Introduction to ML Safety **Homework**

1. A separate model using for understanding reliability is the "Nines of Reliability." An event with a success probability p will have k nines of reliability, where k is defined as

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

Hence

(a) If $p = 90\%, k = 1.0$

(b) If $p = 99\%, k = 2.0$

(c) If $p = 99.9\%, k = 3.0$

A logarithmic scale spaces out distinct levels of reliability more properly. A 1% improvement is more useful the closer to 100%. In terms of nines of reliability, how large is a "mere" 1% improvement from 98% $\rightarrow$ 99% compared to 50% $\rightarrow$ 51%? How many nines of reliability does $p = 100\%$ correspond to? Imagine someone proposes converting all risk measurements into this format; what would be a limitation of this approach?

2. Someone says "first we must quantify the risk more precisely before taking further action." Discuss weaknesses of this proposed strategy.

3. Say land ownership follows a 80/20 Pareto distribution ($\alpha \approx 1.16$), so 80% of the land is owned by 20% of the people. Roughly how much of the land is owned by the top 1% of land owners? (Hint: $(20\%)^3 \approx 1\%$.)

4. Suppose the following power law $y = \frac{1}{4}x^{-3}$ is plotted on a log-log plot. What would be the plotted line's slope?

5. Say a student is taking a mathematics exam. Indicate whether the option describes a known known, known unknown, unknown known, or unknown unknown (use each option only once).

(a) The question for which the student wrote an answer but was uncertain whether their answer was correct.

(b) The student confidently answers $x = 5$ to a question that was accidentally incorrectly written by the instructor, so the question did not even count towards the score and had no correct answer.

(c) The solution to the easiest question on the exam that for which the student knew the answer with certainty.

(d) The question where the student answered the question based on a gut feeling.

6. Someone says Black Monday (1987) is a 20 sigma event. What is the probability of an event that is this extreme or more extreme? Does a Gaussian (thin-tailed) assumption seem reasonable, or does a long tailed assumption seem more reasonable for the events witnessed on Black Monday? Someone says Batman's IQ is 192, and someone says Lex Luthor's IQ is 225. Assuming the average IQ is 100 with a standard deviation of 15, how many sigma is Batman's IQ and Luthor's IQ? What is the probability someone's IQ is greater than or equal to Batman's; the probability someone has an IQ greater than or equal to Lex Luthor's?

7. Recall preventive measures help avoid occurrence of unwanted events, and protective measures limit consequences of realized hazardous events. Given two measures, identify which is preventative or protective.

   (a) boxing helmet; ice pack (hazard: getting hurt in boxing)

   (b) lifeboats on the Titanic; hull of Titanic divided into watertight compartments (Hazard: iceberg)

   (c) eating healthy starting from a young age; chemotherapy (hazard: cancer)

   (d) teaching a child to want to be honest; getting better at catching a child's lie (hazard: lies)

8. "An ounce of prevention is a pound of cure." What does this say about the cost efficacy of preventative and protective measures?

9. Match these words to the following examples:

   - **butterfly effect**: the phenomenon whereby a minute localized change in a complex system can have large effects elsewhere
   - **micro-macro dynamics**: a micro-macro dynamic describes a complex interaction between the micro and macro levels to a system
   - **positive feedback**: a positive feedback loop is a process where one variable increases the quality of another variable which in turn increases the quantity/occurrence of the first
   - **negative feedback**: occurs when some function of the output of a system, process, or mechanism is fed back in a manner that tends to reduce the fluctuations in the output
   - **self-organization**: a process where some form of overall order arises from local interactions between parts of an initially disordered system

(a) When a foot wide snowball rolls down a hill, its surface area increases, and therefore the ball collects more and more snow.

(b) Birds flocking and making structured movement patterns (watch an example here).

(c) A software program instructs a computer's processor to fill its registers, which improves the speed of the software program and allows the software program to change its state.

(d) A waiter got a junior business executive's breakfast order wrong. The executive still ate the breakfast then went on TV for a live interview. The executive belched loud on TV due to the breakfast not being agreeable with his stomach; this made him a laughing stock at his company and caused him to resign.

(e) A person fears socializing due to not being good at it, so they avoid socializing. This makes them comparatively worse than their peers at socializing. This is increases their fear of socializing even more.

(f) A company gets bad reviews. The company's hurt reputation leads top employees to leave. This makes the company worse, leading to even worse reviews.

(g) Making rockets cheaper lead to more space travel, which leads to innovation that makes rockets cheaper.

10. List three sociotechnical factors that end in the word "pressure."

11. Describe a technology that is safe but not reliable. Describe a technology that is reliable but not safe.

12. "For some definitions of safety, one prevents losses due to unintentional actions by benevolent actors. In security, one prevents losses due to intentional actions by malevolent actors. But the key difference of intent usually does not matter in preventing losses (though differences arise with protective mechanisms)." Argue for or against this statement.