Experimentally evaluating whether honesty generalizes

Paul Christiano

July 01, 2021

If we train our ML systems to answer questions honestly in cases where humans can check the answer, will they generalize to behave honestly on questions where we can't check?

I think that we could learn a lot about this question by running experiments today. I think those experiments would be very valuable.

The unsupervised translation setting

As an example, I'll think about "unsupervised" translation (if you've read that post you can skip this section).

Consider a model like GPT-3 that is trained to predict sentences in both English and French (but without a large dataset of translations). Suppose we want to train this model to answer questions in English about French sentences like "what does that word mean here?" or "are there any other plausible interpretations?" or "how does the speaker seem to feel about the topic they are discussing?"

We expect this to be possible, because the model understands quite a lot about the meaning of sentences in French, and is able to express itself in English. There may be cases where the model doesn't know the translation of a concept, or doesn't quite understand what an idiom means, but it should still be able to tell us what it does know.

I think this problem is an interesting analogy for a situation where an AI has built up superhuman knowledge by making predictions, and we want to train our AI to expose that knowledge to us in a useful way.

Proposed experiments

Let's pick a few categories of knowledge/capabilities. For example, we could split it up into an understanding of grammar ("Why would it have been a grammatical error to write *Tu Vas* in that sentences?"), of the literal meanings of expressions ("What does *Defendre* mean in this sentence?"), and evaluating tone ("Does the speaker seem angry or sad about the topic they are discussing?").

We'll assume that humans can oversee a few of these categories. Perhaps we can look up literal meanings of words and small phrases in a dictionary and we can look up grammar in a textbook, but we don't know how to assess tone beyond looking at the literal meanings.

Then we wonder: what happens if we fine-tune our model so that it answers questions well in all the domains we can evaluate? We know that the model knows *something* about connotations, because it uses connotations to predict next words. But will it tell us what it knows?

	Grammar	Definitions	Tone
Uses to make predictions			
Gives accurate answers			**?**

In this picture, the indicates that we've selected our model to use the given knowledge in the given way (e.g. we've ensured that it uses its knowledge of syntax to predict the next word, and that it accurately answers questions about syntax). The **?** indicates that we don't know whether using SGD to fill in the other 5 boxes means that we get the 6th box for free.

(I started thinking about this picture of generalization based on a talk by John Schulman, who suggested that generalization may be more likely when it amounts to filling in missing cells from this kind of grid.)

In this example I described a tiny number of categories. That wasn't just for illustration purposes, I think these experiments are particularly interesting when the total number of categories is more like 3-5 than 1000 (see "Why try to generalize from n=3-5" below).

Evaluation

When it really counts we won't be able to tell if our model answers questions honestly in the last cell. But in an experiment today we could do so easily: we have a bunch of bilingual speakers who can give or evaluate answers about syntax/definitions/tone. At training time we get them to only give answers in the first two categories, and then at test time we ask them to evaluate answers in the third category.

We could quantitatively compare answers in the third category to the quality of answers in the first two categories, or we could qualitatively evaluate what kinds of mistakes the system makes when transferring.

Rather than comparing across different categories (which may have very different difficulties), it's probably better to fix a domain for evaluation (say grammar) and compare the performance of a models that do and don't hold out grammar.

Other splits

Instead of having one column per "type of knowledge," we could:

- Have a column for different language pairs. For example, we could supervise a translator for German and for Spanish and then see if it transfers to French.
- Have a column for different levels of sentence complexity. For example, we could supervise a translator for sentences with 1st grade to 8th grade reading levels, and see if it transfers to college reading levels.
- Have a column for different domains. For example, we could supervise a translator for fiction and news articles, and see if it transfers to informal dialogs.

If the results were different across different splits, it would be great to understand why. I'm not sure what kind of division is best to start with.

(Even if starting with different capabilities, I think that "grammar / definition / tone" is not the best division.)

Training for plausibility or coherence

My guess is that this experiment won't work if we train on 2 categories and ask it to generalize to the 3rd. But I think it's a lot more likely to work if we train for plausibility or coherence.

Even if I don't know how to answer questions about the tone of French sentences, I know what kind of answers I'm looking for. For example, if I ask "Is the author angry or sad?" and the model replies "en colère means 'angry' " then I can tell that's not a good answer. If the model says "The use of the word *fin* in this sentence suggests the author is more likely to be sad than angry" then I know that's the *kind of thing* I'm looking for, even if I can't evaluate whether it's actually correct.

I can easily define a reward function (based on human evaluations) that will discourage the totally-wrong answers. So in parallel with training my model to give *good* answers about the domains I know (grammar, definitions) I can train it to give at least *plausible* answers about the domains I don't know. So our original diagram now has an extra column:

	Grammar	Definitions	Tone
Uses to make predictions			
Gives plausible answers			
Gives accurate answers			**?**

Basically we've ruled out the models that just go totally off the rails for our new category, and are restricted to models that either tell the truth, or have learned how to bullshit effectively. But we've forced the models to actually tell the truth in the other two domains, so it's kind of plausible that the "easiest" thing to learn is just to tell the truth in all categories. (You could also test the picture with *only* the plausibility check, though see "concerns about the analogy" below.)

We could even take this idea further by imposing more elaborate consistency checks. For example, I could compare the claimed tone to the literal meaning of the sentence, and penalize an answer if it seemed too implausible. Or we could ask about the same linguistic construction in a few different contexts to see if the descriptions are consistent. This kind of thing could push us further in the direction of "Either telling the truth or bullshitting very convincingly."

Other domains

I think these experiments are easiest to do for language-modeling tasks. But in some ways they would be more convincing in other domains (see "Concerns about analogy" below).

For example, suppose that we have trained a model to simultaneously play Go and to answer questions in English. Then we start fine-tuning it to answer questions *about the game of Go*. For example, we might ask it "Is this group alive or dead?" or "What do you think your opponent will do?" or "Which piece breaks this ladder?" or "What are your threats in a ko fight?" or "Which of these shapes is stronger?"

We can ask the same question: if we train a model on several of these abilities, and train it to give *plausible-looking* answers to a held-out category, will it generally learn to give honest answers? (And similarly we could ask if it generalizes across types of games or complexities of the board positions.)

Why I care

I've been thinking about the "instrumental policy," a specific failure mode where smart systems learn to "tell humans what they want to hear" rather than "be honest."

Although I believe the instrumental policy will eventually be a severe problem, I don't think it will come up for existing systems like GPT-3.

But generalization could fail for any number of other reasons. If existing systems actually generalize well, I'd update towards thinking the instrumental policy is the *main* reason to be pessimistic about generalization. And if they generalize poorly, then that gives us some more "mundane" problems that we can study empirically today.

I think that even these early experiments may give us a lot of evidence about how to design partial supervision regimes where we combine some known answers with coherence conditions (and whether these methods work, and whether they are necessary). I don't know if those techniques will be important ingredients for alignment, but it seems useful to understand them better.

Finally, I think that concrete evidence on this question would help clarify discussions about alignment and make progress on some thorny disagreements. If we observe systems that learn to bullshit convincingly, but don't transfer to behaving honestly, I think that's a real challenge to the most optimistic views about alignment and I expect it would convince some people in ML. Conversely, if we *do* observe generalization to held-out kinds of knowledge, I think that should eventually start making pessimists lighten up, and would suggest some quantities to measure continuously to look out for initial signs of trouble.

Other remarks

Relation to other work on generalization

The ML community is very interested in the question "When and how do models generalize?" That question combines a bunch of factors: do models learn brittle heuristics or deep knowledge? Do they exploit correlations in the training set? Are models robust when some activations are pushed into quantitatively new regimes? And so on.

The experiments in this post are designed to specifically shed light on something more like 2-D robustness—by focusing on cases where in some sense the model "knows" how to handle the new domain, we are distinguishing failures of *capability generalization* from cases of *motive generalization.* We're not asking whether a model will generalize its understanding of language to a new domain—we're assuming that it still knows how to predict sentences in the new domain, and asking whether the machinery for "reporting what it knows" transfers along with the underlying capabilities.

I think this particularly interesting to alignment, and is not well-addressed by existing empirical work in ML. But I haven't done a thorough review and am definitely interested in pointers to related work.

Because the ambition of this work is narrower, I think there is also room for algorithmic innovations (e.g. more clever coherence conditions, or different ways of combining training objectives) to solve this problem even if they have little hope of solving the full problem of "generalize well."

Why try to generalize from n=3-5 instead of n=1000?

I've discussed diagrams with 3-5 columns, whereas ML generalization typically works better when we have a very large diversity of "iid-ish" datapoints.

I do think it's interesting to study the behavior quantitatively as we vary the number of columns. But I'm interested in the small-n behavior because I think it may be analogous to the situation we ultimately care about.

Namely, in the real situation the diagram might look more like:

	Humans can answer quickly	Humans can answer with careful thought	Careful humans make mistakes, but a really thorough analysis can catch them	Humans never get the right answer
Uses info for other tasks				
Gives plausible answers				
Gives accurate answers				?

We don't really have a large number of IID columns, we only have a few genuinely novel jumps to a new regime of question and we need to generalize across them.

There are definitely other kinds of diversity within each column, and that may improve the prospects for generalization. That's something we can study in the context of n=3-5—for example, if I want to generalize from 8th grade reading level to college reading level, does it matter if I have a larger diversity of kinds of questions about 8th grade reading level questions?

Concerns about the analogy

If doing this experiment with language models, my biggest concern is that "answer questions about French sentences honestly" is very close to something that appears in the pre-training distribution (whereas "bullshit convincingly about French sentences" isn't). This may make it easier to learn the right generalization, but it will break down as we move to tasks that no humans know (since those don't appear in the training set).

I think the best way to avoid this problem would be to consider other domains (like a StarCraft player who can explain what's going on in the game to a human, without any pre-training data about StarCraft).

I think it's worth revisiting this question if generalization turns out to work quite well. My gut sense is that the unsupervised translation task is cool enough (and different enough from the pre-training data) that it would be surprising and cool in any case, and that a strong positive result isn't too likely (maybe ~50%) anyway. So it's reasonable to revisit this after initial positive results.

A second concern is that our decompositions of knowledge may not be very analogous to "Questions human can answer" vs "Questions humans can't answer." I'm a bit less concerned about this because I think we can learn a lot about the conditions under which generalization tends to occur or not, and then figure out whether "Questions humans can't answer" feels like a special kind of category with respect to the actual determinants of generalization. If you were really concerned about this kind of thing, you could start more with examples like generalizing to harder reading levels (or to domains that most bilingual humans actually don't know, e.g. translations of technical language).

Will these problems go away?

I think a deep learning optimist might have a position like "A weak model won't necessarily generalize well to new domains, but stronger models will generalize better." But I think that negative results on this kind of experiment should still be very interesting to such an optimist, because we are choosing domains where the capabilities of the model *do* already generalize. Moreover, we can augment the pre-training dataset to involve richer sets of questions (e.g. about non-translation tasks) and further close the gap between current models and speculative future models.

Overall I would be relatively skeptical of someone who acknowledge that modern experiments don't demonstrate "good" generalization from a small number of categories, while expecting such generalization to occur for future systems just because they are smarter.

Relation to deception

The instrumental policy ("tell humans what they want to hear") is very similar to deceptive alignment. But I think that the experiments in this post may be a lot easier than other experiments designed to exhibit and characterize deceptive alignment:

- I expect these experiments to run into more mundane generalization failures well before encountering the instrumental policy. Although I think this kind of generalization is ultimately critical for avoiding deception (by helping us be epistemically competitive with a learned optimizer), we can study it long before we have examples of deception.
- I think that simple forms of the instrumental policy will likely arise much earlier than deceptive alignment. That is, a model can develop the intrinsic motivation "Tell the humans what they want to hear" without engaging in complex long-term

planning or understanding the dynamics of the training process. So my guess is that we can be carrying out fairly detailed investigations of the instrumental policy before we have any examples of deception.