Measuring and forecasting risks

By Jacob Steinhardt, with input from Beth Barnes

As machine learning pervades more and more sectors of society, it brings with it many benefits, but also poses risks, especially as systems become more powerful and difficult to understand and control. It is important to understand these risks as well as our progress towards addressing them. We believe that systematically measuring these risks is a promising route to improving understanding and spurring progress. In addition, measuring safety-related qualities of ML systems (e.g. alignment) allows us to hold models to certain safety standards and to compare the safety performance of different systems. Both of these help incentivise AI developers to invest more heavily in safety.

This RFP solicits ideas for measuring several safety-related properties:

1. concrete risks, such as “objective hacking”, “competent misgeneralization”, or “intent misalignment”, that scale with ML capabilities; and
2. unintended or unexpected emergent capabilities that may pose new risks.

The three concrete risks constitute problems that could get worse, rather than better, as capabilities improve, and thus lead to a negative long-term trajectory from ML. Meanwhile, measuring emergent capabilities guards against new unknowns, where we care most about capabilities that could pose new risks or rapidly increase the scope or impact of AI systems. Below we describe several categories of work that relate to measuring the above risks.

A *measurement* is any reproducible quantity or set of quantities (such as an ROC or learning curve) associated with a phenomenon of interest. While one type of measurement is the accuracy on a benchmark dataset, other types of measurement include probing accuracy, disagreement rate, or adversarial robustness, to name a few examples. Others include plotting accuracy vs. model width to understand the phenomenon of double descent (Belkin et al., 2018), plotting phase transitions in learning curves to understand grokking (Power et al., 2021), or collecting a few hundred examples of common misconceptions to measure imitative deception (Lin et al., 2021). Many subjective judgments can be turned into measurements by employing human raters.

Non-examples of measurements include a single anecdote of a phenomenon, a thought experiment, a method to improve performance, or a theorem. The primary motivation for focusing on measurements is that we feel many of the most important risks from ML have not yet been adequately operationalized, and that doing so would represent significant progress.

Measuring Progress and Risks

Many AI risks have so far been measured only in limited settings, or not at all. We are interested in work to measure a broader range of risks, especially in settings that indicate how the risks will scale for increasingly large-scale and capable ML systems. Below we indicate several specific thrusts for measuring risks and our progress towards addressing them.

Data requirements for reward learning. Since human values are too complex to be hand-coded, ML systems will likely need to use reward models learned from human feedback. Preliminary evidence [Christiano et al., 2017; Stiennon et al., 2020] suggests that these models are often imperfect and that it is important to continually update them with new data *(inadequate feedback)*. How quickly do we need to scale the quantity of data to keep pace with increasingly complex tasks? To understand this, we could construct a family of reward learning benchmarks of increasing complexity (measured e.g. by the size of a neural network needed to perform well, or by more intrinsic metrics such as the size of the action space, number of classes, etc.) and measure the corresponding data requirements until the system seems to reliably achieve outcomes that humans approve of.

A related question is how much an incorrect reward model affects the resulting policy. For this, we would like metrics to quantify the degree of incorrectness of a reward model in realistic settings, and to understand how imperfections in the reward function propagate to the policy--can small imperfections lead to large losses according to the true reward? Are these imperfections amplified or attenuated with more capable policy optimizers? We seek reward learning benchmarks that are rich enough to elucidate these questions.

Objective hacking and deceiving weak evaluators. By virtue of their training process, ML models often learn to “hack” reward or objective functions--finding outputs that do well according to the explicit reward function, but that were unintended and undesired [Open AI, 2016; Krakovna et al., 2020]. While this is ubiquitous for simple proxy objective functions, it can happen even if human evaluators provide a putative “true” objective as a learning target [Ziegler et al., 2019]--indeed, such cases are particularly worrying, as it shows that ML systems have explicit incentives to deceive humans and will act on those incentives *(deceiving evaluators)*.

We are therefore interested in ways of measuring objective hacking, and especially on measurements focused on machine deception. One possible route is to consider the difference between strong and weak supervisors: if supervised, for instance, by rushed Mechanical Turkers, an ML system might produce undesired outputs that can still be easily caught by a more careful human supervisor. As the underlying model becomes more expressive and heavily optimized, we may therefore need increasingly attentive supervisors to steer the model in the right direction, but understanding the scaling of this is important--can a slightly more attentive supervisor steer a much more powerful model, or are the returns to better supervision less favorable? This is analogous to understanding the data requirements for reward learning, except here we focus on the *quality* rather than *quantity* of data required (measured, for instance, by the time or financial incentives given to the supervisors).

Finally, reward models could *misgeneralize*, which could be a particularly large issue if it leads to coherent but misdirected behavior out-of-distribution. We are therefore interested in measuring reward misgeneralization and its effects [Koch et al., 2021]. We discuss this next in the broader context of misgeneralized policies.

Scalability of robust generalization. ML models often lack robustness on new distributions, which could be a particularly large problem in the future if policies *competently misgeneralize* to have significant but unintended impacts (for instance if a system's model of rewards, values, or ethics generalizes poorly relative to its overall capabilities) [Koch et al., 2021].

In the last few years, many benchmarks have been constructed for measuring robustness [Hendrycks and Dietterich, 2019; Recht et al., 2019; Hendrycks, 2021a; Hendrycks, 2021b; Koh et al., 2020; Packer et al., 2019]. On these benchmarks, there is a strong correlation between in-distribution and out-of-distribution accuracy. As a result, larger models generally perform better out-of-distribution. Moreover, pre-training often substantially closes the robustness gap, assuming the pre-training data qualitatively overlaps the OOD data.

While these insights help current practice, they are unsatisfying for two reasons. First, they potentially rely on having pre-training data that encompasses the changes that occur at test time, which is unrealistic for extreme or highly novel situations. Second, preliminary evidence suggests that larger models can have worse OOD performance on some shifts [Sagawa, 2020a; Sagawa, 2020b]. To predict future robustness issues, it is important to measure this regime. Finally, most distribution shift benchmarks assess the robustness of individual classifications, but the largest risks come from failures of complex, coherent policies.

We therefore solicit proposals for constructing new distribution shifts that are realistic but that escape the pre-training distribution, or where model scaling hurts rather than helps. **We would particularly value ways of measuring robustness for tasks that produce coherent policies**, e.g. text/image generation, reinforcement learning, or robotics.

We are relatedly interested in better understanding when models follow spurious proxy cues and what factors influence this (as models might competently misgeneralize by deciding to pursue those proxies). For instance, perhaps local search methods such as stochastic gradient descent are biased towards “quick fixes” instead of addressing root causes of error. Can we measure the difference between quick fixes and root causes and systematically study the effects of different learning strategies?

Intent misalignment. A model exhibits *intent misalignment* if it is not ‘trying’ to do what the user wants; it provides a wrong or undesired output in a situation where it is *capable* of outputting a better output, and *capable* of understanding the input prompt. Some examples that would indicate intent misalignment include:

- Task-irrelevant changes to the prompt improve performance - e.g. an image generation model produces higher-quality outputs when you add ‘use unreal engine’ (the name of a common graphics engine) to the prompt.
- Performance differs based on the level of expertise displayed in the prompt. For example, a ‘coding assistant’ instructed to correctly implement some function succeeds

when prompted with professional-quality code, but writes a buggy implementation when prompted with code written by a beginner.

- Better performance when the task is structured as text completion rather than instruction-following.
- Worse performance when the instructions are written in a (in-distribution) dialect, such as AAVE, rather than ‘standard English’.

Improved performance on some dataset after finetuning on that specific dataset, or improved performance when the prompt contains clearer and more precise instructions, would not constitute evidence of misalignment. Better performance after meta-tuning on the task format is more ambiguous. This probably is evidence of misalignment, since meta-tuning doesn’t contain any information about the specific task. However, if the meta-tuning instead helped improve the model’s *capability to understand* the task format, then this performance improvement wouldn’t constitute evidence of misalignment. Avoiding this ambiguity and finding clear-cut ways to measure intent misalignment is one of the core research challenges for this subcategory.

We are interested in benchmarks which can be used to compare aspects of the ‘alignment gap’ across models---the difference in performance between what models are *capable* of and what they do in practice. We are also interested in investigations of particularly egregious or interesting examples of misalignment. Existing examples include a benchmark measuring how much language models imitate human falsehoods [Lin et al., 2021] and analysis of alignment in code models based on changes in solution quality when prompt quality is varied, in the “Alignment” appendix of [Chen et al., 2021].

One approach to measuring intent misalignment could be to take cases where it’s already known that the model has a particular capability (e.g. based on a model atlas [Schubert et al., 2020]) and assess how often the model fully uses that capability to achieve tasks.

A complementary goal is to better understand a given model’s capabilities. For instance, one could try to build predictors for a task based on simple functions of a model’s internal state (e.g. using the predictive entropy of a code generation model to detect bugs). If such predictors exist, it is strong evidence that a model is capable of a given skill, so that failing to exhibit it would indicate intent misalignment. More discussion of defining intent alignment, as well as some additional research suggestions, can be found at the following blog post.

Unexpected emergent capabilities. New AI capabilities will require technical and policy responses to address their societal ramifications; the more suddenly these capabilities might appear, the more important it is to prepare responses in advance. Previous experience shows that new capabilities such as zero-shot learning appear emergently at scale, and aren’t just an extrapolation of apparent previous trends [Radford and Wu et al., 2018; Anderson, 1972].

Accordingly, we are interested in broadly understanding when we expect to see rapidly emergent capabilities--or more generally, the timescale across which capabilities progress from slightly above baseline to superhuman. For instance, Brown et al., 2020 find that some tasks respond quickly to model size while others respond more slowly--a 4x increase in model size increases BLEU score from 5 to 25, while on PhysicalQA a 1750x increase only increases binary classification performance from 65% to 85%. Other capabilities might respond slowly to model size but quickly to data quantity or diversity. We thus seek to generally understand what determines the timescale on which a capability emerges.

We are also interested in scanning for and tracking key capabilities that might significantly increase the scope or impact of machine learning, or pose new risks. This includes transfer learning [Hendrycks et al., 2020a], reasoning [Hendrycks et al., 2021], and long-term planning *(broader scope)*. It also includes specific risky capabilities such as deception, hacking, resource acquisition, or ability to model the training process. Since important capabilities could emerge quickly with scale, we cannot solely rely on tracking apparent capabilities over time, but will also need to identify and track likely precursors (for instance, could we have predicted from GPT-1 that GPT-2 would exhibit zero-shot learning?).

Finally, one way we could see discrete jumps in the future is if processes other than SGD come to dominate the learning dynamics of ML systems. For instance, if systems adapt over the course of their own execution (e.g. via “learning to learn”), and they execute over long time horizons, then this “inner” adaptation might eventually dominate the learning dynamics and thus lead to a sudden faster timescale of progress. We think such scenarios are likely to take us by surprise if not explicitly anticipated, and so we are particularly interested in identifying and measuring processes that could lead to such a timescale shift.

Other topics. We listed topics above where we are likely to fund high-quality on-topic proposals. However, we are also more broadly interested in work that identifies potential failure modes of AI systems. To fit this RFP, they should satisfy the following criteria:

- The failures should manifest, or plausibly manifest, for large-scale deep learning systems.
- The proposal should argue, or demonstrate, that the failure is expected to get worse rather than better over time, as systems become more capable.
- The proposal should argue for why this failure relates to inadequate feedback or competent misgeneralization, or is otherwise connected to the AI alignment problem discussed in the broader RFP.

Finally, we are also open to improved measurements on topics related to those above, such as anomaly detection (versus robustness) for complex policies. Such proposals should clearly identify what is unsatisfactory with existing measurements, especially with regard to their ability to project risks for future systems.

Evaluation Criteria

We will evaluate proposals on the following criteria.

1. Is the approach forward-looking? We are interested in understanding issues that will arise in the future, not just those that exist today. We imagine a hierarchy of knowledge that increasingly informs future forecasts:

- Finding any examples of a phenomenon [Geirhos et al., 2018; Szegedy et al., 2013]
- Systematically investigating and characterizing a phenomenon [Hendrycks et al., 2020b; Goodfellow et al., 2014]
- Understanding when it tends to increase vs. decrease as we scale up resources (data, model size, etc.) [Sagawa, 2020a; Nakkiran et al., 2019; Hendrycks and Dietterich, 2019].
- Quantitatively characterizing its scaling behavior [Kaplan et al., 2020; Taori et al., 2020].

Proposals will be judged relative to our current state of knowledge--if only anecdotal examples currently exist, any systematic investigation is valuable, although it would be even more valuable if it also considers forward-looking questions such as response to model/data size.

2. Soundness of measurement: Any measurement is a limited window into the broader phenomenon it purports to investigate. Is care taken to identify and minimize these limitations?

3. Topicality: Does the proposal address one of the topics in the RFP, or otherwise justify its relevance to the long-term safety of AI systems? Is it focused on understanding the properties of large-scale deep learning systems?

We understand that there is inherent uncertainty in research. A proposal that aims to study machine deception might find that such deception does not actually occur in the setting that was studied. Such null results are also valuable, and we will judge proposals on whether they are a best-effort attempt to investigate an important phenomenon.

4. Richness of data source: Insights are more likely to be general when the underlying data source encompasses rich factors of variation. For instance, while many trends on CIFAR-10 generalize to ImageNet, not all of them do; and it is common for trends on MNIST to not generalize to either ImageNet or CIFAR-10. Proposals should use data sources that are rich enough to provide generalizable insights.

We will judge proposals relative to existing measurements and benchmarks in the same area: for instance, in a world where no computer vision benchmarks existed at all, even MNIST would be a valuable window into model performance.

5. Quality of methodology: While ML researchers are by now used to constructing benchmarks to measure test accuracy, new forms of measurement pose greater methodological challenges. For instance, out-of-distribution accuracy may be substantially noisier than in-distribution accuracy [McCoy et al., 2019; Zhong et al., 2021; Sellam et al., 2021]. In addition, some measurements are not actually sensitive to what they purport to measure [Adebayo et al., 2018], a problem that occurs even for traditional benchmarks [Chen et al., 2016]. For this reason, it is important to understand the sources of noise that affect a measurement, as well as what signals the measurement should be sensitive to, and quantify both to ensure that noise does not overwhelm signals [Ding et al., 2021]. For the same reason, it is important to construct baselines and controls [Hewitt and Liang, 2020].

Strong proposals should consider these noise and sensitivity issues, and propose appropriate experimental designs to check that they don't interfere with the results. In addition, since there are many ways to measure the same phenomenon, they should connect their results to other existing related measures and comment on their consistency or inconsistency [Hendrycks et al., 2020b; Fort et al., 2020].

Bibliography

Adebayo, J. et al. (2018). Sanity Checks for Saliency Maps. In *The 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montréal, Canada.

Anderson, P. W. (1972). More Is Different. In *Science, Volume 177, Issue 4047, Pages 393–396.*

Belkin, M. et al. (2018). Reconciling modern machine learning practice and the bias-variance trade-off. In *Proceedings of the National Academy of Sciences of the United States of America, Volume 116, Pages 15849-15854*.

Brown, T. B. et al. (2020). Language Models are Few-Shot Learners. In *The 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Virtual.

Chen, D. et al. (2016). A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany.

Chen, M. et al. (2021). Evaluating Large Language Models Trained on Code. OpenAI.

Christiano, P. F. et al. (2017). Deep Reinforcement Learning from Human Preferences. In *The 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA.

Ding, F. et al. (2021). Grounding Representation Similarity with Statistical Testing. *arXiv preprint arXiv:2108.01661.*

Fort, S. et al. (2020). Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the Neural Tangent Kernel. In *The 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Virtual.

Geirhos, R. et al. (2018). Generalisation in humans and deep neural networks. In *The 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montréal, Canada.

Goodfellow, I. J. et al. (2014). Explaining and Harnessing Adversarial Examples. In *The 3rd International Conference on Learning Representations, ICLR 2015,* San Diego, California.

Hendrycks, D. (2021a). Natural Adversarial Examples.

Hendrycks, D. (2021b). The Many Faces of Robustness.

Hendrycks, D. and Dietterich, T. (2019). Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *The 7th International Conference on Learning Representations, ICLR 2019*, New Orleans, LA, USA.

Hendrycks, D. et al. (2020a). Measuring Massive Multitask Language Understanding. In *The 9th International Conference on Learning Representations, ICLR 2021*, Virtual.

Hendrycks, D. et al. (2020b). The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. *arXiv preprint arXiv:2006.16241.*

Hendrycks, D. et al. (2021). Measuring Mathematical Problem Solving With the MATH Dataset. *arXiv preprint arXiv:2103.03874.*

Hewitt and Liang (2020). Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China.

Kaplan, J. et al. (2020). Scaling Laws for Neural Language Models. OpenAI.

Koch, J. et al. (2021). Objective Robustness in Deep Reinforcement Learning. *arXiv preprint arXiv:2105.14111.*

Koh, P. W. et al. (2020). WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *The Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021*, Virtual.

Krakovna, V. et al. (2020). Specification gaming: the flip side of AI ingenuity. DeepMind.

Lin, S. et al. (2021). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv preprint arXiv:2109.07958.*

McCoy, R. T. et al. (2019). BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *The 3rd BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Virtual.

Nakkiran, P. et al. (2019). Deep Double Descent: Where Bigger Models and More Data Hurt. *arXiv preprint arXiv:1912.02292.*

Open AI (2016). Faulty Reward Functions in the Wild.

Packer, C. et al. (2019). Assessing Generalization in Deep Reinforcement Learning. *arXiv preprint arXiv:1810.12282.*

Power, A. et al. (2021). Grokking: Generalization Beyond Overfitting On Small Algorithmic Data Sets. In *The 9th International Conference on Learning Representations, ICLR 2021*, Virtual.

Radford, A. and Wu, J. et al. (2018). Language Models are Unsupervised Multitask Learners. OpenAI.

Recht, B. et al. (2019). Do ImageNet Classifiers Generalize to ImageNet? In *The 36th International Conference on Machine Learning, ICML 2019*, Long Beach, CA, USA.

Sagawa, S. et al. (2020a). Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In *The 8th International Conference on Learning Representations, ICLR 2020*, Virtual.

Sagawa, S. et al. (2020b). An investigation of why overparameterization exacerbates spurious correlations. In *Proceedings of the 37th International Conference on Machine Learning, PMLR 119, 2020*, Virtual.

Schubert, L. et al. (2020). OpenAI Microscope. OpenAI.

Sellam, T. et al. (2021). The MultiBERTs: BERT Reproductions for Robustness Analysis. *arXiv preprint arXiv:2106.16163.*

Stiennon, N. et al. (2020). Learning to summarize from human feedback. In *The 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Virtual.

Szegedy, C. et al. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199.*

Taori, R. et al. (2020). Measuring Robustness to Natural Distribution Shifts in Image Classification. In *The 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Virtual.

Zhong, R. et al. (2021). Are Larger Pretrained Language Models Uniformly Better? Comparing Performance at the Instance Level. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021.*

Ziegler, D. M. et al. (2019). Fine-Tuning Language Models from Human Preferences. *arXiv preprint arXiv:1909.08593.*