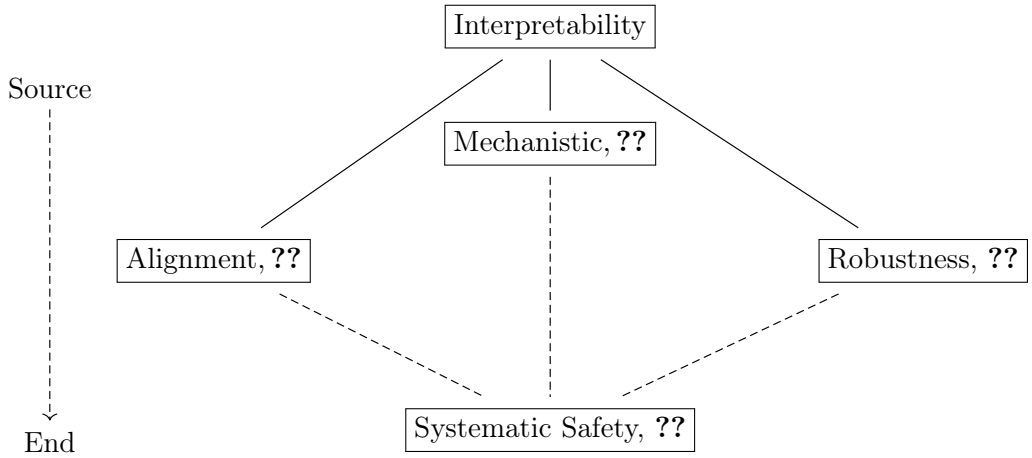
# 1 Safety in AI

## 1.1 A bird eye view of the subject

To organize the article, I categorize the various research areas.




The vertical hierarchy can be thought of as getting from the source (foundational design of the AI systems) to end (their use and place within society). The arrows only mean "contribute". A brief summary of each box is given in, 1.2. Importantly, the whole lifecycle of AI has various stakeholders involved:

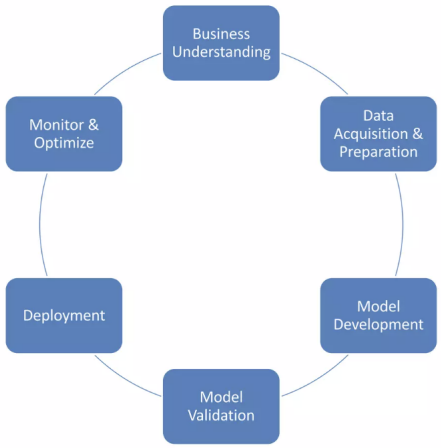



Figure 1: The AI Life Cycle

We refer to [2] for further details.

### 1.2 Overview of each field

We will focus on three of four types of problems, as explained in [4].

- *Robustness*. There are broadly two types of events:
  - *Black swan*: rare extreme, unsual events. Examples of such include the 2010 Flash Crash, [7], or those appearing in autonomous vehicle. In the context of statistics, these are also referred to as *long tail events*. Long tail distributions make the usual arsenal of statistical tools can become less useful. Examples include power law distributions. These are pervasive; see black swan.
  - *Adversaries*: carefully crafted threats, or as Goodfellow et al. describes, " *adversarial examples* are inputs to machine that an attacher has intentionally designed to casue the model to make a mistake." Much literature has focused on "perturbation defense", which was mostly focused from Szegedy et al's, [8]. We discuss more details of this in Sec. **??**.
- *Alignment*, we list a number of main difficulties in alignment.
  - specification. Encoding goals such as judgment, experiences, and happiness is hard. Research is done to assess the language model's ability to evaluate scenarios that could be morally contentions. [3].
  - even when one is able to describe a value to be learned:
    - it can be difficult to optimize, as models can have unintended undesirable secondary objectives. To understand and steer agent's moral values, researchers could dconstruct relevant environments, for instance, Jiminy Cricket, which has 25 text-based games with diverse scenarios, [5].
    - *brittle*, due to reasons as proxy gaming. Further, even with "human approved" proxies, there is the risk of deception from computers. Future systems should strive to address the problem of verification, past attempts include an adversarial process, where during training two agent takes turn making short statements with human judges on the final result, [6].
- *systematic safety*, is the deployment of ML systems in the larger context, including software systems, organizational structure or human society. Two main examples of systematic research are in cybersecurity, such as the defense of potential hackers, and informed decision making.

### 1.3 Homework

*Due: Saturday, November 4th. Homework this week is a reading exercise.* In this week, you will read the following two papers, [4], [1]. For the first paper,

1. Describe a nontrivial strength or weakness of the paper that isn't explicitly mentioned in the readings themselves. (300-400)

2. Summarize the reading. Be sure to read and summarize the contents of each section; do not just describe the overall idea of the paper/article. (400-500)

For the second paper (which is more technical),

3. Write one concept in the article that confuses you/you hope to understand more. Explain why.

# References


[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané, *Concrete problems in ai safety*, 2016.

[2] Marina Danilevsky, Shipi Dhanorkar, Yunyao Li, Lucian Popa, Kun Qian, and Anbang Xu, *Explainability for natural language processing*, Proceedings of the 27th acm sigkdd conference on knowledge discovery & data mining, 2021, pp. 4033–4034.

[3] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt, *Aligning ai with shared human values*, 2023.

[4] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt, *Unsolved problems in ml safety*, 2022.

[5] Dan Hendrycks, Mantas Mazeika, Andy Zou, Sahil Patel, Christine Zhu, Jesus Navarro, Dawn Song, Bo Li, and Jacob Steinhardt, *What would jiminy cricket do? towards agents that behave morally*, 2022.

[6] Geoffrey Irving, Paul Christiano, and Dario Amodei, *Ai safety via debate* (2018).

[7] ANDREI KIRILENKO, ALBERT S. KYLE, MEHRDAD SAMADI, and TUGKAN TUZUN, *The flash crash: High-frequency trading in an electronic market*, The Journal of Finance **72** (2017), no. 3, 967–998.

[8] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, D. Erhan, Ian J. Goodfellow, and Rob Fergus, *Intriguing properties of neural networks*, CoRR **abs/1312.6199** (2013).