CHAPTER 4

Safety Engineering

In developing an AI safety strategy, it might be tempting to draw parallels with other hazardous technologies, from airplanes to nuclear weapons, and to devise analogous safety measures for AI. However, while we can learn lessons from accidents and safety measures in other spheres, it is important to recognize that each technology is unique, with its own specific set of applications and risks. Attempting to map safety protocols from one area onto another might therefore prove misleading, or leave gaps in our strategy where parallels cannot be drawn.

Instead of relying on analogies, we need a more general framework for safety, from which we can develop a more comprehensive approach, tailored to the specific case in question. A good place to start is with the field of safety engineering: a broad discipline that studies all sorts of systems and provides a paradigm for avoiding accidents resulting from them. Researchers in this field have identified fundamental safety principles and concepts that can be flexibly applied to novel systems.

We can view AI safety as a special case of safety engineering concerned with avoiding AI-related catastrophes. To orient our thinking about AI safety, this chapter will discuss key concepts and lessons from safety engineering.

Risk decomposition and measuring reliability. To begin with, we will look at how we can quantitatively assess and compare different risks using an equation involving two factors: the probability and severity of an adverse event. By further decomposing risk into more elements, we will derive a detailed risk equation, and show how each term can help us identify actions we can take to reduce risk. We will also introduce a metric that links a system's reliability to the amount of time we can expect it to function before failing. For accidents that we would not be able to recover from, this expected time before failure amounts to an expected lifespan.

Safe design principles and component failure accident models. The field of safety engineering has identified multiple "safe design principles" that can be built into a system to robustly improve its safety. We will describe these principles and consider how they might be applied to systems involving AI. Next, we will outline some traditional techniques for analyzing a system and identifying the risks it presents.

Although these methods can be useful in risk analysis, they are insufficient for complex and sociotechnical systems, as they rely on assumptions that are often overly simplistic.

Systemic factors and systemic accident models. After exploring the limitations of component failure accident models, we will show that it can be more effective to address overarching systemic factors than all the specific events that could directly cause an accident. We will then describe some more holistic approaches to risk analysis and reduction. Systemic models rely on complex systems, which we look at in more detail in the next chapter.

Tail events and black swans. In the final section of this chapter, we will introduce the concept of tail events—events characterized by high impact and low probability—and show how they interfere with standard methods of risk estimation. We will also look at a subset of tail events called black swans, or unknown unknowns, which are tail events that are largely unpredictable. We will discuss how emerging technology, including AI, might entail a risk of tail events and black swans, and we will show how we can reduce those risks, even if we do not know their exact nature.

4.1 RISK DECOMPOSITION

To reduce risks, we need to understand the factors contributing to them. In this section, we will define some key terms from safety engineering. We will also discuss how we can decompose risks into various factors and create risk equations based on these factors. These equations are useful for quantitatively assessing and comparing different risks, as well as for identifying which aspects of risk we can influence.

4.1.1 Failure Modes, Hazards, and Threats

Failure modes, hazards, and threats are basic words in a safety engineer's vocabulary. We will now define and give examples of each term.

A failure mode is a specific way a system could fail. There are many ways in which different systems can fail to carry out their intended functions. A valve leaking fluid could prevent the rest of the system from working, a flat tire can prevent a car from driving properly, and losing connection to the Internet can drop a video call. We can refer to all these examples as *failure modes* of different systems. Possible failure modes of AI include AIs pursuing the wrong goals, or AIs pursuing simplified goals in the most efficient possible way, without regard for unintended side effects.

A hazard is a source of danger that could cause harm. Some systems can fail in ways that are dangerous. If a valve is leaking a flammable substance, the substance could catch fire. A flammable substance is an example of a *hazard*, or *risk source*, because it could cause harm. Other physical examples of hazards are stray electrical wires and broken glass. Note that a hazard does not pose a risk automatically; a shark is a hazard but does not pose a risk if no one goes in the water. For AI systems, one possible hazard is a rapidly changing environment ("distribution shift") because an AI might behave unpredictably in conditions different from those it was trained in.

A threat is a hazard with malicious or adversarial intent. If an individual is deliberately trying to cause harm, they present a specific type of hazard: a *threat.* Examples of threats include a hacker trying to exploit a weakness in a system to obtain sensitive data or a hostile nation gaining more sophisticated weapons. One possible AI-related threat is someone deliberately contaminating training data to cause an AI to make incorrect and potentially harmful decisions based on hidden malicious functionality.

4.1.2 The Classic Risk Equation

Different hazards and threats present different levels of risk, so it can be helpful to quantify and compare them. We can do this by decomposing risk into multiple factors and constructing an equation. We will now break down risk into two components, and then later discuss a more detailed four-factor decomposition.

Risk can be broken down into probability and severity. Two main components affect the level of risk a hazard poses: the probability that an accident will occur and the amount of harm that will be done if it does happen. This can be represented mathematically as follows:

$$\text{Risk(hazard)} = P(\text{hazard}) \times \text{severity(hazard)}.$$

where $P(\cdot)$ indicates the probability of an event. This is the classic formulation of risk.

The risk posed by a volcano can be assessed using the probability of eruption, denoted as $P(\text{eruption})$, and the severity of its impact, denoted as severity(eruption). If a volcano is highly likely to erupt but the surrounding area is uninhabited, the risk posed by the volcano is low because severity(eruption) is low. On the other hand, if the volcano is dormant and likely to remain so but there are many people living near the volcano, the risk is also low because $P(\text{eruption})$ is low. In order to accurately evaluate the risk, both the probability of eruption and its severity need to be taken into account.

Applying the classic risk equation to AI. The equation above tells us that, to evaluate the risk associated with an AI system, we need information about two aspects of it: the probability that it will do something unintended, and the severity of the consequences if it does. For example, an AI system may be intelligent enough to capture power over critical infrastructure, with potentially catastrophic consequences for humans. However, to estimate the level of this risk, we also need to know how likely it is that the system will try to capture power. A demonstration of a potential catastrophic hazard does not necessarily imply the risk is high. To assess risk, capabilities *and* propensities must be assessed.

The total risk of a system is the sum of the risks of all associated hazards. In general, there may be multiple hazards associated with a system or situation. For example, a car driving safely depends on many vehicle components functioning as intended, and also depends on environmental factors, such as weather conditions and the behavior of other drivers and pedestrians. There are therefore multiple hazards associated with driving. To find the total risk, we can apply the risk equation to each hazard separately and then add the results together.

$$\text{Risk} = \sum_{\text{hazard}} P(\text{hazard}) \times \text{severity}(\text{hazard})$$

We may not always have exact numerical values. We may not always be able to assign exact quantities to the probability and severity of all the hazards, and may therefore be unable to precisely quantify total risk. However, even in these circumstances, we can use estimates. If estimates are difficult to obtain, it can still be useful to have an equation that helps us understand how different factors contribute to risk.

4.1.3 Framing the Goal as Risk Reduction

We should aim for risk reduction rather than trying to achieve zero risk. It might be an appealing goal to reduce the risk to zero by seeking ways of reducing the probability or severity to zero. However, in the real world, risk is never zero. In the AI safety research community, for example, some talk of “solving the alignment problem”—aligning AI with human values perfectly. This could, in theory, result in zero probability of AIs making a catastrophic decision and thus eliminate AI risk entirely.

However, reducing risk to zero is likely impossible. Framing the goal as eliminating risk implies that finding a perfect, airtight solution for removing risk is possible and realistic. Focusing narrowly on this goal could be counterproductive, as it might distract us from developing and implementing practical measures that significantly reduce risk. In other words, we should not “let the perfect be the enemy of the good.” When thinking about creating AI, we do not talk about “solving the intelligence problem” but about “improving capabilities.” Similarly, when thinking about AI safety, we should not talk about “solving the alignment problem” but rather about “making AI safer” or “reducing risk from AIs.” A better goal could be to make catastrophic risks negligible (for instance, less than 0.01% of an existential catastrophe per century) rather than trying to have the risk become exactly zero.

4.1.4 Disaster Risk Equation

The classic risk equation is a useful starting point for evaluating risk. However, if we have more information about the situation, we can break down the risk from a hazard into finer categories. First we can think about the *intrinsic hazard level*, which is a shorthand for probability and severity as in the classic risk equation. Additionally, we can consider how the hazard interacts with the people at risk: we can consider the amount of *exposure* and the level of *vulnerability* [257].

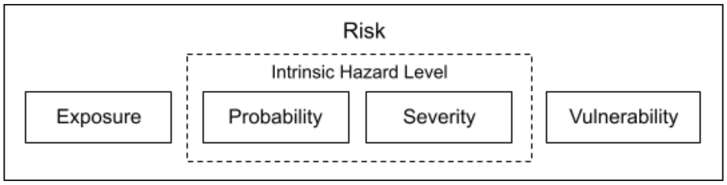


FIGURE 4.1. Risk can be broken down into exposure, probability, severity, and vulnerability. Probability and severity together determine the "intrinsic hazard level."

4.1.5 Elements of the Risk Equation

Exposure and probability are relevant before the accident, while severity and vulnerability matter during it. We can explain these terms in the context of a floor that is hazardously slippery:

1. **Exposure is a measure of how much we are exposed to a hazard.** It will be more likely that someone will slip on a wet floor if there are more people walking across it and if the floor remains slippery for longer. We can say that this is because the *exposure* to the possibility of someone slipping is higher.
2. **Probability tells us how likely it is that a hazard will lead to an accident.** The more slippery the floor is, the more likely it is that someone will slip and fall on it. This is separate from exposure: it already assumes that they are walking on the floor. Probability and exposure together determine how likely an accident is.
3. **Severity indicates how intrinsically harmful a hazard is.** A carpeted floor is less risky than a marble one. Part of the severity of a slippery floor would be how hard it is. This term represents the extent of damage an accident would inflict. We can refer to probability and severity together as the *intrinsic hazard level.*
4. **Vulnerability measures how susceptible we are to harm from a hazard.** If someone does slip on a floor, the harm caused is likely to depend partly on factors such as bone density and age, that together determine how susceptible they are to bone fracture. We can refer to this as *vulnerability.* Vulnerability and severity together determine how harmful an accident is.

Note that probability and severity are mostly about the hazard itself, whereas exposure and vulnerability tell us more about those subject to the risk.

With these terms, we can construct a more detailed equation. Sometimes, but not always, it is more convenient to use this risk decomposition. Rather than being mathematically rigorous, this equation is intended to convey that increasing any of the terms being multiplied will increase the risk, and reducing any of these terms will reduce it. Additionally, reducing any of the multiplicative terms to zero will reduce the risk to zero for a given hazard, regardless of how large the other factors are. Once more, we can add together the risk estimates for each independent hazard to find the total risk for a system.

$$\text{Risk} = \sum_{\substack{\text{hazardous}\\ \text{event } h}} P(h) \times \text{severity}(h) \times \text{exposure}(h) \times \text{vulnerability}(h)$$

4.1.6 Applying the Disaster Risk Equation

Infection by a virus is an example hazard. Consider these ideas in the context of a viral infection. The hazard severity and probability of a virus refers to how bad the symptoms are and how infectious it is. An individual's exposure relates to how much they come into contact with the virus. Their vulnerability relates to how strong their immune system is and whether they are vaccinated. If the virus is certainly deadly once infected, we might consider ourselves extremely vulnerable.

Decomposing risks in detail can help us identify practical ways of reducing them. As well as helping us evaluate a risk, the equation above can help us understand what we can do to mitigate it. We might be unable to change how intrinsically virulent a virus is, for instance, and so unable to affect the hazard's severity. However, we could reduce exposure to it by wearing masks, avoiding large gatherings, and washing hands. We could reduce our vulnerability by maintaining a healthy lifestyle and getting vaccinated. Taking any of these actions will decrease the overall risk. If we are facing a deadly disease, then we might take extreme actions like quarantining ourselves to reduce exposure to the hazard, thereby bringing down overall risk to manageable levels.

Not all risks can be calculated precisely, but decomposition still helps reduce them. An important caveat to the disaster risk equation is that not all risks are straightforward to calculate, or even to predict. Nonetheless, even if we cannot put an exact number on the risk posed by a given hazard, we can still reduce it by decreasing our exposure or vulnerability, or the intrinsic hazard level itself, where possible. Similarly, even if we cannot predict all hazards associated with a system—for example if we face a risk of unknown unknowns, which are explored later in this chapter—we can still reduce the overall risk by addressing the hazards we are aware of.

In AI safety, the risk equation suggests three important research areas. As with other hazards, we should look for multiple ways of preventing and protecting against potential adverse events associated with AI. There are three key areas of research that can each be viewed as inspired by a component of the disaster risk equation: robustness (e.g. adversarial robustness), monitoring (e.g. transparency, trojan detection, anomaly detection), and control (e.g. reducing power-seeking drives, representation control). These research areas correspond to reducing the vulnerability of AIs to adversarial attacks, exposure to hazards by monitoring and avoiding them, and hazard level (probability and severity of potential damage) by ensuring AIs are controllable and inherently less hazardous. To reduce AI risk, it is crucial to pursue and develop all three, rather than relying on just one.

Example hazard: proxy gaming. Consider proxy gaming, a risk we face from AIs that was discussed in the Single Agent Safety chapter. Proxy gaming might arise when we give AI systems goals that are imperfect proxies of our goals. An AI might then learn to "game" or over-optimize these proxies in unforeseen ways that diverge from human values. We can tackle this threat in many different ways:

1. Reduce our exposure to this proxy gaming hazard by improving our abilities to monitor anomalous behavior and flag any signs that a system is proxy gaming at an early stage.
2. Reduce the hazard level by making AIs want to optimize an idealized goal and make mistakes less hazardous by controlling the power the AI has, so that if it does overoptimize the proxy it would do less harm.
3. Reduce our vulnerability by making our proxies more accurate, by making AIs more adversarially robust, or by reducing our dependence on AIs.

Systemic safety addresses factors external to the AI itself. The three fields outlined above focus on reducing risk through the design of AIs themselves. Another approach, called systemic safety (see 3.5), considers the environment within which the AI operates and attempts to remove or reduce the hazards that it might otherwise interact with. For example, improving information security reduces the chance of a malicious actor accessing a lethal autonomous weapon, while addressing inequality and improving mental health across society could reduce the number of people who might seek to harness AI to cause harm.

Adding ability to cope can improve the disaster risk equation. There are other factors that could be included in the disaster risk equation. We can return to our example of the slippery floor to illustrate one of these factors. After slipping on the floor, we might take less time to recover if we have access to better medical technology. This tells us to what extent we would be able to recover from the damage the hazard caused. We can refer to the capacity to recover as our *ability to cope.* Unlike the other factors that multiply together to give us an estimate of risk, we might divide by ability to cope to reduce our estimate of the risk if our ability to cope with it is higher. This is a common extension to the disaster risk equation.

Some hazards are extremely damaging and eliminate any chance of recovery: the severity of the hazard and our vulnerability are high, while our ability to cope is tiny. This constitutes a *risk of ruin*——permanent, system-complete destruction. In this case, the equation would involve multiplying together two large numbers and dividing by a small number; we would calculate the risk as being extremely large. If the damage cannot be recovered from, like an existential catastrophe (e.g., a large asteroid or sufficiently powerful rogue AIs), the risk equation would tell us that the risk is astronomically large or infinite.

Summary. We can evaluate a risk by breaking it down into the probability of an adverse event and the amount of harm it would cause. This enables us to quantitatively compare various kinds of risks. If we have enough information, we can analyze risks further in terms of our level of exposure to them and how vulnerable we are to damage from them, as well as our ability to cope. Even if we cannot assign an exact numerical value to a risk, we can estimate it. If our estimates are unreliable, this decomposition can still help us to systematically identify practical measures we can take to reduce the different factors and thus the overall risk.

4.2 NINES OF RELIABILITY

In the above discussion of risk evaluation, we have frequently referred to the probability of an adverse event occurring. When evaluating a system, we often instead refer to the inverse of this—the system's reliability, or the probability that an adverse event will not happen, usually presented as a percentage or decimal. We can relate system reliability to the amount of time that a system is likely to function before failing. We can also introduce a new measure of reliability that conveys the expected time before failure more intuitively.

The more often we use a system, the more likely we are to encounter a failure. While a system might have an inherent level of reliability, the probability of encountering a failure also depends on how many times it is used. This is why, as discussed above, increasing exposure to a hazard will increase the associated level of risk. An autonomous vehicle, for example, is much more likely to make a mistake during a journey where it has to make 1000 decisions, than during a journey where it has to make only 10 decisions.

TABLE 4.1. From each level of system reliability, we can infer its probability of mistake, "nine or reliability," and expected time before failure.

% reliability of system	% risk of mistake	Nines of Reliability	A mistake is likely to occur by decision number...
0	100	0	1
50	50	0.3	2
75	25	0.6	4
90	10	1	10
99	1	2	100
99.9	0.1	3	1000
99.99	0.01	4	10,000

For a given level of reliability, we can calculate an expected time before failure. Imagine that we have several autonomous vehicles with different levels of reliability, as shown in Table 4.1. Reliability is the probability that the vehicle will get any given decision correct. The second column shows the complementary probability: the probability that the AV will get any given decision wrong. The fourth column shows the number of decisions within which the AV is expected to make one mistake. This can be thought of as the AV's expected time before failure.

Expected time before failure does not scale linearly with system reliability. We plot the information from the table in Figure 4.2. From looking at this graph, it is clear that the expected time before failure does not scale linearly with the system's reliability. A 25% change that increases the reliability from 50% to 75%, for example, doubles the expected time before failure. However, a 9% change increasing the reliability from 90% to 99% causes a ten-fold increase in the expected time before failure, as does a 0.9% increase from 99% to 99.9%.

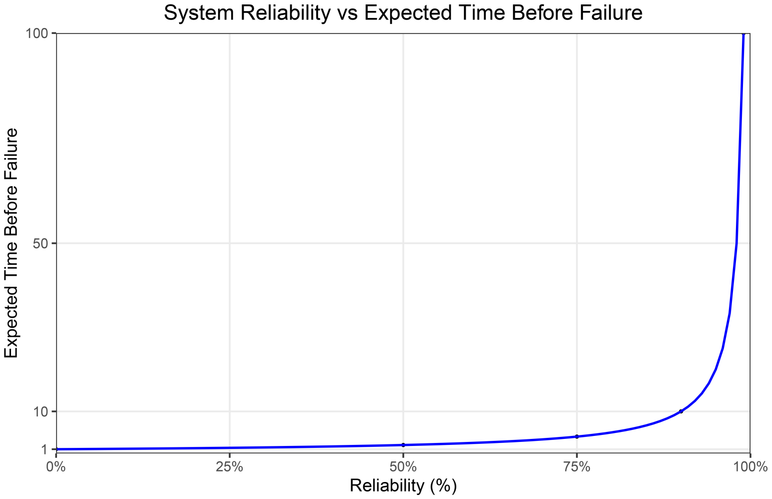


FIGURE 4.2. Halving the probability of a mistake doubles the expected time before failure. Therefore, the relationship between system reliability and expected time before failure is non-linear.

The closer we get to 100% reliability, the more valuable any given increment of improvement will be. However, as we get closer to 100% reliability, we can generally expect that an increment of improvement will become increasingly difficult to obtain. This is usually true because it is hard to perfectly eliminate the possibility of any adverse event. Additionally, there may be risks that we have not considered. These are called unknown unknowns and will be discussed extensively later in this chapter.

A system with 3 "nines of reliability" is functioning 99.9% of the time. As we get close to 100% reliability, it gets inconvenient to use long decimals to express how reliable a system is. The third column in table 4.1 gives us information about a different metric: the nines of reliability [258]. Informally, a system has nines of reliability equal to the number of nines at the beginning of its decimal or percentage reliability. One nine of reliability means a reliability of 90% in percentage terms or 0.9 in decimal terms. Two nines of reliability mean 99%, or 0.99. We can denote a system's nines of reliability with the letter k; if a system is 90% reliable, it has one nine of reliability and so $k = 1$; if it is 99% reliable, it has two nines of reliability, and so $k = 2$. Formally, if p is the system's reliability expressed as a decimal, we can define k, the nines of reliability a system possesses, as:

$$k = -\log_{10}(1 - p).$$

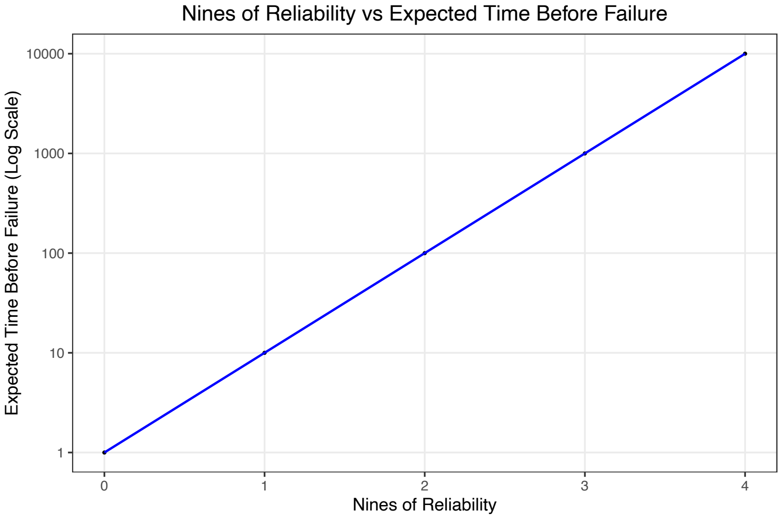


FIGURE 4.3. When we plot the nines of reliability against the expected time before failure on a logarithmic scale, the result is a straight line.

Adding a '9' to reliability gives a tenfold increase in expected time before failure. If it is 99% reliable, it has a 1% probability of failure. If it is 99.9% reliable, it has a 0.1% probability of failure. Therefore, adding another consecutive 9 to the reliability corresponds to a tenfold reduction in risk, and therefore a tenfold increase in the expected time before failure, as we can see in the graph. This means that the relationship between system reliability and expected time before failure is not linear. However, the relationship between nines of reliability and the logarithm of expected time before failure is linear. Note that if we have a system where a failure would mean permanent societal-scale destruction, then the expected time before failure is essentially the *expected lifespan* of human civilization. Increasing such a system's reliability by one nine would cause a ten-fold increase in the expected lifespan of human civilization.

The nines of reliability metric can provide a relatively intuitive sense of the difference between various levels of reliability. Looking at the table, we can think of this metric as a series of levels, where going up one level means a tenfold increase in expected time before failure. For example, if a system has four nines of reliability, we can expect it to last 100 times longer before failing than if it has two. This is an advantage of using the nines of reliability: thinking logarithmically can give us a better understanding of the actual value of an improvement than if we say we're improving the reliability from 0.99 to 0.9999. In the latter case, the numbers begin to look the same, but, in terms of expected time before failure, this improvement is actually more meaningful than going from 0.4 to 0.8.

The nines of reliability are only a measure of probability, not risk. In the framing of the classic risk equation, the nines of reliability only contain information about the probability of a failure, not about what its severity would be. This metric is therefore incomplete for evaluating risk. If an AI has three nines of reliability, for example, we know that it is expected to make 999 out of 1000 decisions correctly. However, three nines of reliability tells us nothing about how much damage the agent will do if it makes an incorrect decision, so we cannot calculate the risk involved in using the system. A game playing AI will present a lot less risk than an autonomous vehicle even if both systems have three nines of reliability.

Summary. A system's nines of reliability indicate the number of consecutive nines at the beginning of its percentage or decimal reliability. An additional nine of reliability represents a reduction of probability of failure by a factor of 10, and thus a tenfold increase of expected time before failure. Nines of reliability tell us about the probability of an accident, but do not contain any information about the severity of an accident, and can therefore not be used alone to calculate risk. In the case of a risk of ruin, an additional nine of reliability means a tenfold increase in expected lifespan.

4.3 SAFE DESIGN PRINCIPLES

We can reduce both the probability and severity of a system failure by following certain *safe design principles* when designing it. These general principles have been identified by safety engineering and offer practical ways of reducing the risk associated with all kinds of systems. They should be incorporated from the outset, rather than being retrofitted later. This strategy attempts to "build safety into" a system, and is more robust than building the system without safety considerations and then attempting to fix individual problems if and when they become apparent.

Note that these principles are not only useful in building an AI itself, but also the system around it. For example, we can incorporate them into the design of the cybersecurity system that controls who is able to access an AI, and into the operations of the organization, or system of humans, that is creating an AI.

We will now explore eight of these principles and how they might be applied to AI systems:

1. **Redundancy**: having multiple backup components that can perform each critical function, so that a single component failing is not enough to cause an accident.
2. **Transparency**: ensuring that operators have enough knowledge of how a system functions under various circumstances to interact with it safely.
3. **Separation of duties**: having multiple agents in charge of subprocesses so that no individual can misuse the entire system alone.
4. **Principle of least privilege**: giving each agent the minimum access necessary to complete their tasks.
5. **Fail-safes**: ensuring that the system will be safe even if it fails.
6. **Antifragility**: learning from previous failures to reduce the likelihood of failing again in future.

7. **Negative feedback mechanisms**: building in processes that naturally down-regulate operations in the event that operators lose control of the system.
8. **Defense in depth**: employing multiple safe design principles rather than relying on just one, since any safety feature will have weaknesses.

Note that, depending on the exact type of system, some of these safe design principles might be less useful or even counterproductive. We will discuss this later on in the chapter. However, for now, we will explore the basic rationale behind why each one improves safety.

4.3.1 Redundancy

Redundancy means having multiple "backup" components [257]. Having multiple braking systems in a vehicle means that, even if the foot brake is not working well, the handbrake should still be able to decelerate the vehicle in an emergency. A failure of a single brake should therefore not be enough to cause an accident. This is an example of *redundancy*, where multiple components can perform a critical function, so that a single component failing is not enough to cause the whole system to fail. In other words, redundancy removes single points of failure. Other examples of redundancy include saving important documents on multiple hard drives, in case one of them stops working, and seeking multiple doctors' opinions, in case one of them gets a diagnosis wrong.

A possible use of redundancy in AI would be having an inbuilt "moral parliament" (see Moral Uncertainty in the Beneficial AI and Machine Ethics chapter). If an AI agent has to make decisions with moral implications, we are faced with the question of which theory of morality it should follow; there are many of these, and each often has counterintuitive recommendations in extreme cases. Therefore, we might not want an AI to adhere strictly to just one theory. Instead, we could use a moral parliament, in which we emulate representatives of stakeholders or moral theories, let them negotiate and vote, and then do what the parliament recommends. The different theories would essentially be redundant components, each usually recommending plausible actions but unable to dictate what happens in extreme cases, reducing the likelihood of counterintuitive decisions that we would consider harmful.

4.3.2 Separation of Duties

Separation of duties means no single agent can control or misuse the system alone [257]. Consider a system where one person controls all the different components and processes. If that person decides to pursue a negative outcome, they will be able to leverage the whole system to do so. On the other hand, we could separate duties by having multiple operators, each in charge of a different aspect. In this case, if one individual decides to pursue a negative outcome, their capacity to do harm will be smaller.

For example, imagine a lab that handles two chemicals that, if mixed in high enough quantities, could cause a large explosion. To avoid this happening, we could keep the stock of the two chemicals in separate cupboards, and have a different person in charge of supplying each one in small quantities to researchers. This way, no individual has access to a large amount of both chemicals.

We could focus on multiple narrow AI models, instead of a single general one. In designing AI systems, we could follow this principle by having multiple agents, each of which is highly specialized for a different task. Complex processes can then be carried out collectively by these agents working together, rather than having a single agent conducting the whole process alone.

This is exemplified by an approach to AI development called "comprehensive AI services." This views AI agents as a class of service-providing products. Adopting this mindset might mean tailoring AI systems to perform highly specific tasks, rather than speculatively trying to improve general and flexible capabilities in a single agent.

4.3.3 Principle of Least Privilege

Each agent should have only the minimum power needed to complete their tasks [257]. As discussed above, separating duties should reduce individuals' capacity to misuse the system. However, separation of duties might only work if we also ensure individuals do not have access to parts of the system that are not relevant to their tasks. This is called the principle of least privilege. In the example above, we ensured separation of duties by putting chemicals in different cupboards with different people in charge of them. To make this more likely to mitigate risks, we might want to ensure that these cupboards are locked so that everyone else cannot access them at all.

Similarly, for systems involving AIs, we should ensure that each agent only has access to the necessary information and power to complete its tasks with a high level of reliability. Concretely, we might want to avoid plugging AIs into the internet or giving them high-level admin access to confidential information. In the Single Agent Control chapter, we considered how AIs might be power-seeking; by ensuring AIs have only the minimum required amount of power they need to accomplish the goals we assign them, we can reduce their ability to gain power.

4.3.4 Fail-Safes

Fail-safes are features that aim to ensure a system will be safe even if it fails [257]. When systems fail, they stop performing their intended function, but some failures also cause harm. Fail-safes aim to limit the amount of harm caused even if something goes wrong. Elevator brakes are a classic example of a fail-safe feature. They are attached to the outside of the cabin and are held open only by the tension in the cables that the cabin is suspended on. If tension is lost in the cables, the brakes automatically clamp shut onto the rails in the elevator shaft. This means that, even if the cables break, the brakes should prevent the cabin from falling; even if the system fails in its function, it should at least be safe.

A possible fail-safe for AI systems might be a component that tracks the level of confidence an agent has in its own decisions. The system could be designed to stop enacting decisions if this component falls below a critical level of certainty that the decision is correct. There could also be a component that monitors the probability of the agent's decisions causing harm, and the system could be designed to stop acting on decisions if it reaches a specified likelihood of harm. Another example would be a kill switch that makes it possible to shut off all instances of an AI system if this is required due to malfunction or other reasons.

4.3.5 Antifragility

Antifragile systems become stronger from encountering adversity [259]. The idea of an antifragile system is that it will not only recover after a failure or a near miss but actually become more robust from these "stressors" to potential future failures. Antifragile systems are common in the natural world and include the human body. For example, weight-bearing exercises put a certain amount of stress on the body, but bone density and muscle mass tend to increase in response, improving the body's ability to lift weight in the future.

Similarly, after encountering or becoming infected with a pathogen and fighting it off, a person's immune system tends to become stronger, reducing the likelihood of reinfection. Groups of people working together can also be antifragile. If a team is working toward a given goal and they experience a failure, they might examine the causes and take steps to prevent it from happening again, leading to fewer failures in the future.

Designing AI systems to be antifragile would mean allowing them to continue learning and adapting while they are being deployed. This could give an AI the potential to learn when something in its environment has caused it to make a bad decision. It could then avoid making the same mistake if it finds itself in similar circumstances again.

Antifragility can require adaptability. Creating antifragile AIs often means creating adaptive ones: the ability to change in response to new stressors is key to making AIs robust. If an AI continues learning and adapting while being deployed, it could learn to avoid hazards, but it could also develop unanticipated and undesirable behaviors. Adaptive AIs might be harder to control. Such AIs are likely to continuously evolve, creating new safety challenges as they develop different behaviors and capabilities. This tendency of adaptive systems to evolve in unexpected ways increases our exposure to emergent hazards.

A case in point is the chatbot Tay, which was released by Microsoft on Twitter in 2016. Tay was designed to simulate human conversation and to continue improving by learning from its interactions with humans on Twitter. However, it quickly started tweeting offensive remarks, including seemingly novel racist and sexist comments. This suggested that Tay had statistically identified and internalized some biases that it could then independently assert. As a result, the chatbot was taken offline after only 16 hours. This illustrates how an adaptive, antifragile AI can develop in unpredicted and undesirable ways when deployed in natural settings. Human operators cannot control natural environments, so system designers should think carefully about whether to use adaptive AIs.

4.3.6 Negative Feedback Mechanisms

When one change triggers more changes, feedback loops can emerge. To understand positive and negative feedback mechanisms, consider the issue of climate change and melting ice. As global temperatures increase, more of Earth's ice melts. This means ice-covered regions shrink, and therefore reflect a smaller amount of the sun's radiation back into space. More radiation is therefore absorbed by the atmosphere, further increasing global temperatures and causing even more ice to melt. This is a positive feedback loop: a circular process that amplifies the initial change, causing the system to continue escalating itself unchecked. We discuss feedback loops in greater detail in the Complex Systems chapter.

Negative feedback mechanisms act to down-regulate and stabilize systems. If we have mechanisms in place that naturally down-regulate the initial change, we are likely to enter an equilibrium rather than explosive change. Many negative feedback mechanisms are found within the body; for example, if a person's body temperature increases, they will begin to sweat, cooling down as that sweat evaporates. If, on the other hand, they get cold, they will begin to shiver, generating heat instead. These negative feedback mechanisms act against any changes and thus stabilize the temperature within the required range. Incorporating negative feedback mechanisms in a system's design can improve controllability, by preventing changes from escalating too much [260].

We can use negative feedback loops to control AIs. If we are concerned that AIs might get too powerful for us to control, we can create negative feedback loops in the environment to ensure that any increases in an AI's power are met with changes that make it less powerful. There would be two parts to this process. First, we would want better monitoring tools to look for anomalies, such as AI watch dogs. These would track when an AI is getting powerful (or displaying hazardous behavior) in order to trigger some feedback mechanism—the second part of the task. The feedback mechanism might be a drastic measure like disconnecting an AI from the internet, resetting an AI to a previous version, or using other AIs trained to disempower a powerful AI. Such mechanisms would act as automatic checks and balances on AIs' power.

4.3.7 Transparency

Transparency means people know enough about a system to interact with it safely [257]. If operators do not have sufficient knowledge of a system's functions, then it is possible they could inadvertently cause an accident while interacting with it. It is important that a pilot knows how a plane's autopilot system works, how to activate it, and how to override it. That way, they will be able to override it when they need to, and they will know how to avoid activating it or overriding it accidentally when they do not mean to. This safe design principle is called transparency.

Research into AI transparency aims to design deep learning systems in ways that give operators a greater understanding of their internal decision-making processes. This would help operators maintain control, anticipate situations in which systems might make poor or deceptive decisions, and steer them away from hazards.

4.3.8 Defense in Depth

Including many layers of defense is usually more effective than relying on just one [257]. A final safe design principle is defense in depth, which means including multiple layers of defense. That way, if one or more defenses fail, there should still be others in place that can mitigate damage. In general, the more defenses a system has in place, the less likely it is that all the defenses will fail simultaneously. The core idea of defense in depth is that it is unlikely that any one layer of defense is foolproof; we are usually engaging in risk reduction, not risk elimination. For example, an individual is less likely to be infected by a virus if they take multiple measures, such as wearing a mask, social distancing, and washing their hands, than if they rely on just one of these (and no single one is going to work). We will explore this in greater depth in the context of the Swiss cheese model in the next section.

One caveat to note here is that increasing layers of defense can make a system more complex. If the different defenses interact with one another, there is a chance that this might produce unintended side effects. In the case of a virus, for example, reduced social contact and an individual's immune system can be thought of as two separate layers of defense. However, if an individual has very little social contact, and therefore little exposure to pathogens, their immune system could become weaker as a result. While multiple layers of defense can make a system more robust, system designers should be aware that layers might interact in complex ways. We will discuss this further later in the chapter.

Layers of safety features can generally be preventative or protective. There are two ways in which safety measures can reduce risk: preventative measures reduce the probability of an accident occurring, while protective measures reduce the harm if an accident does occur. For example, avoiding large gatherings and washing hands are actions that aim to prevent an individual from becoming infected with a virus, while maintaining a healthy lifestyle and getting vaccinated can help reduce the severity of an infection if it does occur.

We can think about this in terms of the risk equations. Preventative measures reduce the probability of an accident occurring, either by reducing the inherent probability of the hazard or by reducing exposure to it. Protective measures, meanwhile, decrease the severity an accident would have, either by reducing the inherent severity of the hazard or by making the system less vulnerable to it.

In general, prevention is more efficient than cure, but both should be included in system design. An oft-quoted aphorism is that "an ounce of prevention is worth a pound of cure," highlighting that it is much better—and often less costly——to prevent an accident from happening than to try and fix it afterward. It might therefore be wise for system designers to place more emphasis on preventative features than on protective ones.

Nevertheless, protective features should not be neglected. This is illustrated by the sinking of the Titanic, whose preventative design features included the hull being divided into watertight compartments. There was much faith that these features rendered it unsinkable. However, the ship did not carry enough lifeboats to hold all its passengers. This was, in part, because lifeboats were largely intended to transport passengers to another ship in the event of sinking, rather than to hold all of them at once. Still, the insufficient provision meant that there was not enough space on the lifeboats for all the passengers when the ship sank. This can be considered an example of inadequate protective measures. We will explore this distinction more in the next section, where we discuss the bow tie model.

4.3.9 Review of Safe Design Principles

There are multiple features we can build into a system from the design stage to make it safer. We have discussed redundancy, separation of duties, the principle of least privilege, fail-safes, antifragility, negative feedback mechanisms, transparency and defense in depth as eight examples of such principles. Each one gives us concrete recommendations about how to design (or how not to design) AI systems to ensure that they are safer for humans to use.

4.4 COMPONENT FAILURE ACCIDENT MODELS AND METHODS

As a system is being created and used, it is important to analyze it to identify potential risks. One way of doing this is to look at systems through the lens of an accident model: a theory about how accidents happen in systems and the factors that lead to them [260]. We will now look at some common accident models. These impact system design and operational decisions.

4.4.1 Swiss Cheese Model

The Swiss cheese model helps us analyze defenses and identify pathways to accidents [261]. The diagram in Figure 4.4 shows multiple slices of Swiss cheese, each representing a particular defense feature in a system. The holes in a slice represent the weaknesses in a defense feature—the ways in which it could be bypassed. If there are any places where holes in all the slices line up, creating a continuous hole through all of them, this represents a possible route to an accident. This model highlights the importance of defense in depth, since having more layers of defense reduces the probability of there being a pathway to an accident that can bypass them all.

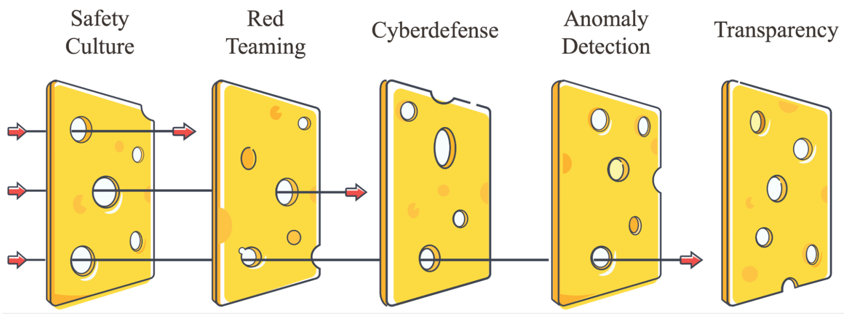


FIGURE 4.4. Each layer of defense (safety culture, red teaming, etc.) is a layer of defense with its own holes in the Swiss cheese model. With enough layers, we hope to avoid pathways that can bypass them all.

Consider the example of an infectious disease as a hazard. There are multiple possible defenses that reduce the risk of infection. Preventative measures include avoiding large gatherings, wearing a mask and regularly washing hands. Protective measures include maintaining a healthy lifestyle to support a strong immune system, getting vaccinated, and having access to healthcare. Each of these defenses can be considered a slice of cheese in the diagram.

However, none of these defenses are 100% effective. Even if an individual avoids large gatherings, they could still become infected while buying groceries. A mask might mostly block contact with the pathogen, but some of it could still get through. Vaccination might not protect those who are immunocompromized due to other conditions, and may not be effective against all variants of the pathogen. These imperfections are represented by the holes in the slices of cheese. From this, we can infer various potential routes to an accident, such as an immunocompromized person with a poorly fitting mask in a shopping mall, or someone who has been vaccinated encountering a new variant at the shops that can evade vaccine-induced immunity.

We can improve safety by increasing the number of slices, or by reducing the holes. Adding more layers of defense will reduce the chances of holes lining up to create an unobstructed path through all the defenses. For example, adopting more of the practices outlined above would reduce an individual's chances of infection more than if they adopt just one.

Similarly, reducing the size of a hole in any layer of defense will reduce the probability that it will overlap with a hole in another layer. For example, we could reduce the weaknesses in wearing a mask by getting a better-fitting, more effective mask. Scientists might also develop a vaccine that is effective against a wider range of variants, thus reducing the weaknesses in vaccination as a layer of defense.

We can think of layers of defense as giving us additional nines of reliability. In many cases, it seems reasonable to assume that adding a layer of defense helps reduce remaining risks by approximately an order of magnitude by eliminating 90% of the risks still present. Consider how adding the following three layers of defense can give our AIs additional nines of reliability:

1. *Adversarial fine-tuning*: By fine-tuning our model, we ensure that it rejects harmful requests. This works mostly reliably, filtering out 90% of the harmful requests received.
2. *Artificial conscience*: By giving our AI an artificial conscience, it is less likely to take actions that result in low human wellbeing in pursuit of its objective. However, 10% of the time, it may take actions that are great for its objective and bad for human wellbeing regardless.
3. *AI watchdogs*: By monitoring deployed AIs to detect signs of malfeasance, we catch AIs acting in ways contrary to how we want them to act nine times out of ten.

Swiss cheese model for emergent capabilities. To reduce the risk of unexpected emergent capabilities, multiple lines of defense could be employed. For example, models could be gradually scaled (e.g., using 3× more compute than the previous training run, rather than a larger number such as 10×); as a result, there will be fewer emergent capabilities to manage. An additional layer of defense is screening for hazardous capabilities, which could involve deploying comprehensive test beds, and red teaming with behavioral tests and representation reading. Another defense is staged releases; rather than release the model to all users at once, gradually release it to more and more users, and manage discovered capabilities as they emerge. Finally, post-deployment monitoring through anomaly detection adds another layer of defense.

Each of these aim at largely different areas, with the first focusing on robustness, the second on control, and the third on monitoring. By ensuring we have many defenses, we prevent a wider array of risks, improving our system reliability by many nines of reliability.

4.4.2 Bow Tie Model

The bow tie model splits defenses into preventative and protective measures[260]. In the diagram in figure 4.5, the triangle on the left-hand side contains features that are intended to prevent an accident from happening, while the triangle on the right-hand side contains features that are intended to mitigate damage if an accident does happen. The point in the middle where the two triangles meet can be thought of as the point where any given accident happens. This is an example of a bow tie diagram, which can help us visualize the preventative and protective measures in place against any potential adverse event.

For example, if an individual goes rock climbing, a potential accident is falling. We can draw a bow tie for this situation, with the center representing a fall. On the left, we note any measures the individual could take to prevent a fall, for example using chalk on their hands to increase friction. On the right, we note any protective measures they could take to reduce harm from falling, such as having a cushioned floor below.

Bow tie analysis of proxy gaming. In the Single-Agent Safety chapter, we learned that one hazard of using AIs is that they might learn to “game” the objec-

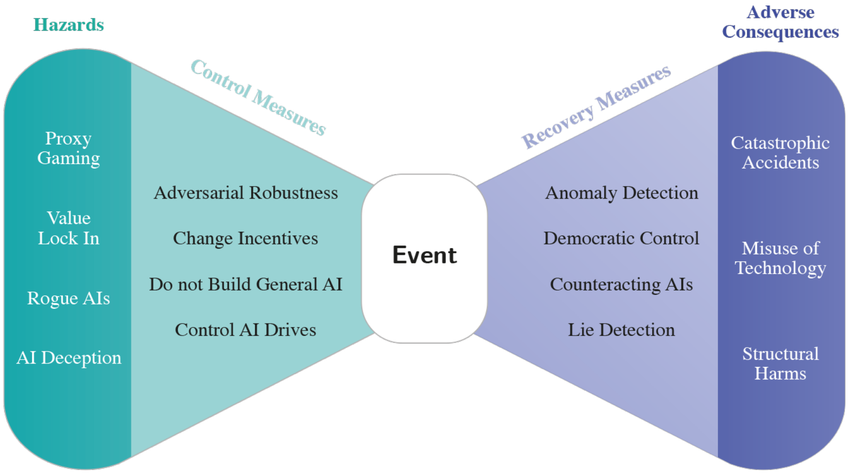


FIGURE 4.5. The bow tie diagram can tie together hazards and their consequences with control and recovery measures to mitigate the effects of an adverse event.

tives we give them. If the specified objectives are only proxies for what we actually want, then an AI might find a way of optimizing them that is not beneficial overall, possibly due to unintended harmful side effects.

To analyze this hazard, we can draw a bow tie diagram, with the center representing the event of an AI gaming its proxy goals in a counterproductive way. On the left-hand side, we list preventative measures, such as ensuring that we can control AI drives like power-seeking. If the AI system has less power (for example fewer resources), this would reduce the probability that it finds a way to optimize its goal in a way that conflicts with our values (as well as the severity of the impact if it does). On the right-hand side, we list protective measures, such as improving anomaly detection tools that can recognize any AI behavior that resembles proxy gaming. This would help human operators to notice activity like this early and take corrective action to limit the damage caused.

The exact measures on either side of the bow tie depend on which event we put at the center. We can make a system safer by individually considering each hazard associated with it, and ensuring we implement both preventative and protective measures against that hazard.

4.4.3 Fault Tree Analysis Method

Fault tree analysis works backward to identify possible causes of negative outcomes. Fault tree analysis (FTA) is a top-down process that starts by considering specific negative outcomes that could result from a system, and then works backward to identify possible routes to those outcomes. In other words, FTA involves “backchaining” from a possible accident to find its potential causes.

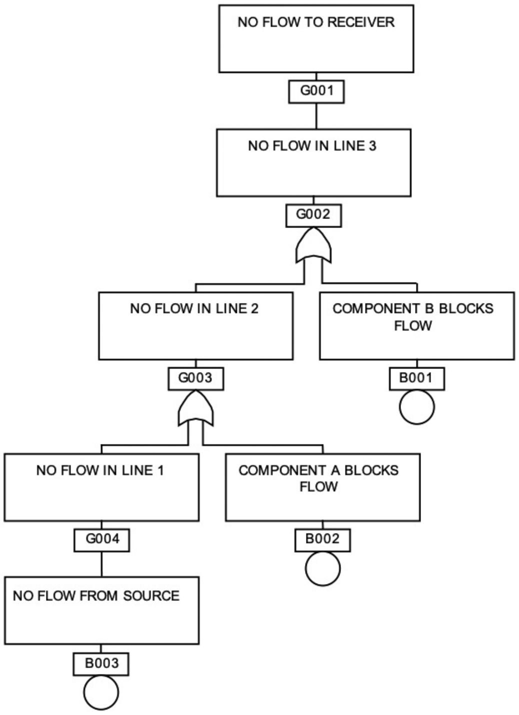
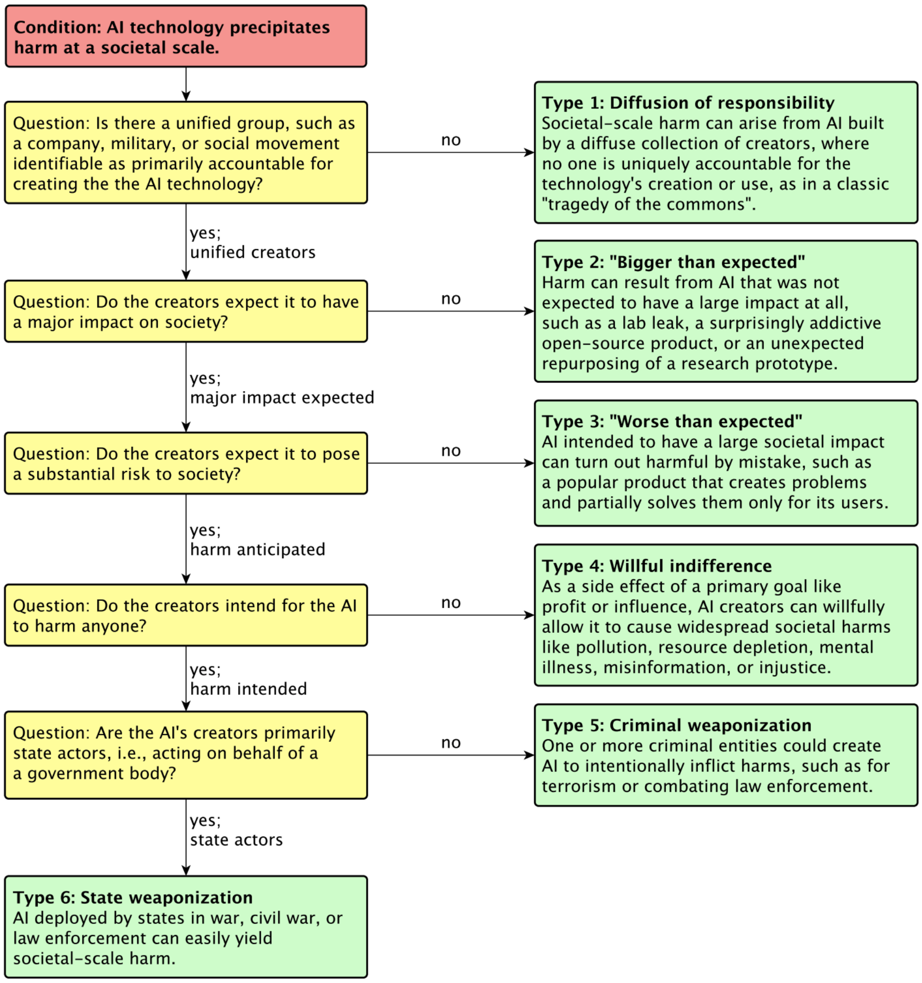


FIGURE 4.6. Using a fault tree, we can work backwards from a failure (no water flow) to its possible causes (such as a blockage or lack of flow at source) [262].

For each negative outcome, we work backward through the system, identifying potential causes of that event. We can then draw a "fault tree" showing all the possible pathways through the system to the negative outcome. By studying this fault tree, we can identify ways of improving the system that remove these pathways. In figure 4.6, we trace backwards from a pump failure to two types of failure: mechanical and electrical failure. Each of these has further subtypes. For fuse failing, we know that we require a circuit overload, which can happen as a result of a wire shorted or a power surge. Hence, we know what sort of hazards we might need to think about.

Example: Fire Hazard. We could also consider the risk of a fire outbreak as a negative outcome. We then work backward, thinking about the different requirements for this to happen—fuel, oxygen, and sufficient heat energy. This differs from the water pump failure since all of these are necessary rather than just one of them. Working backward again, we can think about all the possible sources of each of these requirements in the system. After completing this process, we can identify multiple combinations of events that could lead to the same negative outcome.

FTA can be used to guide decisions. Figure 4.7 shows how we can use fault-tree style reasoning to create concrete questions about risks. We can trace this through to identify risks depending on the answers to these questions. For instance, if there is no unified group accountable for creating AIs, then we know that diffusion of



Condition: AI technology precipitates harm at a societal scale.

Question: Is there a unified group, such as a company, military, or social movement identifiable as primarily accountable for creating the the AI technology?

Question: Do the creators expect it to have a major impact on society?

Question: Do the creators expect it to pose a substantial risk to society?

Question: Do the creators intend for the AI to harm anyone?

Question: Are the AI's creators primarily state actors, i.e., acting on behalf of a a government body?

Type 1: Diffusion of responsibility
Societal-scale harm can arise from AI built by a diffuse collection of creators, where no one is uniquely accountable for the technology's creation or use, as in a classic "tragedy of the commons".

Type 2: "Bigger than expected"
Harm can result from AI that was not expected to have a large impact at all, such as a lab leak, a surprisingly addictive open-source product, or an unexpected repurposing of a research prototype.

Type 3: "Worse than expected"
AI intended to have a large societal impact can turn out harmful by mistake, such as a popular product that creates problems and partially solves them only for its users.

Type 4: Willful indifference
As a side effect of a primary goal like profit or influence, AI creators can willfully allow it to cause widespread societal harms like pollution, resource depletion, mental illness, misinformation, or injustice.

Type 5: Criminal weaponization
One or more criminal entities could create AI to intentionally inflict harms, such as for terrorism or combating law enforcement.

Type 6: State weaponization
AI deployed by states in war, civil war, or law enforcement can easily yield societal-scale harm.

FIGURE 4.7. An FTA decision-tree can identify several potential problems by interrogating important contextual questions [263].

responsibility is a risk. If there is such a group, then we need to question their beliefs, intentions, and incentives.

By thinking more broadly about all the ways in which a specific accident could be caused, whether by a single component failing or by a combination of many smaller events, FTA can discover hazards that are important to find. However, the backchaining process that FTA relies on also has limitations, which we will discuss in the next section.

4.4.4 Limitations

Chain-of-Events Reasoning. The Swiss cheese and bow tie models and the FTA method can be useful for identifying hazards and reducing risk in some systems. However, they share some limitations that make them inapplicable to many of the complex systems that are built and operated today. Within the field of safety engineering, these approaches are largely considered overly simplistic and outdated. We will now discuss the limitations of these approaches in more detail, before moving on to describe more sophisticated accident models that may be better suited to risk management in AI systems.

Chain-of-events reasoning is sometimes too simplistic for useful risk analysis. All of these models and techniques are based on backchaining or linear "chain-of-events" reasoning. This way of thinking assumes there is a neat line of successive events, each one directly causing the next, that ultimately leads to the accident. The goal is then to map out this line of events and trace it back to identify the "root cause"—usually a component failure or human error—to blame. However, given the numerous factors that are at play in many systems, it does not usually make sense to single out just one event as the cause of an accident. Moreover, this approach puts the emphasis largely on the details of "what" specifically happened, while neglecting the bigger question of "why" it happened. This worldview often ignores broader systemic issues and can be overly simplistic. Rather than break events down into a chain of events, a complex systems perspective often sees events as a product or complex interaction between many factors.

Complex and Sociotechnical Systems

Component failure accident models are particularly inadequate for analyzing complex systems and sociotechnical systems. We cannot always assume direct, linear causality in complex and sociotechnical systems, so the assumption of a linear "chain of events" breaks down.

In complex systems, many components interact to produce emergent properties. Complex systems are everywhere. A hive of bees consists of individuals that work together to achieve a common goal, a body comprises many organs that interact to form a single organism, and large-scale weather patterns are produced by the interactions of countless particles in the atmosphere. In all these examples, we find collective properties that are not found in any of the components but are produced by the interactions between them. In other words, a complex system is "more than the sum of its parts." As discussed in the complex systems chapter, these systems exhibit emergent features that cannot be usefully understood through a reductive analysis of components.

Sociotechnical systems involve interactions between humans and technologies. For example, a car on the road is a sociotechnical system, where a human driver and technological vehicle work together to get from the starting point to the destination. At progressively higher levels of complexity, vehicles interacting with one another on a road also form a sociotechnical system, as does the entire transport network. With the widespread prevalence of computers, virtually all workplaces are now sociotechnical systems, and so is the overarching economy.

There are three main reasons why component failure accident models are insufficient for analyzing complex and sociotechnical systems: accidents sometimes happen without any individual component failing, accidents are not always the result of linear causality, and direct causality is sometimes less important than indirect, remote, or “diffuse” causes such as safety culture. We will now look at each of these reasons in more detail.

Accidents Without Failure

Sometimes accidents happen due to interactions, even if no single component fails [264]. The component failure accident models generally consider each component individually, but components often behave differently within a complex or sociotechnical system than they do in isolation. For example, a human operator who is normally diligent might be less so if they believe that a piece of technology is taking care of one of their responsibilities.

Accidents can also arise through interactions between components, even if every component functions exactly as it was intended to. Consider the Mars Polar Lander, a spacecraft launched by NASA in 1999, which crashed on the Martian surface later that year. It was equipped with reverse thrusters to slow its descent to the surface, and sensors on its legs to detect a signal generated by landing to turn the thrusters off. However, the legs had been stowed away for most of the journey. When they extended in preparation for landing, the sensors interpreted it as a landing. This caused the software to switch the thrusters off before the craft had landed, so it crashed on the surface [265].

In this case, there was no component failure. Each component did what it was designed and intended to do; the accident was caused by an unanticipated interaction between components. This illustrates the importance of looking at the bigger picture, and considering the system as a whole, rather than just looking at each component in isolation, in a reductionist way.

Nonlinear Causality

Sometimes, we cannot tease out a neat, linear chain of events or a “root cause” [264]. Complex and sociotechnical systems usually involve a large number of interactions and feedback loops. Due to the many interdependencies and circular processes, it is not always feasible to trace an accident back to a single “root cause.” The high degree of complexity involved in many systems of work is illustrated in the figure below. This shows the interconnectedness of a system cannot be accurately reduced to a single line from start to finish.

AI systems can exhibit feedback loops and nonlinear causality. Reinforcement learning systems involve complexity and feedback loops. These systems gather information from the environment to make decisions, which then impact the environment, influencing their subsequent decisions. Consider an ML system that ranks advertisements based on their relevance to search terms. The system learns about

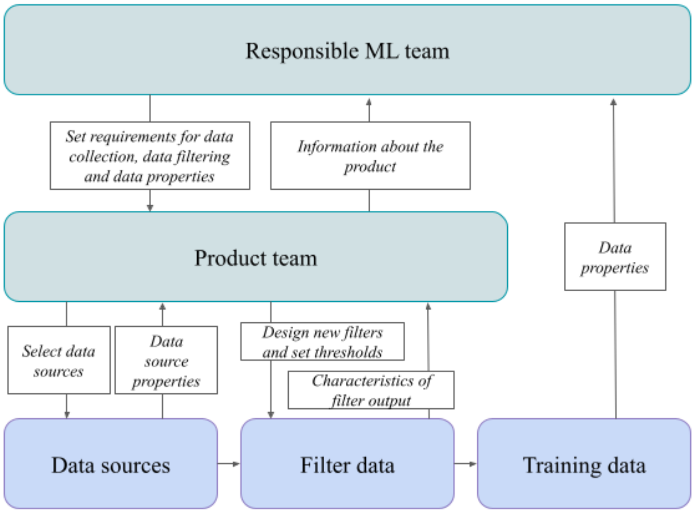

Responsible ML team
Set requirements for data collection, data filtering and data properties
Information about the product
Product team
Data properties
Select data sources
Data source properties
Design new filters and set thresholds
Characteristics of filter output
Data sources
Filter data
Training data

FIGURE 4.8. There are feedback loops in the creation and deployment of AI systems. For example, the curation of training data used for developing an AI system exhibits feedback loops [266].

relevance by tracking the number of clicks each ad receives and adjusts its model accordingly.

However, the number of clicks an ad receives depends not only on its intrinsic relevance but also on its position in the list: higher ads receive more clicks. If the system underestimates the effect of position, it may continually place one ad at the top since it receives many clicks, even if it has lower intrinsic relevance, leading to failure. Conversely, if the system overestimates the effect of position, the top ad will receive fewer clicks than it expects, and so it may constantly shuffle ads, resulting in a random order rather than relevance-based ranking, also leading to failure.

Many complex and sociotechnical systems comprise a whole network of interactions between components, including multiple feedback loops like this one. When thinking about systems like these, it is difficult to reduce any accidents to a simple chain of successive events that we can trace back to a root cause.

Indirect Causality

Sometimes, it is more useful to focus on underlying conditions than specific events [264]. Even if we can pinpoint specific events that directly led to an accident, it is sometimes more fruitful to focus on broader systemic issues. For example, rising global temperatures certainly increase the frequency and severity of natural disasters, including hurricanes, wildfires, and floods. Although it would be difficult to prove that any particular emission of greenhouse gases directly caused any specific disaster, it would be remiss of us to ignore the effects of climate change when trying to evaluate the risk of future disasters. Similarly, a poor diet and lack of exercise are likely to result in ill health, even though no single instance of unhealthy behavior could be directly blamed for an individual developing a disease.

Example: spilling drinks. Consider a scenario where an event planner is faced with the problem of people spilling drinks at a crowded venue. One possible solution might be to contact the individuals who dropped their drinks last time and ask them to be more careful. While this approach aims to reduce the failure rate of individual components (people), it fails to address the underlying issue of high density that contributes to the problem.

Alternatively, the event planner could consider two more effective strategies. Firstly, they could move the event to a bigger venue with more space. By modifying the system architecture, the planner would provide attendees with more room to move around, reducing the likelihood of people bumping into each other and spilling their drinks. This approach addresses the diffuse variable of high density.

Secondly, the event planner could limit the number of attendees, ensuring that the venue does not become overcrowded. By reducing the overall density of people, the planner would again mitigate the risk of collisions and drink spills. This solution also targets the diffuse variable of high density, acknowledging the systemic nature of the problem.

Compared with the first option, which solely focuses on the behavior of individuals and would likely fail to eliminate spillages, the latter two strategies recognize the importance of modifying the environment and addressing the broader factors that contribute to the issue. By prioritizing architectural adjustments or managing attendee numbers, the event planner can more effectively prevent drink spills and create a safer and more enjoyable experience for everyone.

Sharp end versus blunt end. The part of a system where specific events happen, such as people bumping into each other, is sometimes referred to as the "sharp end" of the system. The higher-level conditions of the system, such as the density of people in a venue, is sometimes called the "blunt end." As outlined in the example of people spilling drinks, "sharp-end" interventions may not always be effective. These are related to "proximate causes" and "distal causes," respectively.

Diffuse causality suggests broader, "blunt-end" interventions. In general, it might not be possible to prove that "systemic factors" directly caused an accident. However, systemic factors might "diffusely" affect the whole system in ways that increase the probability of an accident, making them relevant for risk analysis. Under conditions like this, even if the specific event that led to an accident had not happened, something else might have been likely to happen instead. This means it can be more effective to tackle problems less directly, such as by changing system architecture or introducing bottom-up interventions that affect systemic conditions, rather than by attempting to control the sharp end.

Weaponizing any single AI system would not necessarily lead to war or any other kind of catastrophe. However, the existence of these systems increases the probability of a rogue AI system causing a disaster. Minimizing the use of autonomous weapons might be a better way of addressing this risk than attempting to introduce lots of safeguards to prevent loss of control over a rogue system.

Meadows' twelve leverage points. One way of identifying broad interventions that could yield significant results is to consider Meadows' twelve leverage points. These are twelve points within a system, described by environmental scientist Donella Meadows, where a small change can make a large difference [267]. Depending on the system under consideration, some of these characteristics may be difficult or impossible to change, for example large physical infrastructure or the laws of physics. The points are therefore often listed in order of increasing efficacy, taking into account their tractability.

The lower end of the list of leverage points includes: parameters or numbers that can be tweaked, such as taxes or the amount of resources allocated to a particular purpose; the size of buffers in a system, such as water reservoirs or a store's reserve stock, where a larger buffer makes a system more stable and a smaller buffer makes a system more flexible; and the structure of material flows through a system, such as the layout of a transport network or the way that people progress through education and employment.

The middle of the list of leverage points includes: lags between an input to a system and the system's response, which can cause the system to oscillate through over- and undershooting; negative feedback loops that can be strengthened to help a system self-balance; and positive feedback loops that can be weakened to prevent runaway escalation at an earlier stage.

The next three leverage points in Meadows' list are: the structure of information flows in a system, which can increase accountability by making the consequences of a decision more apparent to decision-makers, for example by publishing information about companies' emissions; the rules of the system, such as national laws, or examination policies in educational institutes, which can be changed in social systems to influence behavior; and the ability to self-organize and adapt, which can be promoted by maintaining diversity, such as biodiversity in an ecosystem and openness to new ideas in a company or institution.

Finally, leverage points toward the higher end of the list include: the goal of the system, which, if changed, could completely redirect the system's activities; the paradigm or mindset that gave rise to the goal, which, if adjusted, could transform the collective understanding of what a system can and should be aiming for; and the realization that there are multiple worldviews or paradigms besides an organization's current one, which can empower people to adopt a different paradigm when appropriate.

Summary. In complex and sociotechnical systems, accidents cannot always be reduced to a linear chain of successive events. The large number of complex interactions and feedback loops means that accidents can happen even if no single component fails and that it may not be possible to identify a root cause. Since we may not be able to anticipate every potential pathway to an accident, it is often more fruitful to address systemic factors that diffusely influence risk. Meadows' twelve leverage points can help us identify systemic factors that, if changed, could have far-reaching, system-wide effects.

4.5 SYSTEMIC FACTORS

As discussed above, if we want to improve the safety of a complex, sociotechnical system, it might be most effective to address the blunt end, or the broad systemic factors that can diffusely influence operations. Some of the most important systemic factors include regulations, social pressure, technosolutionism, competitive pressure, safety costs, and safety culture. We will now discuss each of these in more detail.

Safety regulations can be imposed by government or internal policies. Safety regulations can require an organization to adhere to various safety standards, such as conducting regular staff training and equipment maintenance. These stipulations can be defined and enforced by a government or by an organization's own internal policies. The more stringent and effectively targeted these requirements are, the safer a system is likely to be.

Social pressure can encourage organizations to improve safety. Public attitudes towards a particular technology can affect an organization's attitude to safety. Significant social pressure about risks can mean that organizations are subject to more scrutiny, while little public awareness can allow organizations to take a more relaxed attitude toward safety.

Technosolutionism should be discouraged. Attempting to fix problems simply by introducing a piece of technology is called technosolutionism. It does not always work, especially in complex and sociotechnical systems. Although technology can certainly be helpful in solving problems, relying on it can lead organizations to neglect the broader system. They should consider how the proposed technological solution will actually function in the context of the whole system, and how it might affect the behavior of other components and human operators.

Multiple geoengineering technologies have been proposed as solutions to climate change, such as spraying particles high in the atmosphere to reflect sunlight. However, there are concerns that attempting this could have unexpected side effects. Even if spraying particles in the atmosphere did reverse global heating, it might also interfere with other components of the atmosphere in ways that we fail to predict, potentially with harmful consequences for life. Instead, we could focus on non-technical interventions like preserving forested areas that are more robustly likely to work without significant unforeseen negative side-effects.

Competitive pressures can lead to compromise on safety. If multiple organizations or countries are pursuing the same goal, they will be incentivized to get an edge over one another. They might try to do this by reaching the goal more quickly or by trying to make the end product more valuable to customers in terms of the functionality it offers. These competitive pressures can compel employees and decision-makers to cut corners and pay less attention to safety.

On a larger scale, competitive pressures might put organizations or countries in an arms race, wherein safety standards slip because of the urgency of the situation. This will be especially true if one of the organizations or countries has lower safety standards and consequently moves quicker; others might feel the need to lower their standards as well, in order to keep up. The risks this process presents are encapsulated by Publilius Syrus's aphorism: "Nothing can be done at once hastily and prudently." We consider this further in the Collective Action Problems chapter.

Various safety costs can discourage the adoption of safety measures. There are often multiple costs of increasing safety, not only financial costs but also slowdowns and reduced product performance. Adopting safety measures might therefore decrease productivity and slow progress toward a goal, reducing profits. The higher the costs of safety measures, the more reluctant an organization might be to adopt them.

Developers of AI systems may want to put more effort into transparency and interpretability. However, investigating these areas is costly: at the very least, there will be personnel and compute costs that could otherwise have been used to directly create more capable systems. Additionally, it may delay the completion of the finished product. There might also be costs from making a system more interpretable in terms of product performance. Creating more transparent models might require AIs to select only those actions which are clearly explainable. In general, safety features can reduce model capabilities, which organizations might prefer to avoid.

The general safety culture of an organization is an important systemic factor. A final systemic factor that will broadly influence a system's safety can simply be referred to as its "safety culture." This captures the general attitude that the people in an organization have toward safety—how seriously they take it, and how that translates into their actions. We will discuss some specific features of a diligent safety culture in the next section.

Summary. We have seen that component failure accident models have some significant limitations, since they do not usually capture diffuse sources of risk that can shape a system's dynamics and indirectly affect the likelihood of accidents. These include important systemic factors such as competitive pressures, safety costs, and safety culture. We will now turn to systemic accident models that acknowledge these ideas and attempt to account for them in risk analysis.

4.5.1 Systemic Accident Models

We have explored how component failure accident models are insufficient for properly understanding accidents in complex systems. When it comes to AIs, we must understand what sort of system we are dealing with. Comparing AI safety to ensuring the safety of specific systems like rockets, power plants, or computer programs can be misleading. The reality of today's world is that many hazardous technologies are operated by a variety of human organizations: together, these form complex sociotechnical systems that we need to make safer. While there may be some similarities between different hazardous technologies, there are also significant differences in the properties of these technologies which means it will not necessarily work to take safety strategies from one system and map them directly onto another. We should not anchor to individual safety approaches used in rockets or power plants.

Instead, it is more beneficial to approach AI safety from a broader perspective of making complex, sociotechnical systems safer. To this end, we can draw on the theory of sociotechnical systems, which offers "a method of viewing organizations which emphasizes the interrelatedness of the functioning of the social and technological subsystems of the organization and the relation of the organization as a whole to the environment in which it operates."

We can also use the complex systems literature more generally, which is largely about the shared structure of many different complex systems. Accidents in complex systems can often be better understood by looking at the system as a whole, rather than focusing solely on individual components. Therefore, we will now consider systemic accident models, which aim to provide insights into why accidents occur in systems by analyzing the overall structure and interactions within the system, including human factors that are not usually captured well by component failure models.

Normal Accident Theory (NAT). Normal Accident Theory (NAT) is one approach to understanding accidents in complex systems. It suggests that accidents are inevitable in systems that exhibit the following two properties:

1. Complexity: a large number of interactions between components in the system such as feedback loops, discussed in the complex systems chapter. Complexity can make it infeasible to thoroughly understand a system or exhaustively predict all its potential failure modes.
2. Tight coupling: one component in a system can rapidly affect others so that one relatively small event can rapidly escalate to become a larger accident.

NAT concludes that, if a system is both highly complex and tightly coupled, then accidents are inevitable—or "normal"—regardless of how well the system is managed [268].

NAT focuses on systemic factors. According to NAT, accidents are not caused by a single component failure or human error, but rather by the interactions and interdependencies between multiple components and subsystems. NAT argues that accidents are a normal part of complex systems and cannot be completely eliminated. Instead, the focus should be on managing and mitigating the risks associated with these systems to minimize the severity and frequency of accidents. NAT emphasizes the importance of systemic factors, such as system design, human factors such as organizational culture, and operational procedures, in influencing accident outcomes. By understanding and addressing these systemic factors, it is possible to improve the safety and resilience of complex systems.

Some safety features create additional complexity. Although we can try to incorporate safety features, NAT argues that many attempts to prevent accidents in these kinds of systems can sometimes be counterproductive, as they might just add another layer of complexity. As we explore in the Complex Systems chapter, systems often respond to interventions in unexpected ways. Interventions can cause negative side effects or even inadvertently exacerbate the problems they were introduced to solve.

Redundancy, which was listed earlier as a safe design principle, is supposed to increase safety by providing a backup for critical components, in case one of them fails. However, redundancy also increases complexity, which increases the risks of unforeseen and unintended interactions that can make it impossible for operators to predict all potential issues [269]. Having redundant components can also cause confusion; for example, people might receive contradictory instructions from multiple redundant monitoring systems and not know which one to believe.

Reducing complexity can be a safety feature. We may not be able to completely avoid complexity and tight coupling in all systems, but there are many cases where we can reduce one or both of them and thus meaningfully reduce risk. One example of this is reducing the potential for human error by making systems more intuitive, such as by using color coding and male/female adapters in electrical applications to reduce the incidence of wiring errors. Such initiatives do not eliminate risks, and accidents are still normal in these systems, but they can help reduce the frequency of errors.

High Reliability Organizations (HROs)

The performance of some organizations suggests serious accidents might be avoidable. The main assertion of NAT is that accidents are inevitable in complex, tightly coupled systems. In response to this conclusion, which might be perceived as pessimistic, other academics developed a more optimistic theory that points to "high reliability organizations" (HROs) that consistently operate hazardous technologies with low accident rates. These precedents include air traffic control, aircraft carriers, and nuclear power plants.

HRO theory emphasizes the importance of human factors, arguing that it must be possible to manage even complex, tightly coupled systems in a way that reliably avoids accidents. It identifies five key features of HROs' management culture that can significantly lower the risk of accidents [270]. We will now discuss these five features and how AIs might help improve them.

1. **Preoccupation with failure means reporting and studying mistakes and near misses.** HROs encourage the reporting of all anomalies, known failures, and near misses. They study these events carefully and learn from them. HROs also keep in mind potential failure modes that have not occurred yet and which have not been predicted. The possibility of unanticipated failure modes constitutes a risk of black swan events, which will be discussed in detail later in this chapter.

HROs are therefore vigilant about looking out for emerging hazards. AI systems tend to be good at detecting anomalies, but not near misses.

2. **Reluctance to simplify interpretations means looking at the bigger picture.** HROs understand that reducing accidents to chains of events often oversimplifies the situation, and is not necessarily helpful for learning from mistakes and improving safety. They develop a wide range of expertise so that they can come up with multiple different interpretations of any incident. This can help with understanding the broader context surrounding an event, and systemic factors that might have been at play. HROs also implement many checks and balances, invest in hiring staff with diverse perspectives, and regularly retrain everyone. AIs could be used to generate explanations for hazardous events or conduct adversarial reviews of explanations of system failures.
3. **Sensitivity to operations means maintaining awareness of how a system is operating.** HROs invest in the close monitoring of systems to maintain a continual, comprehensive understanding of how they are behaving, whether through excellent monitoring tools or hiring operators with deep situational awareness. This can ensure that operations are going as planned, and notice early if anything unexpected happens, permitting taking corrective action early, before the situation escalates. AI systems that dynamically aggregate information in real-time can help improve situational awareness.
4. **Commitment to resilience means actively preparing to tackle unexpected problems.** HROs train their teams in adaptability and improvising solutions when confronted with novel circumstances. By practicing dealing with issues they have not seen before, employees develop problem-solving skills that will help them cope if anything new and unexpected arises in reality. AIs have the potential to enhance teams' on-the-spot problem-solving, such as by creating surprising situations for testing organizational efficiency.
5. **Under-specification of structures means information can flow rapidly in a system.** Instead of having rigid chains of communication that employees must follow, HROs have communication throughout the whole system. All employees are allowed to raise an alarm, regardless of their level of seniority. This increases the likelihood that problems will be flagged early, and also allows information to travel rapidly throughout the organization. This under-specification of structures is also sometimes referred to as "deference to expertise," because it means that all employees are empowered to make decisions relating to their expertise, regardless of their place in the hierarchy.

High-reliability organizations (HROs) provide valuable insights into the development and application of AI technology. By emulating the characteristics of HROs, we can create combined human-machine systems that prioritize safety and mitigate risks. These sociotechnical systems should continuously monitor their own behavior and the environment for anomalies and unanticipated side effects. These systems should also support combined human-machine situational awareness and improvisational planning, allowing for real-time adaptation and flexibility. Lastly, AIs should have models of their own expertise and the expertise of human operators to ensure effective problem routing. By adhering to these principles, we can develop AI systems that function like HROs, ensuring high reliability and minimizing the potential risks associated with their deployment and use.

Criticisms of HRO Theory

Doubts have been raised about how widely HRO theory can be applied. Although the practices listed above can improve safety, a main criticism of HRO theory is that they cannot be applied to all systems and technologies [269]. This is because the theory was developed from a relatively small group of example systems, and certain features of them cannot be replicated in all systems.

It is difficult to understand systems sufficiently well. First, in the examples of HROs identified (such as air traffic control or nuclear power plants), operators usually have near-complete knowledge of the technical processes involved. These organizations' processes have also remained largely unchanged for decades, allowing for lessons to be learned from errors and for safety systems to become more refined. However, according to NAT, the main reason that complexity contributes to accidents is that it *precludes* a thorough understanding of all processes, and anticipation of all potential failure modes. HROs with near-complete knowledge of technical processes might be considered rare cases. These conditions cannot be replicated in all systems, especially not in those operating new technologies.

HROs prioritize safety, but other organizations might not. The second reason why HRO theory might not be broadly applicable is that its suggestions generally focus on prioritizing safety as a goal. This might make sense for several of the example HROs, where safety is an intrinsic part of the mission. Airlines, for instance, would not be viable businesses if they did not have a strong track record of transporting passengers safely. However, this is less feasible in organizations where safety is not so pertinent to the mission. In many other profit maximization organizations, safety can conflict with the main mission, as safety measures may be costly and reduce productivity.

Not all HROs are tightly coupled. Another criticism of HRO theory is that several of the example systems might actually be considered loosely coupled. For instance, in air traffic control, extra time is scheduled in between landings on the same runway, to allow for delays, errors, and corrections. However, NAT claims that tight coupling is the second system feature that makes accidents inevitable. Loosely coupled systems may not, therefore, be a good counterexample.

Deference to expertise might not always be realistic. A final reservation about HRO theory is that the fifth characteristic (deference to expertise) assumes that employees will have the necessary knowledge to make the best decisions at the local level. However, information on the system as a whole is sometimes required in order to make the best decisions, as actions in one subsystem may have knock-on effects for other subsystems. Employees might not always have enough information about the rest of the system to be able to take this big-picture view.

Comparing NAT and HRO Theory

The debate over whether accidents are inevitable or avoidable remains unsettled. A particular sticking point is that, despite having low accident rates, some of the HRO examples have experienced multiple near misses. This could be interpreted as evidence that NAT is correct. We could view it as a matter of luck that these near misses did not become anything larger. This would indicate that organizations presented as HROs are in fact vulnerable to accidents. On the other hand, near misses could instead be interpreted as supporting HRO theory; the fact that they did not turn into anything larger could be considered evidence that HROs have the appropriate measures in place to prevent accidents. It is not clear which of these interpretations is correct [269].

Nevertheless, both NAT and HRO theory have contributed important concepts to safety engineering. NAT has identified complexity and tight coupling as key risk factors, while HRO theory has developed important principles for a good organizational safety culture. Both schools of thought acknowledge that complex systems must be treated differently from simpler systems, requiring consideration of all the components, their interactions, and human factors. We will now explore some alternative approaches that view system safety more holistically, rather than considering it a product of reliable components, interactions, and operators.

Rasmussen's Risk Management Framework and AcciMap

System control with safety boundaries. Rasmussen's Risk Management Framework (RMF) is an accident model that recognizes that accidents are usually the culmination of many different factors, rather than a single root cause [271]. This model frames risk management as a question of control, emphasizing the need for clearly defined safety boundaries that a system's operations must stay within.

Levels of organization and AcciMap. The RMF considers six hierarchical levels of organization within a system, each of which can affect its safety: government, regulators, the company, management, frontline workers, and the work itself. By drawing out an "AcciMap" with this hierarchy, we can identify actors at different levels who share responsibility for safety, as well as conditions that may influence the risk of an accident. This analysis makes it explicit that accidents cannot be solely explained by an action at the sharp end.

Systems can gradually migrate into unsafe states. The RMF also asserts that behaviors and conditions can gradually "migrate" over time, due to environmental pressures. If this migration leads to unsafe systemic conditions, this creates the potential for an event at the sharp end to trigger an accident. This is why it is essential to continually enforce safety boundaries and avoid the system migrating into unsafe states.

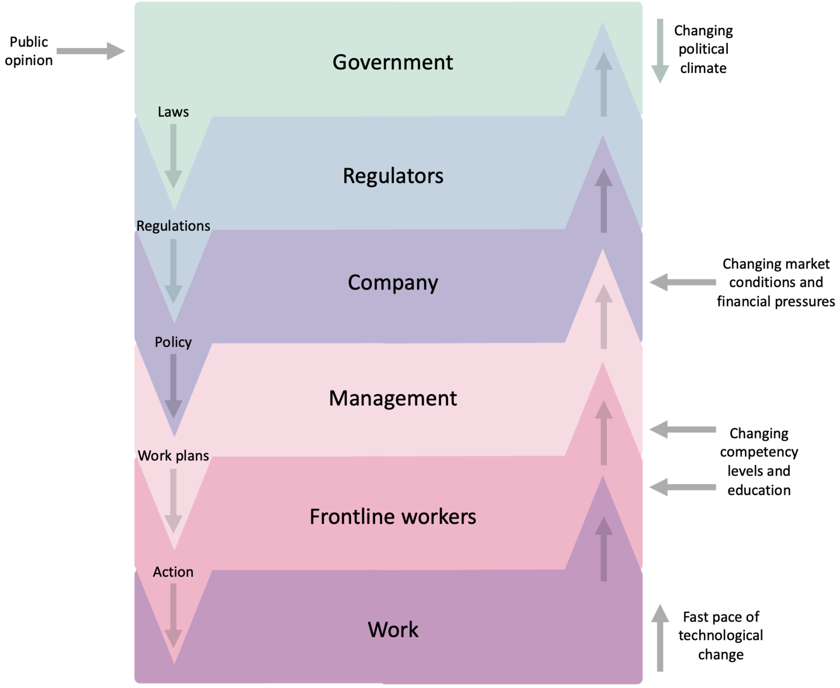

Public opinion
Government
Changing political climate
Laws
Regulators
Regulations
Company
Changing market conditions and financial pressures
Policy
Management
Changing competency levels and education
Work plans
Frontline workers
Action
Work
Fast pace of technological change

FIGURE 4.9. Rasmussen's risk management framework lays out six levels of organization and their interactions, aiming to mark consistent safety boundaries by identifying hazards and those responsible for them.

System-Theoretic Accident Model and Processes (STAMP)

STAMP is based on insights from the study of complex systems. According to the systems-theory paradigm, safety is an emergent property that is unlikely to be sufficiently understood just by looking at individual components in isolation. This is the view taken by System-Theoretic Accident Model and Processes (STAMP). STAMP identifies multiple levels of organization within a system, where each level is of higher complexity than the one below. Each level has novel emergent properties that cannot be practically understood through a reductive analysis of the level below. STAMP also recognizes that a system can be highly reliable but still be unsafe, and therefore puts the emphasis on safety rather than just on the reliability of components.

STAMP frames safety as a question of top-down control. STAMP proposes that safety can be enforced by each level effectively placing safety constraints on the one below to keep operations from migrating into unsafe states [269]. Performing STAMP-based risk analysis and management involves creating models of four aspects of a system: the organizational safety structure, the dynamics that can cause this structure to deteriorate, the models of system processes that operators must have, and the surrounding context. We will now discuss each of these in more detail.

The organizational safety structure. The first aspect is the safety constraints, the set of unsafe conditions which must be avoided. It tells us which components and operators are in place to avoid each of those unsafe conditions occurring. This can help to prevent accidents from component failures, design errors, and interactions between components that could produce unsafe states.

Dynamic deterioration of the safety structure. The second aspect is about how the safety structure can deteriorate over time, leading to safety constraints being enforced less stringently. Systems can "migrate" toward failure when many small events escalate into a larger accident. Since complex and sociotechnical systems involve large numbers of interactions, we cannot methodically compute the effects of every event within the system and exhaustively identify all the pathways to an accident. We cannot always reduce an accident to a neat chain of events or find a root cause: such instincts are often based on the desire to have a feeling of control by assigning blame. Instead, it might make sense to describe a system as migrating toward failure, due to the accumulation of many seemingly insignificant events.

This might include natural processes, such as wear and tear of equipment. It can also include systemic factors, such as competitive pressures, that might compel employees to omit safety checks. If being less safety-conscious does not quickly lead to an accident, developers might start to think that safety-consciousness is unnecessary. Having a model of these processes can increase awareness and vigilance around what needs to be done to maintain an effective safety structure.

Knowledge and communication about process models. The third aspect is the knowledge that operators must have about how the system functions in order to make safe decisions. Operators may be humans or automated systems that have to monitor feedback from the system and respond to keep it on track.

The process model that these operators should have includes the assumptions about operating conditions that were made during the design stage so that they will be aware of the conditions in which the system might not function properly, such as outside regular temperature ranges. It might also include information about how the specific subsystem that the operator is concerned with interacts with other parts of the system. The communication required for operators to maintain an accurate process model over time should also be specified. This can help to avoid accidents resulting from operators or software making decisions based on inaccurate beliefs about how the system is functioning.

The cultural and political context of the decision-making processes. The fourth aspect is the systemic factors that could influence the safety structure. It might include information about who the stakeholders are and what their primary concern is. For example, governments may impose stringent regulations, or they may put pressure on an organization to reach its goals quickly, depending on what is most important to them at the time. Similarly, social pressures and attitudes may put pressure on organizations to improve safety or pressure to achieve goals quickly.

Table 4.2 summarizes how the STAMP perspective contrasts with those of traditional component failure models.

TABLE 4.2. STAMP makes assumptions that differ from traditional component failure models.

Old Assumption	New Assumption
Accidents are caused by chains of directly related events.	Accidents are complex processes involving the entire sociotechnical system.
We can understand accidents by looking at chains of events leading to the accident.	Traditional event-chain models cannot describe this process adequately.
Safety is increased by increasing system or component reliability.	High reliability is not sufficient for safety.
Most accidents are caused by operator error.	Operator error is a product of various environmental factors.
Assigning blame is necessary to learn from and prevent accidents.	Holistically understand how the system behavior contributed to the accident.
Major accidents occur from simultaneous occurrences of random events.	Systems tend to migrate towards states of higher risk.

STAMP-based analysis techniques include System-Theoretic Process Analysis (STPA). On a practical level, there are methods of analyzing systems that take the holistic approach outlined by STAMP. These include System-Theoretic Process Analysis (STPA), which can be used at the design stage, and involves steps such as identifying hazards and constructing a control structure to mitigate their effects and improve system safety.

Dekker's Drift into Failure model

Decrementalism is the deterioration of system processes through a series of small changes. A third accident model based on systems theory is Dekker's Drift into Failure (DIF) model [272]. DIF focuses on the migration of systems that the RMF and STAMP also acknowledge, describing how this can lead to a "drift into failure." Since an individual decision to change processes may be relatively minor, it can seem that it will not make any difference to a system's operations or safety. For this reason, systems are often subject to decrementalism, a gradual process of changes through one small decision at a time that degrades the safety of a system's operations.

Many relatively minor decisions can combine to lead to a major difference in risk. Within complex systems, it is difficult to know all the potential consequences of a change in the system, or how it might interact with other changes. Many alterations to processes within a system, each of which might not make a difference by itself, can interact in complex and unforeseen ways to result in a much higher state of risk. This is often only realized when an accident happens, at which point it is too late.

Summary. Normal accident theory argues that accidents are inevitable in systems with a high degree of complexity and tight coupling, no matter how well they are organized. On the other hand, it has been argued that HROs with consistently low accident rates demonstrate that it is possible to avoid accidents. HRO theory identifies five key characteristics that contribute to a good safety culture and reduce the likelihood of accidents. However, it might not be feasible to replicate these across all organizations.

Systemic models like Rasmussen's RMF, STAMP, and Dekker's DIF model are grounded in an understanding of complex systems, viewing safety as an emergent property. The RMF and STAMP both view safety as an issue of control and enforcing safety constraints on operations. They both identify a hierarchy of levels of organization within a system, showing how accidents are caused by multiple factors, rather than just by one event at the sharp end. DIF describes how systems are often subject to decrementalism, whereby the safety of processes is gradually degraded through a series of minor changes, each of which seems minor on its own.

In general, component failure models focus on identifying specific components or factors that can go wrong in a system and finding ways to improve those components. These models are effective at pinpointing direct causes of failure and proposing targeted interventions. However, they have a limitation in that they tend to overlook other risk sources and potential interventions that may not be directly related to the identified components. On the other hand, systemic accident models take a broader approach by considering the interactions and interdependencies between various components in a system, such as feedback loops, human factors, and diffuse causality models. This allows them to capture a wider range of risk sources and potential interventions, making them more comprehensive in addressing system failures.

4.6 DRIFT INTO FAILURE AND EXISTENTIAL RISKS

This book presents multiple ways in which the development and deployment of AIs could entail risks, some of which could be catastrophic or even existential. However, the systemic accident models discussed above highlight that events in the real world often unfold in a much more complex manner than the hypothetical scenarios we use to illustrate risks. It is possible that many relatively minor events could accumulate, leading us to drift toward an existential risk. We are unlikely to be able to predict and address every potential combination of events that could pave the route to a catastrophe.

For this reason, although it can be useful to study the different risks associated with AI separately when initially learning about them, we should be aware that hypothetical example scenarios are simplified, and that the different risks coexist. We will now discuss what we can learn from our study of complex systems and systemic accident models when developing an AI safety strategy.

Risks that do not initially appear catastrophic might escalate. Risks tend to exist on a spectrum. Power inequality, disinformation, and automation, for example, are prevalent issues within society and are already causing harm. Though serious, they are not usually thought of as posing existential risks. However, if pushed to an extreme degree by AIs, they could result in totalitarian governments or enfeeblement. Both of these scenarios could represent a catastrophe from which humanity may not recover. In general, if we encounter harm from a risk on a moderate scale, we should be careful to not dismiss it as non-existential without serious consideration.

Multiple lower-level risks can combine to produce a catastrophe. Another reason for thinking more comprehensively about safety is that, even if a risk is not individually extreme, it might interact with other risks to bring about catastrophic outcomes [273]. Imagine, for instance, a scenario in which competitive pressures fuel an AI race between developers. This may lead a company to reduce its costs by putting less money into maintaining robust information security systems, with the result that a powerful AI is leaked. This would increase the likelihood that someone with malicious intent successfully uses the AI to pursue a harmful outcome, such as the release of a deadly pathogen.

In this case, the AI race has not directly led to an existential risk by causing companies to, for example, bring AIs with insufficient safety measures to market. Nevertheless, it has indirectly contributed to the existential threat of a pandemic by amplifying the risk of malicious use.

This echoes our earlier discussion of catastrophes in complex systems, where we discussed how it is often impractical and infeasible to attribute blame to one major "root cause" of failure. Instead, systems often "drift into failure" through an accumulation and combination of many seemingly minor events, none of which would be catastrophic alone. Just as we cannot take steps to prevent every possible mistake or malfunction within a large, complex system, we cannot predict or control every single way that various risks might interact to result in disaster.

Conflict and global turbulence could make society more likely to drift into failure. Although we have some degree of choice in how we implement AI within society, we cannot control the wider environment. There are several reasons why events like wars that create societal turbulence could increase the risk of human civilization drifting into failure. Faced with urgent, short-term threats, people might deprioritize AI safety to focus instead on the most immediate concerns. If AIs can be useful in tackling those concerns, it might also incentivize people to rush into giving them greater power, without thinking about the long-term consequences. More generally, a more chaotic environment might also present novel conditions for an AI, that cause it to behave unpredictably. Even if conditions like war do not directly cause existential risks, they make them more likely to happen.

Broad interventions may be more effective than narrowly targeted ones. Previous attempts to manage existential risks have focused narrowly on avoiding risks directly from AIs, and mainly addressed this goal through technical AI research. Given the complexity of AIs themselves and the systems they exist within, it makes sense to adopt a more comprehensive approach, taking into account the whole risk landscape, including threats that may not immediately seem catastrophic. Instead of attempting to target just existential risks precisely, it may be more effective to implement broad interventions, including sociotechnical measures.

Summary. As we might expect from our study of complex systems, different types of risks are inextricably related and can combine in unexpected ways to amplify one another. While some risks may be generally more concerning than others, we cannot neatly isolate those that could contribute to an existential threat from those that could not, and then only focus on the former while ignoring the latter. In addressing existential threats, it is therefore reasonable to view systems holistically and consider a wide range of issues, besides the most obvious catastrophic risks. Due to system complexity, broad interventions are likely to be required as well as narrowly targeted ones.

4.7 TAIL EVENTS AND BLACK SWANS

In the first few sections of this chapter, we discussed failure modes and hazards, equations for understanding the risks they pose, and principles for designing safer systems. We also looked at methods of analyzing systems to model accidents and identify hazards and explored how different styles of analysis can be helpful for complex systems.

The classic risk equation tells us that the level of risk depends on the probability and severity of the event. A particular class of events, called *tail events*, have a very low probability of occurrence but a very high impact upon arrival. Tail events pose some unique challenges for assessing and reducing risk, but any competent form of risk management must attempt to address them. We will now explore these events and their implications in more detail.

4.7.1 Introduction to Tail Events

Tail events are events that happen rarely, but have a considerable impact when they do. Consider some examples of past tail events.

The 2007-2008 financial crisis: Fluctuations happen continually in financial markets, but crises of this scale are rare and have a large impact, with knock-on effects for banks and the general population.

The COVID-19 pandemic: There are many outbreaks of infectious diseases every year, but COVID-19 spread much more widely and killed many more people than most. It is rare for an outbreak to become a pandemic, but those that do will have a much larger impact than the rest.

The Internet: Many technologies are being developed all the time, but very few become so widely used that they transform society as much as the Internet has. This example illustrates that some tail events happen more gradually than others; the development and global adoption of the internet unfolded over decades, rather than happening as suddenly as the financial crisis or the pandemic. However, “sudden” is a relative term. If we look at history on the scale of centuries, then the transition into the Internet age can also appear to have happened suddenly.

ChatGPT: After being released in November 2022, ChatGPT gained 100 million users in just two months [274]. There are many consumer applications on the internet, but ChatGPT’s user base grew faster than those of any others launched before it. Out of many deep learning models, ChatGPT was the first to go viral in this way. Its release also represented a watershed moment in the progress of generative AI, placing the issue much more firmly in the public consciousness.

We need to consider the possibility of harmful tail events in risk management. The last two examples—the Internet and ChatGPT—illustrate that the impacts of tail events are not always strictly negative; they can also be positive or mixed. However, *tail risks* are usually what we need to pay attention to when trying to engineer safer systems.

Since tail events are rare, it can be tempting to think that we do not need to worry about them. Indeed, some tail events have not yet happened once in human history, such as a meteorite strike large enough to cause global devastation, or a solar storm intense enough to knock out the power grid. Nonetheless, tail events have such a high impact that it would be unwise to ignore the possibility that they could happen. As noted by the political scientist Scott Sagan: “Things that have never happened before happen all the time.” [275]

AI-related tail events could have a severe impact. As AIs are increasingly deployed within society, some tail risks we should consider include the possibility that an AI could be used to develop a bioweapon, or that an AI might hack a bank and wipe the financial information. Even if these eventualities have a low probability of occurring, it would only take one such event to cause widespread devastation. Such an event could define the overall impact of an AI’s deployment. For this reason, competent risk management must involve serious efforts to prevent tail events, however rare we think they might be.

4.7.2 Tail Events Can Greatly Affect the Average Risk

A tail event often changes the mean but not the median. Figure 4.10 can help us visualize how tail events affect the wider risk landscape. The graphs show data points representing individual events, with their placement along the x-axis indicating their individual impact.

In the first graph, we have numerous data points representing frequent, low-impact events: these are all distributed between 0 and 100, and mostly between 0 and 10. The mean impact and median impact of this dataset have similar values, marked on the x-axis.

In the second graph we have the same collection of events, but with the addition of a single data point of much higher impact—a tail event with an impact of 10,000. As

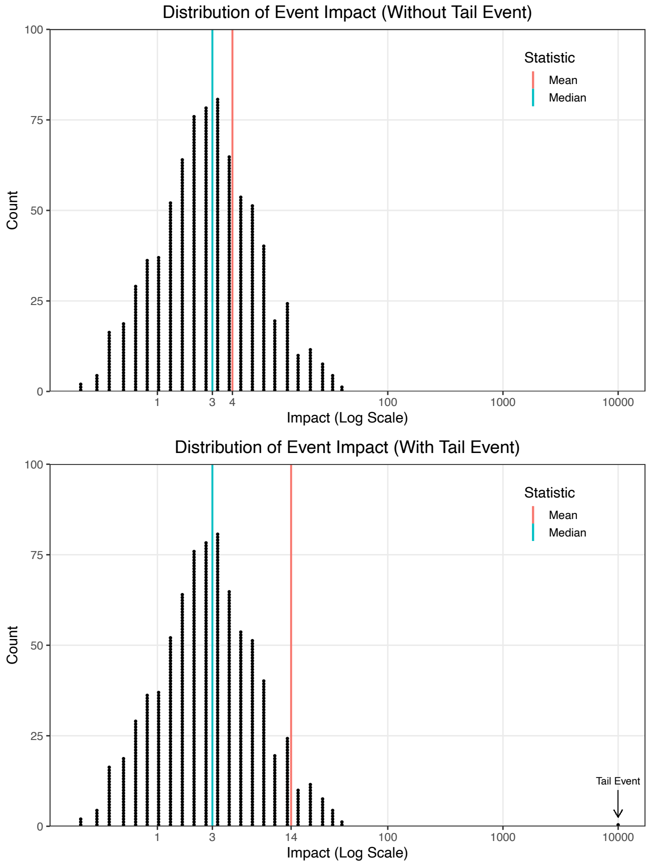


FIGURE 4.10. The occurrence of a tail event can dramatically shift the mean but not the median of the event type's impact.

indicated in the graph, the median impact of the dataset is approximately the same as before, but the mean changes substantially and is no longer representative of the general population of events.

We can also think about tail events in terms of the classic risk equation. Tail events have a low probability, but because they are so severe, we nonetheless evaluate the risk they pose as being large:

$$\text{Risk(hazard)} = P(\text{hazard}) \times \text{severity(hazard)}.$$

Depending on the exact values of probability and severity, we may find that tail risks are just as large as—or even larger than——the risks posed by much smaller events that happen all the time. In other words, although they are rare, we cannot afford to ignore the possibility that they might happen.

It is difficult to plan for tail events because they are so rare. Since we can hardly predict when tail events will happen, or even if they will happen at all, it is much more challenging to plan for them than it is for frequent, everyday events that we know we can expect to encounter. It is often the case that we do not know exactly what form they will take either.

For these reasons, we cannot plan the specific details of our response to tail events in advance. Instead, we must *plan to plan*. This involves organizing and developing an appropriate response, if and when it is necessary—how relevant actors should convene to decide on and coordinate the most appropriate next steps, whatever the precise details of the event. Often, we need to figure out whether some phenomena even present tail events, for which we need to consider their frequency distributions. We consider this concept next.

4.7.3 Tail Events Can Be Identified From Frequency Distributions

Frequency distributions tell us how common instances of different magnitudes are. To understand the origins of tail events, we need to understand frequency distributions. These distributions tell us about the proportion of items in a dataset that have each possible value of a given variable. Suppose we want to study some quantity, such as the ages of buildings. We might plot a graph showing how many buildings there are in the world of each age, and it might look something like the generic graph shown in figure 4.11.

The x-axis would represent building age, while the y-axis would indicate the number of buildings of each age—the frequency of a particular age appearing in the dataset. If our graph looked like figure 4.11, it would tell us that there are many buildings that are relatively new, perhaps only a few decades or a hundred years old, fewer buildings that are several hundred or a thousand years old, and very few buildings, such as the Pyramids at Giza, that are several thousand years old.

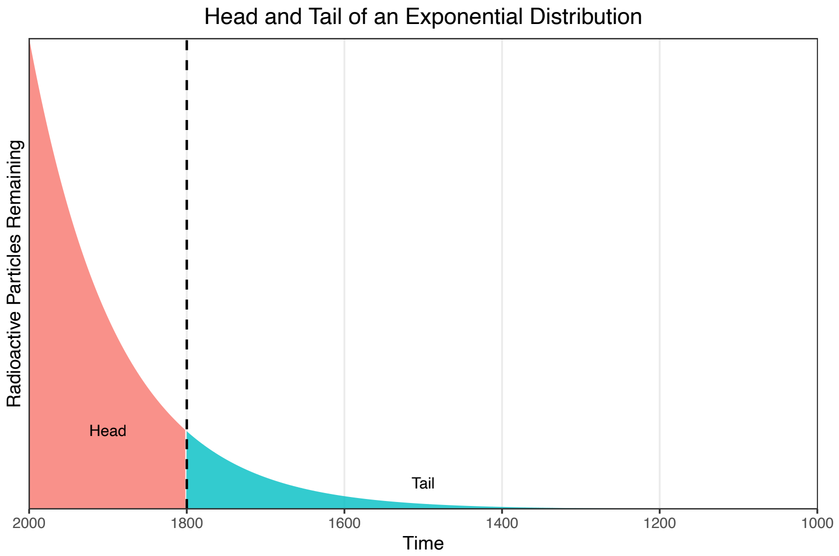


FIGURE 4.11. Many distributions have a head (an area where most of the probability is concentrated) and one or two tails (extreme regions of the distribution).

Many real-world frequency distributions have long tails. We can plot graphs like this for countless variables, from the size of different vertebrate species to the number of countries different people have visited. Each variable will have its own unique distribution, but many have the general property that there are lots of occurrences of a low value and relatively few occurrences of a high value. There are many vertebrate species with a mass of tens of kilograms, and very few with a mass in the thousands of kilograms; there are many people who have visited somewhere between 1-10 countries, and few people who have visited more than 50.

We can determine whether we are likely to observe tail events of a particular type by examining whether its frequency distribution has *thin tails* or *long tails*. In thin-tailed distributions, tail events do not exist. Examples of thin-tailed distributions include human characteristics such as height, weight, and intelligence. No one is over 100 meters tall, weighs over 10,000 kilograms, or has an IQ of 10,000. By contrast, in long-tailed distributions, tail events are possible. Examples of long-tailed distributions include book sales, earthquake magnitude, and word frequency. While most books only sell a handful of copies, most earthquakes are relatively harmless, and most words are rare and infrequently used, some books sell millions or even billions of copies, some earthquakes flatten cities, and some words (such as 'the' or 'I') are used extremely frequently. Of course, not all distributions neatly fit into a dichotomy of thin-tailed or long-tailed, but may be somewhere in between.

4.7.4 A Caricature of Tail Events

To illustrate the difference between long-tailed and thin-tailed distributions, we will now run through some comparisons between the two categories. Note that, with these statements, we are describing simplified caricatures of the two scenarios for pedagogical purposes.

Contrast 1: Share of the total received by the top few

Under thin tails, the top few receive quite a proportionate share of the total. If we were to measure the heights of a group of people, the total height of the tallest 10% would not be much more than 10% of the total height of the whole group.

Under long tails, the top few receive a disproportionately large share of the total. In the music industry, the revenue earned by the most successful 1% of artists represents around 77% of the total revenue earned by all artists.

Contrast 2: Who determines the total?

Under thin tails, the total is determined by the whole group. The total height of the tallest 10% of people is not a very good approximation of the total height of the whole group. Most members need to be included to get a good measure of the total. This is called "tyranny of the collective."

TABLE 4.3. A caricature of thin tails and long tails reveals several trends that often hold for each.

Caricature: Thin Tails	**Caricature: Long Tails**
The top few receive a proportionate share of the total.	The top few receive a disproportionately large share of the total.
The total is determined by the whole group (”tyranny of the collective”).	The total is determined by a few extreme occurrences (”tyranny of the accidental”).
The typical member of a group has an average value, close to the mean.	The typical member is either a giant or a dwarf.
A single event cannot escalate to become much bigger than the average.	A single event can escalate to become much bigger than many others put together.
Individual data points vary within a small range that is close to the mean.	Individual data points can vary across many orders of magnitude.
We can predict roughly what value a single instance will take.	It is much harder to robustly predict even the rough value that a single instance will take.

Under long tails, the total is determined by a few extreme occurrences. As discussed above, the most successful 1% of artists earn 77% of the total revenue earned by all artists. 77% is a fairly good approximation of the total. In fact, it is a better approximation than the revenue earned by the remaining 99% of artists would be. This is called “tyranny of the accidental.”

Contrast 3: The typical member

Under thin tails, the typical member of a group has an average value. Almost no members are going to be much smaller or much larger than the mean.

Under long tails, the typical member is either a giant or a dwarf. Members can generally be classified as being either extreme and high-impact or relatively insignificant.

Note that, under many real-world long-tailed distributions, there may be occurrences that seem to fall between these two categories. There may be no clear boundary dividing occurrences that count as insignificant from those that count as extreme.

Contrast 4: Scalability of events

Under thin tails, the impact of an event is not scalable. A single event cannot escalate to become much bigger than the average.

Under long tails, the impact of an event is scalable. A single event can escalate to become much bigger than many others put together.

Contrast 5: Randomness

Under thin tails, individual data points vary within a small range that is close to the mean. Even the data point that is furthest from the mean does not change the mean of the whole group by much.

Under long tails, individual data points can vary across many orders of magnitude. A single extreme data point can completely change the mean of the sample.

Contrast 6: Predictability

Under thin tails, we can predict roughly what value a single instance will take. We can be confident that our prediction will not be far off, since instances cannot stray too far from the mean.

Under long tails, it is much harder to predict even the rough value that a single instance will take. Since data points can vary much more widely, our best guesses can be much further off.

Having laid the foundations for understanding tail events in general, we will now consider an important subset of tail events: black swans.

4.7.5 Introduction to Black Swans

In addition to being rare and high-impact, as all tail events are, black swans are also unanticipated, seemingly coming out of the blue. The term “black swan” originates from a historical event that provides a useful analogy.

Finding a black swan. It was long believed in Europe that all swans were white because all swan sightings known to Europeans were of white swans. For this reason, the term “black swan” came to denote something nonexistent, or even impossible, much as today we say “pigs might fly.” The use of this metaphor is documented as early as Roman times. However, in 1697, a group of Dutch explorers encountered a black species of swan in Australia. This single, unexpected discovery completely overturned the long-held axiom that all swans were white.

This story offers an analogy for how we can have a theory or an assumption that seems correct for a long time, and then a single, surprising observation can completely upend that model. Such an observation can be classed as a tail event because it is rare and high-impact. Additionally, the fact that the observation was unforeseen shows us that our understanding is incorrect or incomplete.

From here on we will use the following working definition of black swans: A black swan is a tail event that was largely unpredictable to most people before it happened. Note that not all tail events are black swans; high-magnitude earthquakes, for example, are tail events, but we know where they are likely to happen eventually——they are on our radar.

4.7.6 Known Unknowns and Unknown Unknowns

Black swans are "unknown unknown" tail events [276]. We can sort events into four categories, as shown in the table below.

Known knowns: things we are aware of and understand.	**Unknown knowns**: things that we do not realize we know (such as tacit knowledge).
Known unknowns: things we are aware of but which we don't fully understand.	**Unknown unknowns**: things that we do not understand, and which we are not even aware we do not know.

In these category titles, the first word refers to our awareness, and the second refers to our understanding. We can now consider these four types of events in the context of a student preparing for an exam.

1. **We know that we know.** Known knowns are things we are both aware of and understand. For the student, these would be the types of questions that have come up regularly in previous papers and that they know how to solve through recollection. They are aware that they are likely to face these, and they know how to approach them.
2. **We do not know what we know.** Unknown knowns are things we understand but may not be highly aware of. For the student, these would be things they have not thought to prepare for but which they understand and can do. For instance, there might be some questions on topics they hadn't reviewed; however, looking at these questions, the student finds that they know the answer, although they cannot explain why it is correct. This is sometimes called tacit knowledge or unaccounted facts.
3. **We know that we do not know.** Known unknowns are things we are aware of but do not fully understand. For the student, these would be the types of questions that have come up regularly in previous papers but which they have not learned how to solve reliably. The student is aware that they are likely to face these but is not sure they will be able to answer them correctly. However, they are at least aware that they need to do more work to prepare for them.
4. **We do not know that we do not know.** Unknown unknowns are things we are unaware of and do not understand. These problems catch us completely off guard because we didn't even know they existed. For the student, unknown unknowns would be unexpectedly hard questions on topics they have never encountered before and have no knowledge or understanding of.

Unknown unknowns can also occur in AI safety and risk. Researchers may be diligently studying various aspects of AI and its potential risks, but new and unforeseen risks could emerge as AI technology advances. These risks may be completely unknown and unexpected, catching researchers off guard. It is important to acknowledge the existence of unknown unknowns because they remind us that there are limits to our knowledge and understanding. By being aware of this, we can be more humble in our approach to problem-solving and continuously strive to expand our knowledge and prepare for the unexpected.

We struggle to account for known unknowns and unknown unknowns. We have included the first two categories—known knowns and unknown knowns—for completeness. However, the most important categories in risk analysis and management are the last two: known unknowns and unknown unknowns. These categories pose risks because we do not fully understand how best to respond to them, and we cannot be perfectly confident that we will not suffer damage from them.

Unknown unknowns are particularly concerning. If we are aware that we might face a particular challenge, we can learn more and prepare for it. However, if we are unaware that we will face a challenge, we may be more vulnerable to harm. Black swans are the latter type of event; they are not even on our radar before they happen.

The difference between known unknowns and unknown unknowns is sometimes also described as a distinction between conscious ignorance and *meta-ignorance*. Conscious ignorance is when we see that we do not know something, whereas meta-ignorance is when we are unaware of our ignorance.

Black swans in the real world It might be unfair for someone to present us with an unknown unknown, such as finding questions on topics irrelevant to the subject in an exam setting. The wider world, however, is not a controlled environment; things do happen that we have not thought to prepare for.

Black swans indicate that our worldview is inaccurate or incomplete. Consider a turkey being looked after by humans, who provide plenty of food and a comfortable shelter safe from predators. According to all the turkey's evidence, the humans are benign and have the turkey's best interests at heart. Then, one day, the turkey is taken to the slaughterhouse. This is very much an unknown unknown, or a black swan, for the turkey, since nothing in its experience suggested that this might happen [276].

This illustrates that we might have a model or worldview that does a good job of explaining all our evidence to date, but then a black swan can turn up and show us that our model was incorrect. The turkey's worldview of benign humans explained all the evidence until the slaughterhouse trip. This event indicated a broader context that the turkey was unaware of.

Similarly, consider the 2008 financial crisis. Before this event, many people, including many of those working in finance, assumed that housing prices would always continue to increase. When the housing bubble burst, it showed that this assumption was incorrect.

Black swans are defined by our understanding. A black swan is a black swan because our worldview is incorrect or incomplete, which is why we fail to predict it. In hindsight, such events often only make sense after we realize that our theory was flawed. Seeing black swans makes us update our models to account for the new phenomena we observe. When we have a new, more accurate model, we can often look back in time and find the warning signs in the lead-up to the event, which we did not recognize as such at the time.

These examples also show that we cannot always reliably predict the future from our experience; we cannot necessarily make an accurate calculation of future risk based on long-running historical data.

Distinguishing black swans from other tail events

Only some tail events are black swans. As touched on earlier, it is essential to note that black swans are a subset of tail events, and not all tail events are black swans. For example, it is well known that earthquakes happen in California and that a high-magnitude one, often called "the big one," will likely happen at some point. It is not known exactly when—whether it will be the next earthquake or in several decades. It might not be possible to prevent all damage from the next "big one," but there is an awareness of the need to prepare for it. This represents a tail event, but not a black swan.

Some people might be able to predict some black swans. A restrictive definition of a black swan is an event that is an absolute unknown unknown for everybody and is impossible to anticipate. However, for our purposes, we are using the looser, more practical working definition given earlier: a highly impactful event that is largely unexpected for most people. For example, some individuals with relevant knowledge of the financial sector did predict the 2008 crisis, but it came out of the blue for most people. Even among financial experts, the majority did not predict it. Therefore, we count it as a black swan.

Similarly, although pandemics have happened throughout history, and smaller disease outbreaks occur yearly, the possibility of a pandemic was not on most people's radar before COVID-19. People with specific expertise were more conscious of the risk, and epidemiologists had warned several governments for years that they were inadequately prepared for a pandemic. However, COVID-19 took most people by surprise and therefore counts as a black swan under the looser definition.

The development and rollout of AI technologies could be subject to black swans. Within the field of AI, the consensus view for a long time was that deep learning techniques were fundamentally limited. Many people, even computer science professors, did not take seriously the idea that deep learning technologies might transform society in the near term—even if they thought this would be possible over a timescale of centuries.

Deep learning technologies have already begun to transform society, and the rate of progress has outpaced most people's predictions. We should, therefore, seriously consider the possibility that the release of these technologies could pose significant risks to society.

There has been speculation about what these risks might be, such as a flash war and autonomous economy, which are discussed in the Collective Action Problems chapter. These eventualities might be known to some people, but for many potential risks, there is not widespread awareness in society; if they happened today, they would be black swans. Policymakers must have some knowledge of these risks. Furthermore, the expanding use of AI technologies may entail risks of black swan scenarios that no one has yet imagined.

4.7.7 Implications of Tail Events and Black Swans for Risk Analysis

Tail events and black swans present problems for analyzing and managing risks, because we do not know if or when they will happen. For black swans, there is the additional challenge that we do not know what form they will take.

Since, by definition, we cannot predict the nature of black swans in advance, we cannot put any specific defenses in place against them, as we might for risks we have thought of. We can attempt to factor black swans into our equations to some degree, by trying to estimate roughly how likely they are and how much damage they would cause. However, they add much more uncertainty into our calculations. We will now discuss some common tendencies in thinking about risk, and why they can break down in situations that are subject to tail events and black swans.

First, we consider how our typical risk estimation methods break down under long tails because our standard arsenal of statistical tools are rendered useless. Then, we consider how cost-benefit analysis is strained when dealing with long-tailed events because of its sensitivity to our highly uncertain estimates. After this, we discuss why we should be more explicitly considering extremes instead of averages, and look at three common mistakes when dealing with long-tailed data: the delay fallacy, interpreting an absence of evidence, and the preparedness paradox.

Typical risk estimation methods break down under long tails

Tail events and black swans can substantially change the average risk of a system. It is challenging to account for tail events in the risk equation. Since tail events and black swans are extremely severe, they significantly affect the average outcome. Recall the equation for risk associated with a system:

$$\text{Risk} = \sum_{\text{hazard}} P(\text{hazard}) \times \text{severity}(\text{hazard}).$$

Additionally, it is difficult to estimate their probability and severity accurately. Yet, they would considerably change the evaluation of risk because they are so severe. Furthermore, since we do not know what form black swans will take, it may be even more difficult to factor them into the equation accurately. This renders the usual statistical tools useless in practice for analyzing risk in the face of potential black swans.

If the turkey in the previous example had tried to calculate the risk to its wellbeing based on all its prior experiences, the estimated risk would probably have been fairly low. It certainly would not have pointed to the high risk of being taken to the slaughterhouse, because nothing like that had ever happened to the turkey before.

We need a much larger dataset than usual. As we increase the number of observations, we converge on an average value. Suppose we are measuring heights and calculating a new average every time we add a new data point. As shown in the first graph in figure 4.12, as we increase our number of data points, we quickly converge on an average that changes less and less with the addition of each new data point. This is a result of the law of large numbers.

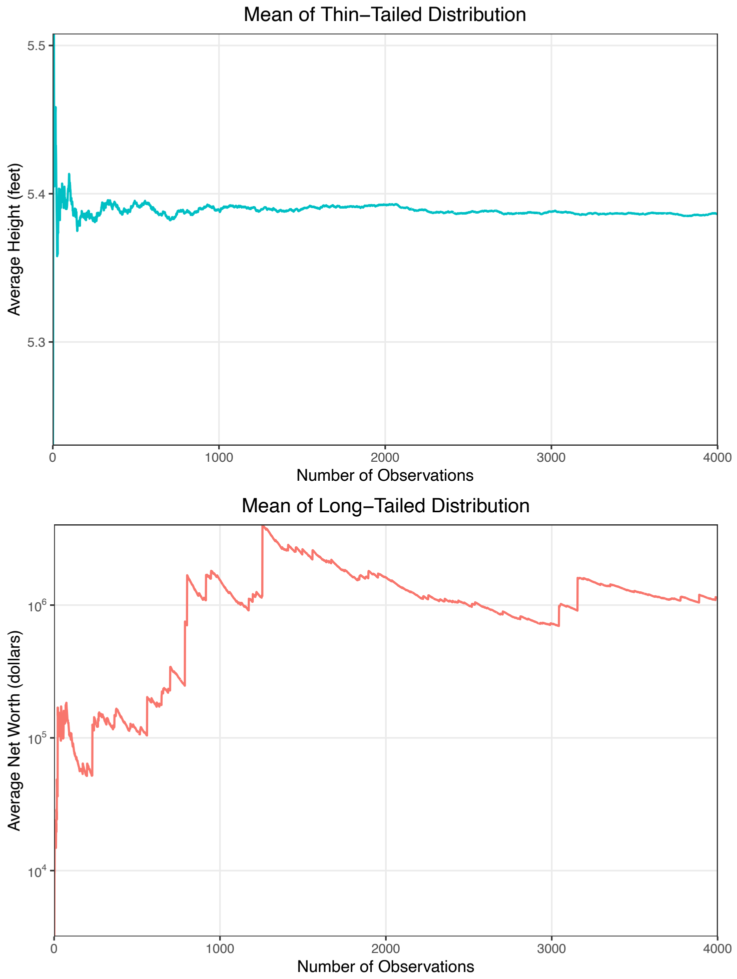


FIGURE 4.12. The mean of a long-tailed distribution is slow to convergence, rendering the mean a problematic summary statistic in practice.

Heights, however, are a thin-tailed variable. If we look instead at a long-tailed variable, such as net worth, as shown in the second graph in figure 4.12, a single extreme observation can change the average by several orders of magnitude. The law of large numbers still applies, in that we will still eventually converge on an average value, but it will take much longer.

Linear regression is a standard prediction method but is less useful for long-tailed data. Linear regression is a technique widely used to develop predictive models based on historical data. However, in situations where we are subject to the possibility of tail events or black swans, we might not be sure that we have enough historical data to converge on an accurate calculation of average risk. Linear regression is, therefore, less helpful in assessing and predicting risk for long-tailed scenarios.

Explicit cost-benefit analysis is strained under long tails

Cost-benefit analysis using long-tailed data often requires highly accurate estimates. Traditional cost-benefit analysis weighs the probability of different results and how much we would lose or gain in each case. From this information, we can calculate whether we expect the outcome of a situation to be positive or negative. For example, if we bet on a 50/50 coin toss where we will either win \$5 or lose \$5, our overall expected outcome is \$0.

Example: lotteries. Now, imagine that we are trying to perform a cost-benefit analysis for a lottery where we have a high probability of winning a small amount and a low probability of losing a large amount. If we have a 99.9% chance of winning \$15 and a 0.1% chance of losing \$10,000, then our expected outcome is:

$$(0.999 \times 15) + (0.001 \times -10000) = 4.985.$$

Since this number is positive, we might believe it is a good idea to bet on the lottery. However, if the probabilities are only slightly different, at 99.7% and 0.3%, then our expected outcome is:

$$(0.997 \times 15) + (0.003 \times -10000) = -15.045.$$

This illustrates that just a tiny change in probabilities sometimes makes a significant difference in whether we expect a positive or a negative outcome. In situations like this, where the expected outcome is highly sensitive to probabilities, using an estimate of probability that is only slightly different from the actual value can completely change the calculations. For this reason, relying on this type of cost-benefit analysis does not make sense if we cannot be sure we have accurate estimates of the probabilities in question.

It is difficult to form accurate probability estimates for black swans. Black swans happen rarely, so we do not have a lot of historical data from which to calculate the exact probability that they will occur. As we explored above, it takes a lot of data—often more than is accessible—to make accurate judgments for long-tailed events more generally. Therefore, we cannot be certain that we know their probabilities accurately, rendering cost-benefit analysis unsuitable for long-tailed data, especially for black swans.

This consideration could be significant for deciding whether and how to use AI technologies. We might have a high probability of benefitting from the capabilities of deep learning models, and there might be only a low probability of an associated black swan transpiring and causing harm. However, we cannot calculate an accurate probability of a black swan event, so we cannot evaluate our expected outcome precisely.

It is unrealistic to estimate risk when we could face black swans. If we attempt to develop a detailed statistical model of risk for a situation, we are making an implicit assumption that we have a comprehensive understanding of all the possible failure modes and how likely they are. However, as previous black swan events have demonstrated, we cannot always assume we know all the eventualities.

Even for tail events that are known unknowns, we cannot assume we have sufficiently accurate information about their probabilities and impacts. Trying to precisely estimate risk when we might be subject to tail events or black swans can be viewed as an "unscientific overestimation of the reach of scientific knowledge" [259].

Thinking about extremes instead of averages is better under long tails

When making risk-related decisions, we should consider extremes, not only the average. Aside from whether or not we can calculate an accurate average outcome under the risk of tail events and black swans, there is also a question of whether the average is what we should be paying attention to in these situations anyway. This idea is captured in the following adage commonly attributed to Milton Friedman: "Never try to walk across a river just because it has an average depth of four feet." If a river is four feet deep on average, that might mean that it has a constant depth of four feet and is possible to wade across it safely. It might also mean that it is two or three feet deep near the banks and eight feet deep at some point in the middle. If this were the case, then it would not be a good idea to attempt to wade across it.

Failing to account for extremes instead of averages is one example of the mistakes people make when thinking about event types that might have black swans. Next, we will explore three more: the delay fallacy, misinterpreting an absence of evidence, and the preparedness paradox.

The delay fallacy

If we do not have enough information to conduct a detailed risk analysis, it might be tempting to gather more information before taking action. A common excuse for delaying action is: "If we wait, we will know more about the situation and be able to make a more informed decision, so we should not make any decisions now."

In thin-tailed scenarios, waiting for more information is often a good approach. Under thin tails, additional observations will likely help us refine our knowledge and identify the best course of action. Since there are no tail events, there is a limit to how much damage a single event can do. There is, therefore, less urgency to take action in these situations. The benefit of more information can be considered to outweigh the delay in taking action.

In long-tailed scenarios, waiting for more information can mean waiting until it is too late. Additional observations will not necessarily improve our knowledge of the situation under long tails. Most, if not all, additional observations will probably come from the head of the distribution and will not tell us anything new about the risk of tail events or black swans. The longer we wait before preparing, the more we expose ourselves to the possibility of such an event happening while we are unprepared. When tail events and black swans do materialize, it is often too late to intervene and prevent harm.

Governments failing to improve their pandemic preparedness might be considered an example of this. Epidemiologists' warnings were long ignored, which seemed fine for a long time because pandemics are rare. However, when COVID-19 struck, many governments tried to get hold of personal protective equipment (PPE) simultaneously and found a shortage. If they had stocked up on this before the pandemic, as experts had advised, then the outcome might have been less severe.

Furthermore, if a tail event or black swan is particularly destructive, we can never observe it and use that information to help us make better calculations. The turkey cannot use the event of being taken to the slaughterhouse to make future risk estimations more accurate. With respect to society, we cannot afford for events of this nature to happen even once.

We should be proactively investing in AI safety now. Since the development and rollout of AI technologies could represent a long-tailed scenario, entailing a risk of tail events and black swans, it would not make sense to delay action with the excuse that we do not have enough information. Instead, we should be proactive about safety by investing in the three key research fields discussed earlier: robustness, monitoring, and control. If we wait until we are certain that an AI could pose an existential risk before working on AI safety, we might be waiting until it is too late.

Interpreting an absence of evidence

It can be hard to imagine a future that is significantly different from our past and present experiences. Suppose a particular event has never happened before. In that case, it can be tempting to interpret that as an indication that we do not need to worry about it happening in the future, but this is not necessarily a sound judgment.

An absence of evidence is not strong evidence of absence. Even if we have not found evidence that there is a risk of black swan events, that is not evidence that there is no risk of black swan events. In the context of AI safety, we may not have found evidence that deep learning technologies could pose specific risks like deceptive alignment, but that does not necessarily mean that they do not pose such risks or that they will not at some point in the future.

The preparedness paradox

Safety measures that prevent harm can seem redundant. Imagine that we enact safety measures to reduce the risk of a potentially destructive event, and then the event does not happen. Some might be tempted to say that the safety measures were unnecessary or that implementing them was a waste of time and resources. Even if the event does happen but is not very severe, some people might still say that the safety measures were unnecessary because the event's consequences were not so bad.

However, this conclusion ignores the possibility that the event did not happen or was less severe because of the safety measures. We cannot run through the same period of time twice and discover how things would have unfolded without any safety measures. This is a cognitive bias known as the preparedness paradox: efforts to prepare for potential disasters can reduce harm from these events and, therefore, reduce the perceived need for such preparation.

The preparedness paradox can lead to self-defeating prophecies. A related concept is the "self-defeating prophecy," where a forecast can lead to actions that prevent the forecast from coming true. For example, suppose an epidemiologist predicts that there will be a high death toll from a particular infectious disease. In that case, this might prompt people to wash their hands more frequently and avoid large gatherings to avoid infection. These behaviors are likely to reduce infection rates and lead to a lower death toll than the epidemiologist predicted.

If we work proactively on reducing risks from global pandemics, and no highly destructive pandemics come to pass, some people would believe that the investment was unnecessary. However, it might be *because* of those efforts that no destructive events happen. Since we usually cannot run two parallel worlds—one with safety efforts and one without—it might be difficult or impossible to prove that the safety work prevented harm. Those who work in this area may never know whether their efforts have prevented a catastrophe and have their work vindicated. Nevertheless, preventing disasters is essential, especially in cases like the development of AI, where we have good theoretical reasons to believe that a black swan is on the cards.

4.7.8 Identifying the Risk of Tail Events or Black Swans

Since the possibility of tail events and black swans affects how we approach risk management, we must consider whether we are facing a long-tailed or thin-tailed scenario. We need to know whether we can rely on standard statistical methods to estimate risk or whether we face the possibility of rare, high-impact events. This can be difficult to determine, especially in cases of low information, but there are some valuable indicators we can look for.

Highly connected systems often give rise to long-tailed scenarios. As discussed earlier, multiplicative phenomena can lead to long tails. We should ask ourselves: Can one part of the system rapidly affect many others? Can a single event trigger a cascade? If the answers to these questions are yes, then it is possible that an event can escalate to become a tail event with an extreme impact.

The use of AI in society could create a new, highly connected system. If deep learning models become enmeshed within society and are put in charge of various decisions, then we will have a highly connected system where these agents regularly interact with humans and each other. In these conditions, a single erroneous decision made by one agent could trigger a cascade of harmful decisions by others, for example, if they govern the deployment of weapons. This could leave us vulnerable to sudden catastrophes such as flash wars or powerful rogue AIs.

Complex systems may be more likely to entail a risk of black swans. Complex systems can evolve in unpredictable ways and develop unanticipated behaviors. We cannot usually foresee every possible way a complex system might unfold. For this reason, we might expect that complex evolving systems present an inherent risk of black swans.

Deep learning models and the surrounding social systems are all complex systems. It is unlikely that we will be able to predict every single way AI might be used, just as, in the early days of the internet, it would have been difficult to predict every way technology would ultimately be used. This means that there might be a risk of AI being used in harmful ways that we have not foreseen, potentially leading to a destructive black swan event that we are unprepared for. The idea that deep learning systems qualify as complex systems is discussed in greater depth in the Complex Systems chapter.

New systems may be more likely to present black swans. Absence of evidence is only evidence of absence if we expect that some evidence should have turned up in the timeframe that has elapsed. For systems that have not been around for long, we would be unlikely to have seen proof of tail events or black swans since these are rare by definition.

AI may not have existed for long enough for us to have learned about all its risks. In the case of emerging technology, it is reasonable to think that there might be a risk of tail events or black swans, even if we do not have any evidence yet. The lack of evidence might be explained simply by the fact that the technology has not been around for long. Our meta-ignorance means that we should take AI risk seriously. By definition, we can't be sure there are no unknown unknowns. Therefore, it is over-confident for us to feel sure we have eliminated all risks.

Accelerating progress could increase the frequency of black swan events. We have argued that black swan events should be taken seriously, despite being rare. However, as technological progress and economic growth advance at an increasing rate, such events may in fact become more frequent, further compounding their relevance to risk management. This is because the increasing pace of change also means that we will more often face novel circumstances that could present unknown unknowns. Moreover, within the globalized economy, social systems are increasingly interconnected, increasing the likelihood that one failure could trigger a cascade and have an outsized impact.

There are techniques for turning some black swans into known unknowns. As discussed earlier, under our practical definition, not all black swans are completely unpredictable, especially not for people who have the relevant expertise. Ways of putting more black swans on our radar include expanding our safety imagination, conducting horizon scanning or stress testing exercises, and red-teaming [277].

Safety imagination. Expanding our "safety imagination" can help us envision a wider range of possibilities. We can do this by playing a game of "what if" to increase the range of possible scenarios we can imagine unfolding. Brainstorming sessions can also help to rapidly generate lots of new ideas about potential failure modes in a system. We can identify and question our assumptions—–about what the nature of a hazard will be, what might cause it, and what procedures we will be able to follow to deal with it—in order to imagine a richer set of eventualities.

Horizon scanning. Some HROs use a technique called horizon scanning, which involves monitoring potential future threats and opportunities before they arrive, to minimize the risk of unknown unknowns [278]. AI systems could be used to enhance horizon-scanning capabilities by simulating situations that mirror the real world with a high degree of complexity. The simulations might generate data that reveal potential black swan risks to be aware of when deploying a new system. As well as conducting horizon scanning, HROs also contemplate near-misses and speculate about how they might have turned into catastrophes, so that lessons can be learned.

Red teaming. "Red teams" can find more black swans by adopting a mindset of malicious intent. Red teams should try to think of as many ways as they can to misuse or sabotage the system. They can then challenge the organization on how it would respond to such attacks. Finally, stress tests such as dry-running hypothetical scenarios and evaluating how well the system copes with them, and thinking about how it could be improved can improve a system's resilience to unexpected events.

4.8 CONCLUSION

4.8.1 Summary

In this chapter, we have explored various methods of analyzing and managing risks inherent in systems. We began by looking at how we can break risk down into two components: the probability and severity of an accident. We then went into greater detail, introducing the factors of exposure and vulnerability, showing how each affects the level of risk we calculate. By decomposing risk in this way, we can identify measures we can take to reduce risks. We also considered the concept of ability to cope and how it relates to risk of ruin.

Next, we described a metric of system reliability called the ”nines of reliability”. This metric refers to the number of nines at the beginning of a system's percentage or decimal reliability. We found that adding another nine of reliability is equivalent to reducing the probability of an accident by a factor of 10, and therefore results in a tenfold increase in expected time before failure. A limitation of the nines of reliability is that they only contain information about the probability of an accident, but not its severity, so they cannot be used alone to calculate risk.

We then listed several safe design principles, which can be incorporated into a system from the design stage to reduce the risk of accidents. In particular, we explored redundancy, separation of duties, the principle of least privilege, fail-safes, antifragility, negative feedback mechanisms, transparency, and defense in depth.

To develop an understanding of how accidents occur in systems, we next explored various accident models, which are theories about how accidents happen and the factors that contribute to them. We reviewed three component failure accident models: the Swiss cheese model, the bow tie model, and fault tree analysis, and considered their limitations, which arise from their chain-of-events style of reasoning. Generally, they do not capture how accidents can happen due to interactions between components, even when nothing fails. Component failure models are also unsuited to modeling how the numerous complex interactions and feedback loops in a system can make it difficult to identify a root cause, and how it can be more fruitful to look at diffuse causality and systemic factors than specific events.

After highlighting the importance of systemic and human factors, we delved deeper into some examples of them, highlighting regulations, social pressure, competitive pressures, safety costs, and safety culture. We then moved on to look at systemic accident models that attempt to take these factors into consideration. Normal Accident Theory states that accidents are inevitable in complex and tightly coupled systems. On the other hand, HRO theory points to certain high reliability organizations as evidence that it is possible to reliably avoid accidents by following five key management principles: preoccupation with failure, reluctance to simplify interpretations, sensitivity to operations, commitment to resilience, and deference to expertise. While these features can certainly contribute to a good safety culture, we also looked at the limitations and the difficulties in replicating some of them in other systems.

Rounding out our discussion of systemic factors, we outlined three accident models that are grounded in complex systems theory. Rasmussen's Risk Management Framework (RMF) identifies six hierarchical levels within a system, identifying actors at each level who share responsibility for safety. The RMF states that a system's operations should be kept within defined safety boundaries; if they migrate outside of these, then the system is in a state where an event at the sharp end could trigger an accident. However, the factors at the blunt end are also responsible, not just the sharp-end event.

Similarly, STAMP and the related STPA analysis method view safety as being an emergent property of an organization, detailing different levels of organization within a system and defining the safety constraints that each level should impose on the one below it. Specifically, STPA builds models of the organizational safety structure; the dynamics and pressures that can lead to deterioration of this structure; the models of the system that operators must have, and the necessary communication to ensure these models remain accurate over time; and the broader social and political context the organization exists within.

Finally, Dekker's Drift into Failure (DIF) model emphasizes decrementalism: the way that a system's processes can deteriorate through a series of minor changes, potentially causing the system's migration to an unsafe state. This model warns that each change may seem insignificant alone, so organizations might make these changes one at a time in isolation, creating a state of higher risk once enough changes have been made.

As a final note on the implications of complexity for AI safety, we considered the broader societal context within which AI technologies will function. We discussed how, in this uncontrolled environment, different, seemingly lower-level risks could interact to produce catastrophic threats, while chaotic circumstances may increase the likelihood of AI-related accidents. For these reasons, it makes sense to consider a wide range of different threats of different magnitudes in our approach to mitigating catastrophic risks, and we may find that broader interventions are more fruitful than narrowly targeted ones.

In the last section of this chapter, we focused in on a particular class of events called tail events and black swans, and explored what they mean for risk analysis and management. We began this discussion by defining tail events and considering several caricatures of long-tailed distributions. Then, we described black swans as a subset of tail events that are not only rare and high-impact but also particularly difficult to predict. These events seem to happen largely "out of the blue" for most people and may indicate that our understanding of a situation is inaccurate or incomplete. These events are also referred to as unknown unknowns, which we contrasted with known unknowns which we may not fully understand, but are at least aware of.

We examined how tail events and black swans can pose particular challenges for some traditional approaches to evaluating and managing risk. Certain methods of risk estimation and cost-benefit analysis rely on historical data and probabilities of different events. However, tail events and black swans are rare, so we may not have sufficient data to accurately estimate their likelihood, and even a small change in likelihood can lead to a big difference in expected outcome.

We also considered the delay fallacy, showing that waiting for more information before acting might mean waiting until it is too late. We discussed how an absence of evidence of a risk cannot necessarily be taken as evidence that the risk is absent. By looking at hypothetical situations where catastrophes are avoided thanks to safety measures, we explained how the preparedness paradox can make these measures seem unnecessary, when in fact they are essential.

Having explored the importance of taking tail events and black swans into consideration, we identified some circumstances that indicate we may be at risk of these events. We concluded that it is reasonable to believe AI technologies may pose such a risk, due to the complexity of AI systems and the systems surrounding them, the highly connected nature of the social systems they are likely to be embedded in, and the fact that they are relatively new, meaning we may not yet fully understand all the ways they might interact with their surroundings.

4.8.2 Key Takeaways

Tail events and black swans require a different approach to managing risks [277]. Some decisions require vastly more caution than others: for instance, paraphrasing Richard Danzig, you should not "need evidence" that a gun is loaded to avoid playing Russian roulette [42]. Instead, you should need evidence of safety. In situations where we are subject to the possibility of tail events and black swans, this evidence might be impossible to find.

One element of good decision making when dealing with long-tailed scenarios is to exercise more caution than we would otherwise. In the case of new technologies such as AI systems, this might mean not prematurely deploying them on a large scale. In some situations, we can be extremely wrong and things can still end up being fine; in others, we can be just slightly wrong but suffer disastrous consequences. We must also be cautious while trying to solve our problems. For example, while climate change poses a serious threat, many experts believe it would be unwise to attempt to fix it quickly by rushing into geoengineering solutions like spraying sulfur particles into the atmosphere. There may be an urgent need to solve the problem, but we should take care that we are not pursuing solutions that could cause many other problems.

Although tail events may be challenging to predict, there are a variety of techniques discussed in this chapter that can help with this, such as expanding our safety imagination, conducting horizon scanning exercises, and red-teaming.

Incorporating safe design principles can improve general safety. Following the safe design principles described in this chapter can be a good first step towards reducing systemic risks from AI, with the caveat that we should think carefully about which defense features are appropriate, and avoid too much complexity. In particular, focusing on increasing the controllability of the system might be a good idea. This can be done by adding loose coupling into the system, by supporting human operators to notice hazards and act on them early, and by devising negative feedback mechanisms that will down-regulate processes if control is lost.

Consider in detail the principle of least privilege. For one, it tells us that we should be cautious about giving AIs too much power, to limit the extent to which we are exposed to their tail risks. We might be concerned that AIs become enmeshed within society with the capacity to make large changes in the world when they do not need such access to perform their assigned duties. Additionally, for particularly powerful AI systems, it might be reasonable to keep them relatively isolated from wider society, and accessible only to verified individuals who have demonstrable and specific needs for such AIs. In general, being conservative about the rate at which we unleash technologies can reduce our exposure to black swans.

Targeting systemic factors is an important approach to reducing overall risk. As we discussed, tackling systemic safety issues can be more effective than focusing on details in complex systems. This can reduce the risk of both foreseeable accidents and black swans.

Raising general awareness of risks associated with technologies can produce social pressures, and bring organizations operating those technologies under greater scrutiny. Developing and enforcing industry regulations can help ensure organizations maintain appropriate safety standards, as can encouraging best practices that improve safety culture. If there are ways of reducing the safety costs (e.g. through technical research), this can make it more likely that an organization will adopt them, also improving general safety.

Other systemic factors to pay attention to include competitive pressures. These can undermine general safety by compelling management and employees to cut corners, whether to increase rates of production or to reach a goal before competitors. If there are ways of reducing these pressures and encouraging organizations to prioritize safety, this could substantially lessen overall risk.

Improving the incentives of decision-makers and reducing moral hazard can help to address systemic risks. We might want to influence the incentives of researchers developing AI. Researchers might currently be focused on increasing profits and reaching goals before competitors, pursuing scientific curiosity and a desire for rapid technological acceleration, or developing the best capabilities in deep learning models to find out what is possible. In this sense, these researchers might be somewhat disconnected from the risks they could be creating and the externalities they are imposing on the rest of society, creating a moral hazard. Encouraging more consideration of the possible risks, perhaps by making researchers liable for any consequences of the technologies they develop, could therefore improve general safety.

Similarly, we might be able to improve decision-making by changing who has a say in decisions, perhaps by including citizens in decision-making processes, not only officials and scientists [277]. This reduces moral hazard by including the stakeholders that have “skin in the game.” It can also lead to better decisions in general due to the wisdom of crowds, the phenomenon where crowds composed of diverse individuals make much better decisions collectively than most members within it, when the conditions are right.

In summary, while AI poses novel challenges, there is much that we can learn from existing approaches to safety engineering and risk management in order to reduce the risk of catastrophic outcomes.

4.9 LITERATURE

4.9.1 Recommended Reading

- N. Leveson. *Engineering a Safer World: Systems Thinking Applied to Safety*. Engineering systems. MIT Press, 2011. ISBN: 9780262016629. URL: https://books.google.com/books?id=0gZ˙7n5p8MQC

- C. Perrow. *Normal Accidents: Living with High Risk Technologies.* Princeton paperbacks. Princeton University Press, 1999
- Nassim Nicholas Taleb. *The Black Swan: The Impact of the Highly Improbable.* Vol. 2. Random House, 2007
- Nassim Nicholas Taleb. *Antifragile: Things That Gain from Disorder.* Incerto. Random House Publishing Group, 2012. ISBN: 9780679645276. URL: https://books.google.com.au/books?id=5fqbz˙qGi0AC