AI safety HW1

David McSharry

1 Hazard analysis

1. 0.301 vs 0.0088. The difference is much greater closer to 100 percent. $p = 1 \implies k = \infty$. A limitation would be that decimals in these units are much less intuitive and this could lead to confusion.

2. In the case that the risk is higher than you thought, waiting to determine this would mean you leave yourself at a higher risk for much longer than if you addressed is straight away. Even if the risk is lower, reducing the risk in first place is still a positive.

3. $0.2 * 0.2 * 0.2 = 0.01$ so the top 0.01 own 0.99 of all the land

4. -3

5. known unknown, unknown unknown, known known, unknown known

6. 2.754×10^{-89} (from wolfram alpha). Very unlikely lol. Long tailed seems reasonable given how unlikely this is in a Gaussian distribution (there are less than 10^{83} atoms in the universe), given reasonable priors about whether the distribution is Gaussian or long tailed our posterior should be long tailed. 4.4×10^{-10} and 3.9×10^{-17} respectively (using wolfram again).

7. (protective, preventative), (protective, preventative), (preventative,protective), (preventative, protective)

8. This says that prevention is *normally* more cost effective on the margin.

9. a - positive feedback, b - self-organization, c - micromacro, d - butterfly effect, e - negative feedback, f - neg feedback, g - pos feedback.

10. social pressures, productivity pressures, competition pressures.

11. Cheap battery powered toys, nuclear reactor

12. I would argue that removing the distinction does more harm than good. A method for preventing losses via security could be to hire a security guard, and a method for preventing losses via safety could be to give employees free coffee. These methods do not work visa versa and so the distinction is useful for describing the ways in which losses can incur. You can generalize these two methods to preventing risk or something if you want a catch all term to describe preventing loss.

2 Robustness

1. $\|x - x_{adv}\|_p = (\sum_i |x - x_{adv}|^p)^{1/p}$, and $x - x_{adv} = \epsilon \implies \|x - x_{adv}\|_p = (d \times |\epsilon|^p)^{1/p} = d^{1/p} \times \epsilon$ so for integer d greater than 0 this is bigger for $p = 1$.
2. beta(1,1) unifrom, beta(5,5) peak in center, beta(0,5,0.5) peaks at corners.
3. a - false, b - true, c - false, d - true
4. c - true
5. a - false, b - false, c - true, d - true
6. a - false, b - true, c - false, d - true
7. a - false, b - true, c - false, d - false, e - false
8. a - false, b -true, c - true, d - true, e - true