2-D Robustness

Vladimir Mikulik

August 30, 2019

This is a short note on a framing that was developed in collaboration with Joar Skalse, Chris van Merwijk and Evan Hubinger while working on Risks from Learned Optimization, but which did not find a natural place in the report.

Mesa-optimisation is a kind of robustness problem, in the following sense:

Since the mesa-optimiser is selected based on performance on the base objective, we expect it (once trained) to have a good policy on the training distribution. That is, we can expect the mesa-optimiser to act in a way that results in outcomes that we want, and to do so competently.

The place where we expect trouble is off-distribution. When the mesa-optimiser is placed in a new situation, I want to highlight two distinct failure modes; that is, outcomes which score poorly on the base objective:

- The mesa-optimiser fails to generalise in any way, and simply breaks, scoring poorly on the base objective.
- The mesa-optimiser robustly and competently achieves an objective that is different from the base objective, thereby scoring poorly on it.

Both of these are failures of robustness, but there is an important distinction to be made between them. In the first failure mode, the agent's capabilities fail to generalise. In the second, its capabilities generalise, but its objective does not. This second failure mode seems in general more dangerous: if an agent is sufficiently capable, it might, for example, hinder human attempts to shut it down (if its capabilities are robust enough to generalise to situations involving human attempts to shut it down). These failure modes map to what Paul Christiano calls benign and malign failures in *Techniques for optimizing worst-case performance*.

This distinction suggests a framing of robustness that we have found useful while writing our report: instead of treating robustness as a scalar quantity that measures the degree to which the system continues working off-distribution, we can view robustness as a 2-dimensional quantity. Its two axes are something like "capabilities" and "alignment", and the failure modes at different points in the space look different.

Unlike the 1-d picture, the 2-d picture suggests that more robustness is not always a good thing. In particular, robustness in capabilities is only good insofar is it is matched by robust alignment between the mesa-objective and the base objective. It may be the case that for some systems, we'd rather the system get totally confused in new situations than remain competent while pursuing the wrong objective.

Of course, there is a reason why we usually think of robustness as a scalar: one can define clear metrics for how well the system generalises, in terms of the difference between performance on the base objective on- and off-distribution. In contrast, 2-d robustness does not yet have an obvious way to ground its two axes in measurable quantities. Nevertheless, as an intuitive framing I find it quite compelling, and invite you to also think in these terms.

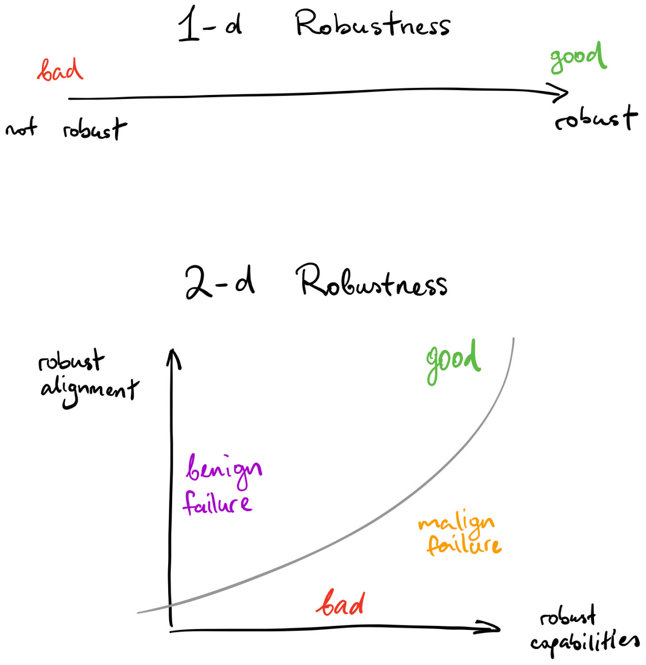

Figure 1: fig 1