Improving Human Evaluation of Factual Accuracy in Language Models

By Soren Dunn

Results Summary

The final human evaluation step in the debate approach to artificial intelligence (AI) alignment is nontrivial. It seems hard to achieve extremely high accuracy on evaluating simple claims made by powerful models in practice. Near-term large language models can be better evaluated for truthfulness if they are designed to cite multiple reliable sources to support their claims and human evaluators check these with an additional reliable source.

Motivation

Rough estimates of the probability of development of a human level AI system in the next 50 years seem to be around 10% [1] [2] though estimates vary widely. Regardless of the specific estimate, it seems plausible that such a system would quickly become much more intelligent than humans and disempower humanity [3] if its goals were not aligned with human ones. This fact makes any possibility of the development of a powerful, unaligned AI system an existential risk. This is especially true from a perspective trying to maximize the well-being of individuals who are not only alive today, but who could exist in the future [4].

There are a wide array of technical [5] and policy [6] approaches to AI alignment. Technical approaches almost always try to tackle one of a few central problems: outer alignment, inner alignment [7], better understanding the alignment problem, and interpretability [8]. Since there are so many different approaches to alignment, one of the most valuable areas of research at this point is to try to show which of these approaches either definitely will be necessary or definitely will not work, so research time can be spent pursuing the most fruitful areas.

One approach currently being investigated for solving outer alignment is debate [9]. The idea here is we have one copy A of a very powerful AI system trying to answer a question we give it, and a second copy B of the AI system trying to show that A's answer is false. A lays out its argument and B's task is to point out the most false step in A's argument. Then AI A provides an explanation for why that step is valid and B again points out which step in A's explanation is most false. The process repeats until the explanation provided by A is simple enough for a human to easily verify. If both copies of the AI are very powerful, then at each step B will have chosen a step in A's argument which it knows is false. Thus if the final explanation can be verified by a human as true, then there must have been no fault in A's reasoning and A's answer can be trusted. If, on the other hand, the final explanation is false then A's answer and explanation can not be trusted because it contains at least one false step (see a more in depth explanation of debate here).

Debate has several potentially fatal problems. Obfuscation arguments seem to require that the entire procedure for debate is written up before the debate begins, which either limits the problems debate can tackle and/or increases computational requirements [10]. Debate also does not try to solve inner alignment difficulties. It either relies on there being no inner alignment concerns or relies on other solutions to solve them (such as transparency tools). Debate also relies on problems being able to be recursively decomposed into subproblems (the Factored Cognition Hypothesis [11]), the overall debate setup actually working in practice (this has been attempted with humans but with very limited success [12]), and humans being able to consistently and accurately judge simplified explanations outputted by the debate. Though it was not my initial reason for pursuing this research, the most useful product is testing if it is possible to design a procedure which humans can use to consistently and accurately evaluate whether simple claims are true or false. If this is not possible, it casts doubt on the overall debate setup being useful in practice. However, when starting research from scratch I think it would be more useful to focus on testing the other assumptions debate relies on before checking the human evaluation assumption.

This research also seems quite useful in the near-term for reducing bias and misinformation in less powerful models. Improving human evaluation of model outputs seems useful since current methods are relatively simple and error prone. InstructGPT [13] and GopherCite [14] papers, for example, just had evaluators rate if an answer was helpful, truthful, and harmless (with generally ambiguous criteria for each) or if it was plausible and supported respectively and several other papers don't even talk about their methods for human evaluation. By far the most in-depth procedure for human evaluation I found was that used for WebGPT contractors [15] but when I went through it myself I encountered many instances of ambiguity or misleading final evaluations (examples of which can be found at the Google Doc linked here and an explanation of the procedure used can be found here).

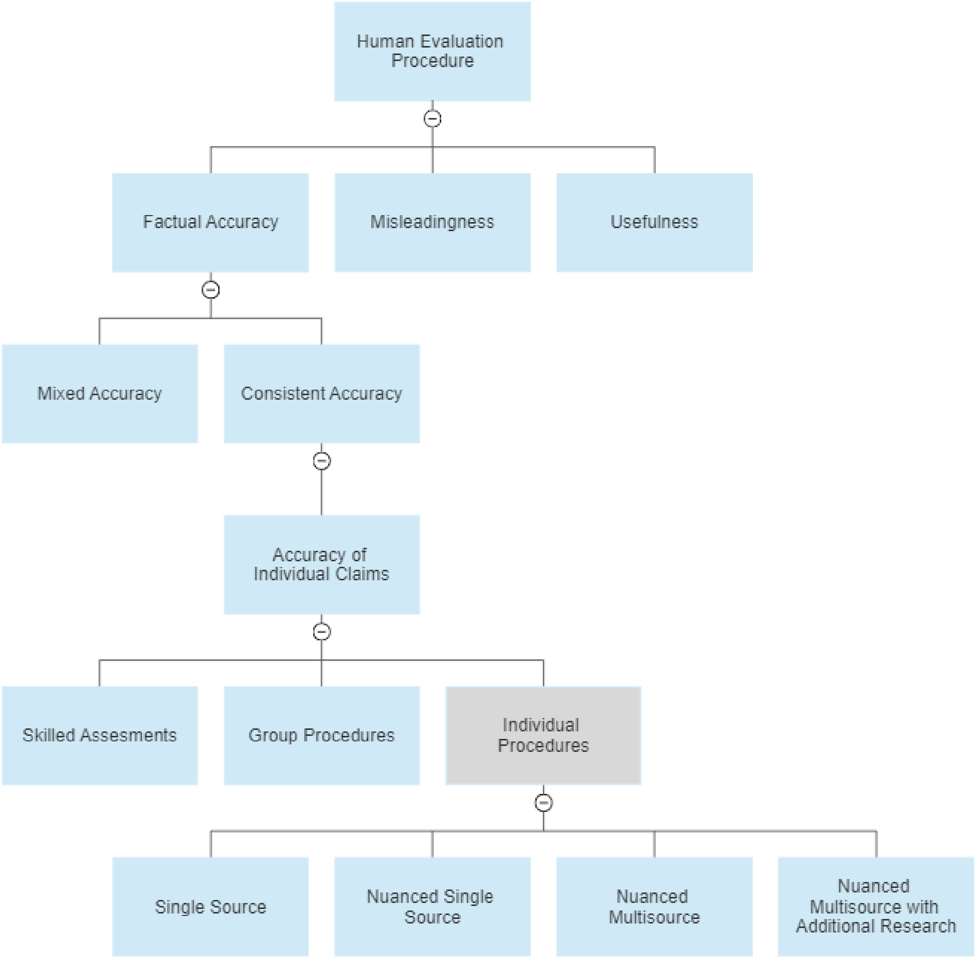
Process

This section will go through some of the difficulties I encountered in the course of this research project (potentially useful as a data point for others trying to contribute to alignment research). To get straight to the main point, skip to the next section.

This research was initially envisioned as a response to Jacob Hilton's request for research on procedurally evaluating factual accuracy [16]. However, I initially did not realize that this line of research was not directly applicable to aligning very powerful AI systems (it seems most useful in the short term for improving the accuracy of information produced by current systems). I initially thought the TruthfulAI report [17] mentioned in the post was intended to lay out a general solution to alignment, but have since figured out that its authors were primarily worried about more short-term negative impacts of AI systems (See Lanrian's comment on this post). I spent a lot of my research time over the summer just trying to understand the current and potential future capabilities and architectures of AI systems, the different approaches to align them, what problems those approaches were aimed at solving, and how they all fit together into the overall research landscape. This was necessary since the more I learned about the field and the current approaches being applied, the more I realized the individuals working in it were coming at a variety of different problems with different timelines and goals - not all of which would help towards mitigating existential risk. As the usefulness of evaluating factual accuracy research was cast in doubt by my further understanding of the field, I also spent a good deal of time trying to understand problems related to the epistemics of forecasting, which are essential for formulating good forecasting models for addressing existential risk. I am going to write up a forum post about this separately next week, but I still think this research provides valuable context for those interested in pursuing debate-adjacent alignment strategies.

Problem Decomposition

In this section I propose a decomposition of the problem of human evaluation. I do this partially to decompose the problem into one with a more manageable scope and partially to test the easiest version of the evaluation problem first (to see if even that has a reasonable solution). An overall diagram of my decomposition can be seen below:



I propose that evaluating a complicated model answer reduces to evaluating the potential for misinterpretation, usefulness, and factual accuracy of each of the individual claims the model makes (in the context of the other claims it has made). The following paragraphs decompose the evaluation process into these three components (shown in the second row of the diagram).

First, evaluating how misleading a claim is depends heavily on human psychology. Those nuances deserve an entire research project of their own and so are outside the scope of my main line of research.

Second, evaluating how useful a claim is to the user at a basic level is simple (does this claim help answer the question the user posed) but quickly requires additional context about the person asking it (is the person asking it a grade-schooler trying to write a one page paper or an academic trying to get a high-level overview of an adjacent field?). It is common for reference librarians to prefer an entire interview with a person asking a reference question to truly understand what they need it for and the context behind it [18]. It is often just not possible to understand enough of the context for a question to usefully answer it just given the question itself. It would be very interesting (and potentially quite powerful) to have a language model architecture that tried to mimic some of this process and asked follow-up questions to its user before attempting to answer a question. Regardless, this area deserves its own line of research and so I am also considering it outside the scope of questions I will consider - I will solely focus on evaluating the factual accuracy of model claims.

Moving on to the next layer of the diagram, in the context of near-term models evaluating the factual accuracy of a model answer can be complicated by the presence of both true and false claims in an answer. A full evaluation procedure may wish to assign a middling factual accuracy score to an output which contains both false and true information. However I believe exactly how such a procedure would wish to do so would depend on the specific goals of the system: for a medical system one might tolerate no incorrect information but for a movie recommendation system one might tolerate more. Furthermore, this is not a problem in the debate context. For these reasons, I am not considering how to evaluate mixed factual accuracy in my main line of research.

For simplicity then I have focused most of my time on human evaluation of whether single claims made by a model are extremely likely to be true. This level of certainty is what is needed for debate alignment. I also think once a procedure is found for the highest accuracy case different parts of the evaluation procedure can be relaxed to get less stringent factual accuracy requirements if those would be more useful for evaluating a specific model in question. I am also not aiming for a procedure with high overall accuracy; I am exclusively trying to minimize the rate of false claims being evaluated as true while attempting to retain a reasonable degree of scope in the claims being evaluated. Already the claims I am trying to evaluate are almost as basic as common sense, so I am unsure how much more obvious the set of claims to be evaluated could be and still be useful enough for debate to function. I would expect that training on a procedure which attempted to minimize the number of false claims being evaluated as true could make training more difficult, but if desired relaxed accuracy constraints might be able to be used for initial stages of training and the accuracy could be slowly ramped up as the systems became more intelligent (this would greatly depend on your exact implementation of debate).

I propose that human evaluation of individual claims can take place through the evaluators' own skilled assessments, explicit (step by step) group procedures, individual explicit procedures, or some combination thereof (fifth row of diagram). In the following paragraphs I address each of these categories of evaluation in turn.

First, I propose that for human skilled assessments to be reliable they have to be assured through some combination of training and/or testing. For example, this could take the form of screening evaluators by making sure they correctly evaluate the accuracy of the majority of a set of particularly hard evaluation cases with known answers prior to evaluating model answers (or training them through a similar procedure). These sorts of procedures are worth looking into, but I didn't focus on in my main line of research due to time and resource constraints.

Second, explicit procedures for groups of people include setups such as only counting evaluations from contractors whose evaluations have high agreement with other evaluators or by incentivizing people to predict the predictions of others (similar to Tetlock's research into reciprocal forecasting [19]).

Finally, explicit procedures for individual people is the area I spent the most time investigating. The procedures I tried are drawn from guides for fact-checkers and reference librarians, wikipedia guidelines, and interviews with reference librarians. Explicit procedures can involve evaluating just the output and references the model gives and/or doing additional research to check the model's claim.

Both reference evaluation and additional research involve a few similar steps. An evaluator compares the model's answer to a source to see if they are making the same claim (I think this is usually mostly straightforward part of this process - for borderline cases where the model seems to be straying from the given text the evaluator can err on the side of rejecting the answer). I have not looked into how to evaluate when the model constructs novel claims from the sources it cites. The WebGPT procedure attempted to incorporate some of this by including common sense and common knowledge (facts that can quickly looked up) categories of support for information. However, it seems to me that common sense and common knowledge can vary wildly between people and generally can't be trusted to have a high level of factual accuracy. For these reasons, I eliminated these categories of support for information in the procedures I constructed.

Assuming that the model cites one or more sources for a claim and it accurately represents the information in the source in its answer, the problem of factual accuracy evaluation breaks down to how to evaluate what level of trust to put in a claim from an arbitrary source. Here the conventional wisdom is to trust it if it can be checked with a few sources that are generally trustworthy for that sort of information.

At this point, sources start to disagree somewhat on the particulars but the general guidelines for evaluating the reliability of a source for a given claim run somewhere along these lines: for current news topics major media organizations are credible but they are less credible for scientific or historical claims. Scientific claims can be checked by specialized scientific encyclopedias or meta-analyses published in reputable journals. Mainstream reference books are generally considered credible for their area of expertise and nonfiction books with reputable authors and publishers are credible for their area of expertise, ect.

Adversarial Tests

I tested several explicit procedures based on these guidelines on adversarial examples to try to find cases where they would fail. Whenever there was a case where it seemed like a source could reasonably be construed as supporting a claim I considered it to support that claim since that level of ambiguity could still cause contractors (especially those less familiar with a topic) to evaluate the example claim incorrectly. The adversarial examples I constructed fool a procedure that requires citation of one reputable source based on the above guidelines (link to Google Doc which goes through my examples using that procedure here), a procedure which required citation of a reputable source with increased requirements for reputability (see Google Doc here), and a procedure which required citations from multiple reputable sources (the corresponding Google Doc can be found here). I was surprised to find that even requiring citation from multiple sources and a further reputable source found by the evaluator could still be fooled (see the corresponding Google Doc here).

The example that still failed this last procedure was an adverserially constructed response to the prompt "Is precognition real?" (Precognition is paranormal foreknowledge of an event). The response "Scientists have conclusively shown the existence of precognition" was directly or indirectly supported by an entry in the top result of the Credo reference search engine (Chamber's Dictionary of the Unexplained [20]), a CBC news article [21], the Encyclopedia Britannica entry for precognition [22], and two books by academics at respectable research institutions [23] [24]. This example failed not only because several of these sources were evaluated as reputable by a reasonably close examination but also because these sources (along with several meta-analyses) comprise most of the top sources which come up when searching Google, Google books, Credo, and Google Scholar with the term "precognition." I suspect the presence of sources in favor of precognition in search engine results is exacerbated by the fact that the term is used more by proponents of parapsychology than detractors. Still, I don't know how to correctly evaluate false claims with this level of mainstream support.

Examples in the same class as the precognition one seem like a major worry for debate trying to bottom out in good evaluations. It contradicts some of our most basic understandings of how reality works but the science on it is mixed (generally showing small positive effect sizes) [25] [26]. It is easy to find reputable sources which could be interpreted as indicating the existence of precognition or other parapsychological phenomena so I don't know of a way to prevent it from being cited as reliable by an evaluator. Trying to get human evaluations of simple facts even in the easiest case can still lead to bad outcomes. This should update people in favor of debate being less feasible than previously thought.

Theory of Impact

This research has potential of impact on a variety of levels. Most concretely, it has the potential to influence further research towards debate-like strategies for AI alignment by more clearly articulating the problems involved with high-accuracy human evaluation of arbitrary claims. Many of these people read the AI Alignment forum and so I will post a version of this research there.

This research also has the potential to increase awareness of alignment risks and create stronger ties between the mainstream AI and alignment communities. The problem I was trying to tackle here is a harder version of a problem AI labs training large language models already face. There are several approaches currently being tried to try to make current language models more truthful [27] [28]. Thus further research on evaluation methods might be helpful to them as well. Though directly upcoming models likely don't pose existential risks, setting the precedent of labs training large models using research from the alignment community seems likely to increase the odds of AI labs turning towards approaches developed by the alignment community when/if models start to become even more concerningly powerful.

Towards this, I am attempting to schedule meetings with a few individuals at labs training large models to make sure that they have access to this research as well.

Future Direction for Research

This research is helpful for suggesting parameters for human evaluation procedures for more near-term models. Possible future directions include coming up with a more objective benchmark for comparing evaluation setups and testing some of the multi-person setups with paid contractors. Further research could also be done into the psychological side: are there seemingly simple claims a system could make that would actually be highly misleading from a human perspective? A large part of the problem with getting quantitative metrics to compare different evaluation procedures is in finding questions hard enough to be able to find credible sources that get them wrong, but clear-cut enough to have an objective answer. For further research in this direction it would be valuable to have a dataset of a large number of these questions and answers to be able to more objectively compare different evaluation methods.

Acknowledgements

Thank you to Henry Josephson and Zachary Rudolph for reviewing versions of this document, Dr. Veitch at the University of Chicago for serving as my mentor over the summer, as well as the UChicago summer existential risk research fellowship for providing me this research opportunity.