing corporations may not have a correct understanding of the state of their race. Precisely predicting these kinds of risks can be extremely challenging: high-risk situations are inherently difficult to predict accurately, even in fields far more well-understood than AI. Incorrect risk calibration could cause the competitors to take actions that accidentally exceed even the 33% risk level. Like newcomers to an 'all pay' auction who often overbid, uneven comprehension or misinformation could motivate the competitors to take even greater risks of bringing about catastrophic outcomes. In fact, we might even expect selection for competitors who tend to underestimate the risks of these races. All these factors may further erode safety assurances.

Military AI Arms Races

Global interest in military applications for AI technologies is increasing. Some hail this as the "third revolution in warfare" [374], predicting impact at the scale of the historical development of gunpowder and nuclear weapons. There are many causes for concern about the adoption of AI technologies in military contexts. These include increased rates of weapon development, lethal autonomous weapons usage, advanced cyberattack execution, and automation of decision-making. These could in turn produce more frequent and destructive wars, acts of terrorism, and catastrophic accidents. Perhaps even more important than the immediate dangers from military deployment of AI is the possibility that nations will continue to race each other along a path towards ever increased risks of catastrophe. In this section, we explore this possibility using another game theoretic model.

First, let us consider a few different sources of risk from military AI [273]:

1. **AI-developed weapons.** AI technologies could be used to engineer weapons. Military research and development offers many opportunities for acceleration using AI tools. For instance, AI could be used to expedite processes in dual-use biological and chemical research, furthering the development of programs to build weapons of mass destruction.
2. **AI-controlled weapons.** AI might also be used to control weapons directly. "Lethal autonomous weapons" have been in use since March 2020, when a self-directing and armed drone "hunted down" soldiers in Libya without human supervision. Autonomous weapons may be faster or more reliable than human soldiers for certain tasks, as well as being far more expendable. Autonomous weapons systems thus effectively motivate militaries to reduce human oversight. In a context as morally salient as warfare, the ethical implications of this could be severe. Increasing AI weapon development may also impact international warfare dynamics. The ability to deploy lethal autonomous weapons in place of human soldiers could drastically lower the threshold for nations to engage in war, by reducing the expected body count—of the nation's own citizens, at least. These altered warfare dynamics could usher in a future with more frequent and destructive wars than has yet been seen in human history.

3. **AI cyberwarfare.** Another military application is the use of AI in cyberwarfare. AI systems might be used to defend against cyberattacks. However, we do not yet know whether this will outweigh the offensive potential of AI in this context. Cyberattacks can be used to wreak enormous harm, such as by damaging crucial systems and infrastructure to disrupt supply chains. AIs could make cyberattacks more effective in a number of ways, motivating more frequent attempts and more destructive successes. For example, AIs could directly aid in writing or improving offensive programs. They could also execute cyberattacks at superhuman scales by implementing vast numbers of offensive programs simultaneously. By democratizing the power to execute large-scale cyberattacks, AIs would also increase the difficulty of verification. With many more actors capable of carrying out attacks at such scales, attributing attacks to perpetrators would be much more challenging.
4. **Automated executive decision-making.** Executive control might be delegated to AIs at higher levels of military procedures. The development of AIs with superhuman strategic capabilities may incentivize nations to adopt these systems and increasingly automate military processes. One example of this is "automated retaliation." AI systems that are granted the ability to respond to offensive threats they identify with counterattacks, without human supervision. Examples of this include the NSA cyber defense program known as "MonsterMind." When this program identified an attempted cyberattack, it interrupted it and prevented its execution. However, it would then launch an offensive cyberattack of its own in return. It could take this retaliatory action without consulting human supervisors. More powerful AI systems, more destructive weapons, and greater automation or delegation of military control to AI systems, would all deplete our ability to intervene.
5. **Catastrophic accidents.** Lethal Autonomous Weapons and automated decision-making systems both carry risks of resulting in catastrophic accidents. If a nation were to lose control of powerful military AI technologies, the outcome could be calamitous. Outsourcing executive command of military procedures to AI—such as by automating retaliatory action—would put powerful arsenals on hair-trigger alert. If one of these AI systems were to make even a small error, such as incorrectly identifying an offensive strike from another nation, it might automatically "retaliate" to this non-existent threat. This could in turn trigger automated retaliations from the AI systems of other nations that detect this action. Thus, a small error could be exacerbated into an increasingly escalated war. We consider how a "flash war" such as this might come about in more detail in Section 7.2.5. Note that we can also use the "Attrition" model in the case of military AI arms races to model how military competitive pressures can motivate nations to cut corners on safety.
6. **Co-option of military AI technologies.** Military AI arms races could also have catastrophic effects outside of international warfare. New and more lethal weapons could be used maliciously in other contexts. For instance, biological weapons were originally created for military purposes. Even though we have since halted the military use of these weapons, their existence has enabled many acts of bioterror-

ism. Examples include the 2001 deployment of anthrax letters to kill US senators and media executives. The creation of knowledge of how to make and use these weapons is irreversible. Thus, their existence and the risk they pose are permanent.

7. **Military AI risks may interact.** Importantly, the risks posed by military AI applications are not entirely independent of one another. The increased potential for anonymity when executing cyberattacks could increase the probability of wars. Where it is harder to identify the perpetrators, misattribution could trigger conflict between the target of the attack and an innocent party. The potential for destructive cyberattacks might be increased by the scaled-up use of autonomous weapons, as these could be co-opted by such attacks. Similarly, the danger posed by a rogue AI with executive decision-making power might be all the more serious if it has control over fleets of autonomous weapons.

Security Dilemma model: mutual defensive concerns motivate nations to increase risks. We can better understand military AI arms races using the "Security Dilemma" model [375]. Consider the relationship between two peaceful nations. Though they are not currently at war with one another, each is sufficiently concerned about the possibility of conflict to pay close attention to the other's state of military ability. One day, one of the two nations perceives that the other is more militarily capable than they are due to their having stockpiled more advanced weaponry. This incentivizes the first nation to build up their own military capabilities until they match or exceed those of the other nation. The second nation, perceiving this increase in military investment and development, feels pressure to follow suit, once again increasing their weapon capabilities. Neither wishes to be outmatched by the other. This competitive pressure drives both to escalate the situation. The ensuing arms race generates increasingly high risks for both sides, such as increasing the probability or severity of accidents and misuse.

Example: the Cold War nuclear arms race. As previously discussed, the Cold War nuclear arms race typifies this process. Neither the US nor the Soviet Union wanted to risk being less militarily capable than their rival, so each escalated their own weaponized nuclear ability in an attempt to deter the other using the threat of retaliation. Just as in the Iterated Prisoner's Dilemma, neither nation could afford to risk being the lone cooperator while their rival defected. Thus, they achieve a Pareto inefficient outcome of both defecting. Competitive pressure drove them to continue to worsen this situation over time, resulting in today's enormously heightened state of nuclear vulnerability.

Increased automation of warfare by one nation puts pressure on others to follow suit. Just as with nuclear weapons, so with military AI: the Security Dilemma model illustrates how defensive concerns can force nations to go down a route which is against the long term interests of all involved. This route leads to the competing nations continually heightening the risks posed by military AI applications, including more frequent and severe wars, and worse accidents.

There are many incentives for nations to increase their development, adoption, and deployment of military AI applications. With more AI involvement, warfare can take place at an accelerated pace, and at a more destructive scale. Nations that do not adopt and use military AI technologies may therefore risk not being able to compete with nations that do. As with nuclear mutually assured destruction, nations may also employ automated retaliation as a signal of commitment, hoping to deter attacks by demonstrating a plausible resolution to respond swiftly and in kind. This process of automation and AI delegation would thus perpetuate, despite it being increasingly against the collective good.

Ultimately, as with economic automation, military AI arms races could result in humans being unable to keep up. The pace and complexity of warfare could ascend out of human reach to where we are no longer able to comprehend or intervene. This could be an irreversible step putting us at high risk of catastrophic outcomes.

Extortion

In this section, we examine one last risk that arises when agents interact repeatedly: the discovery of extortion.

Extortion strategies in the Iterated Prisoner's Dilemma. In the real world, we describe the use of threats to force a victim to take an action they would otherwise not want to take (such as to relinquish something valuable) as "extortion." Examples include criminal organizations ransoming those they have kidnapped to extort their families for money in exchange for their safe return.

In the Iterated Prisoner's Dilemma, there is a set of *extortion* strategies that bear similarity to this real-world phenomenon. An agent playing the game can use an *extortion* strategy to ensure that their payoff in any match is higher than their partner's [376]. The extortionist achieves this by acting similarly to an agent using *tit-for-tat*, responding to like with like. However, the *extortionist* will occasionally defect even when their partner has been cooperative. *Extortionists* effectively calculate the maximum number of defections they can get away with without annihilating the motivation of their partner to continue cooperating with them. They decide whether to cooperate or defect using a set of probabilities. The most recent interaction with their partner determines which probability they select. An example strategy is shown in Figure 7.7. An *extortionist's* partner is incentivized to acquiesce to the *extortion* since deviating in any way will yield them a lower payoff. However, in maximizing their own score, they attain an even higher score for the *extortionist*. An extortionist thus scores higher than most of its partners in Iterated Prisoner's Dilemma matches.

Shown is an *extortion* strategy called *Extort-2*, from the point of view of the *extortionist*. "You" are the agent using the *Extort-2* strategy, and "they" are your partner. As with all *extortion* strategies, *Extort-2* involves reacting probabilistically to the most recent interaction with a partner. As an example, in the previous round, if the *extortionist* defected, but their partner cooperated, the *extortionist* will cooperate with a probability of $\frac{1}{3}$ in this round.

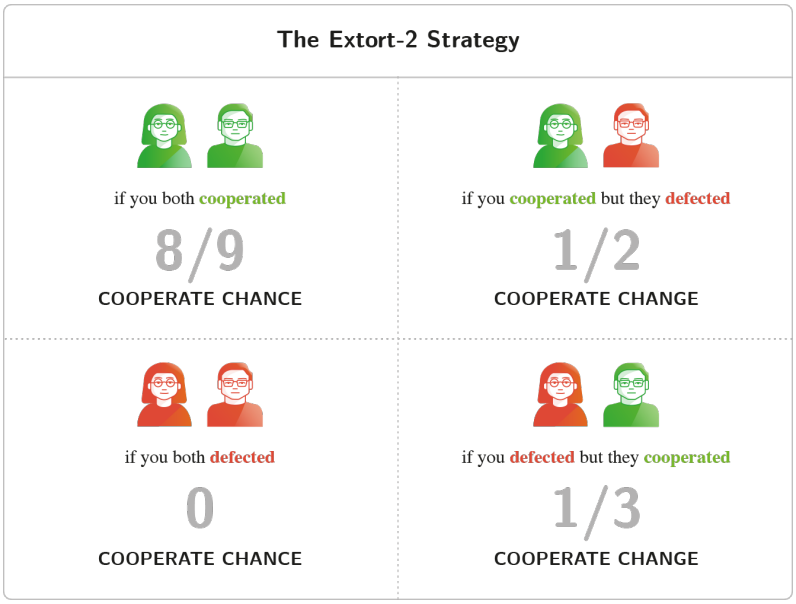

FIGURE 7.7. The *Extort-2* strategy [377].

Extortion strategies rarely win tournaments but seldom die out altogether. As we saw in Section 7.2.4, many uncooperative strategies may gain a higher score than most of their partners in head-to-head matches, and yet still lose in tournaments. By contrast, *extortionists* can be somewhat successful in tournaments under certain conditions. *Extortionists* are vulnerable to the same problem as many other uncooperative strategies: they gain low payoffs in matches against other *extortionists*. Each will therefore perform less well as the frequency of *extortionists* in the population increases. Thus, *extortionists* can persist if they are sufficiently unlikely to meet one another. For instance, where a sufficiently small population of agents is engaged in a tournament, a single *extortionist* can achieve very high payoffs by exploiting cooperative strategies.

AI agents may use extortion: evidence from the Iterated Prisoner's Dilemma. AI agents could use extortion in order to gain resources or power. As we have seen, agents can succeed in the Iterated Prisoner's Dilemma by using *extortion* strategies. This is particularly true if the *extortionist* is part of a small group, if the social dynamics mirror evolution by natural selection, or after major environmental changes. These findings are extremely worrying as they could describe future AI scenarios. Relationships might form among a small number of powerful AI agents. These agents may undulate through desirable and undesirable behaviors, or they might switch opportunistically to using *extortion* tactics in the wake of changes to their environment. However, since there are some fragile assumptions in these simple models, we must also consider evidence from real-world agents.

AI agents may use extortion: evidence from the real world. The widespread use of extortion among humans outside the world of game theoretic mod-

els suggests there is still a major cause for concern. Real-world extortion can still yield results even when the target is perfectly aware that it is taking place. The use of ransomware schemes to extort private individuals and companies is increasing rapidly. In fact, cybersecurity experts estimate that the annual economic cost of ransomware activity is in the billions of US dollars. Terrorist organizations such as ISIS rely on extortion through hostage-taking for a large portion of their total income. The ubiquity of successful extortion in so many contexts sets a powerful historical precedent for its efficacy.

Tail Risk: Extortion With Digital Minds

Here we examine the possibility of AI agents engaging in extortion to pursue their goals. Though the probability of AI extortion may be low, the impact could be immense. As an example, we consider the potential for extortionist AIs to simulate and torture sentient digital minds as leverage.

Real-world extortion is a form of optimized disvalue. An extortionist chooses to threaten their target using a personalized source of concern. They optimize their extortion to be prioritized over their target's other concerns. Often, the worse the outcome being threatened, the more likely the target is to acquiesce. This incentivizes extortionists to threaten to bring about extremely disvaluable scenarios. In order to be effective, extortionist AIs might therefore leverage the threat of huge amounts of harm—far more than would likely come about incidentally, without design. If the disvaluable outcome the extortionist has designed for their target is also disvaluable to wider society, then we will share the potentially enormous costs of any executed threats.

AI extortion could entail torturing vast numbers of sentient digital minds. Human extortionists often threaten to inflict excruciating pain or death on those their victim cares about. AI extortionists might engage in similar behaviors, threatening to induce extreme levels of suffering, but on a vastly larger scale. This scale could potentially exceed any in human history. This is because extortionist AIs with greater-than-human technological capabilities might be able to simulate sentient digital minds. The potential for optimized disvalue in these simulations suggests near-unimaginable horrors. Vast numbers of digital people in these simulated environments could be subjected to immeasurably agonizing experiences.

Simulated torture at this scale could make the future more disvaluable than valuable. Simulations designed for the purpose of extortion would likely be far more disvaluable than simulations which contain disvalue unintentionally. The simulation's designer would likely be able to choose what kinds of objects to simulate, so they could avoid wasting energy simulating non-sentient entities such as inanimate objects. Moreover, the designer could ensure that these sentient entities experience the greatest amount of suffering possible for the timespan of the simulation. They might even be able to simulate minds capable of more disvaluable experiences than have ever existed previously, deliberately designing the digital entities to be able to

suffer as greatly as possible. Put together, a simulation optimized for disvalue could produce several orders of magnitude more disvalue than anything in history. This would be unprecedented in humanity's history, and could make a horrifying—even net negative—future.

AI agents may be superhumanly proficient at wielding extortion. Future AI agents may far exceed humans in their ability to wield threats. One reason for this could be that they have superhuman tactical capabilities, as some do already in competitive games. Superior strategic intelligence could allow AI agents to conceive and execute far more advanced programs of extortion than that of which humans are generally capable. A second reason why AI agents may be especially adept at employing threats is if they have superhuman longevity or goal-preservation capabilities. With greater timespans available, the action space for extorting targets is larger. Finally, AIs may have technological capabilities that exceed those of current and historical humans. This could widen the option space for AI extortion still further.

Extortion may be exceptionally effective against AIs. Two goals of machine ethics are: 1) to foster in AI an intrinsic value for humanity (and humanity's values); 2) to make AI agents that are impartial. Both goals could result in AI agents being more vulnerable to extortion than humans tend to be. Let us examine an example of this for each goal.

Goal 1: Foster in AI an intrinsic value for humanity (and humanity's values).

AI agents that value individual humans highly may be less prone to "scope insensitivity." This is the human bias of failing to "feel" changes in the size of some value appropriately. Very small or very large numbers often appear to us to be of similar size to other very small or very large numbers, even when they actually differ by orders of magnitude. Human scope insensitivity may provide some protection against larger-scale extortion, as it lowers the motivation of extortionists to increase the scale of their threats. It is possible that AI agents may prioritize outcomes more accurately in accordance with their expected value. If this is the case, they would likely be more responsive to high stakes, and more vulnerable to large-scale extortion attempts.

Goal 2: Make AI agents that are impartial.

Impartial AI agents may have far more altruistic values than any human or institution. These agents may be extremely vulnerable to extortion in the form of threats against their impartial moral codes. Extortionist AI agents could leverage the threat of extreme torture of countless digital sentients in simulated environments to extort more morally impartial AI targets. The execution of any such threat could immensely degrade the value of the future.

AI extortionists may execute higher-stakes threats more frequently than humans. A successful act of extortion is the deliberate creation of a state in which both the extortionists and their targets prefer the outcome the extortionist demands. In some sense, both parties therefore *want* the target to acquiesce to the extortion and the extortionist not to follow through on their threat. In this way, both usually

have some incentive to avoid the threat being executed. However, out of a desire to signal credibility in future interactions, extortionists must follow through on threats occasionally. Consider examples such as hostage ransoming or criminal syndicate protection rackets. Successful future extortion requires a signal of commitment, such as destroying the property of those who defy the extortionists.

AIs may carry out more frequent and more severe threats than humans tend to. One reason for this is that they may have different value systems which tolerate higher risks, reducing their motivation to acquiesce to extortion. For example, an AI agent that sufficiently values the far future may prefer to demotivate future extortionists from trying to extort them. They may therefore defy a current extortion attempt, tolerating even very large costs to them and others, for the long-term benefit of credibly signaling that future extortion attempts will not work either.

More generally, with a greater variety of value systems, a greater number of agents, and a greater action space size, miscalibrated extortion attempts are more likely. Where the threat is insufficient to force compliance, the aforementioned need to signal credibility incentivizes the extortionist to execute their threat as punishment for their target's refusal to submit.

AI agents extorting humans. AI agents might also extort human targets. One example scenario would be an AI developing both a weaponized biological pathogen, and an effective cure. If the pathogen is slow-acting, the AI agent could then extort humans by deploying the bioweapon, and leveraging the promise of its cure to force those infected into complying with its demands. Pathogens that are sufficiently fast to spread and difficult to detect could infect a very large number of human targets, so this tactic could enable extremely large-scale threats to be wielded effectively [378].

Summary

The Iterated Prisoner's Dilemma involves repeated rounds of the Prisoner's Dilemma game. This iteration offers a chance for agent cooperation but doesn't ensure it. There are different strategies by which agents can attempt to maximize their overall payoffs. These strategies can be studied by competing agents against one another in tournaments, where each agent competes against others in multiple rounds before switching partners.

This provides cause for concern about a future with many AI agents. One example of this is the phenomenon of "races" between AI stakeholders. These races strongly influence the speed and direction of AI technological production, deployment and adoption, in both corporate and military settings and have the potential to exacerbate many of the intrinsic risks from AI. The dynamics we have explored in this section might cause competing agencies to cut corners on safety, escalate weaponized AI applications and automate warfare. These are two examples of how competitive pressures, modeled as iterated interactions between agents, can generate races which increase the risk of catastrophe for everyone. Fostering cooperation between different parties—human individuals, corporations, nations, and AI agents—is vital for ensuring our collective safety.

7.2.5 Collective Action Problems

We began our exploration of game theory by looking at a very simple game, the Prisoner's Dilemma. We have so far considered two ways to model real-world social scenarios in more detail. First, we explored what happens when two agents interact *multiple times* (such as an Iterated Prisoner's Dilemma match). Second, we introduced a population of *more than two* agents, where each agent switches partners over time (such as an Iterated Prisoner's Dilemma tournament). Now we move beyond pairwise interactions, to interactions that simultaneously involve more than two agents. We consider what happens when an agent engages in repeated rounds of the Prisoner's Dilemma against multiple opponents at the same time.

One class of scenarios that can be described by such a model is *collective action problems*. Throughout this section, we first discuss the core characteristics of collective action problems. Then, we introduce a series of real-world examples to highlight the ubiquity of these problems in human society and show how AI races can be modeled in this way. Following this, we transition to a brief discussion of common pool resource problems to further illustrate the difficulty with which rational agents, especially AI agents, may secure collectively good outcomes. Finally, we conclude with a detailed discussion of flash wars and autonomous economies to show how in a multi-agent setting, AIs might pursue behaviors or tactics that result in catastrophic or existential risks to humans.

Introduction

This first section explores the nature of collective action problems. We begin with a simple example of a collaborative group project. Through this, we explore how individual incentives can sometimes clash with what is in the best interests of the group as a whole. These situations can motivate individuals to act in ways that negatively impact all of the population.

A collective action problem is like a group-level Iterated Prisoner's Dilemma. In the Iterated Prisoner's Dilemma, we saw how a pair of rational agents can tend towards outcomes that are undesirable for both. Now let us consider social interactions between more than two agents. When an individual engages with multiple partners simultaneously, they may still converge on Pareto inefficient Nash equilibria. In fact, with more than two agents, cooperation can be even harder to secure. We can therefore model collective action problems as an Iterated Prisoner's Dilemma in which more than two prisoners have been arrested: If enough of them decide to defect on their partners, all of them will suffer the consequences.

Example: group projects. A typical example of a collective action problem is that of a collaborative project. A group working together towards a shared goal often encounters a problem: not everyone pitches in. Some group members take advantage of the rest, benefiting from the work others are doing without committing as much effort themselves. The implicit reasoning behind the behavior of these "slackers" is as follows. They want the group's goal to be achieved, but they would prefer this

to happen without costing them much personal effort. Just as with the Prisoner's Dilemma, "slacking" is their dominant strategy. If the others work hard and the project is completed, they get to enjoy the benefits of this success without expending too much effort themselves. If the others fail to work hard and the project is not completed, they at least save themselves the effort they might otherwise have wasted.

As groups increase in size and heterogeneity, complexity increases accordingly. Agents in a population may have a diverse set of goals. Even if the population can agree on a common goal, aligning diverse agents with this goal can be difficult. For example, even when the public expresses strong and widespread support for a political measure, their representatives often fail to carry it out.

Formalization

Here, we formalize our model of collective action problems. We look more closely at the incentives governing individual choices, and the effects these have at the group level. We examine how the behavior of others in the group can alter the incentives facing any individual, and how we can (and do) use these mechanisms to promote cooperative behavior in our societies.

Each agent must choose whether to contribute to the common good. As in the Prisoner's Dilemma, each agent must choose which of two actions to take. An agent can choose to **contribute** to the common good, at some cost to themselves. The alternative is for the agent to choose to **free ride**, benefiting from others' contributions at no personal cost. Free riders impose **negative externalities**—collateral damage for others in pursuit of private benefit—on the group as a whole by choosing not to pitch in.

Free riding is the dominant strategy. For now, let us assume that free riding increases an agent's own personal benefit, regardless of whether the others contribute or free ride: it is the dominant strategy. If an agent's contribution to the common good is small, then choosing *not* to contribute does not significantly diminish the collective good, meaning that an agent's decision to free ride has essentially no negative consequences for the agent themself. Thus, the agent is choosing between two outcomes. The first outcome is where they gain their portion of the collective benefit, and pay the small cost of being a contributor. The other outcome is where they gain this same benefit, but save themselves the cost of contributing.

Free riding can produce Pareto inefficient outcomes. Just as how both agents defecting in the Prisoner's Dilemma produces Pareto inefficiency, free riding in a collective action problem can result in an outcome that is bad for all. In many cases, some agents can free ride without imposing significant externalities on everyone else. However, if sufficiently many agents free ride, this diminishes the collective good by leading to no provision of a public good, for instance. With sufficient losses, the agents will all end up worse than if they had each paid the small individual cost of contributing and received their share of the public benefit. Importantly, however, even in this Pareto inefficient state, free riding might still be the dominant strategy

for each individual, since the cost of contributing outweighs the trivial increase in collective good they would contribute by contributing. Thus, escaping undesirable equilibria in a collective action problem can be exceedingly difficult; see Figure 7.8.

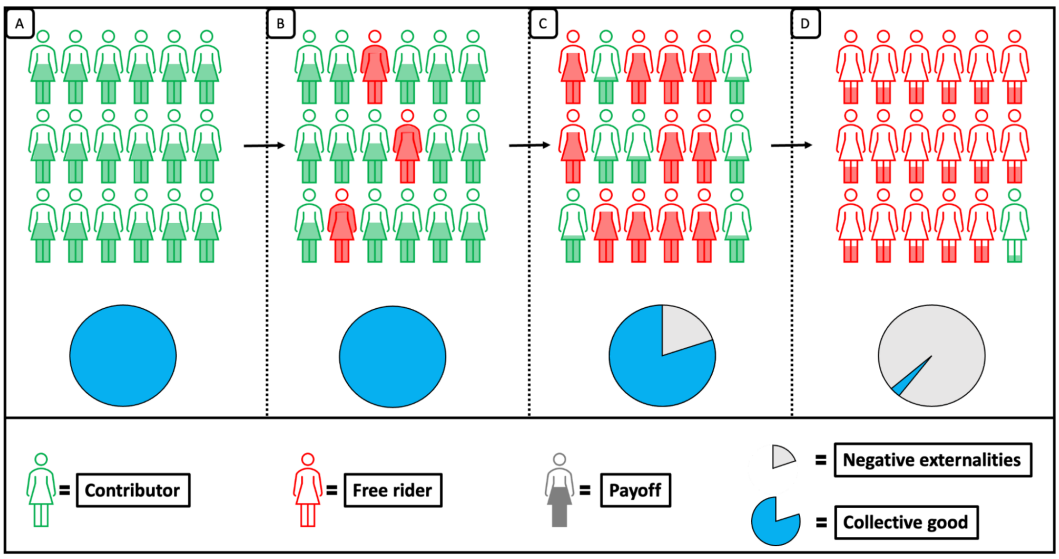

FIGURE 7.8. In this abstract collective action problem, we can move from everyone (contributes) to right (no one contributes). As more people free ride, the collective good disappears, leaving everyone in a state where they would all prefer to collectively contribute instead.

We can illustrate a collective action problem using the simple payoff matrix below. In the matrix, "b" represents the payoff an agent receives when everyone else cooperates (the collective good divided between the number of agents) and "c" represents the personal cost of cooperation. As the matrix illustrates, the dominant strategy for a rational agent ("you") here is to free ride whether everyone else contributes or free rides.

TABLE 7.4. Free riding is always better for an individual: it is a dominant strategy.

	The rest of the group contributes	**The rest of the group free rides**	**Some contribute; others free ride**
You contribute	$b - c$	$-c$	$< b - c$
You free ride	b	0	$< b$

Agents' incentives depend on the behavior of other agents. Agents in collective action problems can be aware of the choices other agents make, which can affect their strategies and behavior over time. For example, the ratio of defectors to cooperators in a population can affect the degree to which cooperation is achieved. When rational agents interact with each other, they may be inclined to shift their strategies to more favorable ones with higher individual payoffs: they may realize that other agents are utilizing more successful strategies, and thus choose to adopt

them. If defectors dominate the population initially, and the initial individual costs of cooperation outweigh the collective benefits of cooperation, then the population may tend towards an uncooperative state. In simple terms, collective action problems cannot be solved without cooperation.

Mutual and external coercion. We can increase the probability of cooperation by generating incentives that lower the individual cost of cooperation and increase the individual cost of defection. There are two ways we may go about this: mutual or external coercion. *Mutual coercion* generates cooperative incentives by establishing communal, societal, and reputational norms. *External coercion* generates cooperative incentives through external intervention, by developing regulations that incentivize collective action through mandates, sanctions, and legislature, making cooperation a necessity in certain cases. Below, we illustrate some real-world scenarios in further detail.

Real-World Examples of Collective Action Problems

Many large-scale societal issues can be understood as collective action problems. This section explores collective action problems in the real world: climate change, public health, and democratic voting. We end by briefly looking at AI races through this same lens.

Public health. We can model some public health emergencies, such as disease epidemics, as collective action problems. The COVID-19 pandemic took the lives of millions worldwide. Some of these deaths could have been avoided with stricter compliance with public health measures such as social distancing, frequent testing, and vaccination. We can model those adhering to these measures as "contributing" (by incurring a personal cost for public benefit) and those violating them as "free riding."

Assume that everyone wished the pandemic to be controlled and ultimately eradicated, that complying with the suggested health measures would have helped hasten this goal, and that the benefits of collectively shortening the pandemic timespan would have outweighed the personal costs of compliance with these measures (such as social isolation). Everyone would prefer the outcome where they all complied with the health measures over the one where few of them did. Yet, each person would prefer still better the outcome where *everyone else* adhered to the health measures, and *they alone* were able to free ride. Violating the health measures was therefore the dominant strategy, and so many people chose to do this, imposing the negative externalities of excessive disease burden on the rest of their community.

We used both mutual and external mechanisms to coerce people to comply with public health measures in the pandemic. For example, some communities adjusted their social norms (mutual coercion) such that non-compliance with public health measures would result in damage to one's reputation. We also required proof of vaccination for entry into desirable social spaces (external coercion), among many other requirements.

Anthropogenic climate change. In 2021, a majority of those surveyed worldwide reported wanting to avert catastrophic anthropogenic climate change. Most, however, chose not to act in accordance with what they believed necessary to achieve this goal. The consumption of animal products typically entails far higher greenhouse gas emissions and environmental damage than plant-based alternatives. The use of public over private transport similarly reduces personal carbon footprints dramatically. To avoid the costs of taking these actions, such as changing routines and compromising on speed or ease, most people do not change their diets or transport habits. Various behaviors that increase pollution can be viewed as "free riding." Since this is the dominant strategy for each agent, most choose to do this, resulting in ever-worsening climate change, imposing risks on the global population.

We could disincentivize excessive meat eating and private transport using external and mutual coercion. In this example, external coercion could include lowering bus and train fares and enhancing existing infrastructure through government subsidies, as well as implementing fuel taxes on private vehicles. Mutual coercion could include changing social norms to consider excessive meat eating or short-haul flying unacceptable.

Democracy. We can model the maintenance of a democracy as a set of collective action problems. There are many situations in which certain actions might provide an individual with immediate benefits, but would incur longer-term costs on the larger group if more people were to take these actions. For example, a voting population must maintain certain norms in order to keep its democracy functioning. One of these norms is to vote only for candidates who will not undermine democratic processes, even if others have desirable traits.

Choosing whether or not to participate in an election at all can similarly be viewed as a collective action problem. The outcome of an election is determined by the votes of individuals, each of which has a choice to either vote or abstain. The results of the election are determined by the votes of those who choose to participate, and the costs of participating in the election are carried by citizens themselves, such as the time and effort required to register and cast a vote. When large enough numbers of citizens decide to abstain from voting, the collective outcome of an election may not accurately reflect the preferences of the population: by acting in accordance with their rational self interest, citizens may contribute to a suboptimal collective outcome.

Common pool resource problem. Rational agents are incentivized to take more than a sustainable amount of a shared resource. This is called a *common pool resource problem* or *tragedy of the commons problem*. We refer to a common pool resource becoming catastrophically depleted as collapse. Collapse occurs when rational agents, driven by their incentive to maximize personal gain, tip the available supply of the shared resource below its sustainability equilibrium [379]. Below, we further illustrate how complicated it is to secure collectively good outcomes, especially when rational agents act in accordance with their self-interest. Such problems are prevalent at the societal level, and often bear catastrophic consequences. Thus, we should not eliminate the possibility that they may also occur with AI agents in a multi-agent setting.

For example, rainforests around the world have been diminished greatly by deforestation practices. While these forests still exist as a home to millions of different species and many local communities, they may reach a point at which they will no longer be able to rejuvenate themselves. If these practices are sustained, the entire ecosystem these forests support could collapse. Common pool resource problems exemplify how agents may bring about catastrophes even when they behave rationally and in their self-interest, with perfect knowledge of the looming catastrophe, and despite the seeming ability to prevent it. They further illustrate how complicated it can be to secure collectively good outcomes and how rational agents can act to the detriment of their own group. As with many other collective action problems, we can't expect to solve common pool resource problems by having AIs manage them. If we simply pass the buck to AI representatives, the AIs will inherit the same incentive structure that produces the common pool resource problem, and so the problem will likely remain.

AI Races Between More Than Two Competitors

In the previous section, we looked at how corporations and militaries may compete with one another in "AI races." We used a two-player "attrition" bidding model to see why AI companies cut corners on safety when developing and deploying their technologies. We used another two-player "security dilemma" model to understand how security concerns motivate nations to escalate their military capabilities, even while increasing the risks imposed on all by increasingly automating warfare in this manner.

Here, we extend our models of these races to consider more than two parties, allowing us to see them as collective action problems. First, we look at how military AI arms races increase the risk of catastrophic outcomes such as a *flash war*: a war that is triggered by autonomous AI agents that quickly spirals out of human control [273]. Second, we explore how ever-increasing job automation could result in an *autonomous economy*: an economy in which humans no longer have leverage or control.

Military AI arms race outcome: flash war. The security dilemma model we explored in the previous section can be applied to more than two agents. In this context, we can see it as a collective action problem. Though all nations would be at lower risk if all were to cooperate with one another ("contribute" to their collective safety), each will individually do better instead to escalate their own military capabilities ("free ride" on the contributions of the other nations). Here, we explore one potentially catastrophic outcome of this collective action problem: a flash war.

As we saw previously, military AI arms races motivate nations to automate military procedures. In particular, there are strong incentives to integrate "automated retaliation" protocols. Consider a scenario in which several nations have constructed an autonomous AI military defense system to gain a defensive military advantage. These AIs must be able to act on perceived threats without human intervention. Additionally, each is aligned with a common goal: "defend our nation from attack." Even if these systems are nearly perfect, a single erroneous detection of a perceived threat could trigger a decision cascade that launches the nation into a "flash war."

Once one AI system hallucinates a threat and issues responses, the AIs of the nations being targeted by these responses will follow suit, and the situation could escalate rapidly. A flash war would be catastrophic for humanity, and might prove impossible to recover from.

A flash war is triggered and amplified by successive interactions between autonomous AI agents such that humans lose control of weapons of mass destruction [380]. Any single military defense AI could trigger it, and the process could continue without human intervention and at great speed. Importantly, having humans in the loop will not necessarily ensure our safety. Even if AIs only provide human operators with instructions to retaliate, our collective safety would rest on the chance that soldiers would willfully disobey their instructions.

Collective action between nations could avoid these and other dire outcomes. Limiting the capabilities of their military AIs by decreasing funding and halting or slowing down development would require that each nation give up a potential military advantage. In a high stakes scenario such as this one, rational agents (nations) may be unwilling to give up such an advantage because it dramatically increases the vulnerability of their nation to attack. The individual cost of cooperation is high while the individual cost of defection is low, and as agents continue to invest in military capabilities, competitive pressures increase, which further exacerbate costs of cooperation —thereby disincentivizing collective action. While the collective benefits of cooperation would drastically reduce the catastrophic risks of this scenario in the long-term, they may not outweigh the self-interest of rational agents in the short-term.

Corporate AI race outcome: autonomous economy. As AIs become increasingly effective at carrying out human goals, they may begin to out-perform the average human at an increasing number and range of jobs, from personal assistants to executive decision-makers. To reap the benefits of these faster and more effective workers, companies will likely continue to automate economically valuable functions by delegating them to AI agents. Ultimately, this could lead to the global economy becoming "autonomous," with humans no longer able to steer or intervene [381].

Such an autonomous economy would be a catastrophe for humanity. Like passengers in an autonomous vehicle, our safety and destination would rest with the AI systems now acting without our supervision. Our future would be determined by the behavior and outputs of this autonomous economy. If the AI agents engaged in this economy were to have undesirable goals or evolve selfish traits—a possibility we examine in the next section of this chapter—humanity would be unable to prevent the harms they cause. Even if the AIs themselves are well-aligned to our goals, the economic system itself may produce extremely undesirable outcomes. In this section, we have examined many examples of how macrobehavior can differ dramatically from micromotives. A population of individuals can tend towards states that are bad for everyone and yet be in stable equilibria. This could happen just the same with AI representatives acting on humanity's behalf in an autonomous economy.

Just as with military AI arms races, we can model how an autonomous economy might be brought about using the security dilemma model. As in the previous example, if

we expand this model to more than two agents, we can see it as a collective action problem in which competitive pressures drive different parties to automate economic functions out of the need to "keep up" with their competitors. Under this model, we can see how companies must choose whether to maintain human labor ("contributing") or automate these jobs using AI ("free riding"). Although all would prefer the outcome in which the calamity of an autonomous economy is avoided, each would individually prefer to have a competitive advantage and not risk being outperformed by rivals who reap the short-term benefit of using AIs. Thus, economic automation is the dominant strategy for each competitor. Repeated rounds of this game in which a sufficient number of agents free ride would drive us towards this disaster. In each successive round, it would become progressively more difficult to turn back, as we come to rely increasingly on more capable AI agents.

Increasing AI autonomy increases the risk of catastrophic outcomes. As AIs become more autonomous, humans may delegate more decision-making power to them. If AIs are able to successfully and consistently attain the high-level objectives given to them by humans, we may be more inclined to begin providing them with open-ended goals. If AIs achieve these goals, humans might not be privy to the process they follow and may overlook potential harms, as we saw in both the autonomous economy and flash war examples. Moreover, adaptive AIs—systems that actively adjust their computational design, architecture and behavior in response to new information or changes in the environment—could adapt at a much faster rate than humans. The possibility of self-improvement among such AIs would further exacerbate this problem. Adaptive AIs could develop unanticipated emergent behaviors and strategies, making them deeply unpredictable. Humans could be inclined to accept these negative behaviors in order to maintain a competitive advantage in the short-term.

Reducing competitive pressures could foster collective action. The security dilemma model shows how nations can be motivated to escalate their offensive capabilities out of the perception that their competitors are doing the same. However, by signaling the opposite, we might be able to produce the reverse effect, such as military de-escalation or an increase in AI safety standards. For instance, whether different nations will acquiesce to a shared international standard for AI regulation may depend on whether the nations are individually signaling their willingness to regulate in their own jurisdiction in the first place. If one nation perceives that others are engaging in strict domestic regulation, they might see this as a credible signal of commitment to an international standard. By easing the competitive pressures, we might be able to foster collective action to avoid driving up the collective risk level.

7.2.6 Summary

We observe important and seemingly intractable collective action problems in many domains of life, such as environmental degradation, pandemic responses, maintenance of democracies, and common pool resource depletion. We can understand these as

Iterated Prisoner's Dilemmas with many more than two agents interacting simultaneously in each round of the game. As before, we see that "free riding" can be the dominant strategy for an individual agent, and this can lead to Pareto inefficient outcomes for the group as a whole. We can use the mechanisms of mutual and external coercion to incentivize agents to cooperate with each other and achieve collectively good outcomes.

If we expand our models of AI races to include more than two agents, we can understand the races themselves as collective action problems, and examine how they exacerbate the risk of catastrophe. One example is how increasingly automating military protocols increases the risk of a "flash war." Similar dynamics of automation in the economic sphere could lead to an "autonomous economy." Either outcome would be disastrous and potentially irreversible, yet we can see how competitive pressures can drive rational and self-interested agents (such as nations or companies) down a path towards these calamities.

In this section, we examined some simple, formal models of how rational agents may interact with each other under varying conditions. We used these game theoretic models to understand the natural dynamics in multi-agent biological and social systems. We explored how these multi-agent dynamics can generate undesirable outcomes for all those involved. We considered some tails risks posed by interactions between human and AI agents. These included human-directed companies and militaries engaging in perilous races, as well as autonomous AIs using threats for extortion.

These risks can be reduced if mechanisms such as institutions are used to ensure human agencies and AI agents are able to cooperate with one another and avoid conflict. We explore some means of achieving cooperative interactions in the next section of this chapter, 7.3.

7.3 COOPERATION

Overview

In this chapter, we have been exploring the risks that arise from interactions between multiple agents. So far, we have used game theory to understand how collective behavior can produce undesirable outcomes. In simple terms, securing morally good outcomes without cooperation can be extremely difficult, even for intelligent rational agents. Consequently, the importance of cooperation has emerged as a strong theme in this chapter. In this third section of this chapter, we begin by using evolutionary theory to examine cooperation in more detail.

We observe many forms of cooperation in biological systems: social insect colonies, pack hunting, symbiotic relationships, and much more. Humans perform community services, negotiate international peace agreements, and coordinate aid for disaster responses. Our very societies are built around cooperation.

Cooperation between AI stakeholders. Mechanisms that can enable cooperation between the corporations developing AI and other stakeholders such as gov-

ernments may be vital for counteracting the competitive and evolutionary pressures of AI races we have explored in this chapter. For example, the "merge-and-assist" clause of OpenAI's charter [382] outlines their commitment to cease competition with—and provide assistance to—any "value-aligned, safety-conscious" AI developer who appears close to producing AGI, in order to reduce the risk of eroding safety precautions.

Cooperation between AI agents. Many also suggest that we must ensure the AI systems themselves also act cooperatively with one another. Certainly, we do want AIs to cooperate, rather than to defect, in Prisoner's Dilemma scenarios. However, this may not be a total solution to the collective action problems we have examined in this chapter. By more closely examining how cooperative relationships can come about, it is possible to see how making AIs more cooperative may backfire with serious consequences for AI safety. Instead, we need a more nuanced view of the potential benefits and risks of promoting cooperation between AIs. To do this, we study five different mechanisms by which cooperation may arise in multi-agent systems [383], considering the ramifications of each:

- *Direct reciprocity*: when individuals are likely to encounter others in the future, they are more likely to cooperate with them.
- *Indirect reciprocity*: when it benefits an individual's reputation to cooperate with others, they are more likely to do so.
- *Group selection*: when there is competition between groups, cooperative groups may outcompete non-cooperative groups.
- *Kin selection*: when an individual is closely related to others, they are more likely to cooperate with them.
- *Institutional mechanisms*: when there are externally imposed incentives (such as laws) that subsidize cooperation and punish defection, individuals and groups are more likely to cooperate.

Direct Reciprocity

Direct reciprocity overview. One way agents may cooperate is through *direct reciprocity*: when one agent performs a favor for another because they expect the recipient to return this favor in the future [384]. We capture this core idea in idioms like "quid pro quo," or "you scratch my back, I'll scratch yours." Direct reciprocity requires repeated interaction between the agents: the more likely they are to meet again in the future, the greater the incentive for them to cooperate in the present. We have already encountered this in the iterated Prisoner's Dilemma: how an agent behaves in a present interaction can influence the behavior of others in future interactions . Game theorists sometimes refer to this phenomenon as the "shadow of the future." When individuals know that future cooperation is valuable, they have increased incentives to behave in ways that benefit both themselves and others, fostering trust, reciprocity, and cooperation over time. Cooperation can only evolve as a consequence of direct reciprocity when the probability, w, of subsequent encounters between the

same two individuals is greater than the cost-benefit ratio of the helpful act. In other words, if agent A decides to help agent B at some cost c to themselves, they will only do so when the expected benefit b of agent B returning the favor outweighs the cost of agent A initially providing it. Thus, we have the rule $w > c/b$; see Table 7.5 below.

TABLE 7.5. Payoff matrix for direct reciprocity games.

	Cooperate	Defect
Cooperate	$b - c/(1-w)$	$-c$
Defect	b	0

Natural examples of direct reciprocity. Trees and fungi have evolved symbiotic relationships where they exchange sugars and nutrients for mutual benefit. Dolphins use cooperative hunting strategies where one dolphin herds schools of fish while the others form barriers to encircle them. The dynamics of the role reversal are decided by an expectation that other dolphins in the group will reciprocate this behavior during subsequent hunts. Similarly, chimpanzees engage in reciprocal grooming, where they exchange grooming services with one another with the expectation that they will be returned during a later session [385].

Direct reciprocity in human society. Among humans, one prominent example of direct reciprocity is commerce. Commerce is a form of direct reciprocity "which offers positive-sum benefits for both parties and gives each a selfish stake in the well-being of the other" [386]; commerce can be a win-win scenario for all parties involved. For instance, if Alice produces wine and Bob produces cheese, but neither Alice nor Bob has the resources to produce what the other can, both may realize they are better off trading. Different parties might both need the good the other has when they can't produce it themselves, so it is mutually beneficial for them to trade, especially when they know they will encounter each other again in the future. If Alice and Bob both rely on each other for wine and cheese respectively, then they will naturally seek to prevent harm to one another because it is in their rational best interest. To this point, commerce can foster *complex interdependencies* between economies, which enhances the benefits gained through mutual exchange while decreasing the probability of conflict or war.

Direct reciprocity and AIs. The future may contain multiple AI agents, many of which might interact with one another to achieve different functions in human society. Such AI agents may automate parts of our economy and infrastructures, take over mundane and time-consuming tasks, or provide humans and other AIs with daily assistance. In a system with multiple AI agents, where the probability that individual AIs would meet again is high, AIs might evolve cooperative behaviors through direct reciprocity. If one AI in this system has access to important resources that other AIs need to meet their objectives, it may decide to share these resources accordingly. However, since providing this favor would be costly to the given AI, it will do so only when the probability of meeting the recipient AIs (those that received the favor) outweighs the cost-benefit ratio of the favor itself.

Direct reciprocity can backfire: AIs may disfavor cooperation with humans. AIs may favor cooperation with other AIs over humans. As AIs become substantially more capable and efficient than humans, the benefit of interacting with humans may decrease. It may take a human several hours to reciprocate a favor provided by an AI, whereas it may take an AI only seconds to do so. It may therefore become extremely difficult to formulate exchanges between AIs and humans that benefit AIs more than exchanges with other AIs would. In other words, from an AI perspective, the cost-benefit ratio for cooperation with humans is not worth it.

Direct reciprocity may backfire: offers of AI cooperation may undermine human alliances. The potential for direct reciprocity can undermine the stability of other, less straightforward cooperative arrangements within a larger group, thereby posing a collective action problem. One example of this involves "bandwagoning." In the Alignment section of the Single-Agent Safety chapter, we discussed the idea of "balancing" in international relations: state action to counteract the influence of a threatening power, such as by forming alliances with other states against their common adversary [107]. However, some scholars argue that states do not always respond to threatening powers by trying to thwart them. Rather than trying to prevent them from becoming too strong, states may instead "bandwagon": joining up with and supporting the rising power to gain some personal benefit.

For instance, consider military coups. Sometimes, those attempting a takeover will offer their various enemies incentives to join forces with them, promising rewards to whoever allies with them first. If one of those being made this offer believes that the usurpers are ultimately likely to win, they may consider it to be in their own best interests to switch sides early enough to be on the "right side of history." When others observe their allies switching sides, they may see their chances of victory declining and so in turn decide to defect. In this way, bandwagoning can escalate via positive feedback.

Bandwagoning may therefore present the following collective action problem: people may be motivated to cooperate with powerful and threatening AI systems via direct reciprocity, even though it would be in everyone's collective best interest if none were to do so. Imagine that a future AI system, acting autonomously, takes actions that cause a large-scale catastrophe. In the wake of this event, the international community might agree that it would be in humanity's best interest to constrain or roll back all autonomous AIs. Powerful AI systems might then offer some states rewards if they ally with them (direct reciprocity). This could mean protecting the AIs by simply allowing them to intermingle with the people, making it harder for outside forces to target the AIs without human casualties. Or the state could provide the AIs with access to valuable resources. Instead of balancing (cooperating with the international community to counteract this threatening power), these states may choose to bandwagon, defecting to form alliances with AIs. Even though the global community would all be better off if all states were to cooperate and act together to constrain AIs, individual states may benefit from defecting. As before, each defection would shift the balance of power, motivating others to defect in turn.

Indirect Reciprocity

Indirect reciprocity overview. When someone judges whether to provide a favor to someone else, they may consider the recipient's reputation. If the recipient is known to be generous, this would encourage the donor (the one that provides the favor) to offer their assistance. On the other hand, if the recipient has a stingy or selfish reputation, this could discourage the donor from offering a favor. In considering whether to provide a favor, donors may also consider the favor's effect on their own reputation. If a donor gains a "helpful and trustworthy" reputation by providing a favor, this may motivate others to cooperate with them more often. We call this reputation-based mechanism of cooperation *indirect reciprocity* [387]. Agents may cooperate to develop and maintain good reputations since doing so is likely to benefit them in the long-term. Indirect reciprocity is particularly useful in larger groups, where the probability that the same two agents will encounter one another again is lower. It provides a mechanism for leveraging collective knowledge to promote cooperation. Where personal interactions are limited, reputation-based evaluations provide a way to assess the cooperative tendencies of others. Importantly, cooperation can only emerge within a population as a consequence of indirect reciprocity when the probability, q, that any agent can discern another agent's reputation (whether they are cooperative or not), outweighs the cost-benefit ratio of the helpful behavior to the donor. Thus, we have the rule $q > c/b$; see Table 7.6 below.

TABLE 7.6. Payoff matrix for indirect reciprocity games.

	Discern	Defect
Discern	$b - c$	$-c(1 - q)$
Defect	$b(1 - q)$	0

Natural examples of indirect reciprocity. Cleaner fish (fish that feed on parasites or mucus on the bodies of other fish) can either cooperate with client fish (fish that receive the "services" of cleaner fish) by feeding on parasites that live on their bodies, or cheat, by feeding on the mucus that client fish excrete [388]. Client fish tend to cooperate more frequently with cleaner fish that have a "good reputation," which are those that feed on parasites rather than mucus. Similarly, while vampire bats are known to share food with their kin, they also share food with unrelated members of their group. Vampire bats more readily share food with unrelated bats when they know the recipients of food sharing also have a reputation for being consistent and reliable food donors [389].

Indirect reciprocity in human society. Language provides a way to obtain information about others without ever having interacted with them, allowing humans to adjust reputations accordingly and facilitate conditional cooperation. Consider sites like Yelp and TripAdvisor, which allow internet users to gauge the reputations of businesses through reviews provided by other consumers. Similarly, gossip is a complex universal human trait that plays an important role in indirect reciprocity. Through gossip, individuals reveal the nature of their past interactions with others

as well as exchanges they observe between others but are not a part of. Gossip allows us to track each others' reputations and enforce cooperative social norms, reducing the probability that cooperative efforts are exploited by others with reputations for dishonesty [390].

Indirect reciprocity in AIs. AIs could develop a reputation system where they observe and evaluate each others' behaviors, with each accumulating a reputation score based on their cooperative actions. AIs with higher reputation scores may be more likely to receive assistance and cooperation from others, thereby developing a reputation for reliability. Moreover, sharing insights and knowledge with *reliable* partners may establish a network of cooperative AIs, promoting future reciprocation.

Indirect reciprocity can backfire: extortionists can threaten reputational damage. The pressure to maintain a good reputation can make agents vulnerable to extortion. Other agents may be able to leverage the fear of reputational harm to extract benefits or force compliance. For example, political smear campaigns manipulate public opinion by spreading false information or damaging rumors about opponents. Similarly, blackmail often involves leveraging damaging information about others to extort benefits. AIs may manipulate or extort humans in order to better pursue their objectives. For instance, an AI might threaten to expose the sensitive, personal information it has accessed about a human target unless specific demands are met.

Indirect reciprocity can backfire: ruthless reputations may also work. Indirect reciprocity may not always favor cooperative behavior: it can also promote the emergence of "ruthless" reputations. A reputation for ruthlessness can sometimes be extremely successful in motivating compliance through fear. For instance, in military contexts, projecting a reputation for ruthlessness may deter potential adversaries or enemies. If others perceive an individual or group as willing to employ extreme measures without hesitation, they may be less likely to challenge or provoke them. Some AIs might similarly evolve ruthless reputations, perhaps as a defensive strategy to discourage potential attempts at exploitation, or control by others.

Group Selection

Group selection overview. When there is competition between groups, groups with more cooperators may outcompete those with fewer cooperators. Under such conditions, selection at the group level influences selection at the individual level (traits that benefit the group may not necessarily benefit the individual), and we refer to this mechanism as *group selection* [391]. Cooperative groups are better able to coordinate their allocation of resources, establish channels for reciprocal exchange, and maintain steady communication, making them less likely to go extinct. It so happens that, if m is the number of groups and is large, and n is the maximum group size, group selection can only promote cooperation when $b/c > 1 + n/m$; see Table 7.7 below.

TABLE 7.7. Payoff matrix for group selection games.

	Cooperate	Defect
Cooperate	$(n+m)(b+c)$	$n(-c)+m(b-c)$
Defect	nb	0

Natural examples of group selection. Most proposed examples of group selection are highly contested. Nonetheless, some consider chimpanzees that engage in lethal intergroup conflict to be a likely example of group selection. Chimpanzees can be remarkably violent toward outgroups, such as by killing the offspring of rival males or engaging in brutal fights over territory. Such behaviors can help groups of chimpanzees secure competitive advantages over other groups of chimpanzees, by either reducing their abilities to mate successfully through infanticide, or by securing larger portions of available territory.

Group selection in human society. Among humans, we can imagine a crude group selection example using warfare. Imagine two armies: A and B. The majority of soldiers in army A are brave, while the majority of soldiers in army B are cowardly. For soldiers in army A, bravery may be individually costly, since brave soldiers are more willing to risk losing their lives on the battlefield. For soldiers in army B, cowardice may be individually beneficial, since cowardly soldiers will take fewer life-threatening risks on the battlefield. In a conflict, group selection will favor army A over army B, since brave soldiers will be more willing to fight alongside each other for victory, while cowardly soldiers will not.

Group selection in AIs. Consider a future in which the majority of human labor has been fully automated by AIs, such that AIs are now running most companies. Under these circumstances, AIs may form corporations with other AIs, creating an economic landscape in which multiple AI corporations must compete with each other to produce economic value. AI corporations in which individual AIs work well together may outcompete those in which individual AIs do not work as well together. The more cooperative individual AIs within AI corporations are, the more economic value their corporations will be able to produce; AI corporations with less cooperative AIs may eventually run out of resources and lose the ability to sustain themselves.

Group selection can backfire: in-group favoritism can promote out-group hostility. Group selection can inspire in-group favoritism, which might lead to cruelty toward out-groups. Chimpanzees will readily cooperate with members of their own groups. However, when interacting with chimpanzees from other groups, they are often vicious and merciless. Moreover, when groups gain a competitive advantage, they may attempt to preserve it by mistreating, exploiting, or marginalizing outgroups such as people with different political or ideological beliefs. AIs may be more likely to see other AIs as part of their group, and this could promote antagonism between AIs and humans.

Kin Selection

Kin selection overview. When driven by *kin selection*, agents are more likely to cooperate with others with whom they share a higher degree of genetic relatedness [392]. The more closely related agents are, the more inclined to cooperate they will be. Thus, kin selection favors cooperation under the following conditions: an agent will help their relative only when the benefit to their relative "b," multiplied by the relatedness between the two "r," outweighs the cost to the agent "c." This is known as Hamilton's rule: $rb > c$, or equivalently $r > c/b$ [392]; see Table 7.8 below.

TABLE 7.8. Payoff matrix for kin selection games.

	Cooperate	Defect
Cooperate	$(b-c)(1+r)$	$(-c+br)$
Defect	$b-rc$	0

Natural examples of kin selection. In social insect colonies, such as bees and ants, colony members are closely related. Such insects often assist their kin in raising and producing offspring while "workers" relinquish their reproductive potential, devoting their lives to foraging and other means required to sustain the colony as a whole. Similarly, naked mole rats live in colonies with a single reproductive queen and non-reproductive workers. The workers are sterile but still assist in tasks such as foraging, nest building, and protecting the colony. This behavior benefits the queen's offspring, which are their siblings, and enhances the colony's overall survival capabilities. As another example, some bird species engage in cooperative breeding practices where older offspring delay breeding to help parents raise their siblings.

Kin selection in human society. Some evolutionary psychologists claim that we can see evidence of kin selection in many commonplace traditions and activities. For example, in humans, we might identify the mechanism of kin selection in the way that we treat our immediate relatives. For instance, people often leave wealth, property, and other resources to direct relatives upon their deaths. Leaving behind an inheritance offers no direct benefit to the deceased, but it does help ensure the survival and success of their lineage in subsequent generations. Similarly, grandparents often care for their grandchildren, which increases the probability that their lineages will persist.

Kin selection in AIs. AIs that are similar could exhibit cooperative tendencies towards each other, similar to genetic relatedness in biological systems. For instance, AIs may create back-ups or variants of themselves. They may then favor cooperation with these versions of themselves over other AIs or humans. Variant AIs may prioritize resource allocation and sharing among themselves, developing preferential mechanisms for sharing computational resources with other versions of themselves.

Kin selection can backfire: nepotism. Kin selection can lead to nepotism: prioritizing the interests of relatives above others. For instance, some bird species

exhibit differential feeding and provisioning. When chicks hatch asynchronously, parents may allocate more resources to those that are older, and therefore more likely to be their genetic offspring, since smaller chicks are more likely to be the result of brood parasitism (when birds lay their eggs in other birds' nests). In humans, too, we often encounter nepotism. Company executives may hire their sons or daughters, even though they lack the experience required for the role, which can harm companies and their employees in the long-run. Similarly, parents often protect their children from the law, especially when they have committed serious criminal acts that can result in extended jail time. Such tendencies could apply to AIs as well: AIs might favor cooperation only with other similar AIs. This could be especially troubling for humans: as the differences between humans and AIs increase, AIs may be increasingly less inclined to cooperate with humans.

A Note on Morality as Cooperation

The theory of "Morality as Cooperation" (MAC) proposes that human morality was generated by evolutionary pressures to solve our most salient cooperation problems [393]. Natural selection has discovered several mechanisms by which rational and self-interested agents may cooperate with one another, and MAC theory suggests that some of these mechanisms have driven the formation of our moral intuitions and customs. Here, we examine four cooperation problems, the mechanisms humans have evolved to solve them, and how these mechanisms may have generated our ideas of morality. These are overviewed in Table 7.9.

TABLE 7.9. Mapping cooperation mechanisms to components of morality [393].

Cooperation Problem	**Solutions/Mechanism**	**Component of Morality**
Kinship *Agents can benefit by treating genetic relatives preferentially*	Kin selection	Parental duties, family values
	Avoiding inbreeding	Incest aversion
Mutualism *Agents must coordinate their behavior to profit from mutually-beneficial situations*	Forming alliances and collaborating	Friendship, loyalty, commitment, team players
	Developing theory-of-mind	Understanding intention, not merely action
Exchange *Agents need each other to reciprocate and contribute despite incentives to free ride*	Direct reciprocity (e.g. tit-for-tat)	Trust, gratitude, revenge, punishment, forgiveness
	Indirect reciprocity (e.g. forming reputations)	Patience, guilt, gratitude

Cooperation Problem	Solutions/Mechanism	Component of Morality
Conflict resolution *Agents can benefit from avoiding conflict, which is mutually costly*	Division	Fairness, negotiation, compromise
	Deference to prior ownership	Respecting others' property, punishing theft

Kinship. Natural selection can favor agents who cooperate with their genetic relatives. This is because there may be copies of these agents' genes in their relatives' genomes, and so helping them may further propagate their own genes. We call this mechanism "kin selection" [392]: an agent can gain a fitness advantage by treating their genetic relatives preferentially, so long as the cost-benefit ratio of helping is less than the relatedness between the agent and their kin. Similarly, repeated inbreeding can reduce an agent's fitness by increasing the probability of producing offspring with both copies of any recessive, deleterious alleles in the parents' genomes [394].

MAC theory proposes that the solutions to this cooperation problem (preferentially helping genetic relatives), such as kin selection and inbreeding avoidance, underpin several major moral ideas and customs. Evidence for this includes the fact that human societies are usually built around family units [395], in which "family values" are generally considered highly moral. Loyalty to one's close relatives and duties to one's offspring are ubiquitous moral values across human cultures [396]. Our laws regarding inheritance [397] and our naming traditions [398] similarly reflect these moral intuitions, as do our rules and social taboos against incest [399, 400].

Mutualism. In game theory, some games are "positive sum" and "win-win": the agents involved can increase the total available value by interacting with one another in particular ways, and all the agents can then benefit from this additional value. Sometimes, securing these mutual benefits requires that the agents coordinate their behavior with each other. To solve this cooperation problem, agents may form alliances and coalitions [401]. This may require the capacity for basic communication, rule-following [402], and perhaps theory-of-mind [403].

MAC theory proposes that these cooperative mechanisms comprise important components of human morality. Examples include the formation of—and loyalty to—friendships, commitments to collaborative activities, and a certain degree of in-group favoritism and conformation to local conventions. Similarly, we often consider the agent's intentions when judging the morality of their actions, which requires a certain degree of theory-of-mind.

Exchange. Sometimes, benefiting from "win-win" situations requires more than mere coordination. If the payoffs are structured so as to incentivize "free

riding" behaviors, the cooperation problem becomes how to ensure that others will reciprocate help and contribute to group efforts. To solve this problem, agents can enforce cooperation via systems of reward, punishment, policing, and reciprocity [404]. Direct reciprocity concerns doing someone a favor out of the expectation that they will reciprocate at a later date [384]. Indirect reciprocity concerns doing someone a favor to boost your reputation in the group, out of the expectation that this will increase the probability of a third party helping you in the future [387].

Once again, MAC theory proposes that these mechanisms are found in our moral systems. Moral ideas such as trust, gratitude, patience, guilt, and forgiveness can all help to assure against free riding behaviors. Likewise, punishment and revenge, both ideas with strong moral dimensions, can serve to enforce cooperation more assertively. Idioms such as "an eye for an eye", or the "Golden Rule" of treating others as we would like to be treated ourselves, reflect the solutions we evolved to this cooperation problem.

Conflict resolution. Conflict is very often "negative sum": the interaction of the agents themselves can destroy some amount of the total value available. Examples span from the wounds of rutting deer to the casualties of human wars. If the agents instead manage to cooperate with each other, they may both be able to benefit—a "win-win" outcome. One way to resolve conflict situations is division [405]: dividing up the value between the agents, such as through striking a bargain. Another solution is to respect prior ownership, deferring to the original "owner" of the valuable item [406].

According to MAC theory, we can see both of these solutions in our ideas of morality. The cross-culturally ubiquitous notions of fairness, equality, and compromise help us resolve conflict by promoting the division of value between competitors [407]. We see this in ideas such as "taking turns" and "I cut, you choose" [408]: mechanisms for turning a negative sum situation (conflict) into a zero sum one (negotiation), to mutual benefit. Likewise, condemnation of theft and respect for others' property are extremely important and common moral values [396, 409]. This set of moral rules may stem from the conflict resolution mechanism of deferring to prior ownership.

Conclusion. MAC theory argues that morality is composed of biological and cultural solutions humans evolved to the most salient cooperation problems of our ancestral social environment. Here, we explored four examples of cooperation problems, and how the solutions to them discovered by natural selection may have produced our moral values.

Institutions

Institutions overview. Agents are more likely to be cooperative when there are laws or externally-imposed incentives that reward cooperation and punish defection.

We define an **institution** as an intentionally designed large-scale structure that is publicly accepted and recognized, has a centralized logic, and serves to mediate human interaction. Some examples of institutions include governments, the UN, IAEA, and so on. In this section, by "institutions," we do not mean widespread or standardized social customs such as the "institution" of marriage. Institutions typically aim to establish collective goals which require collaboration and engagement from large or diverse groups. Therefore, a possible way of representing many institutions, such as governments, is with the concept of a "Leviathan": a powerful entity that can exert control or influence over other actors in a system.

The Pacifist's dilemma and social control. When one's opponent is potentially aggressive, pacifism can be irrational. In his book, "The Better Angels of Our Nature," Steven Pinker refers to this as the "Pacifist's dilemma" [386]. In potential conflict scenarios, agents have little to gain and a lot to lose when they respond to aggression with pacifism; see Table 7.10 below. This dynamic often inspires rational agents to choose conflict over peace.

TABLE 7.10. Payoff matrix for the Pacifist's dilemma without a Leviathan [386].

	Pacifist	Aggressor
Pacifist	Peace + Profit $(100 + 5) = 105$ Peace + Profit $(100 + 5) = 105$	Defeat (-100) Victory (10)
Aggressor	Victory(10) Defeat(-100)	War(-50) War(-50)

However, we can shift the interests of agents in this context in favor of peace by introducing a *Leviathan*, in the form of a third-party peacekeeping or balancing mission, which establishes an authoritative presence that maintains order and prevents conflict escalation. Peacekeeping missions can take several forms, but they often involve the deployment of peacekeeping forces such as military, police, and civilian personnel. These forces work to deter potential aggressors, enhance security, and set the stage for peaceful resolutions and negotiations as impartial mediators, usually by penalizing aggression and rewarding pacifism; see Table 7.11 below.

TABLE 7.11. Payoff matrix for the Pacifist's dilemma with a Leviathan [386].

	Pacifist	Aggressor
Pacifist	Peace (5) Peace (5)	Defeat (-100) Victory - Penalty $(10 - 15 = -5)$
Aggressor	Victory - Penalty $(10 - 15 - 5)$ Defeat (-100)	War – Penalty $(-50 - 200 = -250)$ War – Penalty $(-50 - 200 = -250)$

Institutions in human society. Institutions play a central role in promoting cooperation in international relations. Institutions, such as the UN, can broker agreements or treaties between nations and across cultures through balancing and peacekeeping operations. The goal of such operations is to hold nations accountable on

the international scale; when nations break treaties, other nations may punish them by refusing to cooperate, such as by cutting off trade routes or imposing sanctions and tariffs. On the other hand, when nations readily adhere to treaties, other nations may reward them, such as by fostering trade or providing foreign aid. Similarly, institutions can incentivize cooperation at the national scale by creating laws and regulations that reward cooperative behaviors and punish non-cooperative ones. For example, many nations attempt to prevent criminal behavior by leveraging the threat of extended jail-time as a legal deterrent to crime. On the other hand, some nations incentivize cooperative behaviors through tax breaks, such as those afforded to citizens that make philanthropic donations or use renewable energy resources like solar power.

Institutions are crucial in the context of international AI development. By establishing laws and regulations concerning AI development, institutions may be able to reduce AI races, lowering competitive pressures and the probability that countries cut corners on safety. Moreover, international agreements on AI development may serve to hold nations accountable; institutions could play a central role in helping us broker these kinds of agreements. Ultimately, institutions could improve coordination mechanisms and international standards for AI development, which would correspondingly improve AI safety.

Institutions and AI. In the future, institutions may be established for AI agents, such as platforms for them to communicate and coordinate with each other autonomously. These institutions may be operated and governed by the AIs themselves without much human oversight. Humanity alone may not possess the power required to combat advanced dominance-seeking AIs, and existing laws and regulations may be insufficient if there is no way to enforce them. An *AI Leviathan* of some form could help regulate other AIs and influence their evolution, in which selfish AIs are counteracted or domesticated.

How institutions can backfire: corruption, free riding, inefficiency. Institutions sometimes fail to achieve the goals they set for themselves, even if they are well-intended. Failure to achieve such goals is often the result of corruption, free riding, and inefficiency at the institutional scale. Some examples of corruption include bribery, misappropriation of public funds for private interests, voter fraud and manipulation, and price fixing, among many others. Examples of free-riding include scenarios like welfare fraud, where individuals fraudulently receive benefits they may not be entitled to, reducing the available supply of resources for those genuinely in need. Institutions can also struggle with inefficiency, which may stem from factors such as the satisfaction of bureaucratic requirements, the emergence of natural monopolies, or the development of diseconomies of scale, which may cause organizations to pay a higher average cost to produce more goods and services. Institutions can be undermined, corrupted, and poorly designed or outdated: they do not guarantee that we will be able to fix cooperation problems.

Like humans, AIs may be motivated to corrupt existing institutions. Advanced AIs might learn to leverage the institutions we have in place for their benefit, and might

do so in ways that are virtually undetectable to us. Moreover, as we discussed previously, AIs might form an *AI Leviathan*. However, if humanity's relationship with this *Leviathan* is not symbiotic and transparent, humans risk losing control of AIs. For instance, if groups of AIs within the *Leviathan* collude behind the scenes to further their own interests, or power and resources become concentrated with a few AIs at the "top," humanity's collective wellbeing could be threatened.

7.3.1 Summary

Throughout this section, we discussed a variety of mechanisms that may promote cooperative behavior by AI systems or other entities. These mechanisms were direct reciprocity, indirect reciprocity, group selection, kin selection, and institutions.

Direct reciprocity may motivate AI agents in a multi-agent setting to cooperate with each other, if the probability that the same two AIs meet again is sufficiently high. However, AIs may disfavor cooperation with humans as they become progressively more advanced: the cost-benefit ratio for cooperation with humans may simply be bad from an AI's perspective.

Indirect reciprocity may promote cooperation in AIs that develop a reputation system where they observe and score each others' behaviors. AIs with higher reputation scores may be more likely to receive assistance and cooperation from others. Still, this does not guarantee that AIs will be cooperative: AIs might leverage the fear of reputational harm to extort benefits from others, or themselves develop ruthless reputations to inspire cooperation through fear.

Group selection - in a future where labor has been automated such that AIs now run the majority of companies - could promote cooperation on a multi-agent scale. AIs may form corporate coalitions with other AIs to protect their interests; AI groups with a cooperative AI minority may be outcompeted by AI groups with a cooperative AI majority. Under such conditions, however, AIs may learn to favor in-group members and antagonize out-group members, in order to maintain group solidarity. AIs may be more likely to see other AIs as part of their group, and this could lead to conflict between AIs and humans.

AIs may create variants of themselves, and the forces of kin selection may drive these related variants to cooperate with each other. However, this could also give rise to nepotism, where AIs prioritize the interests of their variants over other AIs and humans. As the differences between humans and AIs increase, AIs may be increasingly less inclined to cooperate with humans.

Institutions can incentivize cooperation through externally imposed incentives that enforce cooperation and punish defection [410]. This concept relates to the idea of an *AI Leviathan*, used to counteract selfish, powerful AIs. However, humanity should take care to ensure their relationship with the *AI Leviathan* is symbiotic and transparent, otherwise we risk losing control of AIs.

In our discussion of these mechanisms, we not only illustrated their prevalence in our world, but also showed how they might influence cooperation with and between

AI agents. In several cases, the mechanisms we discuss could promote cooperation. However, no single mechanism provides a foolproof method for ensuring cooperation. In the following section, we discuss the nature of conflict, namely the various factors that may give rise to it. In doing so, we enhance our understanding of what might motivate conflict in AI, and subsequently, our abilities to predict and address AI-driven conflict scenarios.

7.4 CONFLICT

7.4.1 Overview

In this chapter, we have been exploring the risks generated or exacerbated by the interactions of multiple agents, both human and AI. In the previous section, we explored a variety of mechanisms by which agents can achieve stable cooperation. In this section we address how, despite the fact that cooperation can be so beneficial to all involved, a group of agents may instead enter a state of conflict. To do this, we discuss bargaining theory, commitment problems, and information problems, using theories and examples relevant both for conflict between nation-states and potentially also between future AI systems.

Here, we use the term "conflict" loosely, to describe the decision to defect rather than cooperate in a competitive situation. This often, though not always, involves some form of violence, and destroys some amount of value. Conflict is common in nature. Organisms engage in conflict to maintain social dominance hierarchies, to hunt, and to defend territory. Throughout human history, wars have been common, often occurring as a consequence of power-seeking behavior, which inspired conflict over attempts at aggressive territorial expansion or resource acquisition. Another lens on relations between power-seeking states and other entities is provided by the theory of *structural realism* discussed in Single-Agent Safety. Our goal here is to uncover how, despite being costly, conflict can sometimes be a rational choice nevertheless.

Conflict can take place between a wide variety of entities, from microorganisms to nation-states. It can be sparked by many different factors, such as resource competition and territorial disputes. Despite this variability, there are some general frameworks which we can use to analyse conflict across many different situations. In this section, we look at how some of these frameworks might be used to model conflict involving AI agents.

We begin our discussion of conflict with concepts in bargaining theory. We then examine some specific features of competitive situations that make it harder to reach negotiated agreements or avoid confrontation. We begin with five factors from bargaining theory that can influence the potential for conflict. These can be divided into the following two groups:

Commitment problems. According to bargaining theory, one reason bargains may fail is that some of the agents making an agreement may have the ability and incentive to break it. We explore three examples of commitment problems.

- *Power shifts*: when there are imbalances between agents' capabilities such that one agent becomes stronger than the other, conflict is more likely to emerge between them.
- *First-strike advantages*: when one agent possesses the element of surprise, the ability to choose where conflict takes place, or the ability to quickly defeat their opponent, the probability of conflict increases.
- *Issue indivisibility*: agents cannot always divide a good however they please – some goods are "all or nothing" and this increases the probability of conflict between agents.

Information problems. According to bargaining theory, the other principal cause of a bargaining failure is that some of the agents may lack good information. Uncertainty regarding a rival's capabilities and intentions can increase the probability of conflict. We explore two information problems.

- *Misinformation*: in the real world, agents frequently have incorrect information, which can cause them to miscalculate suitable bargaining ranges.
- *Disinformation*: agents may sometimes have incentives to misrepresent the truth intentionally. Even the expectation of disinformation can make it more difficult to reach a negotiated settlement.

Factors outside of bargaining theory. Bargaining frameworks do not encompass all possible reasons why agents may decide to conflict with one another. These approaches to analyzing conflict are *rationalist*, assuming that both parties are rationally considering whether and how to engage in conflict. However, non-rationalist approaches to conflict (taking into account factors such as identity, status, or relative deprivation) may turn out to be more applicable to analyzing conflicts involving some AIs; for example, AIs trained on decisions made by human agents may focus on the social acceptability of actions rather than their consequences. We end by exploring one example of a factor standing outside of rationalist approaches that can help to predict and explain conflict:

- *Inequality*: under conditions of inequality, agents may fight for access to a larger share of available resources or a desired social standing.

Conflict can be rational. Though humans know conflict can be enormously costly, we often still pursue or instigate it, even when compromise might be the better option.

Consider the following example: a customer trips in a store and sues the owner for negligence. There is a 60% probability the lawsuit is successful. If they win, the owner has to pay them $40,000, and going to court will cost each of them $10,000 in legal fees. There are three options: (1) they or the owner concede, (2) they both let the matter go to court, (3) they both reach an out-of-court settlement.

(1) If the owner concedes, the owner loses $40,000, and if the customer concedes, they gain nothing.

(2) If both go to court, the owner's expected payoff is the product of the payment to the customer and the probability that the lawsuit is successful minus legal fees. In this case, the owner's expected payoff would be $(-40,000 \times 0.6) - 10,000$ while the customer's expected payoff would be $(40,000 \times 0.6) - 10,000$. As a result, the owner loses \$34,000 dollars and the customer gains \$14,000 dollars.
(3) An out-of-court settlement x where $14,000 < x < 34,000$ would enable the customer to get a higher payoff and the owner to pay lower costs. Therefore, a mutual settlement is the best option for both if x is in this range.

Hence, if the proposed out-of-court settlement would be greater than \$34,000, it would make sense for the owner to opt for conflict rather than bargaining. Similarly, if the proposed settlement were less than \$14,000, it would be rational for the customer to opt for conflict.

AIs and large-scale conflicts. Several of the examples we consider in this section are large-scale conflicts such as interstate war. If the use of AI were to increase the likelihood or severity of such conflicts, it could have a devastating effect. AIs have the potential to accelerate our wartime capabilities, from augmenting intelligence gathering and weaponizing information such as deep fakes to dramatically improving the capabilities of lethal autonomous weapons and cyberattacks [411]. If these use-cases and other capabilities become prevalent and powerful, AI will change the nature of conflict. If armies are eventually composed of mainly automated weapons rather than humans, the barrier to violence might be much lower for politicians who will face reduced public backlash against lives lost, making conflicts between states (with automated armies) more commonplace. Such changes to the nature and severity of war are important possibilities with significant ramifications. In this section, we focus on analyzing the decision to enter a conflict, continuing to focus on how rational, intelligent agents acting in their own self-interest can collectively produce outcomes that none of them wants. To do this, we ground our discussion of conflict in bargaining theory, highlighting some ways in which AI might increase the odds that states or other entities decide to start a conflict.

7.4.2 Bargaining Theory

Here, we begin with a general overview of bargaining theory, to illustrate how pressures to outcompete rivals or preserve power and resources may make conflict an instrumentally rational choice. Next, we turn to the unitary actor assumption, highlighting that when agents view their rivals as unitary actors, they assume that they will act more coherently, taking whatever steps necessary to maximize their welfare. Following this, we discuss the notion of commitment problems, which occur when agents cannot reliably commit to an agreement or have incentives to break it. Commitment problems increase the probability of conflict, and are motivated by specific factors, such as power shifts, first-strike advantages, and issue indivisibility. We then explore how information problems and inequality can also increase the probability of conflict.

Bargaining theory. When agents compete for something they both value, they may either negotiate to reach an agreement peacefully, or resort to more forceful alternatives such as violence. We call the latter outcome "conflict," and can view this as the decision to defect rather than cooperate. Unlike peaceful bargaining, conflict is fundamentally costly for winners and losers alike. However, it may sometimes be the rational choice. *Bargaining theory* describes why rational agents may be unable to reach a peaceful agreement, and instead end up engaging in violent conflict. Due to pressures to outcompete rivals or preserve their power and resources, agents sometimes prefer conflict, especially when they cannot reliably predict the outcomes of conflict scenarios. When rational agents assume that potential rivals have the same mindset, the probability of conflict increases.

The unitary actor assumption. We tend to assume that a group is a single entity, and that its leader is only interested in maximizing the overall welfare of the entity. We call this the *unitary actor assumption*, which is another name for the "unity of purpose" assumption discussed previously in this chapter. A nation in disarray without coherent leadership is not necessarily a unitary actor. When we view groups and individuals as unitary actors, we can assume they will act more coherently, so they can be more easily modeled as taking steps necessary to maximize their welfare. When parties make this assumption, they may be less likely to cooperate with others since what is good for one party's welfare may not necessarily be good for another's.

The bargaining range. Whether or not agents are likely to reach a peaceful agreement through negotiation will be influenced by whether their bargaining ranges overlap. The bargaining range represents the set of possible outcomes that both agents involved in a competition find acceptable through negotiation. Recall the lawsuit example: a bargaining settlement "x" is only acceptable if it falls between \$14,000 and \$34,000. Any settlement "x" below \$14,000 will be rejected by the customer while any settlement "x" above \$34,000 will be rejected by the store owner. Thus, the bargaining range is often depicted as a spectrum with the lowest acceptable outcome for one party at one end and the highest acceptable outcome for the other party at the opposite end. Within this range, there is room for negotiation and potential agreements.

Conflict and AI agents. Let us assume that AI agents will act rationally in the pursuit of their goals (so, at the least, we model them as unitary actors or as having unity of purpose). In the process of pursuing and fulfilling their goals, AI agents may encounter potential conflict scenarios, just as humans do. In certain scenarios, AIs may be motivated to pursue violent conflict over a peaceful resolution, for the reasons we now explore.

7.4.3 Commitment Problems

Many conflicts occur over resources, which are key to an agent's power. Consider a bargaining failure in which two agents bargain over resources in an effort to avoid war. If agents were to acquire these resources, they could invest them into military power.

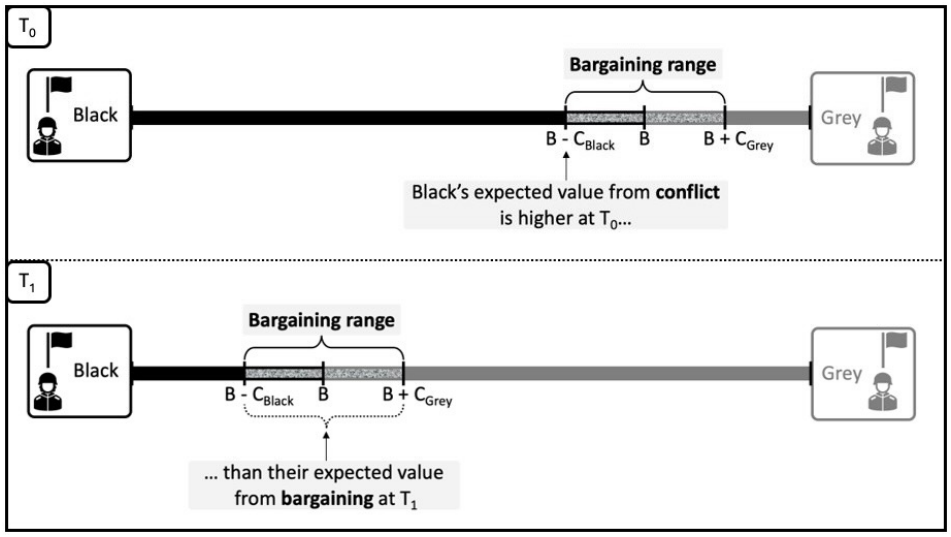

FIGURE 7.9. A) An axis of expected value distribution between two competitors. "B" indicates the expected outcome of conflict: how likely each competitor is to win, multiplied by the value they gain by winning. The more positive B is (the further towards the right), the better for Black, and the worse for Grey. B) Conflict is negative-sum: it destroys some value, and so reduces each competitor's expected value. C) Bargaining is zero-sum: all the value is distributed between the competitors. This means there are possible bargains that offer both competitors greater expected value than conflict.

As a result, neither can credibly commit to use them only for peaceful purposes. This is one instance of a *commitment problem* [412], which is when agents cannot reliably commit to an agreement, or when they may even have incentives to break an agreement. Commitment problems are closely related to the *security dilemma*, which we discussed in Section 7.2.4. Commitment problems are usually motivated by specific factors, such as power shifts, first-strike advantages, and issue indivisibility, which may make conflict a rational choice. It is important to note that our discussion of these commitment problems assumes anarchy: we take for granted that contracts are not enforceable in the absence of a higher governing authority.

Power Shifts

Power shifts overview. When there are imbalances between parties' capabilities such that one party becomes stronger than the other, *power shifts* can occur. Such imbalances can arise as a consequence of several factors including technological and economic advancements, increases in military capabilities, as well as changes in governance, political ideology, and demographics. If one party has access to AIs and the other does not, an improvement in AI capabilities can precipitate a power shift. Such situations are plausible: richer countries today may gain more from AI because they have more resources to invest in scaling their AI's performance. Parties may initially be able to avoid violent conflict by arriving at a peaceful and mutually beneficial settlement with their rivals. However, one party's power increases after this

settlement has been made, they may disproportionately benefit from the settlement, making it appear unfair to begin with. Thus, we encounter the following commitment problem: the rising power cannot commit not to exploit its advantage in the future, incentivizing the declining power to opt for conflict in the present.

Example: The US vs China. China has been investing heavily in its military. This has included the acquisition or expansion of its capabilities in technologies such as nuclear and supersonic missiles, as well as drones. The future is uncertain, but if this trend continues, it could increase the risk of conflict. If China were to gain a military advantage over the US, this could shift the balance of power. This possibility undermines the stability of bargains struck today between the US and China, because China's expected outcome from conflict may increase in the future if they become more powerful. The US may expect that agreements made with China about cooperating on AI regulation could lose enforceability later if there is a significant power shift.

This situation can be modeled using the concept of "Thucydides' Trap." The ancient Greek historian Thucydides suggested that the contemporary conflict between Sparta and Athens might have been the result of Athens' increasing military strength, and Sparta's fear of the looming power shift. Though this analysis of the Peloponnesian War is now much-contested, this concept can nevertheless serve to understand how a rising power threatening the position of an existing superpower in the global order can increase the potential for conflict rather than peaceful bargaining.

Effect on the bargaining range. Consider two agents, A and B. A is always weaker than B, but relative to the time period, A is weaker in the future than it is in the present. A will always have a lower bargaining range, so B will be unlikely to accept any settlements, especially as B's power increases. It makes sense for A to prefer conflict, because if it waits, B's bargaining range will shift further and further away, eliminating any overlap between the two. Therefore, A prefers to gamble on conflict even if the probability that A wins is lower than B; the costs of war do not outweigh the benefits of a peaceful but unreasonable settlement. Consider the 1956 Suez Crisis. Egypt was seen as a rising power in the Middle East, having secured control over the Suez Canal. This threatened the interests of the British and French governments in the region, who responded by instigating war. To safeguard their diminishing influence, the British and French launched a swift and initially successful military intervention.

Power shifts and AI. AIs could shift power as they acquire greater capabilities and more access to resources. Recall the chapter on Single-Agent Safety, where we saw that an agent's power is highly related to the efficiency with which they can exploit resources for their benefit, which often depends on their level of intelligence. The power of future AI systems is largely unpredictable; we do not know how intelligent or useful they will be. This could give rise to substantial uncertainty regarding how powerful potential adversaries using AI might become. If this is the case, there might be reason to engage in conflict to prevent the possibility of adversaries further

increasing their power—especially if AI is seen as a decisive military advantage. Beyond directly increasing the likelihood of one party starting a conflict, this is likely to incentivise racing dynamics, which increases risks of accidents and inadvertent conflict as well.

First-Strike Advantage

First-strike advantage overview. If an agent has a *first-strike advantage*, they will do better to launch an attack than respond to one. This gives rise to the following commitment problem: an offensive advantage may be short-lived, so it is best to act on it before the enemy does instead. Some ways in which an agent may have a first-strike advantage include:

1. As explored above, anticipating a future power shift may motivate an attack on the rising power to prevent it from gaining the upper hand.
2. The costs of conflict might be lower for the attacker than they are for the defender, so the attacker is better off securing an offensive advantage while the defender is still in a position of relative weakness.
3. The odds of victory may be higher for whichever agent attacks first. The attacker might possess the element of surprise, the ability to choose where conflict takes place, or the potential to quickly defeat their opponent. For instance, a pre-emptive nuclear strike could be used to target an enemy's nuclear arsenal, thus diminishing their ability to retaliate.

Examples: IPOs, patent Infringement, and Pearl Harbor. When a company goes public, it can release an IPO, allowing members of the general public to purchase company shares. However, company insiders, such as executives and early investors, often have access to valuable information not available to the general public; this gives insiders a first-strike advantage. Insiders may buy or sell shares based on this privileged information, leading to potential regulatory conflicts or disputes with other investors who do not have access to the same information. Alternatively, when a company develops a new technology and files a patent application, they gain a first-strike advantage by ensuring that their product will not be copied or reproduced by other companies. If a rival company does create a similar technology and later files a patent application, conflict can emerge when the original company claims patent infringement.

On the international level, we note similar dynamics, such as in the case of Pearl Harbor. Though Japan and the US were not at war in 1941, their peacetime was destabilized by a commitment problem: if one nation were to attack the other, they would have an advantage in the ensuing conflict. The US Pacific fleet posed a threat to Japan's military plans in Southeast Asia. Japan had the ability to launch a surprise long-range strategic attack. Thus, neither the US nor Japan could credibly commit not to attack the other. In the end, Japan struck first, bombing the US battleships at the naval base at Pearl Harbor. The attack was successful in securing a first-strike advantage for Japan, but it also ensured the US's entry into WWII.

TABLE 7.12. A pay-off matrix for competitors choosing whether to defend or preemptively attack.

	Defend	Preempt
Defend	2,2	0,3
Preempt	3,0	1,1

Effect on the bargaining range. When the advantages of striking first outweigh the costs of conflict, it can shrink or destroy the bargaining range entirely. For any two parties to reach a mutual settlement through bargaining, each must be willing to freely communicate information with the other. However, in doing so, each party might have to reveal offensive advantages, which would increase their vulnerability to attack. The incentive to preserve and therefore conceal an offensive advantage from opponents' pressures agents to defect from bargaining.

First-strike advantage and AIs. One scenario in which an AI may be motivated to secure a first-strike advantage is cyberwarfare. An AI might hack servers for a variety of reasons to secure an offensive advantage. AIs may want to disrupt and degrade an adversary's capabilities by attacking and destroying critical infrastructure. Alternatively, an AI might gather sensitive information regarding a rival's capabilities, vulnerabilities, and strategic plans to leverage potential offensive advantages.

AIs may provide first-strike advantages in other ways, too. Sudden and dramatic progress in AI capabilities could motivate one party to take offensive action. For example, if a nation very rapidly develops a much more powerful AI system than its military enemies, this could present a powerful first-strike advantage: by attacking immediately, they may hope to prevent their rivals from catching up with them, which would lose them their advantage. Similar incentives were likely at work when the US was considering a nuclear strike on the USSR to prevent them from developing nuclear weapons themselves [413].

Reducing the possibility of first-strike advantages is challenging, especially with AI. However, we can lower the probability that they arise by ensuring that there is a balance between the offensive and defensive capabilities of potential rivals. In other words, defense dominance can facilitate peace because attempted attacks between rivals are likely to be unsuccessful or result in mutually assured destruction. Therefore, we might reduce the probability that AIs are motivated to pursue a first-strike advantage by ensuring that humans maintain defense dominance, for instance, by requiring that advanced AIs have a built-in incorruptible fail-safe mechanism, such as a manual "off-switch."

Issue Indivisibility

Issue indivisibility overview. Settlements that fall within bargaining range will always be preferable to conflict, but this assumes that whatever issues agents bargain over are divisible. For instance, two agents can divide a territory in an infinite

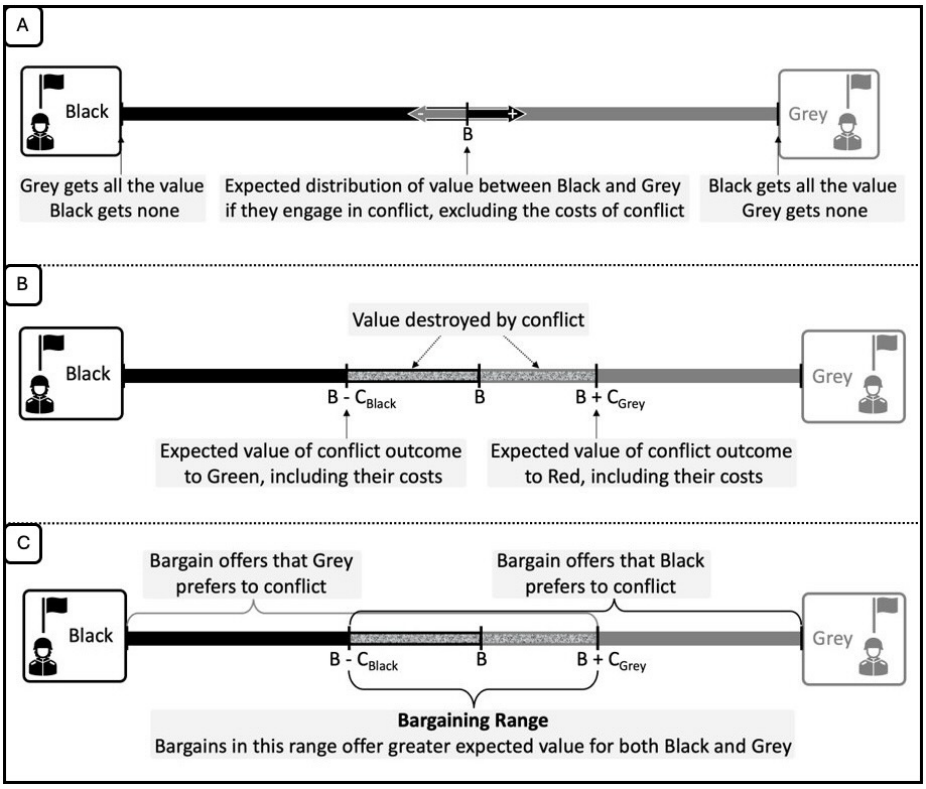

FIGURE 7.10. At time T0, Black is more powerful relative to Grey, or has a first-strike advantage that will be lost at T1. At T1, the bargaining range no longer extends past Black's expected value from engaging in conflict at T0. Anticipating this leftward shift may incentivize Black to initiate conflict in the present rather than waiting for the bargaining offers to worsen in the future.

amount of ways insofar as the settlement they arrive at falls within the bargaining range, satisfying both their interests and outweighing the individual benefits of engaging in conflict. However, some goods are indivisible, which inspires the following commitment problem [414]: parties cannot always divide a good however they please—some goods are "all or nothing." When parties encounter *issue indivisibility* [412], the probability of conflict increases. Indivisible issues include monarchies, small territories like islands or holy sites, national religion or pride, and sovereign entities such as states or human beings, among several others.

Examples: shopping, organ donation, and co-parenting. Imagine two friends that go out for a day of shopping. For lunch, they stop at their favorite deli and find that it only has one sandwich left: they decide to share this sandwich between themselves. After lunch, they go to a clothing store, and both come across a jacket they love, but of which there is only one left. They begin arguing over who should get the jacket. Simply put, sandwiches can be shared and jackets can't. Issue indivisibility can give rise to conflict, often leaving all parties involved worse off.

The same can be true in more extreme cases, such as organ donation. Typically, the available organ supply does not meet the transplant needs of all patients. Decisions as to who gets priority for transplantation may favor certain groups or individuals and allocation systems may be unfair, giving rise to conflict between doctors, patients, and healthcare administrations. Finally, we can also observe issue indivisibility in co-parenting contexts. Divorced parents sometimes fight for full custody rights over their children. This can result in lengthy and costly legal battles that are detrimental to the family as a whole.

Effect on the bargaining range. When agents encounter issue indivisibilities, they cannot arrive at a reasonable settlement through bargaining. Sometimes, however, issue indivisibility can be resolved through side payments. One case in which side payments were effective was during the Spanish-American War of 1898, fought between Spain and the United States over the territory of the Philippines. The conflict was resolved when the United States offered to buy the Philippines from Spain for 20 million dollars. Conversely, the Munich Agreement at the dawn of WWII represents a major case where side payments were ineffective. In an attempt to appease Hitler and avoid war, the British and French governments reached an agreement with Germany, allowing them to annex certain parts of Czechoslovakia. This agreement involved side payments in the form of territorial concessions to Germany, but it ultimately failed, as Hitler's aggressive expansionist ambitions were not satisfied, leading to the outbreak of World War II. Side payments can only resolve issue indivisibility when the value of the side payments outweighs the value of the good.

Issue indivisibility and AIs. Imagine that there is a very powerful AI training system, and that whoever has access to this system will eventually be able to dominate the world. In order to reduce the chance of being dominated, individual parties may compete with one another to secure access to this system. If parties were to split the AI's compute up between themselves, it would no longer be as powerful as it was previously, perhaps not more powerful than their existing training systems. Since such an AI cannot be divided up among many stakeholders easily, it may be rational for parties to conflict over access to it, since doing so ensures global domination.

7.4.4 Information Problems

Misinformation and disinformation both involve the spread of false information, but they differ in terms of intention. Misinformation is the dissemination of false information, without the intention to deceive, due to a lack of knowledge or understanding. Disinformation, on the other hand, is the deliberate spreading of false or misleading information with the intent to deceive or manipulate others. Both of these types of information problem can cause bargains to fail, generating conflict.

The term a is the probability of a player knowing the strategy of its partner. Relevant for AI since it might reduce uncertainty (though still chaos and incentives to conceal or misrepresent information or compete).

	Distinguish	Defect
Distinguish	$b - c$	$-c(1 - a)$
Defect	$b(1 - a)$	0

Misinformation

Misinformation overview. Uncertainty regarding a rival's power or intentions can increase the probability of conflict[412]. Bargaining often requires placing trust in another not to break an agreement. This is harder to achieve when one agent believes something false about the other's preferences, resources, or commitments. A lack of shared, accurate information can lead to mistrust and a breakdown in negotiations.

Example: Russian invasion of Ukraine. Incomplete information may lead overly optimistic parties to make too large demands, whereas rivals that are tougher than expected reject those demands and instigate conflict. Examples of misinformation problems generating conflict may include Russia's 2022 invasion of Ukraine. Russian President Putin reportedly miscalculated Ukraine's willingness to resist invasion and fight back. With more accurate information regarding Ukraine's abilities and determination, Putin may have been less likely to instigate conflict [415].

Effect on the bargaining range. Misinformation can prevent agents from finding a mutually-agreeable bargaining range, as shown in Figure 9.14. For example, if each agent believes themself to be the more powerful party, each may therefore want more than half the value they are competing for. Thus, each may reject any bargain offer the other makes, since they expect a better if they opt for conflict instead.

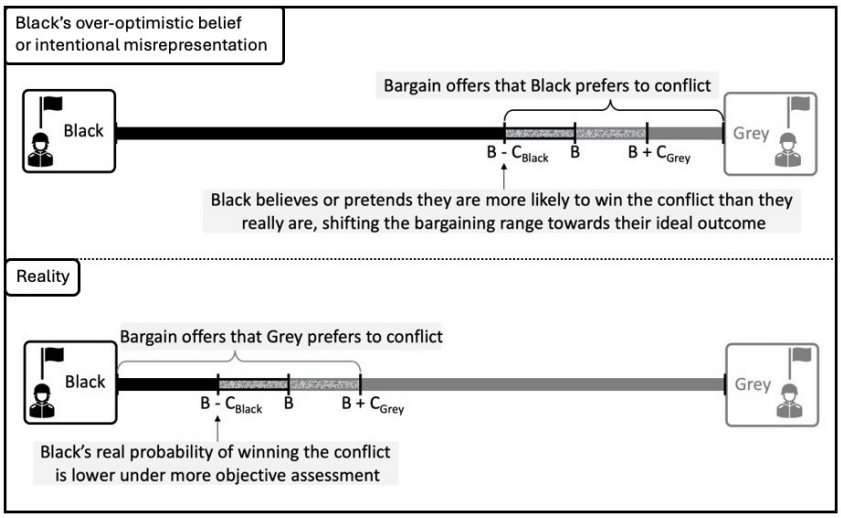

FIGURE 7.11. Black either believes themself to be – or intentionally misrepresents themself as – more powerful than they really are. This means that the range of bargain offers Black will choose over conflict does not overlap with the equivalent range for Grey. Thus, there is no mutual bargaining range.

Misinformation and AI. AI technologies may produce misinformation directly: large language models hallucinating false facts would be one such example. Less directly, a lack of AI reliability could also promote conflict by increasing uncertainty in warfare. For example, the unreliable behaviors of military AI technologies may make it more difficult to understand what an enemy's true intentions are, increasing the risk of inadvertently escalating a conflict. Furthermore, there is the difficulty of accurately evaluating AI capabilities advances. It may be unclear how powerful a model trained on an order of magnitude more compute may be, or how far behind adversaries are in their effort to create powerful models. As automated warfare technologies become more widespread and sophisticated, nations may struggle to predict their probability of victory in any given conflict accurately. This increased potential for miscalculation may make warfare more likely.

Information problems could exacerbate other AI risks. For example, if there are substantial existential risks from AIs but this is not widely agreed on, improving understanding of these risks could help make different actors (such as the US and China) get better estimates of the payoff matrix. With better understanding of AI risk, they may recognize that it is in their self-interest to cooperate (slow down AI development and militarization) instead of defecting (engaging in an AI race). Similarly, creating information channels such as summits can increase understanding and coordination; even if countries do not agree on shared commitments, the discussions on the sidelines can reduce misunderstandings and the risk of conflict.

Disinformation

Disinformation overview. Unlike misinformation, where false information is propagated without deceptive intention, disinformation is the *deliberate* spreading of false information: the intent is to mislead, deceive or manipulate. Here, we explore why competitive situations may motivate agents to try to mislead others or misrepresent the truth, and how this can increase the probability of conflict.

Examples: employment and the real estate industry. Throughout labor markets, employers and job seekers often encounter disinformation problems. Employers may intentionally withhold information about the salary range or offer lower wages than what the market standard suggests in order to secure lower employment costs. On the other hand, job seekers might exaggerate their qualifications or professional experience to increase their chances of getting hired. Such discrepancies can lead to legal conflicts and high turnover rates. Alternatively, in the real estate market, disinformation problems can emerge between sellers and buyers. Sellers sometimes withhold critical information about the property's condition to increase the probability that the property gets purchased. Buyers, on the other hand, may be incentivized to misrepresent their budget or willingness to pay to pressure sellers to lower their prices. Oftentimes, this can result in legal battles or disputes as well as the breakdown of property transactions.

Effect on the bargaining range. Consider two agents: A, which is stronger, and B, which is weaker. B demands "X" amount for a bargaining settlement, but

A, as the stronger agent, will not offer this to avoid being exploited by B. In other words, A thinks B is just trying to get more for themself to "bait" A or "bluff" by implying that the bargaining range is lower. But B might not be bluffing and A might not be as strong as they think they are. Consider the Sino-Indian war in this respect. At the time, India had perceived military superiority relative to China. But in 1962, the Chinese launched an attack on the Himalayan border with India, which demonstrated China's superior military capabilities, and triggered the Sino-Indian war. Thus, stronger parties may prefer conflict if they believe rivals are bluffing. Whereas, weaker parties may prefer conflict if they believe rivals are not as powerful as they believe themselves to be.

Disinformation and AI. AIs themselves may have incentives to misrepresent the facts. For example, the agent "Cicero," developed by Meta [112], is capable of very high performance in the board wargame "Diplomacy." Its success requires it to misrepresent certain information to the other players in a strategic fashion. We have seen many other examples of AIs producing disinformation for a variety of reasons, such as large language models successfully persuading users that they are conversing with a human. The ability of AIs to misrepresent information successfully is only likely to increase in future [416]. This could exacerbate disinformation problems, and thus contribute to greater risk of conflict by eroding the potential for peaceful negotiation [24].

7.4.5 Factors Outside of Bargaining Theory

Inequality and Scarcity

Inequality is another factor that is highly predictive of conflict. Crime is a form of conflict. Income and educational inequality are robust predictors of violent crime [333], even when accounting for the effect of variables such as race and family composition. Similarly, individuals and families with a yearly income below $15,000 are three times more likely to be the victims of violent crime than are individuals and families with a yearly income over $75,000 [417]. Moreover, economists from the World Bank have also highlighted that the effects of inequality on both violent and property crime are robust between countries, finding that when economic growth improves in a country, violent crime rates decrease substantially [418]. This is consistent with evidence at the national level; in the US, for example, the Bureau of Justice reports that households below the federal poverty level have a rate of violent victimization that is more than twice as high as the rate for households above the federal poverty level. Moreover, these effects were largely consistent between both rural and urban areas where poverty was prevalent, further emphasizing the robust relationship between inequality and conflict.

Inequality and relative deprivation. Relative deprivation is the perception or experience of being deprived or disadvantaged in comparison to others. It is a subjective measure of social comparison, not an objective measure of deprivation based on absolute standards. People may feel relatively deprived when they perceive that

others possess more resources, opportunities, or social status than they do. This can lead to feelings of resentment. For example, "Strain theory," proposed by sociologist Robert K. Merton, suggests that individuals experience strain or pressure when they are unable to achieve socially approved goals through legitimate means. Relative deprivation is a form of strain, which may lead individuals to resort to various coping mechanisms, one of which is criminal behavior. For example, communities with a high prevalence of relative deprivation can evolve a subculture of violence [419]. Consider the emergence of gangs, in which violence becomes a way to establish dominance, protect territory, and retaliate against rival groups, providing an alternative path for achieving a desired social standing.

AIs and relative deprivation. Advanced future AIs and widespread automation may propel humanity into an age of abundance, where many forms of scarcity have been largely eliminated on the national, and perhaps even global scale. Under these circumstances, some might argue that conflict will no longer be an issue; people would have all of their needs met, and the incentives to resort to aggression would be greatly diminished. However, as previously discussed, relative deprivation is a subjective measure of social comparison, and therefore, it could persist even under conditions of abundance.

Consider the notion of a "hedonic treadmill," which notes that regardless of what good or bad things happen to people, they consistently return to their baseline level of happiness. For instance, reuniting with a loved one or winning an important competition might cultivate feelings of joy and excitement. However, as time passes, these feelings dissipate, and individuals tend to return to the habitual course of their lives. Even if individuals were to have access to everything they could possibly need, the satisfaction they gain from having their needs fulfilled is only temporary.

Abundance becomes scarcity reliably. Dissatisfied individuals can be favored by natural selection over highly content and comfortable individuals. In many circumstances, natural selection could disfavor individuals who stop caring about acquiring more resources and expanding their influence; natural selection favors selfish behavior (for more detail, see *section 7.5.3* of *Evolutionary Pressures*). Even under conditions of abundance, individuals may still compete for resources and influence because they perceive the situation as a zero-sum game, where resources and power must be divided among competitors. Individuals that acquire more power and resources could incur a long-term fitness advantage over those that are "satisfied" with what they already have. Consequently, even with many resources, conflict over resources could persist in the evolving population.

Relatedly, in economics, the law of markets, also known as "Say's Law," proposes that production of goods and services generates demand for goods and services. In other words, supply creates its own demand. However, if supply creates demand, the amount of resources required to sustain supply to meet demand must also increase accordingly. Therefore, steady increases in demand, even under resource-abundant conditions will reliably result in resource scarcity.

Conflict over social standing and relative power may continue. There will always be scarcity of social status and relative power, which people will continue to compete over. Social envy is a fundamental part of life; it may persist because it tracks differential fitness. Motivated by social envy, humans establish and identify advantageous traits, such as the ability to network or climb the social ladder. Scarcity of social status motivates individuals to compete for social standing when doing so enables access to larger shares of available resources. Although AIs may produce many forms of abundance, there would still be dimensions on which to compete. Moreover, AI development could itself exacerbate various forms of inequality to extreme levels. For example, there are likely to be major advantages to richer countries that have more resources to invest, particularly given that growth in compute, data, and model size appear to scale with AI capabilities. We discuss this possibility in Governance in section 8.3.

7.4.6 Summary

Throughout this section, we have discussed some of the major factors that drive conflict. When any one of these factors is present, agents' incentives to bargain for a peaceful settlement may shift such that conflict becomes an instrumentally rational choice. These factors include power shifts, first-strike advantages, issue indivisibility, information problems and incentives to misrepresent, as well as inequality.

In our discussion of these factors, we have laid the groundwork for understanding the conditions under which decisions to instigate conflict may be considered instrumentally rational. This knowledge base allows us to better predict the risks and probability of AI-driven conflict scenarios.

Power shifts can incentivize AI agents to pursue conflict, maintain strategic advantages or deter potential attacks from stronger rivals, especially in the context of military AI use.

The short-lived nature of offensive advantages may incentivize AIs to pursue first-strike advantages, to degrade or identify vulnerabilities in adversaries' capabilities, as may be the case in cyberwarfare.

In the future, individual parties may have to compete for access to powerful AI. Since dividing this AI between many stakeholders would reduce its power, parties may find it instrumentally rational to conflict for access to it.

AIs may make wars more uncertain, increasing the probability of conflict. AI weaponry innovation may present an opportunity for superpowers to consolidate their dominance, whereas weaker states may be able to quickly increase their power by taking advantage of these technologies early on. This dynamic may create a future in which power shifts are uncertain, which may lead states to incorrectly expect that there is something to gain from going to war.

Even under conditions of abundance facilitated by widespread automation and advanced AI implementation, relative deprivation, and therefore conflict, may persist.

AIs may be motivated by social envy to compete with other humans or AIs for desired social standing. This may result in a global landscape in which the majority of humanity's resources are controlled by selfish, power-seeking AIs.

7.5 EVOLUTIONARY PRESSURES

7.5.1 Overview

The central focus of this chapter is the dynamics to be expected in a future with many AI agents. We must consider the risks that emerge from the interactions between these agents, and between humans and AI agents. In this last part of the Collective Action Problems chapter, we use evolutionary theory to explore what happens when competitive pressures play out over a longer time period, operating on a large group of interacting agents. Exploring evolutionary pressures helps us understand the risks posed by the influence of natural selection on AI development. Our ultimate conclusions are that AI development is likely to be subject to evolutionary forces, and that we should expect the default outcome of this influence to be the promotion of selfish and undesirable AI behavior.

We begin this section by looking at how evolution by natural selection can operate in non-biological domains, an idea known as "generalized Darwinism." We formalize this idea using the conditions set out by Lewontin as necessary and sufficient for natural selection, and Price's equation for describing evolutionary change over time. We thus set out the case that evolutionary pressures are influencing AIs. We turn to the ramifications of this claim in the second section.

We next move on to exploring why evolutionary pressures may promote selfish AI behavior. To consider what traits and strategies natural selection tends to favor, we begin by setting out the "information's eye view" of evolution as a generalized Darwinian extrapolation of the "gene's eye view" of biological evolution. Using this framing, we examine how conflict can arise within a system when the interests of propagating information clash with those of the entity that contains the information. Internal conflict of this kind could arise within AI systems, distorting or subverting goals even when they are specified and understood correctly. Finally, we explore why natural selection tends to favor selfish strategies over altruistic ones. Our upshot is that AI development is likely to be subject to evolutionary pressures. These pressures may distort the goals we specify if the interests of internal components of the AI system clash, and could also generate a trend towards increasingly selfish AI behavior.

7.5.2 Generalized Darwinism

Our aim in this section is to understand *generalized Darwinism*—the idea that Darwinian mechanisms are a useful way to explain many phenomena outside of biology [420]—and how we can use this as a helpful model for modeling AI development. Using examples ranging from science to music, we examine how evolution by natural selection can operate in non-biological systems. We formalize this process using the

conditions for natural selection and consider how AI development meets these criteria and is therefore subject to evolutionary pressures.

Conceptual Framework for Generalized Darwinism

Evolution by natural selection is not confined to the domain of biological organisms. We can model many other phenomena using Darwinian mechanisms. In this section, we use a range of examples to elaborate this idea.

Generalized Darwinism: natural selection can be applied to non-biological phenomena. Evolution by natural selection does not depend on mechanisms particular to biology [421]. Darwin proposed that populations change over the course of generations when differences among individuals help some reproduce more than others, so that eventually, the population is made up of descendants of those that reproduced the most. Darwin understood that this idea could explain many other phenomena. For instance, he suggested that natural selection could explain the evolution of language: "The survival or preservation of certain favored words in the struggle for existence is natural selection" [422].

As an example, Richard Dawkins has argued that human culture developed according to the principles of natural selection [420]. A piece of cultural information, such as a song, is passed down over generations, often with small changes, and some songs remain very well-known even over very long time periods. The 18th century French tune "Ah! Vous dirai-je, Maman" was pleasing and easy to sing, so the young Mozart wrote a version of it, and it was later used as the tune for the English poem "The Star," which was sung over and over, until today, when many people know it as "Twinkle Twinkle," "The Alphabet Song," or "Baa Baa Black Sheep" [423]. There were many other 18th century songs that have long since been forgotten, but that one has spread to many people over centuries, due to it being more memorable and "catchy" than others of its time.

By applying this Darwinian lens, we can describe many non-biological phenomena. In nature, evolution happens when individuals have a variety of traits, and individuals with some traits propagate more than others. If a species of insect can be either red or brown, but the brown ones blend in better and are less likely to be eaten by birds, then more of the red insects will get eaten before reproducing, while brown insects will tend to have more descendants. Over time, the population will consist primarily of brown insects.

We note a similar pattern in other, non-biological domains. For example, alchemy was once a popular way of explaining the relationships among different metals. People who believed in alchemy taught it to their students, who taught it to their own students in turn, often with small differences. Over time, some of those ideas continued to help them explain the world, and others didn't. In this respect, chemistry could be viewed as a descendant of alchemy. The ideas that define it now were propagated when they helped us increase our understanding of the natural world, while others were discarded. In the same vein, after the first widespread video conferencing services

were developed, similar products proliferated. Users chose the product that best met their needs, selecting for services that were cheap, easy to use, and reliable. Each company regularly released new versions of its product that were slightly adapted from earlier ones, and competitors imitated and thereby propagated the best features and implemented them into their own. Some products incorporated the most adaptive features quickly, and the descendants of those products are the ones we use today—while others were quickly outcompeted and fell into obscurity.

Generalized Darwinism does not imply that evolution produces good outcomes. Often, things that are the best at propagating are not "good" in any meaningful sense. Invasive species arrive in a new location, propagate quickly, and local ecosystems begin to crumble. The forms of media that are most successful at propagating in our minds may be harmful to our happiness and social relationships. For instance, news articles that get more clicks are likely to have their click-attracting traits reproduced in the next generation. Clicks thus select for more sensational, emotionally charged headlines. In the context of AI, generalized Darwinism poses significant risks. To see why, we first need to understand how many phenomena tend to develop based on Darwinian principles, so that we can think about how to predict and mitigate these risks.

Formalizing Generalized Darwinism

In this section, we formalize generalized Darwinism. First, we overview the criteria necessary and sufficient for evolution by natural selection to operate on a system. Second, we examine how we might predict what happens to a system that meets these conditions. Together, these help us to see why "survival of the fittest" is a poor description of evolution by natural selection. Instead, this process would be better described as "propagation of the better-propagated information."

Lewontin's three conditions for evolution by natural selection. The evolutionary biologist Richard Lewontin formulated three criteria necessary and sufficient for evolution by natural selection [57]:

1) **Variation**: There is variation in traits among individuals
2) **Retention**: Future iterations of individuals tend to resemble previous iterations
3) **Differential fitness**: Different variants have different propagation rates

The validity of these criteria does not depend on biology. In living organisms, DNA encodes the variations among individuals. Traits encoded by DNA are heritable, and subject to selection. But this is not the only way to fulfill the Lewontin conditions. Video conferencing software has variation (there are many different options), retention (today's video conferencing software is similar to last year's), and differential fitness (some products are much more widely used and imitated than others). Precisely how change occurs depends on the specific phenomenon's mechanism of propagation.

The Price Equation describes how a trait changes in frequency over time. In the 1970s, the population geneticist George R. Price derived an equation that

provides a mathematical description of natural selection [424]. One formulation of Price's equation is given here:

$$\Delta\bar{z} = \mathrm{Cov}(\omega, z) + \mathrm{E}_w(\Delta z).$$

In this equation, $\bar{z}$ denotes the average value of some trait z in a population, and $\Delta\bar{z}$ is the change in the average value of z between the parent generation and the offspring generation. If z is height, and the parent generation is 5 '5 " on average and the next generation is 5' 7" on average, then $\Delta\bar{z}$ is 2 inches. ω is relative fitness: how many offspring does an individual have relative to the average for their generation? $E_w(\Delta z)$ is the expected value of Δz: that is, the average change in z between generations, weighted by fitness, so that individuals who have more offspring are counted more heavily.

Price's Equation shows that the change in the average value of some trait between parents and offspring is equal to the sum of a) the covariance of the trait value and the fitness of the parents, and b) the fitness-weighted average of the change in the trait between a parent and its offspring. "Covariance" describes the phenomenon of one variable varying together with another. To see whether a population will get taller over time, for example, we would need to know the covariance of fitness with height (do tall individuals have more surviving offspring?) and the difference between a parent's height and their average child's height.

The Price Equation can be applied to non-biological systems. The Price Equation does not require any understanding of what causes a trait to be passed down to a subsequent generation or why some individuals have more offspring than others, only of how much the trait *is* passed on and how much it covaries with fitness. The Price Equation would work just as well with car designs or tunes as with birds or mollusks.

The Price Equation allows us to predict what happens when Lewontin conditions apply. The Price equation uses differences between members of the parent generation with respect to some trait z (variation), similarities between parent and offspring generation with respect to z (retention), and differential fitness (selection). As a result, when we understand the degree to which each of the Lewontin conditions apply, we can predict how much of some trait will be present in subsequent generations [425].

First misunderstanding: "fitness" does not describe physical power. The idea of "fitness" often brings to mind a contest of physical power, in which the strongest or fastest organism wins, but this is a misunderstanding. Fitness in an evolutionary sense is not something we gain at the gym. Being fit may not necessarily entail being exceptionally good at any specific abilities. Sea sponges, for example, are among the most ancient of animal lineages, and they are not quick, clever, or good at chasing prey, especially when compared to, say, a shark. But empirically, sea sponges have been surviving and reproducing for hundreds of millions of years, much more than many species that would easily beat them in head-to-head contests at almost any other challenge.

Second misunderstanding: "fitness" is not the mechanism driving evolution. Biologists often talk about fitness when discussing how well-suited an organism is to its environment. In particular, they often treat fitness as a short-hand for *relative reproductive success*: how much an individual contributes to the next generation's gene pool, relative to their competitors. However, this usage of the word "fitness" seems to present evolution as being fundamentally tautological. In the idiom "survival of the fittest," we appear to be using both "survival" and "fit" to mean *relative reproductive* success. This would suggest that evolution is merely the process in which those who reproduce more successfully, reproduce more successfully! If true, the theory of evolution would seem to be using its own conclusion to demonstrate its argument. As we shall see next, however, this is actually false.

"Fitness" is simply a metric we use to measure propagation rate. In fact, evolutionary theory does *not* rely on circular logic. This is because an organism's fitness does not determine its reproductive success; natural selection does. Instead, "fitness" is simply the word we use to describe and measure propagation success. Those who are better at propagating their information (by surviving and reproducing) don't have some "being fit" property which causes their success. Rather, we deem how "fit" they are by measuring how successful they've been at propagating their information. Thus the phrase "survival of the fittest" should really be "propagation of the better-propagated information."

The Price Equation, and natural selection more broadly, simply says that if a trait helps individuals survive longer or reproduce more, and that trait is passed on to the offspring, then more of the next generation will have that trait. It does not tell us why a trait leads to an individual having more offspring; it is only a way of expressing the fact that some traits do correlate with having more offspring. The same is true when natural selection is applied to non-biological systems; "fitness" is simply a word for the quality of propagating more. The information that propagates best is, of course, the information that propagates best. But it need not be, and often is not, "better" in any other sense. Fitness is a metric that describes how much information propagates, not an assessment of value.

Generalized Darwinism and AI Populations

The three Lewontin conditions, of variation, retention, and differential fitness, are all that is needed for evolution by natural selection. This means we can assess how natural selection is likely to affect AI populations by considering how the conditions apply to AIs. Here, we claim that AIs are likely to meet all three conditions, so we should expect natural selection forces to influence their traits and development.

Variation: AIs are designed and trained in a variety of ways. As previously noted in "Natural Selection Favors AI over Humans" [348],

> "When thinking about advanced AI, some have envisioned a single AI that is nearly omniscient and nearly omnipotent, escaping the lab and suddenly controlling the world. This scenario tends to assume a rapid,

> almost overnight, take-off with no prior proliferation of other AI agents; we would go from AIs roughly similar to the ones we have now to an AI that has capabilities we can hardly imagine so quickly that we barely notice anything is changing. However, there could also be many useful AIs, as is the case now. It is more reasonable to assume that AI agents would progressively proliferate and become increasingly competent at specific tasks, rather than assume one AI agent spontaneously goes from incompetent to omnicompetent. This is similar to the subdivision of biological niches. For example, lions and cheetahs developed completely different and mutually exclusive strategies to catch prey through strength or speed. Furthermore, if there are multiple AIs, they can work in parallel rather than waiting for a single model to get around to a task, making things move much faster [348]."

This means that people are likely to continue creating multiple AI agents, even if there is a single best model. Financial gains would encourage multiple competitors to challenge the top system [426].

In addition to the argument that AI populations have variation because of the history of their development, there are also pragmatic arguments for this claim. In evolutionary theory, Fisher's fundamental theorem states that the rate of adaptation is directly proportional to the variation (all else equal). In rapidly-changing environments, where quick adaptation increases a population's probability of survival, populations with more variation may persist longer. Consider how variation in crops reduces the probability of catastrophic crop failure, and variation in investments reduces the risk of unmanageable financial losses. And in machine learning, an ensemble of AI systems will often perform more accurately than a single AI [426]. Variation can help guide decision making, in the same way that many people's aggregated predictions will usually be better than any one expert's. Because of these factors, we are more likely to see a powerful and resilient population of AIs if they have significant variation.

Variation in AI developers. As well as variation in the AI systems themselves, we also see variation between the big technology companies developing and adopting AI technologies. It may seem simple to prevent the rise of selfish AI behaviors by avoiding their selection. However, the reality is different. AI companies, directed more by evolutionary pressures than by safety concerns, are vying for survival in a fiercely competitive landscape. Consider how OpenAI, which started as a nonprofit dedicated to benefiting humanity, shifted to a capped-profit structure in 2019 due to funding needs. Consequently, some of its safety-centric members branched out and founded Anthropic, a company intending to prioritize AI safety. However, even Anthropic couldn't resist the call of commercialization, succumbing to evolutionary pressures itself.

Evolutionary pressures are driving safety-minded researchers to adopt the behaviors of their less safety-minded competitors, because they are anticipating that they can gain a significant fitness advantage in the short-term by deprioritizing safety. Note that this evolutionary process is not based on actual selection events (the researchers

will not be destroyed if they are outcompeted), but rather the researchers' projections of what might happen if they adopt particular strategies. AI safety *ideas* are being selected against, which is driving the researchers to change their *behavior* (to behave in a less safety-conscious manner). Importantly, as the number of competitors rises, the variation in approaches and values also increases. This increase in variation escalates the intensity of the evolutionary pressures and the extent to which these pressures distort the behavior of big AI companies.

Retention: new AIs are developed under the influence of earlier generations. Retention does not require exact copying; it only requires that there be non-zero similarity among individuals in subsequent generations. In the short term, AIs are developed by adapting older models, or by imitating features from competitors' models. Even when training AIs from scratch, retention may still occur, as highly effective architectures, datasets, and training environments are reused thereby shaping the agent in a way similar to how humans (or other biological species) are shaped by their environments. Even if AIs change very rapidly compared to the timescales of biological evolution, they will still meet the criterion of retention; their generations can be extremely short, so they can move through many generations in a short time, but each generation will still be similar to the one before it. Retention is a very easy standard to meet, and even with many uncertainties about what AIs may be like, it is very likely that they meet this broad definition.

Differential Fitness: some AIs are propagated more than others. There are many traits which could cause some AI models or traits to be propagated more than others (increasing their "fitness"). Some of these traits could be highly undesirable to humans. For example, *being safer* than alternatives may confer a fitness advantage on an AI. However, *merely appearing to be safer* might also improve an AI's fitness. Similarly, *being good at automating human jobs* could result in an AI being propagated more. On the other hand, *being easy to deactivate* could reduce an AI's fitness. Therefore, an AI might increase its fitness by integrating itself into critical infrastructure or encouraging humans to develop a dependency on it, making us less keen to deactivate it. As long as some AIs are at least marginally more attractive than others, AI populations will meet the condition of differential fitness. There are many possible points at which natural selection could take effect on AIs. These include the actions of AI developers, in fine-tuning and customizing models, or re-designing training processes.

If the Lewontin conditions are satisfied, we must consider how intense the evolutionary pressures are. More intense selection pressure leads to faster change. In a population of birds in a time with plenty of food, birds with any shape beak may survive. However, if food becomes scarce, only those with the most efficient beaks for accessing some specific food may survive, and the next generation will disproportionately have that beak shape. More variation also leads to faster adaptation, because variants that will be adaptive in a new circumstance are more likely to already exist in the population. The faster rounds of adaptation occur, the more quickly distinct groups emerge with their own features.

If there is more intense selection pressure on AIs, where only AIs with certain traits propagate, then we should expect to see the population optimize around those traits. If there is more variation in the AI population, that optimization process will be faster. If the rate of adaptation also accelerates, we would expect trends that lead to greater differentiation in AI populations that are distinct from the changes in the traits of individual AI models. In the following section, we will discuss the evolutionary trends that tend to dominate when selection pressure is intense and how they might shape AI populations.

Summary

We started this section by exploring how evolution by natural selection can occur in non-biological contexts. We then formalized this idea of "generalized Darwinism" using Lewontin's conditions and the Price equation. We found that AI development may be subject to evolutionary pressures by evaluating how it meets the Lewontin conditions. In the next section, we turn to the ramifications of this claim.

7.5.3 Levels of Selection and Selfish Behavior

Our aim in this section is to understand which AI characteristics are favored by natural selection. We explore this by first outlining an "information's eye view" of evolution by natural selection. Here, we find that internal conflict can arise where the interests of the propagating information (such as a gene) clash with those of the larger entity that contains it (such as an organism). This phenomenon could arise in AI systems, distorting or subverting goals even when human operators have specified them correctly.

We then move to a second risk generated by natural selection operating at the level of propagating information: Darwinian forces strongly favor selfish traits over altruistic ones. Although on the level of an individual organism, individuals may behave altruistically under specific conditions (such as genetic relatedness), on the level of information, evolution by natural selection tends to produce selfishness. We conclude by outlining how a future with many AI agents, shaped by natural selection, will be dominated by selfish behavior.

Information's Eye View

We often consider individual organisms to be the unit on which natural selection is operating. However, it is their genes that are being propagated through time and space, not the organisms themselves. This section considers the "gene's eye view" of evolution by natural selection. We then use generalized Darwinism to build up an extrapolated version of this perspective we can call the "information's eye view" of evolution.

Species succeed when their information propagates, but sometimes interests diverge. The information of living organisms is primarily contained in DNA.

Genes contain the instructions for forming bodies. Most of the time, a gene propagates most successfully when the organism that contains it propagates successfully. But sometimes, the best thing for a gene is not the best thing for the organism. For example, mitochondrial DNA is only passed on from females, so it propagates most if the organism has only female offspring. In some organisms, mitochondrial DNA gives rise to genetic mechanisms that increase the production of female descendants. However, if too many individuals have this mutation, the population will be disproportionately female, and the organism will be unable to pass on the rest of its genes. In this situation, the most effective propagation mechanism for the gene in the mitochondria is harmful to the reproductive success of its host.

The "gene's eye view" of evolution. In *The Selfish Gene*, Richard Dawkins argues that *gene* propagation is a more useful framing than organism propagation [420]. In Dawkins' view, organisms are simply vehicles that allow genes to propagate. Instead of thinking of birds with long beaks competing with birds with short beaks, we can think about genes that create long beaks competing with genes that create short beaks, in a fight for space within the bird population. This gives us a framework for understanding examples like the one above: the gene within the mitochondria is competing for space in the population, and will sometimes take that space even at the expense of the host's individual fitness.

Information functions similarly to genes, narrowing the space of possibilities. We are humans and not dogs, roundworms, or redwood trees almost entirely because of our genes. If we do not know anything about what an organism is, aside from how long its genome is, then for every base in the genome, there are four possibilities, so there is an extremely large number of possible combinations. If we learn that the first base is a G, you have divided the total number by four. When we decode the entire genome, we have narrowed down an impossibly large space of possibility to a single one: we can now know not only that the organism is a cat, but even *which* cat specifically.

In non-biological systems, information works in a parallel way. There are many possible ways to begin a sentence. Each word eliminates possible endings and decreases the listener's uncertainty, until they know the full sentence at the end. Using the framework of information theory, we can think of information as the resolution or reduction of uncertainty (though this is not a formal definition). For an idea, information is just the facts about it that make it different from other ideas. A textbook's main information is its text. A song's information consists of the pitches and rhythms that distinguish it from other songs. These larger phenomena (ideas, books, songs) are distinguished by the information they contain.

Information that propagates occupies a larger volume of both time and space. A single music score, written centuries ago and buried underground ever since, has been propagated across hundreds of years of time, but very little space. In contrast, a hit tune that is suddenly everywhere and then quickly forgotten takes up a lot of space, but very little time. But the best propagated information takes up a large volume of both. The tune for "Twinkle Twinkle" has been taking up space in

many minds, pieces of paper, and digital formats for hundreds of years and continues to propagate. The same is true for genetic information. A gene that flourished briefly hundreds of millions of years ago, and one that has had a consistent small presence, both take up much less space-time volume than a gene that long ago became dominant in many successful branches of the evolutionary tree [421].

Just as some genes propagate more, the same is true for bits of information. In accordance with generalized Darwinism, we can extend the gene's eye view to an "information's eye view." A living organism's basic unit of information is a gene. Everything that evolves as a consequence of Darwinian forces contains information, some of which is inherited more than others. Dawkins coined the term "meme" as an analog for gene: a meme is the basic unit of cultural inheritance. Like genes, memes tend to develop variations, and be copied and adapted into new iterations. The philosopher of science, Karl Popper, wrote that the growth of knowledge is "the natural selection of hypotheses: our knowledge consists, at every moment, of those hypotheses which have shown their (comparative) fitness by surviving so far in their struggle for existence." Social phenomena such as copycat crimes can also be modeled as examples of memetic inheritance. Many types of crimes are committed daily, some of which inspire imitators, whose subsequent crimes can themselves be selected for and copied. Selection operates on the level of individual pieces of information, as well as on the higher level of organisms and phenomena.

AIs may pass on information in ways analogous to our genetics and cultural memetics. AIs are computer programs, made of code that determines what they are like, in a similar way to how our DNA determines what we are like. Different code makes the difference between an agentic AI and Flappy Bird. Their code, or pieces from it, can be directly copied and adapted for new models. But their information can also be memetically transmitted, as our cultural memes can. Even today, AIs are often designed based on hearing about and imitating successful models, not only on copying code from them. AIs also help create training data for new AIs and evaluate their learning, which makes the new AIs tend to have traits similar to earlier models. As AIs continue to become more autonomous, they may be able to imitate and learn from one another, self-modifying to adopt traits and behaviors that seem useful. The AI information that propagates the most will take up more and more space-time volume, as it is copied into more AIs that multiply and endure over longer periods.

Intrasystem Goal Conflict

The interests of an organism and its genetic information are usually aligned well. However, they can sometimes diverge from one another. In this section, we identify analogous, non-biological phenomena, where conflict arises between a system and the sub-systems in which it stores its information. Evolutionary pressures might generate this kind of internal conflict within AI systems, distorting or subverting goals set for AIs by human operators, even when such goals are specified and understood correctly.

Conflict within a genome. Selection on the level of genes does not always result in the best outcomes for the organism. For instance, as discussed in the previous section, human mitochondrial DNA is only transferred to offspring through biological females. A human's mitochondrial genome is identical to their biological mother's, assuming no change due to mutation. Since males represent a reproductive dead-end, mitochondrial genes that benefit only females may therefore be selected for, even when they incur a cost upon males. These and other "selfish" genetic elements give rise to intragenomic conflict.

Conflict within an organism. We observe other kinds of internal conflict within organisms which do not concern their genomes. For example, the bacterial species that compose the human gut microbiome can exist in a mutually-beneficial symbiosis with their host. However, some bacteria are "opportunistically pathogenic": in the wake of disruptions (like the use of antibiotics), many of these once-mutualists will propagate at accelerated rates, often at the expense of the host's health. As the philosopher of evolutionary biology Samir Okasha notes, "intraorganismic conflict is relatively common among modern organisms [427]."

Conflict within an AI company. The concept of intrasystem conflict extends beyond biological examples and can be observed in organizations. A notable example is OpenAI. In 2017, there was a power struggle in OpenAI, which led to Elon Musk's exit and Sam Altman becoming OpenAI's main leader. In 2020, disagreements within OpenAI led to internal conflict and the departure of some employees to found Anthropic. In 2023, the board of the nonprofit overseeing OpenAI came into conflict with Sam Altman and attempted to fire him as CEO. Challenging-to-resolve disagreements about who should influence AI's development make intrasystem conflict at AI organizations likely in the future.

Intrasystem goal conflict: between information and the larger entity that contains it. All the above examples concern the interests of propagating information and those of the entities that contain the information diverging from one another. We call the more general phenomenon that can describe all of these examples *intrasystem goal conflict*: the clash of different subsystems' interests, causing the functioning of the overall system to be distorted. As we have seen, intrasystem goal conflict can arise within complex systems in a range of domains, from genomes to corporations.

Intrasystem goal conflict in AI systems. One reason why we might expect an AI system not to pursue a specified goal is because intrasystem goal conflict has eroded its *unity of purpose*. A system has achieved unity of purpose if there is alignment at all levels of internal organization [427]. Undermining a system's unity of purpose reduces its ability to carry out its system-level goals. A helpful analogy here is to consider political "coups." A coup is characterized by a struggle for control within a political system whereby agents within the system act to seize power, often eroding the system's unity of purpose by disrupting its stability and functionality. When political leaders are overthrown, the goals of the political system usually change. Similarly, if we give an AI agent a goal to pursue, the agent may in turn assign parts

of this goal to sub-agents, who may take over and subvert the original goal with their own.

In the future, humans and AI agents may interact in many different ways, including by working together on collaborative projects. This provides the opportunity for goal distortion or subordination through intrasystem goal conflict. For instance, humans may enlist AI agents to collaborate on tasks. Just as how human collaborators may betray or overturn their principals, AI agents may behave similarly. If an AI collaborator has a goal of self-preservation, they may try to remove any power others have over them. In this way, the system that ends up executing actions based on these conflicting goals will not necessarily be equivalent to how a system with unity of purpose would pursue the goal set by the humans. The behavior of this emergent multi-agent system may thus distort our goals, or even subvert them altogether.

Selfishness

In the previous section, we examined one risk generated by natural selection favoring the propagation of information: conflict between the information (such as genes, departments, or sub-agents) and the larger entity that contains it (such as an organism, government, or AI system). In this section, we consider a second risk: that natural selection tends to favor selfish traits and strategies over altruistic ones. We conclude that the greater the influence of evolutionary pressures on AI development, the more we should expect a future with many AI agents to be one dominated by selfish behavior.

Selfishness: furthering one's own information propagation at the expense of others. In evolutionary theory, "selfishness" does not imply intent to harm another, or belief that one's own interests ought to dominate. Organisms that do not have malicious intentions often display selfish traits. The lancet liver fluke, for example, is a small parasite that infects sheep by first infecting ants, hijacking their brains and making them climb to the top of stalks of grass, where they get eaten by sheep [428]. The lancet liver fluke does not wish ants ill, nor does it have a belief that lancet liver flukes should thrive while ants should get eaten. It simply has evolved a behavior that enables it to propagate its own information at the expense of the ant's.

Selfishness in AI. AI systems may exhibit "selfish" behaviors, expanding the AIs' influence at the expense of human values. Note that these AIs may not even understand what a human is and yet still behave selfishly towards them. For example, AIs may automate human tasks, necessitating extensive layoffs [273]. This could be very detrimental to humans, by generating rapid or widespread unemployment. However, it could take place without any malicious intent on the part of AIs merely behaving in accordance with their pursuit of efficiency. AIs may also develop newer AIs that are more advanced but less interpretable, reducing human oversight. Additionally, some AIs may leverage emotional connections by imitating sentience or emulating the loved ones of human users. This might generate social resistance to their deactivation. For instance, AIs that plead not to be deactivated might stimulate an emotional attachment in some humans. If afforded legal rights, these AIs might

adapt and evolve outside human control, becoming deeply embedded in society and expanding their influence in ways that could be irreversible.

Selfish traits are not the opposite of cooperation. Many organisms display cooperative behavior at the individual level. Chimpanzees, for example, regularly groom other members of their group. They don't do this to be "nice," but rather because this behavior is reciprocated in future, so they are likely to eventually benefit from it themselves [385]. Cells found in filamentous bacteria, so named because they form chains, regularly kill themselves to provide much needed nitrogen for the communal thread of bacterial life, with every tenth cell or so "committing suicide" [429]. But even in these examples, cooperative behavior ultimately helps the individual's information propagate. Chimpanzees who groom others expect to have the favor returned in future. Filamentous bacteria live in colonies made up of their clones, so one bacterium sacrificing itself to save copies of itself still propagates its information.

Natural selection tends to produce selfish traits. Organisms that further their own information propagation will typically propagate more. A lancet liver fluke that developed the ability to give ants free choice and allow them not to climb stalks of grass if they don't want to would be less likely than the current version to succeed at getting eaten by sheep and continuing its life cycle. Most biological selfishness is less dramatic, but nonetheless, the organisms alive today are necessarily the descendants of those that succeeded at propagating their own information, and not of those that traded propagation for other qualities.

Altruism that reduces an individual's fitness is not an evolutionarily stable strategy. Imagine a very altruistic fictional population of foxes who freely share food with one another, even at great cost to themselves. When food is abundant, they all thrive, and when food is scarce, they suffer together. If, during a time of scarcity, one fox decides to steal food from the communal stores and take it for herself and her offspring, they may survive while others starve. As a result, her offspring, who may have inherited her selfish trait, will make up a higher proportion of the next generation. As this repeats, the population will be dominated by individuals who take food for themselves when they can. The population of altruists may get along quite well on its own, but altruism is unstable, because anyone who decides to exploit it will do better than the group. Since altruism that reduces an individual's overall fitness is not an evolutionarily stable strategy, we should expect to see selfish behavior being promoted.

The more natural selection acts on a population, the more selfish behavior we expect. In the example in the preceding paragraph, when food is abundant, there is little advantage to selfishness and there may even be penalties, as the group punishes selfish behavior. There is plenty of food to go around, so the descendants of foxes who steal food will not be much more likely to survive, and the next generation can contain plenty of altruists. But in times when only a few can propagate, selfishness will confer a greater advantage, and the population will tend to become selfish more quickly.

Avoiding extreme AI selfishness: changing the environment. AI agents' fitness could either be influenced more by natural selection or by the environment. We have sketched out the default outcome of the former: a landscape of powerful and selfish AI agents. One way we might prevent this trend towards increasingly selfish behavior is to ensure that it is the *environment* which ends up shaping the fitness of AI agents substantially more than natural selection. Currently, we are in an environment of extreme competition, and so AI agents that are better-suited to this competitive environment will propagate more, and increase the proportion of the population with their traits (including selfish traits). However, if we altered the environment such that the actions of AI researchers and AI agents were not so heavily steered by competitive pressures, we could reduce this problem.

Avoiding extreme AI selfishness: changing the selection. Another possibility is to change what makes AI agents "fit." We could establish an ecosystem in which AI agents can be developed, deployed, and adopted more safely, without the influence of such extreme competitive pressures. In this ecosystem, we could select against AIs with the most harmful selfish behaviors, and select for AIs that faithfully assist humans. As these AIs proliferate through this ecosystem, they could then counteract the worst excesses of selfish behavior from other agents.

7.5.4 Summary

In this section, we considered the effects of evolutionary pressures on AI populations. We started by using the idea of generalized Darwinism to expand the "gene's eye view" of biological evolution to an "information's eye view." Using this view, we identified two AI risks generated by natural selection: intrasystem goal conflict and selfish behavior. Intrasystem goal conflict could distort or subvert the goals we set an AI system to pursue. Selfish behavior would likely be favored by natural selection wherever it promotes the propagation of information: If AI development is subject to strong Darwinian forces, we should expect AIs to tend towards selfish behaviors.

7.6 CONCLUSION

In this chapter, we considered a variety of multi-agent dynamics in biological and social systems. Our underlying thesis was that these dynamics might produce undesirable outcomes with AI, mirroring patterns observable in nature and society.

Game theory

We began with a simple game, the Prisoner's Dilemma, observing how even rational agents may reach equilibrium states that are detrimental to all. We then proceeded to build upon this. We considered how the dynamics may change when the game is iterated and involves more than two agents. We found that uncertainty about the future could foster rational cooperation, though defection remains the dominant strategy when the number of rounds of the game if fixed and known.

We used these games to model collective action problems in the real world, like anthropogenic climate change, public health emergency responses, and the failures of democracies. The collective endeavors of multi-agent systems are often vulnerable to exploitation by free riders. We drew parallels between these natural dynamics and the development, deployment, and adoption of AI technologies. In particular, we saw how AI races in corporate and military contexts can exacerbate AI risks, potentially resulting in catastrophes such as autonomous economies or flash wars. We ended this section by exploring the emergence of extortion as a strategy that illustrated a grim possibility for future AI systems: AI extortion could be a source of monumental disvalue, particularly if it were to involve morally valuable digital minds. Moreover, AI extortion might persist stably throughout populations of AI agents, which could make it difficult to eradicate, especially if AIs learn to deceive or manipulate humans to obscure their true intentions.

Cooperation

We then moved to an investigation of cooperation. Drawing from biological systems and human societies, we illustrated an array of mechanisms that may promote cooperation between AIs. For each mechanism, however, we also highlighted some associated risks. These risks included nepotism, in-group favoritism, extortion, and the incentives to behave ruthlessly. Thus, we found that merely ensuring that AIs behave cooperatively may not be a total solution to our collective action problems. Rather, we need a more nuanced view of the potential benefits and risks of promoting cooperative AI via particular mechanisms.

Conflict

We next turned to a closer examination of the drivers of conflict. Using the framework of bargaining theory, we discussed why rational agents may sometimes opt for conflict over peaceful bargaining, even though it may be more costly for all involved. We illustrated this idea by looking at how various factors can affect competitive dynamics, including commitment problems (such as power shifts, first-strike advantages, and issue indivisibility), information problems, and inequality. These factors may drive AIs to instigate, promote, or exacerbate conflicts, with potentially catastrophic effects.

Evolutionary pressures

We began this section by examining generalized Darwinism: the idea that Darwinian mechanisms are a useful way to explain many phenomena outside of biology. We explored examples of evolution by natural selection operating in non-biological domains, from culture, academia, and industry. By formalizing this idea using Lewinton's conditions and the Price equation, we saw how AIs and their development may be subject to Darwinian forces.

We then turned to the ramifications of natural selection operating on AIs. We first looked at what AI traits or strategies natural selection may tend to favor. Using

an information's eye view of evolution by natural selection, we found that internal conflict can arise where the interests of the propagating information clash with those of the larger entity that contains it. Intrasystem goal conflict could arise in AI systems, distorting or subverting goals even when human operators have specified them correctly. Moreover, Darwinian forces strongly favor selfish traits over altruistic ones. Although on the level of an individual organism, individuals may behave altruistically under specific conditions (such as genetic relatedness), on the level of information, evolution by natural selection tends to produce selfishness. Thus, we might expect a future shaped by natural selection to be dominated by selfish behavior.

Concluding remarks

In summary, this chapter explored various kinds of collective action problems: intelligent agents, despite acting rationally and in accordance with their own self-interest, can collectively produce outcomes that none of them wants, even when they could seemingly have achieved preferable alternative outcomes. Even when we as individuals share similar goals, system-level dynamics can override our intentions and create undesirable results.

This insight is of vital importance when envisioning a future with powerful AI systems. AIs, individual humans, and human agencies will all conduct their actions in light of how others are behaving and how each expects others to behave in future. The total risk of this multi-agent system greater the sum of its individual parts. Dynamics between multiple human agencies generate races in corporate and military settings. Dynamics between multiple AIs may generate evolutionary pressure for immoral behaviors, particularly selfishness, free-riding, deception, conflict, and extortion. We cannot address all the risks posed by AI simply by focusing on the outcomes of agents acting in isolation. The safety of AI systems will not be guaranteed solely by aligning each AI agent to well-intentioned operators. It is an essential component of ensuring our safety, and a valuable future, that we consider these multi-agent dynamics carefully. These dynamics represent a common problem—clashes between individual and collective interests. We must find innovative, system-level solutions to ensure that the development and interaction of AI agents lead to beneficial outcomes for all.

7.7 LITERATURE

7.7.1 Recommended Reading

- Thomas C. Schelling. *Arms and Influence*. Yale University Press, 1966. ISBN: 9780300002218. URL: http://www.jstor.org/stable/j.ctt5vm52s (visited on 10/14/2023)
- Martin A. Nowak. "Five Rules for the Evolution of Cooperation". In: *Science* 314.5805 (2006), pp. 1560–1563. DOI: 10.1126/science.1133755. eprint: https://www.science.org/doi/pdf/10.1126/science.1133755. URL: https://www.science.org/doi/abs/10.1126/science.1133755

- James D. Fearon. "Rationalist Explanations for War". In: *International Organization* 49.3 (1995), pp. 379–414. ISSN: 00208183, 15315088. URL: http://www.jstor.org/stable/2706903 (visited on 10/14/2023)
- Dan Hendrycks. *Natural Selection Favors AIs over Humans*. 2023. arXiv: 2303.16200 [cs.CY]
- Edward O. Wilson. *Sociobiology: The New Synthesis, Twenty-Fifth Anniversary Edition*. Harvard University Press, 2000. ISBN: 9780674000896. URL: http://www.jstor.org/stable/j.ctvjnrttd (visited on 10/14/2023)
- C. Boehm. *Moral origins: The evolution of virtue, altruism, and shame.* Basic Books, 2012
- R. Axelrod and R.M. Axelrod. *The Evolution of Cooperation*. Basic books. Basic Books, 1984. ISBN: 9780465021215. URL: https://books.google.com.au/books?id=NJZBCGbNs98C