Get insights on the latest developments in AI delivered to your inbox

Devising ML Metrics

new subfields, we must define the metrics that correlate with progress on the problems we care about. Formalizing these metrics into benchmarks will be crucial to capturing the attention of

AI Safety, Ethics, and Society

Representation Engineering: a New Way of Understanding Models

inherently located on the edge of chaos, since their design requires the researcher to effectively concretize a nebulous notion into a single number.

Clear Evaluation

Perhaps the most important quality of a benchmark is having clear evaluation. To

loss of 1.8 mean?"), and recognizing that there is substantial implicit wisdom in precedents is necessary for creating benchmarks. Non-imitative tendencies (similarly, overconfidence in the reach of one's own intellect)

About us Our work AI Risk Resources Contact Careers Donate

- Include clear floors and ceilings for the benchmark. Many benchmarks use accuracy, mainly because the benchmark has a clear floor and ceiling. If metrics do not have clear

divided into many different subcategories that are qualitatively different. Some researchers instinctively resist averaging performance across the subcategories into an overall "average" metric. This is a very bad idea, because having a get used. In some cases, researchers do not want to report eight different numbers, and would rather report just one. A person's short-term memory can't store a dozen subtly different

removing barriers to entry:

- Avoid expanding many modalities at once. A benchmark combining RL, NLP, and CV is unlikely to be used, because few researchers have skills in one should be wary of doing this unless necessary.

- Make good software packages and codebases that people can easily use without much training. This means

This website uses cookies

We use cookies to personalise content and ads, to provide social media features and to analyse our traffic. We also share information about your use of our site with our social media, advertising and analytics partners who may combine it with other information that you've provided to them or that they've collected from your use of their services.

Necessary

Preferences

Statistics

Marketing

Show details >

progress on. Instead, zoom in on the structures that are most difficult for current systems, and remove parts of the benchmark that are already solved.

- As mentioned in our last post, using human feedback in evaluations of a and for academics requires IRB approval. This creates major barriers to entry. More broadly, using human feedback indicates that the central problem has not been shaped into a

The most difficult part of designing a benchmark is concretizing a nebulous idea into a metric. Doing this first requires thinking of an idea to be tested, which is difficult in itself, but thinking of how to concretize it is even more difficult. The aim broader problem and simple enough that it can be concretized and improved. The task requires foresight: where will the machine learning field go next? What is becoming possible that previously wasn't?

that the benchmark is being designed for. This likely means doing research in the area, talking to others doing research, and absorbing what it's like to use a benchmark. It's not wise to try to swoop into a field without knowing anything about how researchers in the field approach their the community you aim to mobilize. An additional benefit is that listening to researchers' intuitions on a given problem might give ideas for how to concretize it. One way of doing this is to collect their

expectations. From an outside view, it is quite difficult to design a good metric, or else it would be easy to write highly impactful papers.

The internet has a vast amount of data that can be collected. If it appears necessary makes sense to spend more time scouring the internet (note that this is different for applications projects, such as self-driving cars). If there is nothing relevant on the internet, this may indicate that the idea

different writers and so on is very useful because it allows for making progress in a number of dimensions at once. Beware of believing that there are more dimensions than they are: for instance, procedurally generated data may appear to have many dimensions ("we have infinitely many there are not many dimensions to it. Adding a random number generator to choose the coefficient for a particular piece of data only adds a single new dimension, not infinitely many. For this

necessary to maintain the area to make sure that researchers are doing things correctly. Any benchmark is of course susceptible to being gamed (some more than others): researchers can create methods that exploit some peculiarity of the benchmark rather than make progress reviewers might recognize that the approach is gaming the benchmark. However, sometimes reviewers don't know any better, and that's why it's often necessary to reduce the effect by

baseline for OOD detection, and benchmarks for robustness (ImageNet-C) and large language models (MMLU, MATH).

Thomas Woodside contributed to drafting this post in 2022 when he was CAIS's first is *here*.

This website uses cookies

We use cookies to personalise content and ads, to provide social media features and to analyse our traffic. We also share information about your use of our site with our social media, advertising and analytics partners who may combine it with other information that you've provided to them or that they've collected from your use of their services.

Necessary

Preferences

Statistics

Marketing

Show details

CAIS is an AI safety non-profit. Our mission is to reduce societal-scale risks from artificial intelligence.

About us Our work AI Risk Resources Contact Careers Donate

OUR WORK

View All Work

Statement on AI Risk

OUR MISSION

About Us

2023 Impact Report

GET INVOLVED

Donate

Contact Us

Credits

Website by Osborn Design Works