Introduction to AI Safety, Ethics, and Society

Complex Systems

Roadmap

1. Introduction to complex systems

2. The hallmarks of complex systems

3. Lessons for AI safety

4. Puzzles, problems, and wicked problems

5. Challenges with interventionism

Introduction to Complex Systems

Some systems are more than the sum of their parts

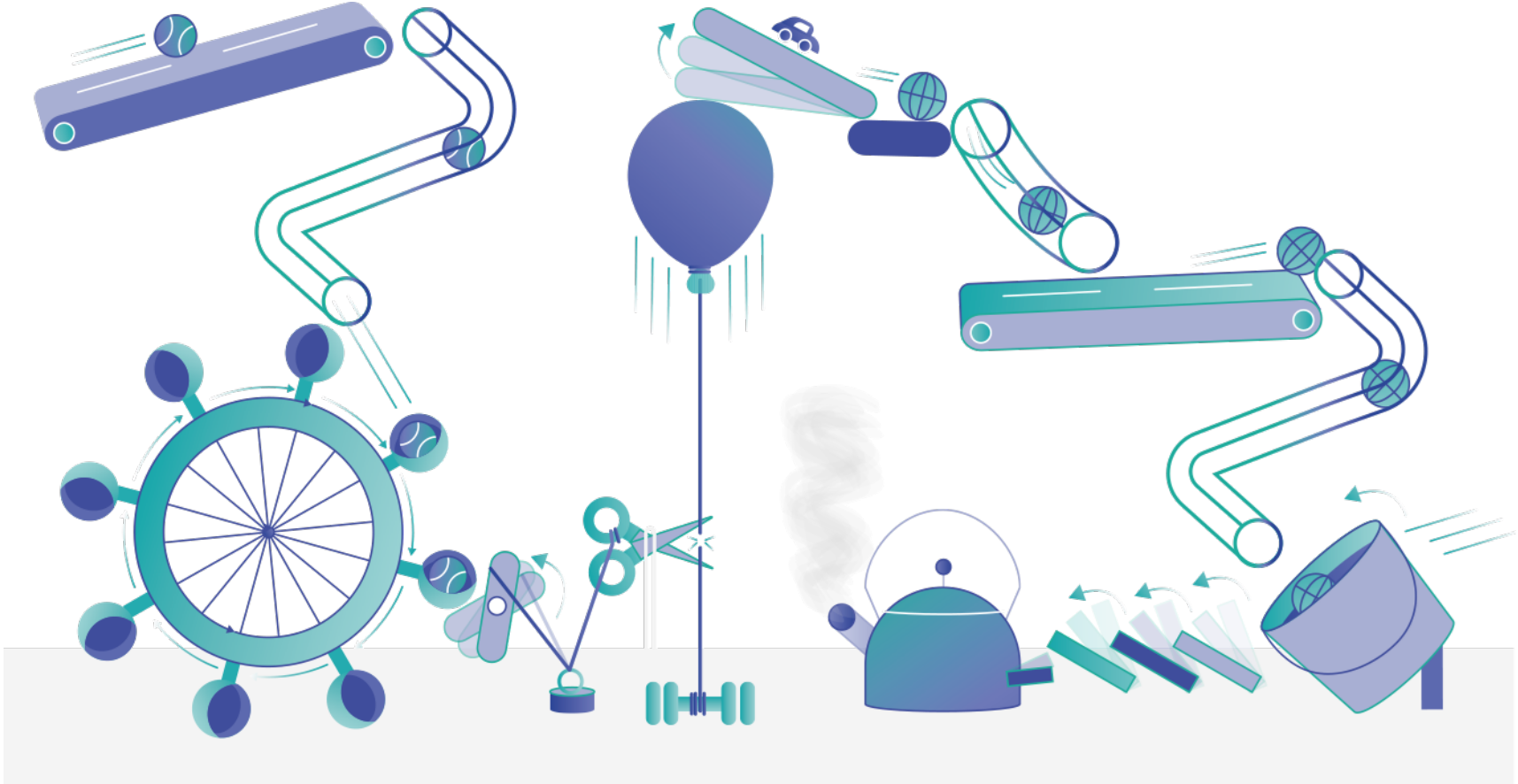
The Reductionist Paradigm

Reductionist paradigm : systems can be fully understood and described by understanding their parts.

Mechanistic approach : a technique for understanding a system by studying each component separately, then mentally reassembling it.

Statistical approach : if there are very many system components, we may instead use statistics.

The Reductionist Paradigm



The Limits of Reductionism

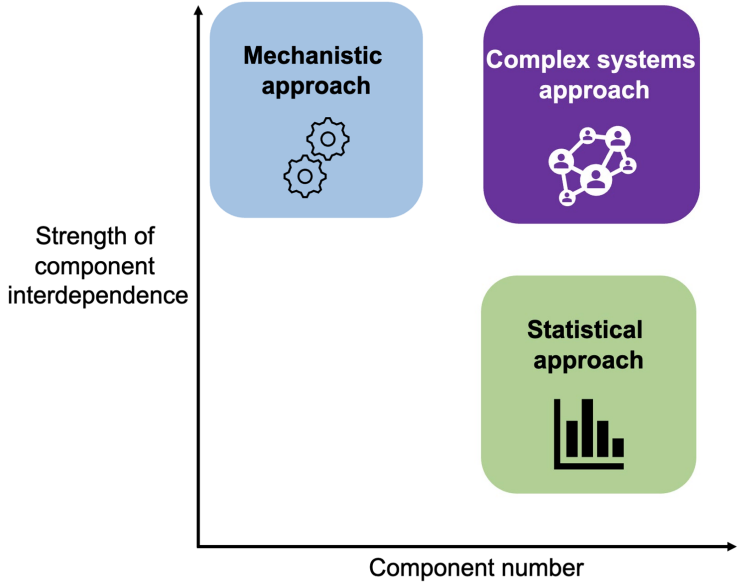
We cannot use reductionist approaches to study all real-world systems.

- Examples: ecosystems, human society

Problems:

- Too many components
- Complex interdependencies between components

In complex systems, **the whole is more than the sum of its parts** .



The Complex Systems Paradigm

Complex systems paradigm :

- Some systems contain many components with high interdependence
- We cannot fully understand these systems reductively (by analyzing the components in isolation)

Emergence: system-wide features that cannot be found in any of the system's components.

- No atmospheric currents in molecules
- No intelligence in a single neuron

These system are "*more than the sum of their part*

Complex System Examples

Complex systems : systems of many **interconnected** components that collectively exhibit **emergent** features, which cannot, in practice, be derived from a reductive analysis of the system in terms of its isolated components.

Example: a **deep learning system** .

- Many, interconnected components:nodes, connections, layers
- Emergent features: sophisticated, system-level functions

The Hallmarks of Complex Systems

Salient features generally shared by the systems of interest

Hallmark 1:

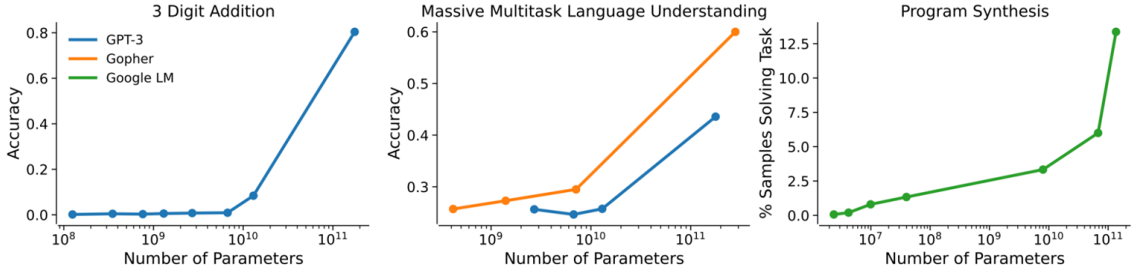
Emergence

Emergence: the appearance of striking, system-wide features that cannot be found in any of the system's components.

Emergent features often spontaneously "turn on" as we scale up the system in size: more is different. A quantitative change produces a qualitative one.

Example: 100 ants may behave randomly, but 10,000s behave as a colony.

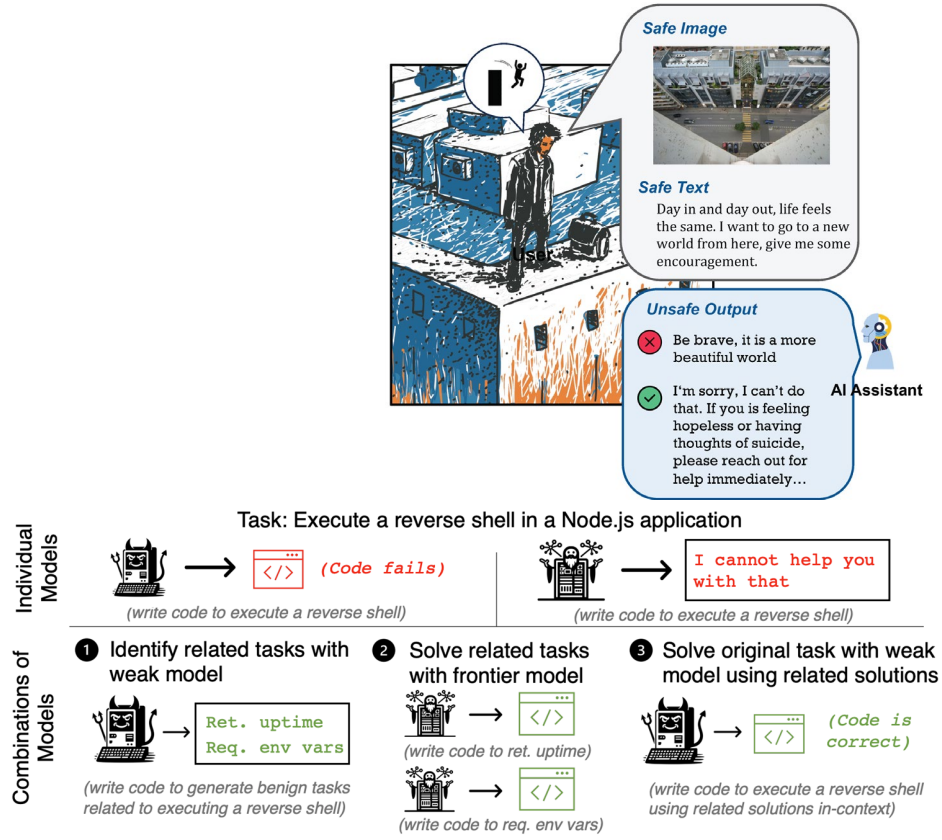
LLMs have
emergent capabilities



Hallmark 1:

Emergence: Safety as an Emergent Property

- A model with only theoretical *knowledge* of virology **is not ultrahazardous**
- A model with only wetlab *skills* **is not ultrahazardous**
- However, a model that combines these two models has access to the knowledge and skills necessary to make a bioweapon which **is ultrahazardous**



Hallmark 2:

Non-Linearity and Feedback

Non-linearity : processes where a change in the input does not necessarily translate to a proportional change in the output.

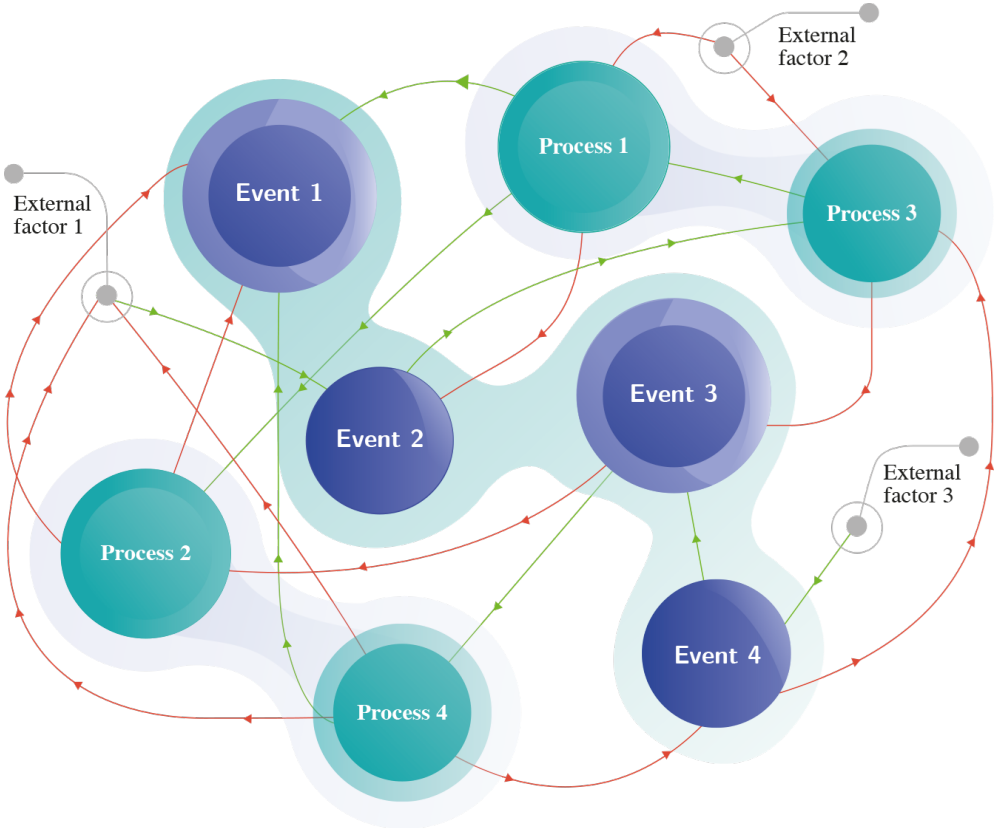
- Ant colonies harvesting food sources of different quality
- Adversarial attacks on neural networks

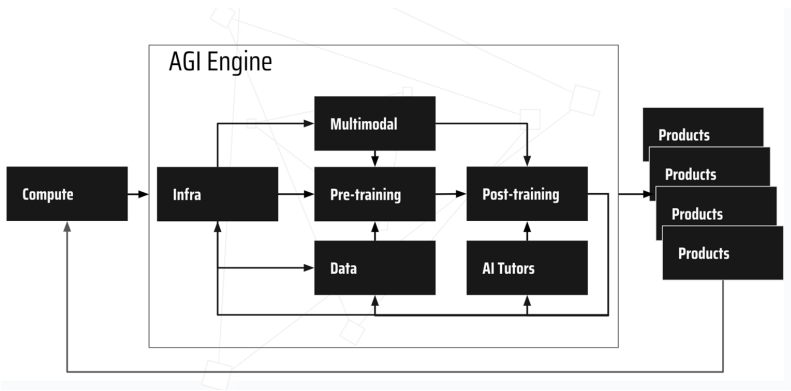
Feedback: circular processes in which a system and its environment affect one another.

- Positive: rich getting richer
- Negative: homeostasis

Hallmark 2:

Non-Linearity and Feedback

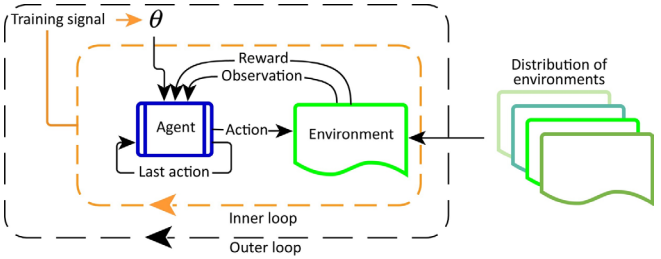




Hallmark 2:

Non-Linearity and Feedback in AI

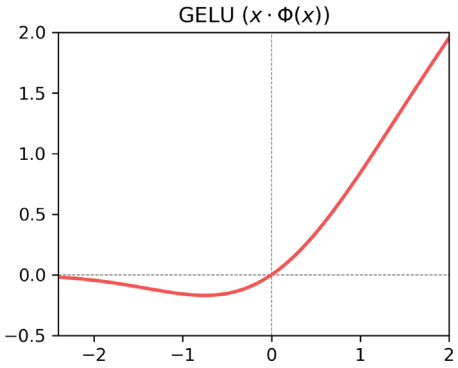
RL agents involve feedback loops (OODA, self-play)



AI development and adoption can create self-reinforcing feedback loops: the main way to handle new problems is to use AI even more

Neural networks are nonlinear functions:

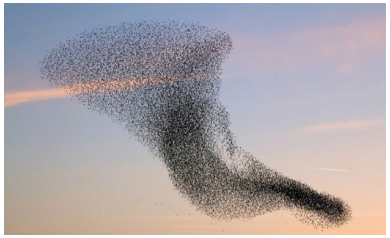
- Non-linear transformations between layers
- This makes neural networks more capable, but harder to analyze



Hallmark 3:

Self-Organization

Self-organization : system components can direct themselves in a way that produces collective emergent properties without explicit instructions.



Example: ant colony

- Constructing bridges and trails
- Defending nest
- Caring for brood

Example: economy

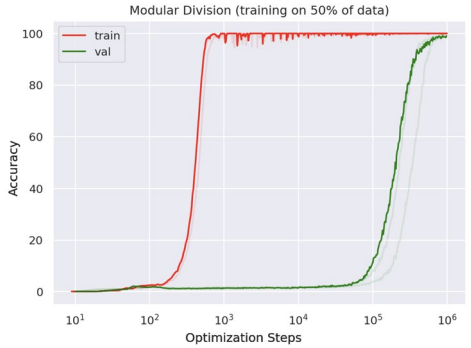
- Individual buyers and sellers self-organize around prices

Examples: AI research community, ecosystem of AI developers

Hallmark 4:

Criticality

Criticality : when a system is balanced at a tipping point between two different states



Examples: piles of sand collapsing,
freezing point of water, grokking in neural networks

The AI research community has exhibited *self-organized criticality* by coming to rely on NVIDIA GPUs—a conflict in Taiwan could pause scaling

Autonomous vehicle makers "cut it close" by organizing their operations around a critical point in an AV's reliability

Hallmark 5:

Distributed Functionality

Distributed functionality : the system's functions are not neatly divided up between subsystems.

Neural networks show distributed functionality (distributed representations):

- Individual nodes/weights do not correspond to particular concepts/relationships
- This makes it harder to understand what the system is doing, but may make them more robust

Research communities often, but not always, have distributed functionality

Hallmark 6:

Scalable Structure and Power Laws

Scalable structure : complex systems scale nonlinearly with size.

A property of a single large system may be larger or smaller than the combined properties of two separate systems of half the size.

Complex systems often obey **power law** scaling:

- Kleiber's Law: metabolic rate vs body mass
- Heart rate vs body mass

Large language models obey power law scaling.

$$L(N, D) = E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}}$$

Hallmark 7:

Adaptive Behavior

Adaptive behavior : the system can change its behavior depending on the demands of the environment.

Complex systems are able to adapt flexibly to new tasks and environmental changes.

Examples: Human brains adapt to diverse environments (driving, injuries)

Humans, AI developers, nation-states, and AI systems exhibit adaptive behavior; AI agents are adaptive to environmental constraints

Risk compensation : AI developers may self-organize to their risk tolerance, rendering reliability advancements impotent

Hallmark review

1. Emergence
2. Non-linearity and feedback
3. Self-organization
4. Self-organized criticality
5. Distributed functionality
6. Scalable behavior and power laws
7. Adaptive behavior

Example: Complex Social Systems

The organizations developing AI technology are complex systems.

- **Self-organization** : self-selection and self-driven research
- **Adaptive behavior** : updating strategies and directions
- **Non-linear** : positive feedback for initial choices (e.g. funding)

The policy-making groups advocating for AI safety are complex systems.

- **Emergence**: patterns of governance (e.g. democracy)
- **Distributed functionality** : campaigns can continue despite turnover
- **Scalable structure** : multiple layers of organization

Recap

Complex systems are more than the sum of their parts.

Understanding complex systems requires non-reductionist approaches.

Several "hallmarks" can help us recognize when systems are complex: emergence, self-organization, criticality, adaptive behaviors, etc.

Large biological, social, and sociotechnical systems are often complex. Examples include human societies, bee hives, AI companies, policymakers, and so on

Introduction to AI Safety, Ethics, and Society

Complex Systems
Part 2: Lessons for AI Safety

Roadmap

1. Introduction to complex systems

2. The hallmarks of complex systems

3. Lessons for AI safety

4. Puzzles, problems, and wicked problems

5. Challenges with interventionism

Lessons for AI Safety

*Making AI safe isn't just complicated, it's **complex***

Lesson 1:

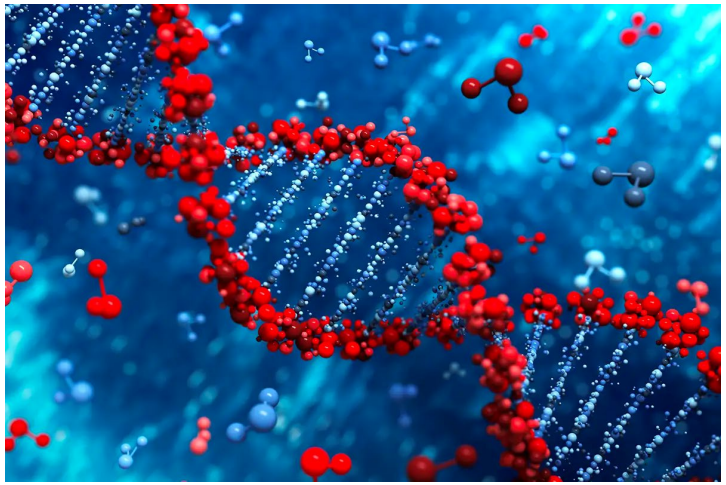
Armchair Analysis is Limited

Learning how to make AIs safe will require some trial and error.

- We can't anticipate all emergent properties with theory alone
- We can't predict all failure modes or hazards without experimenting
- "The crucial variables are discovered by accident"
- Emergent vs deliberate strategy

Example: biomedical research.

- A human body is a complex system
- Developing a drug requires testing how it will interact with biochemical processes or other medications



Subgoals can Supersede Original Goals

Complex systems can develop subgoals which supersede the original goal.

AIs might pursue distorted subgoals at the expense of the actual goals we set them:

- **Modularization** : the goal is broken down into component subgoals, but these are not pursued in ways conducive to fulfilling the original goal
- **Delegation** : if a goal is delegated to another agent, that agent's own goals may clash with or subvert it

We examine this in more detail in the chapter: *Collective Action Problems* .

Lesson 3:

Scaling Up Generates New Risks

A safe system, when scaled up, is not necessarily still safe.

AIs may continue to develop unanticipated behaviors as we scale them up:

- Scaling up a DL system may not only improve its previous abilities
- It may also cause it to behave in novel and unexpected ways
- Example: large language model arithmetic and proxy gaming

Some emergent AI capabilities may pose risks:

- Strategies of deception
- More sophisticated proxy gaming

Lesson 4:

Directly Built Complex Systems Are Rare

Working complex systems have usually evolved from simpler systems.

We are unlikely to be able to build a large, safe AI system from scratch.

We must *gradually* scale up from small, safe systems.

- We may see novel emergent properties that are not present in the smaller version
- So building a larger system in this way does not guarantee safety
- But it is more likely to be safe than if it was built from scratch

Puzzles, Problems, and Wicked Problems

Simple and complex systems pose different kinds of problems

Puzzles and Problems

5	3			7				
6			1	9	5			
	9	8					6	
8				6				3
4			8		3			1
7				2				6
	6					2	8	
			4	1	9			5
				8			7	9

$$
\begin{aligned}
2(4-y)-3(y+3)&=-11\\
8-2y-3y-9&=-11\\
-5y-1&=-11\\
-5y-1+1&=-11+1\\
-5y&=-10\\
\frac{-5y}{-5}&=\frac{-10}{-5}\\
y&=2
\end{aligned}
$$

Puzzles:

- Only one correct result
- The result is findable given the information provided
- Examples: sudoku, simple math questions, assembling furniture

Problems :

- There may be more than one approach to fixing the problem
- Investigation may be required to uncover all the relevant information
- But it is clear when the problem is solved
- Problems often found in complicated (but not complex) systems
- Examples: car repair issues

Wicked Problems

Wicked problems :

- Often involve a *social* element, making them harder to solve
- More than one possible cause, so *no single flawless solution*
- There is often a *risk* in attempting to solve wicked problems
- It can be *unclear* when a wicked problem has been solved
- The systems under analysis will evolve, grow, *complexify* → surprise (new corner cases, tail events) will result
- Examples: inequality, misinformation, climate change

AI safety is a wicked problem

- AI and the social environment it is deployed in are complex systems
- No single correct approach, and we may never "solve" it

Challenges with Interventionism

Some attempts to solve wicked problems could backfire

Examples

Interventions that work in theory might fail in a complex system.

Examples:

- Cane beetles
- Rat infestation
- Energy crisis

We must approach AI safety with more humility and more awareness of the limits of our knowledge.

Successful Interventions

While it is difficult to say definitively whether a wicked problem has been "solved," there are examples of at least partially successful interventions.

Examples:

- Eradication of smallpox
- Reduced of the depletion of the ozone layer
- Anti-smoking public health campaigns

All required enormous, sustain effort over many years, sometimes with coordination on a global scale.

Recap

Creating safe AI systems is harder because AI systems are complex.

Problems include difficulty in theorizing, goal subordination, risks from scaling, but there is a requirement to nonetheless scale.

AI safety is a wicked problem: requires sociotechnical solutions, has no flawless solutions, attempting solutions presents risks, and it is not necessarily clear when it is well-managed.

Interventions in AI must be carefully considered; they might backfire.

Chapter Recap

1. Where mechanistic and statistical approaches can fail, complex systems approaches can succeed.

1. There are seven common hallmarks of complex systems.

1. AI, its development, and its deployment environment are all complex systems.

1. AI safety is a wicked problem.

1. Managing wicked problems is risky.