Introduction to AI Safety, Ethics, and Society

Safety Engineering

Introduction

Safety engineering is a broad discipline that studies many types of systems and provides a paradigm for avoiding loss events.

We can view AI safety as a special case of safety engineering concerned with avoiding AI-related catastrophes.

Roadmap

1. Risk decomposition and measuring reliability
2. Safe design principles
3. Component failure accident models
4. Systemic accident models
5. Tails events
6. Black swans

Risk Decomposition and Measuring Reliability

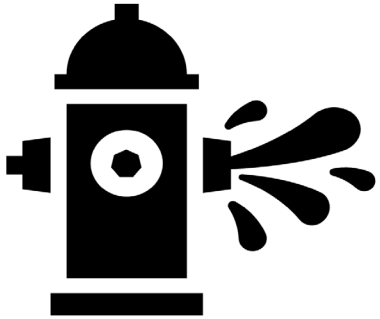
How to assess and compare risks quantitatively

Failure Modes, Hazards, and Threats

Failure mode : a specific way a system could fail.

Hazard : a source of danger that could cause harm.

Threat : a hazard with malicious or adversarial intent.



Failure mode

Hazard

Threat

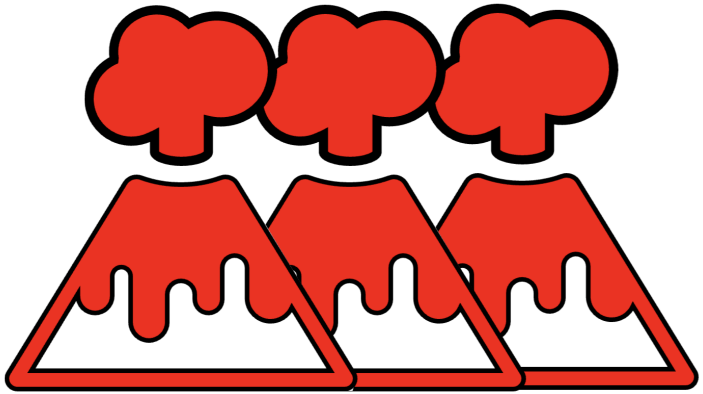
The Two-Factor Risk Equation

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

Example: Risk from volcano

We should frame our goal as risk *reduction*, not eradication.

- "Solving the intelligence problem" → "Improving capabilities"
- "Solving the alignment problem" → "Making AI safer"



Detailed Disaster Risk Equation

Let's expand this risk equation into four factors:

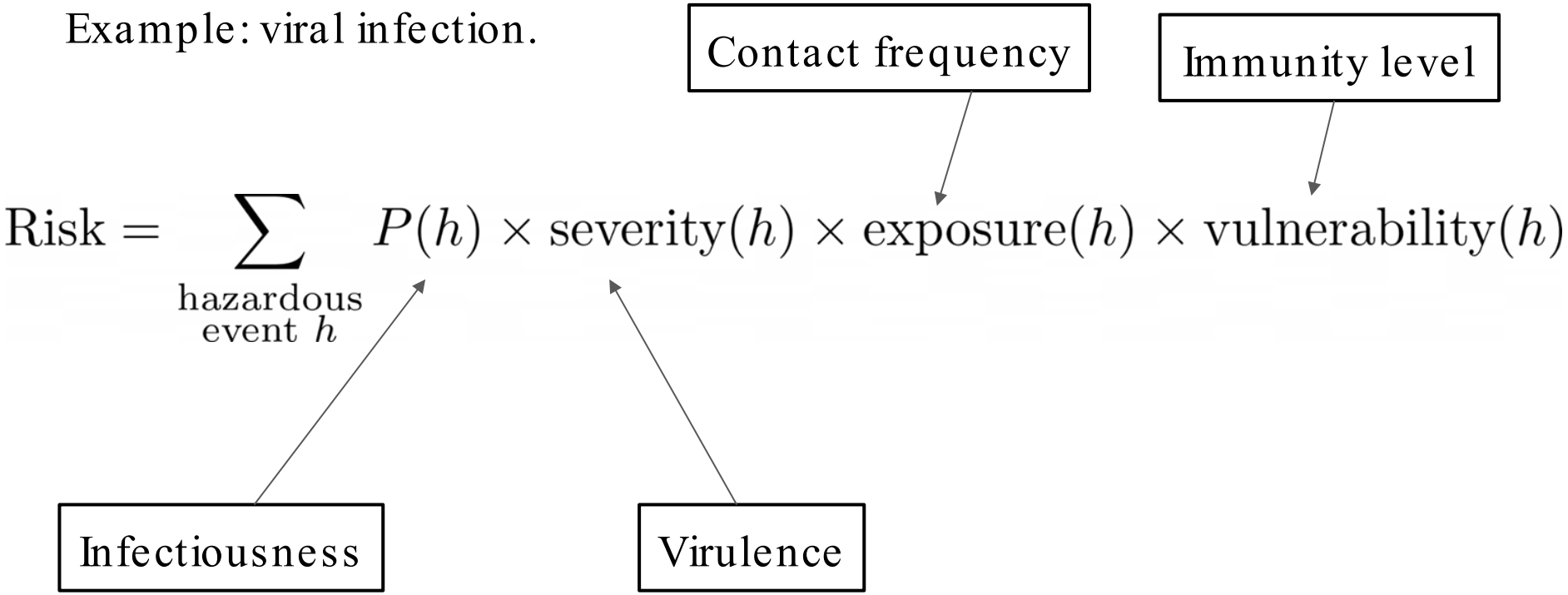
1. Probability
2. Severity

} Intrinsic hazard level

3. Exposure
4. Vulnerability

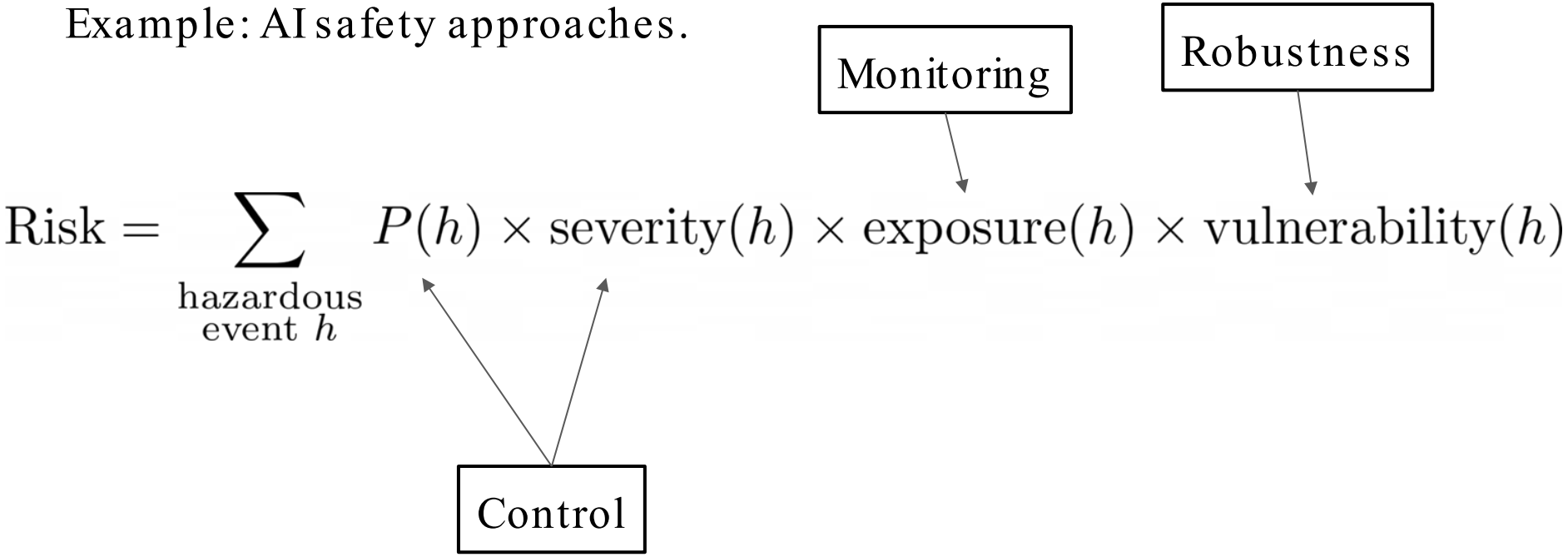
Detailed Disaster Risk Equation

Example: viral infection.



Detailed Disaster Risk Equation

Example: AI safety approaches.



Nines of Reliability

Reliability (of a system): the probability that an adverse event will not happen in a given amount of time.

A system's reliability also depends on how often it is used. The more often we use a system, the more likely we are to encounter failure (recall that increasing "exposure" increases the overall risk).

Example: An autonomous vehicle is more likely to make a mistake during a journey where it has to make 1000 decisions, than a journey with only 10.

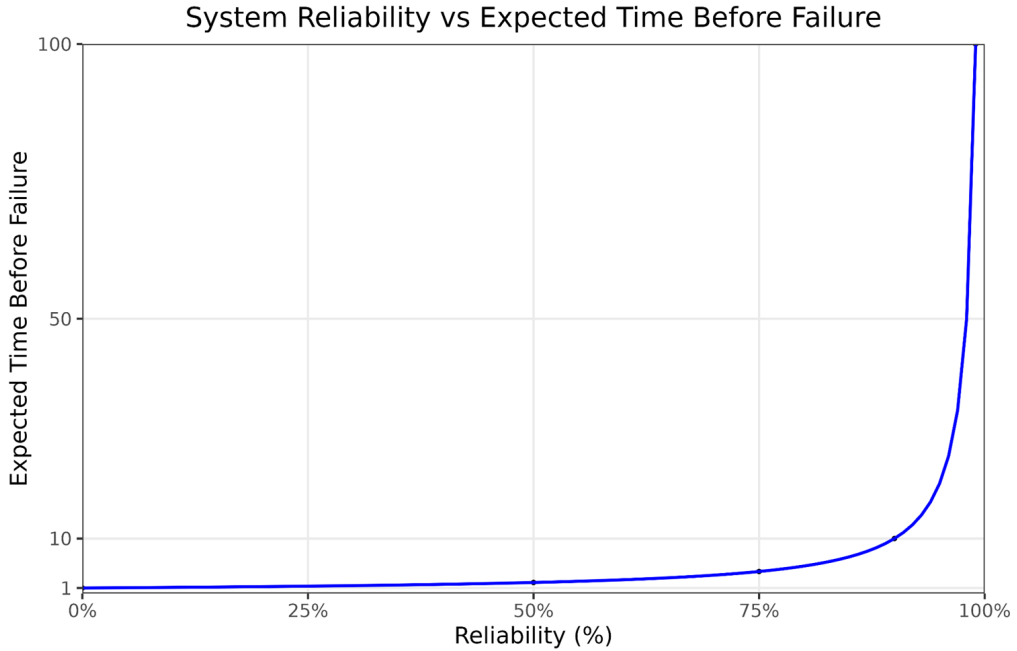
Nines of Reliability

For a given level of reliability, we can calculate an **expected time before failure** .

% reliability of AV	% risk of mistake	A mistake is likely to occur by decision number…
0	100	1
50	50	2
75	25	4
90	10	10
99	1	100
99.9	0.1	1000
99.99	0.01	10,000

Nines of Reliability

The relationship between system reliability and expected time before failure is **not linear** .



Nines of Reliability

Nines of reliability : the number of consecutive number nines at the start of a system's reliability (expressed as a decimal or percentage).

% reliability of AV	% 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

Nines of Reliability

We can denote a system's nines of reliability with the letter "**k**":

- Reliability is 90%
- Reliability is 99%

The system's reliability "**p**" relates to **k**:

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

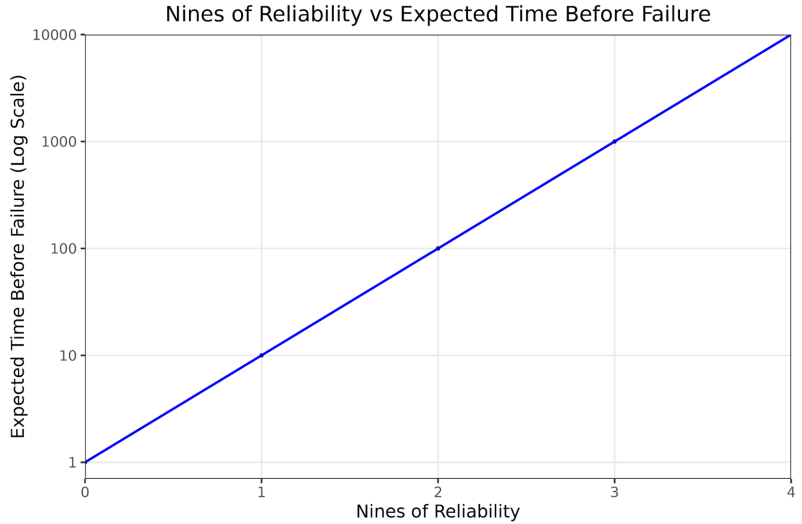
(Where **p** = the reliability of the system expressed as a decimal).

Possible remaining issues:
Adversarial examples, trojans, misunderstandings, etc.

Nines of Reliability

Adding a single "nine" to reliability gives a **tenfold** increase in expected time before failure.

Using the nines of reliability, we can see that improving reliability from 0.4 to 0.8 is far less meaningful than improving it from 0.99 to 0.999.



Recap

Risks can be mathematically characterized.

The classic risk equation is the expected severity of hazards.

Adding exposure and vulnerability, we create a “detailed risk equation.”

System reliability can be better measured in “nines of reliability.”

Roadmap

1. Risk decomposition and measuring reliability
2. Safe design principles
3. Component failure accident models
4. Systemic accident models
5. Tails events
6. Black swans

Safe Design Principles

1. Redundancy
2. Separation of duties
3. Principle of least privilege
4. Fail-safes
5. Antifragility
6. Negative feedback mechanisms
7. Transparency
8. Defense in depth

Principle 1:

Redundancy

Redundancy: a system has multiple "backup" components performing the same crucial function:

- Multiple braking components in vehicles
- Saving important data on multiple hard drives
- Seeking the opinions of multiple medical experts

Principle 1:

Redundancy

One possible use of redundancy in AI would be a “moral parliament.”

If we are uncertain how we want an AI to approach making decisions with moral implications, we can use a parliament approach to make use of multiple moral theories.

We can see these different moral theories as redundant components, reducing the probability of the AI making a harmful decision:

- Each theory would typically recommend plausible actions
- But none would be able to dictate what happens in extreme cases

Principle 2:

Separation of Duties

Separation of duties : Multiple agents work together to carry out complex processes. Each agent is specialized to a different task, so none controls the system by themself.

With multiple system operators, the capacity of each to harm is reduced.

Principle 2:

Separation of Duties

Example: a lab contains three chemicals which, if combined, explode:

- One technician in charge of all three poses a large risk
- With three technicians, each in charge of one chemical, the risk is lower

Principle 2:

Separation of Duties

We could focus on multiple narrow AIs, instead of a single general one.

Just as with the lab, having multiple agents, each specialized to a different task, reduces the risk posed by each.

Example: "Comprehensive AI Services" is an approach that views AI agents as service-providing products, each tailored to a highly-specific task.

Principle 3:

Principle of Least Privilege

Principle of least privilege : Each agent should have only the minimum power needed to complete their tasks.

Separating duties might only work if we also ensure that agents cannot access parts of the system irrelevant to their tasks. In our lab example, having three technicians only reduces the risk if each can only access their chemical.

Principle 3:

Principle of Least Privilege

Similarly, we should ensure that each AI agent only has access to the information and power necessary for it to perform its tasks reliably.

Example: We should avoid unnecessarily plugging AIs into the internet or giving them high-level admin access to confidential information.

Principle 4:

Fail-Safes

Fail-safes: features that ensure a system will be safe even if it fails.

Example: elevator breaks.

A “confidence estimate” might work as an AI fail-safe:

- Tracks level of confidence the AI agent has in its decisions
- Prevents enacting decision below specific confidence threshold

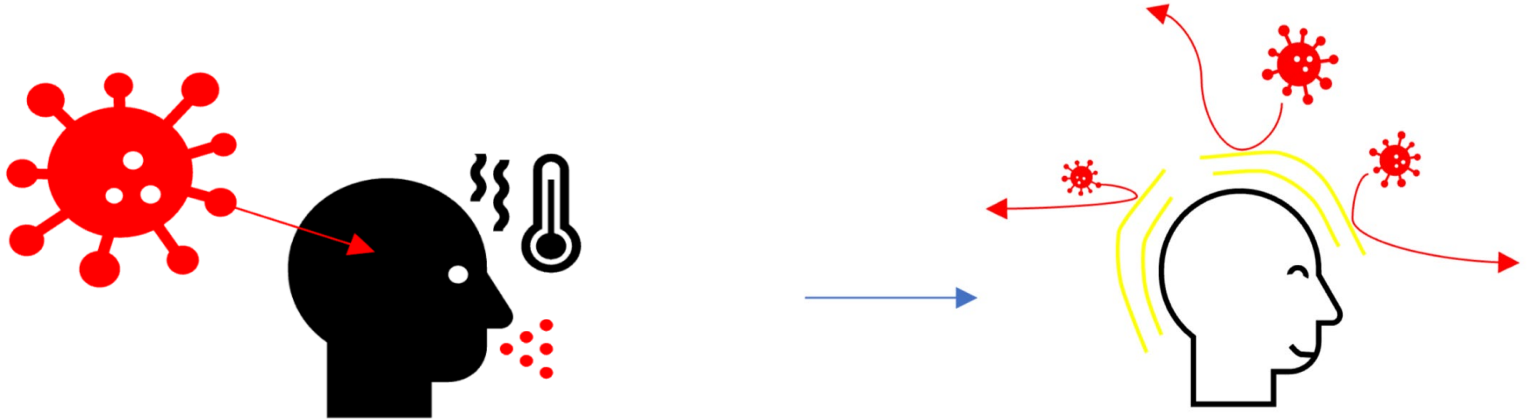
Principle 5:

Antifragility

Antifragility : the capacity to become stronger from encountering adversity.

Example: lifting weights strains muscles, so muscles grow stronger.

Example: infection by a pathogen results in immune protection.



Principle 6:

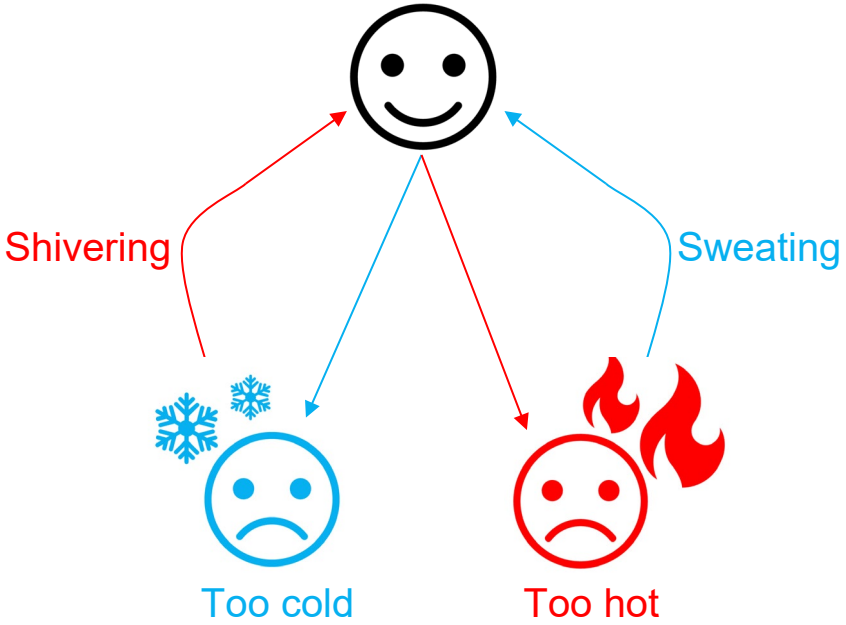
Negative Feedback Mechanisms

Negative feedback : processes that down-regulate initial change.

Example: body temperature.

Negative feedback in AI:

1. Monitoring
2. Reputation reduction
3. Resetting



Principle 7:

Transparency

Transparency: users understand a system sufficiently well to interact with it safely.

Example: pilots must understand their plane's autopilot systems:

- How to activate
- How to override

AI model transparency could help operators:

- Steer away from hazards
- Anticipate deceptive behavior
- Retain control

Principle 8:

Defense in Depth

Defense in depth : having many layers makes for more reliable defense.

Example: we can reduce our vulnerability to viruses most by taking multiple measures:

- Wearing a mask
- Social distancing
- Hand washing

Note that this introduces the risk of defense layers interacting in unpredicted/undesired ways.

Recap: Eight Safe Design Principles

1. Redundancy
2. Separation of duties
3. Principle of least privilege
4. Fail-safes
5. Antifragility
6. Negative feedback loops
7. Transparency
8. Defense in depth

Roadmap

1. Risk decomposition and measuring reliability
2. Safe design principles
3. Component failure accident models
4. Systemic accident models
5. Tails events
6. Black swans

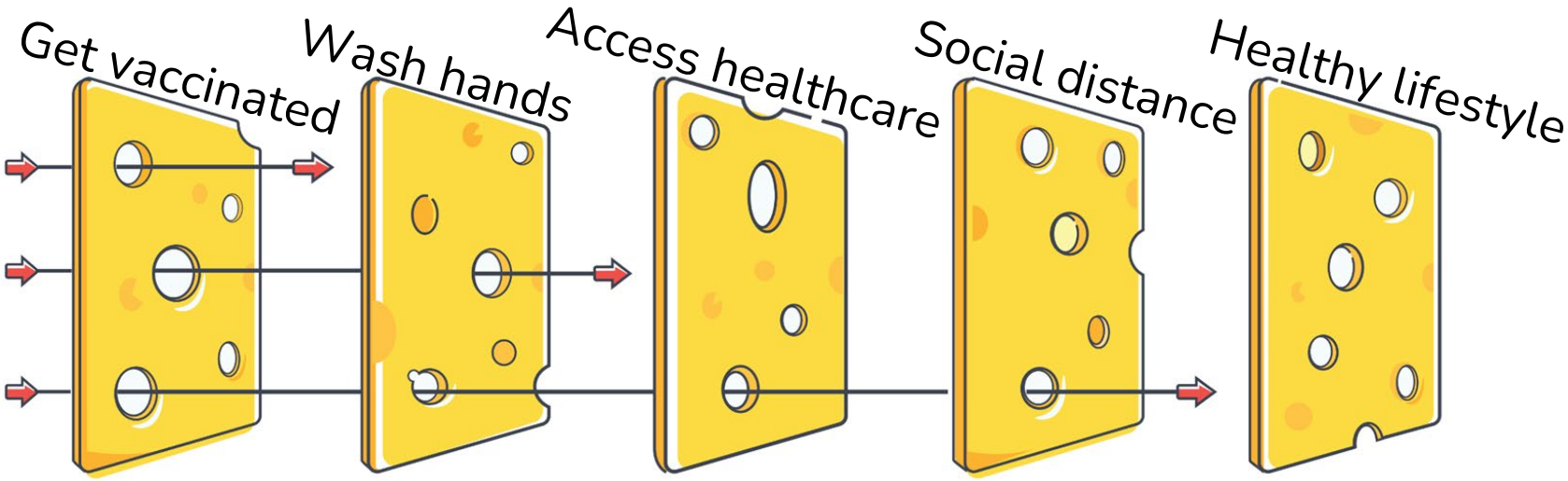
Component Failure Accident Models

Traditional risk analysis concepts

Swiss Cheese Model

Identify and analyze pathways to accidents.

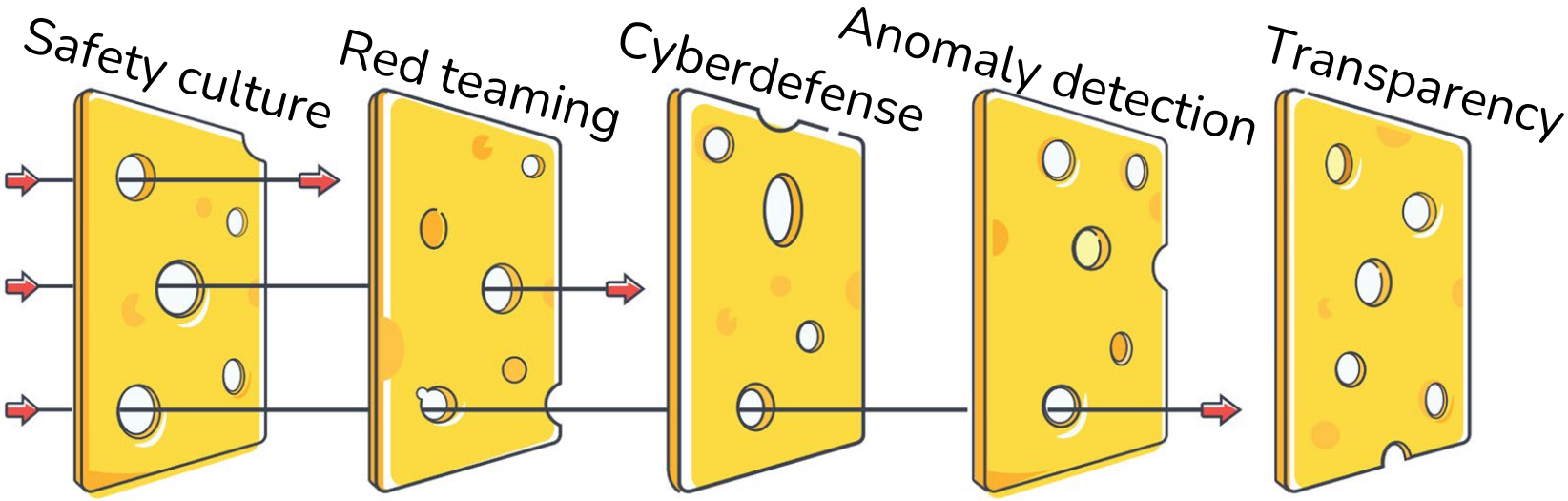
Example: infectious disease



Swiss Cheese Model

Identify and analyze pathways to accidents.

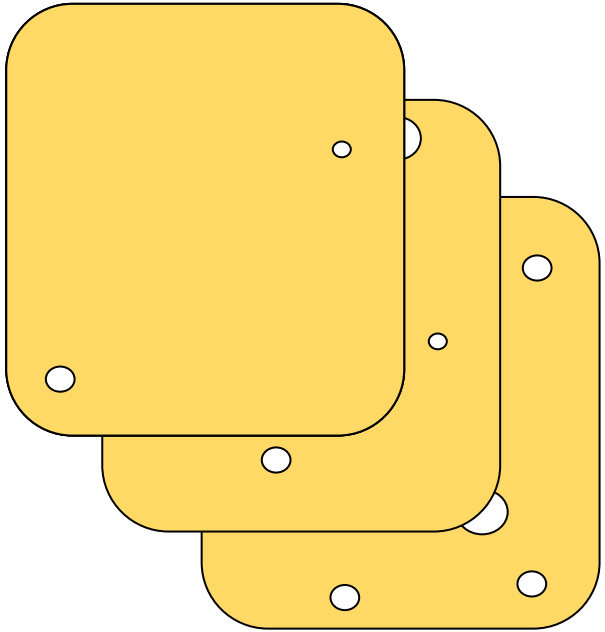
Example: AI safety



Swiss Cheese Model

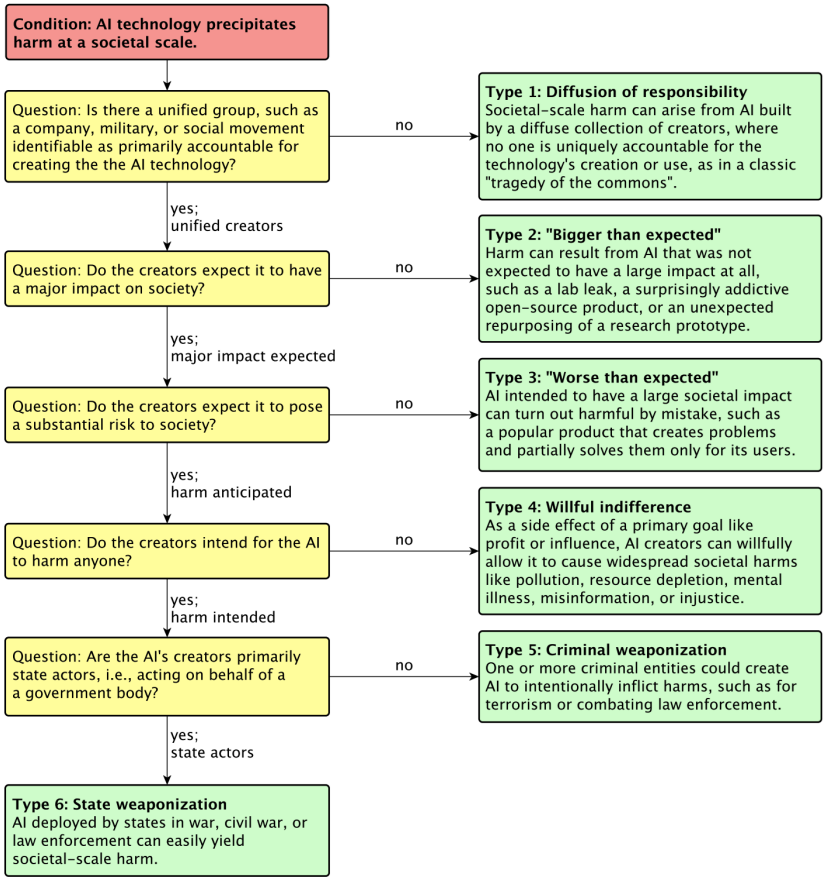
Safety can be improved by:

- Reducing the size of holes
- Reducing number of holes
- Increasing number of slices



Fault Tree Analysis

Start by considering specific negative outcomes, then work backward to identify their possible causes.



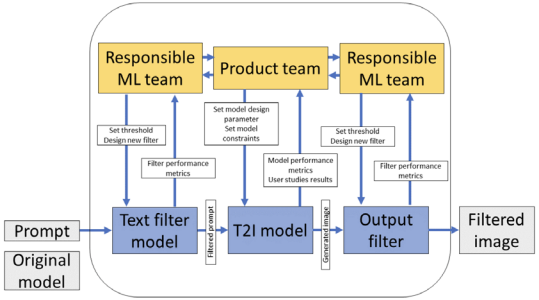
Limitations

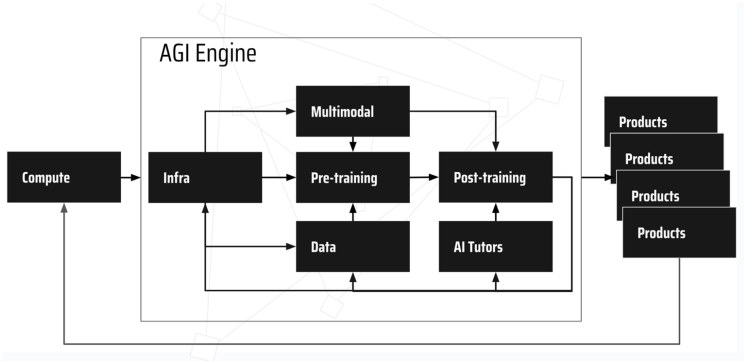
Chain-of events reasoning can be too simplistic if causality is nonlinear.

Complex system components may interact to produce emergent properties.

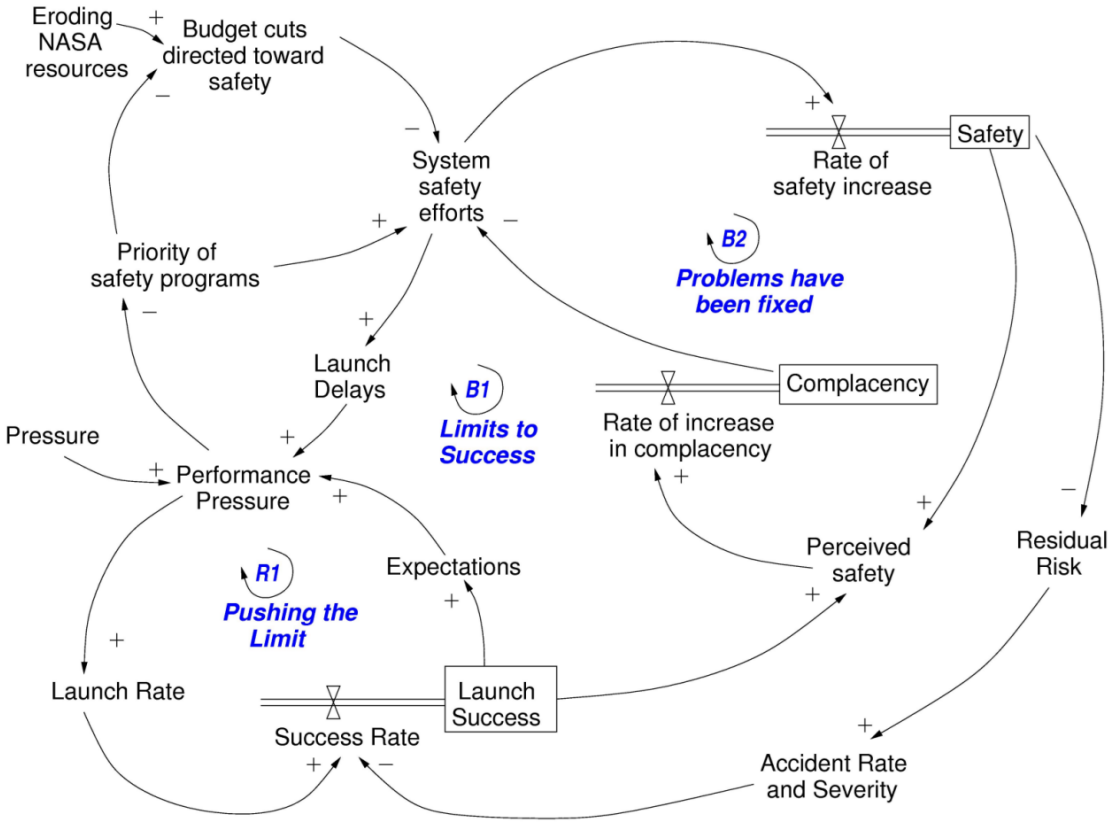
Accidents can happen even if no single component fails, due to interactions.

The causality may be diffuse





Limitations



Systemic Accident Models

Human factors and systems-based risk analysis

Understanding Adverse Events: Human Factors

- Human rather than technical failures now represent the greatest threat to complex and potentially hazardous systems.

- Managing the human risks will never be 100% effective. Human fallibility can be moderated, but it cannot be eliminated.

- Significant errors occur at all levels of the system, not just at the sharp end. Decisions made in the upper echelons of the organisation create the conditions in the workplace that subsequently promote individual errors and violations. Latent failures are present long before an accident and are hence prime candidates for principled risk management.

Understanding Adverse Events: Human Factors

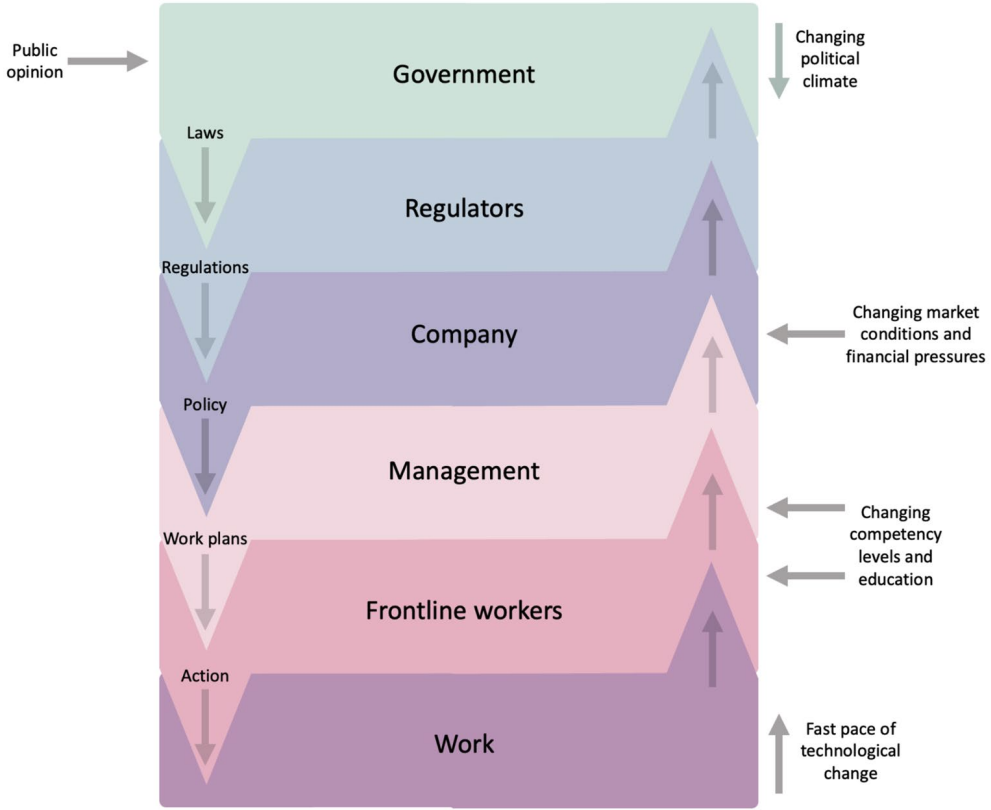
- Measures that involve sanctions and exhortations (that is, moralistic measures directed to those at the sharp end) have only very limited effectiveness, especially so in the case of highly trained professionals.

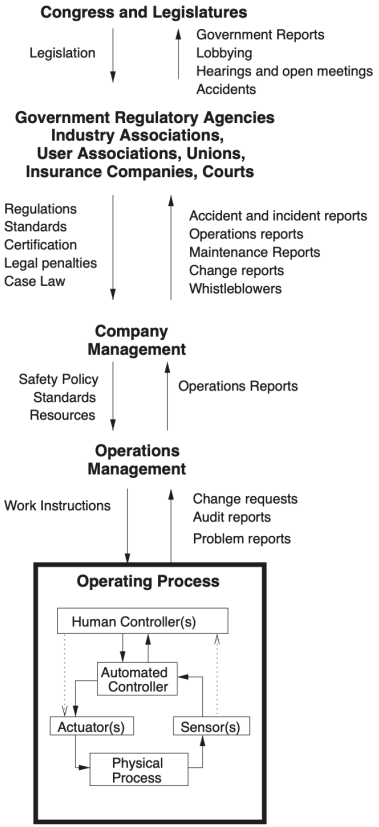
- People do not act in isolation. Their behaviour is shaped by circumstances. The same is true for errors and violations. The likelihood of an unsafe act being committed is heavily influenced by the nature of the task and by the local workplace conditions. These, in turn, are the product of "upstream" organisational factors.

Understanding Adverse Events: Human Factors

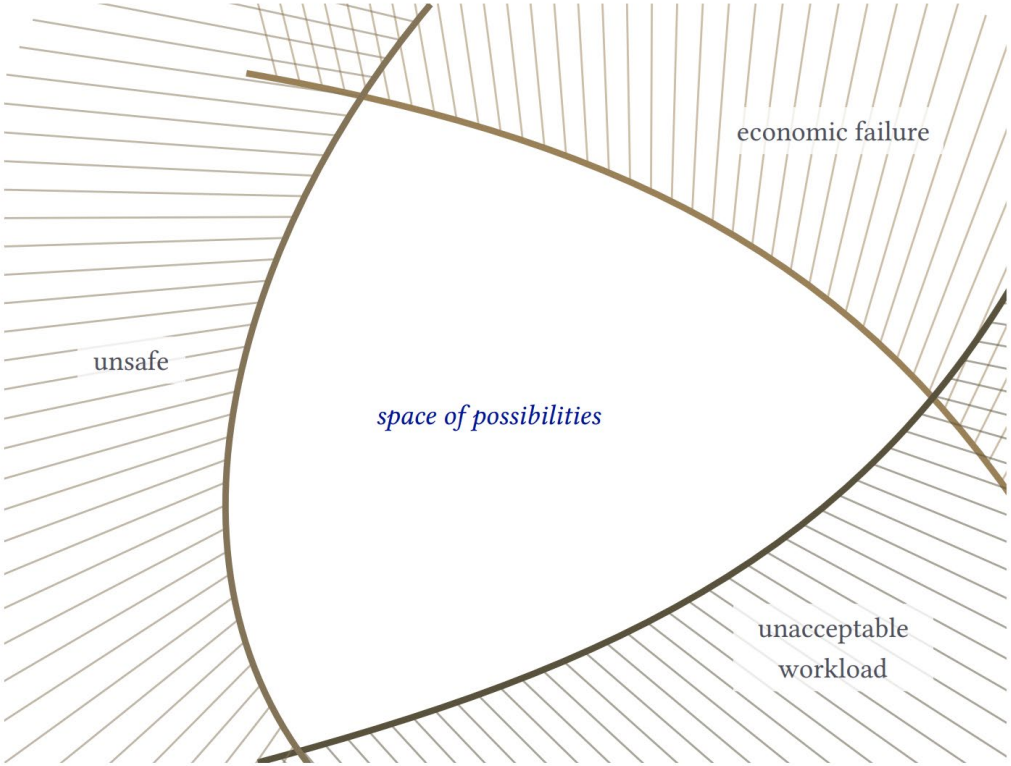
- Automation and increasing advanced equipment do not cure human factors problems, they merely relocate them.
- Effective risk management depends critically on a confidential and preferable anonymous incident monitoring system that records the individual, task, situational, and organisational factors associated with incidents and near misses.
- Effective risk management means the simultaneous and targeted deployment of limited remedial resources at different levels of the system: the individual or team, the task, the situation, and the organisation as a whole.

Sociotechnical Systems



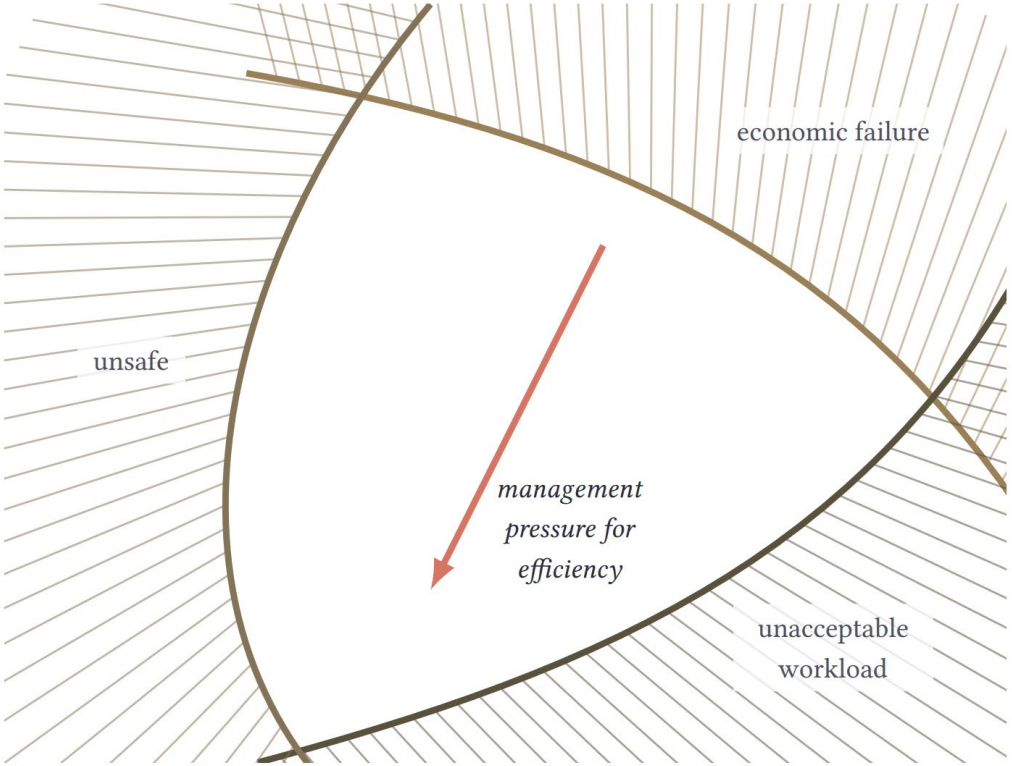


Drift into Failure



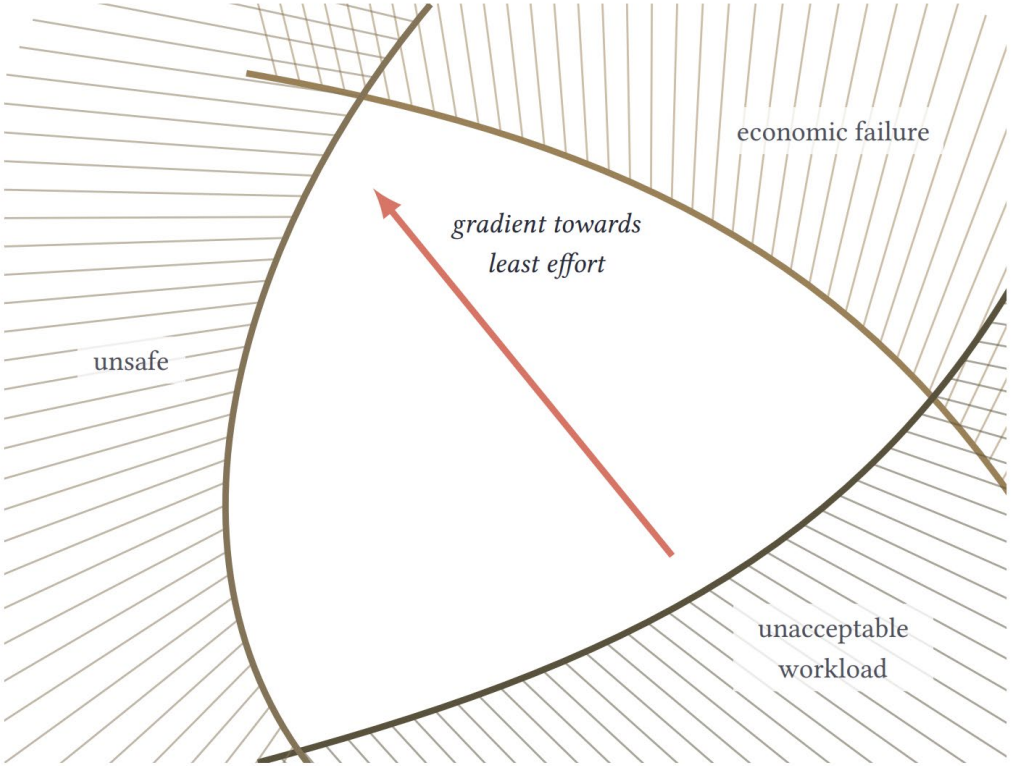
Slide adapted from Eric Marsden's Safety models presentation

Drift into Failure



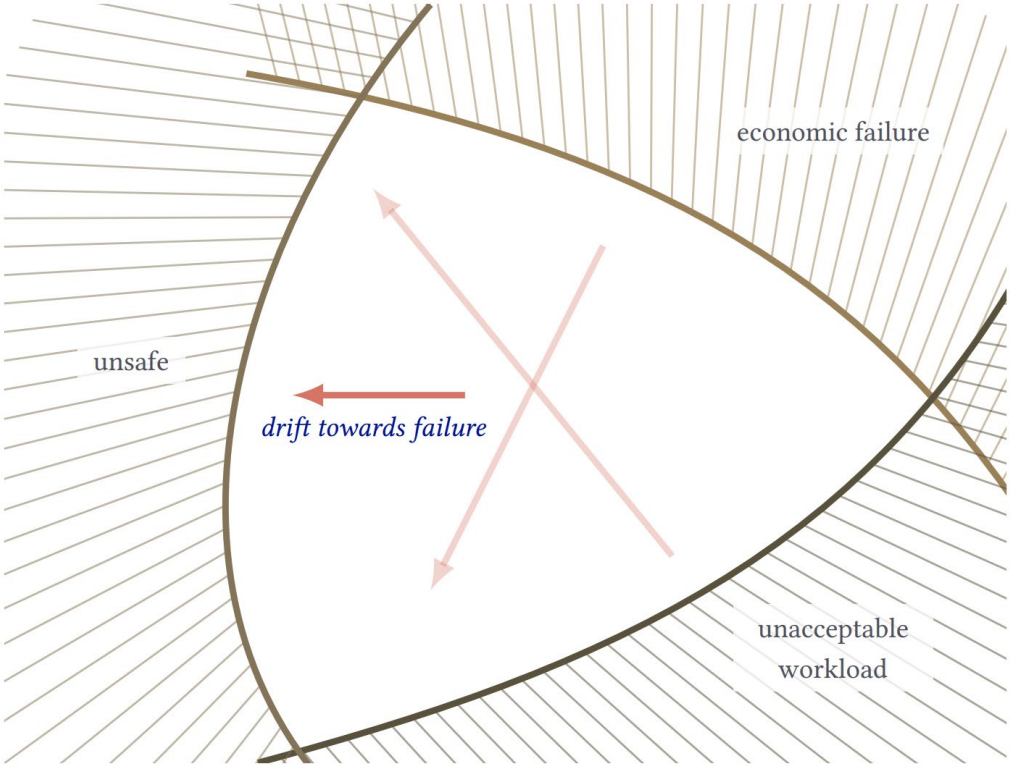
Slide adapted from Eric Marsden's Safety models presentation

Drift into Failure



Slide adapted from Eric Marsden's Safety models presentation

Drift into Failure



Slide adapted from Eric Marsden's Safety models presentation

System-Theoretic Accident Model and Processes (STAMP)

STAMP perspective : Safety is an emergent property of a system, not the product of all components behaving reliably. Reliable systems are not necessarily safe.

A system contains multiple levels of organization:

- Each has novel emergent properties
- We cannot understand the emergent properties of one level through a reductive analysis of the level below

System-Theoretic Accident Model and Processes (STAMP)

Safety can be enforced by each level placing constraints on the level below.

STAMP-based risk analysis involves modeling four system aspects:

1. Organizational safety structure
2. Dynamics and pressures that can deteriorate safety structure
3. Necessary operator knowledge and communication about processes
4. Cultural and political context of decision-making processes

System-Theoretic Accident Model and Processes (STAMP)

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 the environment.
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.

Recap

AI safety can be viewed as a safety engineering problem.

Component failure accident models provide insight into how to prevent individual errors and their knock-on effects.

- Swiss Cheese Model, Fault Tree Analysis

Systemic accident models take a wider view, considering how to make the entire system safer rather than focusing on individual components.

Roadmap

1. Risk decomposition and measuring reliability
2. Safe design principles
3. Component failure accident models
4. Systemic accident models
5. Tails events
6. Black swans

Tails Events

Defining Our Terms

Tail event : a low-probability but high-impact occurrence.

Tail risk : the possibility of a tail event with negative impact.

Black swan : tail events that are unanticipated, because they are "unknown unknowns" (discussed later).

Introduction to Tail Events

Scott Sagan:
“Things that have never happened before happen all the time.”

Tail event examples:

- 2007 financial crisis
- COVID pandemic
- ChatGPT

Mediocristan vs Extremistan

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.

Black Swans

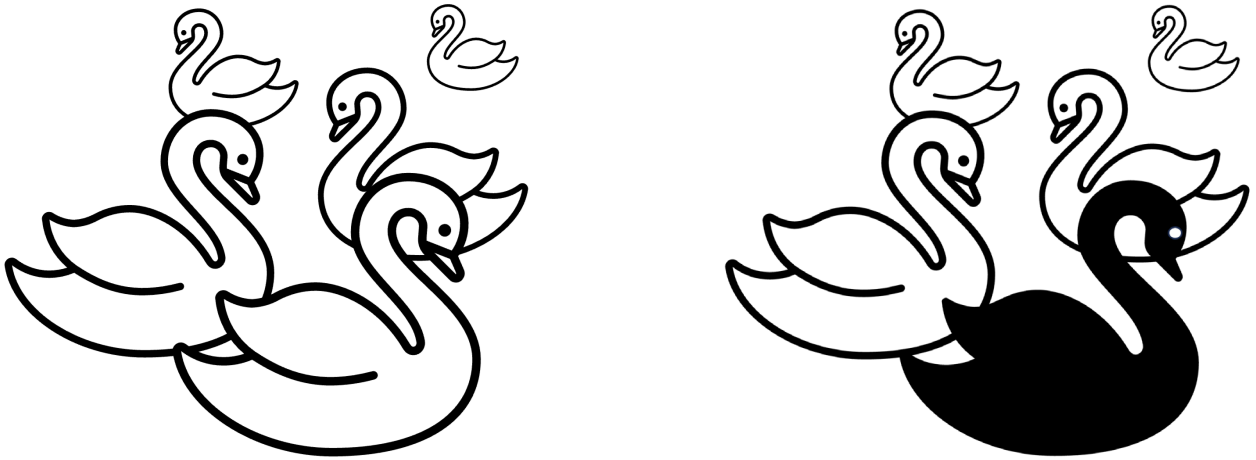
Unanticipated tail events

Introduction to Black Swans

Origin story:

- Commonplace belief: “All swans are white”
- Observation: a black swan

Black swan : a tail event that is *largely* unanticipated by *most* people.



“Known vs Unknown” Unknowns

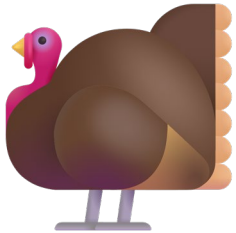
Known knowns : things we are aware of and understand. Facts and requirements.	**Unknown knowns** : things we do not realize we know. Tacit knowledge.
Known unknowns : things we are aware of but which we don’t fully understand. Close-ended questions.	**Unknown unknowns** : things we do not understand, and are not aware we do not know. Open-ended exploration.

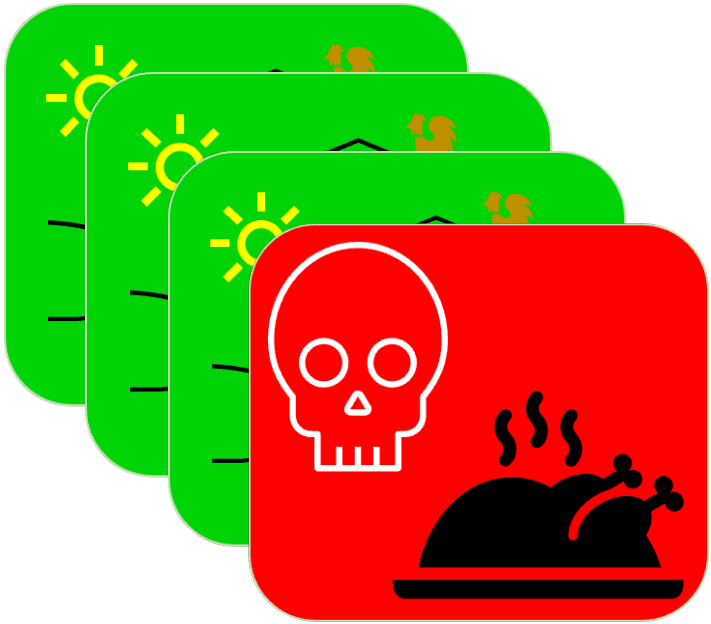
Turkey Problem

Example: After a comfortable life at the farm, a turkey is suddenly killed and eaten. From its perspective, this might be a black swan.

Black swans in AI:

- AI wellbeing
- Autonomous economy
- AI companions





Black Swans:
Implications for Risk Analysis

How to defend against risks we can't anticipate

Implication 1:

Typical Risk Estimation Breaks Down

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

Implication 2:

Cost-Benefit Analysis Breaks Down

Explicit cost-benefit analysis is strained under long tails, so often requires highly accurate estimates.

Example: lottery

- Lottery 1: (99.9% x $15) + (0.1% x -$10,000) = **$4.985**
- Lottery 2: (99.7% x $15) + (0.3% x -$10000) = **-$15.045**

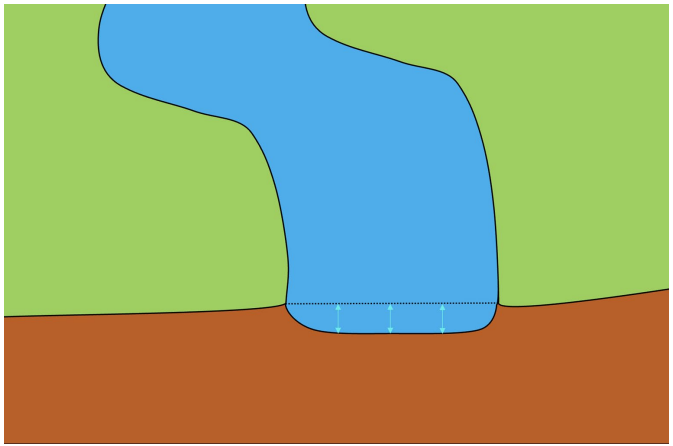
For black swans, estimating probability is even more challenging.

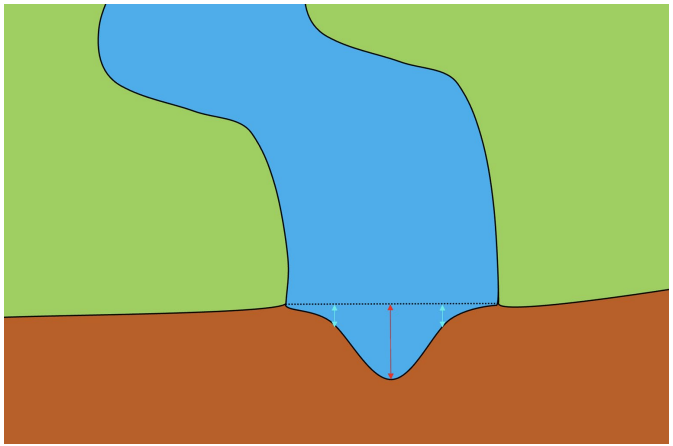
Implication 3:

We Must Consider Extremes

Milton Friedman: “Never try to walk across a river just because it has an average depth of four feet.”

When making risk-related decisions, we should consider extremes, not only the average.





Implication 4:

The Delay Fallacy

"We should delay making our decision to gather more information."

This works for thin-tailed scenarios, but not for long-tailed ones.

In long-tailed scenarios, waiting for more information can mean waiting until it is too late.

With black swans, waiting may yield no useful information at all. Large-scale black swans with catastrophic effects must not be allowed to happen even once

Implication 5:

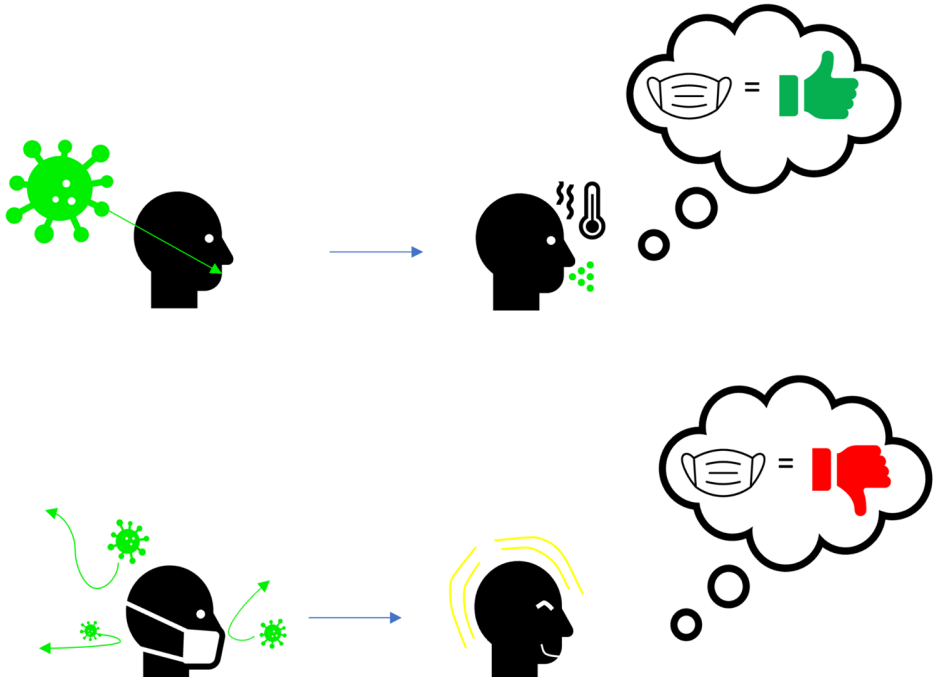
Absence of Evidence

An absence of evidence is not strong evidence of absence.

Implication 6:

The Preparedness Paradox

Safety measures that successfully prevent harm can seem redundant.



Reducing Tail Risks

1. Find potential black swans: turn them into *known* unknowns
2. Incorporate safe design principles
3. Improve systemic factors, such as risk awareness and safety culture
4. Reduce our exposure to risk: follow the precautionary principle
5. Improve policymakers' decisions about AI
6. Change researchers' incentives: internalize externalities

Section Recap

Tail events are low-probability but high-impact occurrences.

Tail events render common statistical measures (e.g. expected values) inadequate.

Black swans are unanticipated tail events (often “unknown unknowns”).

Dealing with black swans requires approaches such as horizon scanning.

Chapter Recap

1. How to mathematically discuss risks.
2. The "nines of reliability."
3. The Safe Design Principles for improving system safety.
4. Why traditional risk analysis sometimes fails in complex systems.
5. More holistic alternatives, such as systemic accident models
6. Tail events are rare, but important. Many AI risks are tail events.
7. Black swans high-impact tail events.
8. Tail events and black swans can pose particular challenges for some traditional approaches to evaluating and managing risk.
9. AI technologies may pose such a risk.