1) Does the trajectory of AI capability development match that of biological evolution?
    a) The order in which capabilities arise in machine intelligence is likely to have policy and safety implications. If early biological evolution for intelligence appeared sufficiently similar to capability development in deep learning (or machine learning more broadly), it may be possible to extrapolate dynamics from later biological evolution to future machine learning progress. On the biological side, this analysis would require assessing the behaviors of minimally cognitive organisms (such as plants) up through relatively sophisticated ones (such as bees). It is unlikely that these agents could easily be assessed in an ability-centric way, rather than a task-centric way. However, this is also true for artificial agents. Instead, the frequencies of complex behaviors of different sorts could be measured against biological scaling factors such as neuron count (and other brain-level characteristics), evolutionary time, selective pressure, development length of individual organisms, etc. Some of these factors could be considered directly analogous to AI-relevant variables such as model size, the amount of compute that went into training a model, the amount of research time invested in a problem, etc. Likewise, AI tasks could be categorized and graded according to complexity, then plotted against AI-relevant scaling factors. Comparison between these two trajectories would likely end up being somewhat subjective, as the tasks machine learning models are trained for do not map very well to those biological organisms evolved for.
    b) Since different orderings of AI capabilities are likely to imply different danger scenarios, having a sense of the probabilities of these orderings will allow us to focus more preparation on higher probability scenarios.
    c) The impact of this analysis depends firstly on its tractability, and secondly on the degree to which we would be able to make use of such results. One downside of the methodology described is that it relies on observational analysis of already published work, which, in the case of biological organisms, is likely to report a fairly superficial level of detail, or may simply not exist for many domains of behavior. It may also be difficult to estimate scaling parameters such as selective pressure or amount of research time invested for many instances of behavior. And while it is likely that some kind of comparison could be made between two relatively complete trajectories (or conversely a more definitive statement that the trajectories are not comparable), there is a decent chance that no compelling case could be made for either comparability or incomparability. Many of the tractability concerns could be assessed relatively early on, and the project could be abandoned or modified if it did not seem fruitful. As to the utility of the results, this is an open question and would not likely be addressed directly by this project, but would require further working building on the findings by myself or others.
2) How tractable is long-term forecasting?
    a) Evidence from the Good Judgment Project suggests that otherwise high-performing human forecasters are not able to make accurate predictions beyond a timescale of two years (in the domain of geopolitics). However, the

success of many important endeavors depends on the ability of a decision-maker to make accurate predictions on multi-year or multi-decade timescales. I would first build a dataset of predictions from the past 200 years that attempted to forecast outcomes five or more years away in a wide variety of domains. In addition to conventional predictions, I would also consider “implicit” predictions, such as investments, marriages, corporate decisions, etc. From this data I would extract a number of strategies and sort predictions according to their choice of strategy, then determine which strategies were successful on various timescales and in various domains. I would also look for correspondence with short-term forecasting strategies.

b) Understanding the strategies used for long-term forecasts and their success rates in different domains could help improve institutional decision-making, for example by adjusting levels of confidence in domains with especially good or especially poor records of long-range forecasting, or in different forecasting strategies.

c) I think a successful version of this project would be unambiguously useful, but its usefulness would be necessarily hard to measure concretely on a short timescale. In lieu of concrete measurement, readers in important decision-making roles could be surveyed on how much the output changed their approach to longterm forecasting. The larger issue is, again, tractability. The above approach assumes the availability of a large amount of rich historical data, which may not actually exist; however, this would become apparent early in the project. It may also be difficult to define strategies in a valid and useful way, and to identify them once defined. Additionally, although easier to analyze, it’s likely that many publicly-made historical predictions were not honest representations of the predictor’s beliefs. Conversely, “implicit” predictions are more strongly incentivized to be correct, but are also fuzzier.