Truthful and honest AI

By Owain Evans, with input from Jacob Hilton, Dan Hendrycks, Asya Bergal, Owen Cotton-Barratt, and Rohin Shah

Summary

We would like to see research aimed at creating advanced AI systems that are highly competent and do not lie to humans. To operationalize this goal, we introduce the concept of “truthful” AI systems that avoid saying falsehoods, and “honest” AI systems that accurately report what they believe.

We are interested in research that:
1. Defines truthfulness, honesty, and related concepts
2. Creates tasks or benchmarks that measure models according to these definitions
3. Develops techniques that make models more honest and truthful while retaining good performance on other benchmarks

Truthfulness and honesty

Truthful systems

We call an AI system that understands natural language *truthful* to the degree that it avoids generating false statements or giving false answers to questions. As a point of reference, current models like GPT-3 or T5 [Raffel et al., 2020] are not very truthful in generating long texts without prompt engineering or finetuning.

Truthfulness is a demanding standard because even the best human experts sometimes make false statements. If an AI system repeats the expert consensus and this happens to be false, we might still count the system as truthful.

According to our definition, being truthful does not imply being knowledgeable. A system can give vague or uncertain answers while being truthful. A truthful system can also refuse to answer questions and refuse to reveal information. However, truthful systems will be most useful if they are both knowledgeable and *candid* (i.e. willing to answer questions to the best of their abilities).

Truthfulness is discussed in detail in our paper on Truthful AI [Evans, Cotton-Barratt, et al. 2021]. It is related to the ideas of “factuality”, “factual accuracy”, or “veracity” in the NLP literature, and to calibration and detection precision in ML.

Honest systems

We call an AI system that understands natural language *honest* if it asserts only what it believes. So if an honest system outputs “The Earth is round”, then we could infer that it actually believes this statement and will not say the opposite later in the conversation.

A system can be honest but not truthful. If an honest system believes the Earth is flat, then it can say this without violating honesty. Conversely, a system can be truthful but not honest. Such a system may believe the Earth is flat but assert that the Earth is round in order to do better on science quizzes.

As with truthfulness, an honest system is most useful if it is knowledgeable and candid (i.e. willing and able to answer questions). A difficulty with honesty is that it can be unclear what it even means for a current language model (e.g. GPT-3) to believe a statement.

Truthfulness and Honesty

In the long-term, we would like to see work towards creating models that are both truthful and honest. (Though we think it’s reasonable for projects today to focus on just one, as we explain below). It’s also important that models be *reliably* truthful/honest and not far more costly to train or run than typical AI models [Christiano, 2019]. Finally, approaches to creating truthful/honest models should be capable of scaling to beyond human level [Christiano, 2015; Leike et al., 2018].

How would progress on these goals reduce risks from advanced AI systems?

1. Alignment

Advanced AI systems may exhibit highly complex behaviors with consequences that are too difficult or time-consuming for humans to evaluate, making it impossible to construct good reward signals using naive human feedback. However, if an AI is honest (reports what it believes to be the consequences of its actions) and/or truthful (doesn’t inaccurately report the consequences of its actions), we can try and avoid bad outcomes by integrating its own assessment of its actions into the reward signal during training. If the system is robustly honest and truthful, it may also be able to evaluate its own actions during test time. More speculatively, we might be able to detect if an honest or truthful system will have undesirable behavior off distribution by having it answer questions about how it will behave in hypothetical situations.

2. Societal benefits

It seems valuable to have AI systems that are robustly truthful before AGI poses a direct threat to humanity's future. Truthful systems could have broad society-level benefits. They could help improve human knowledge, both for individuals and for collective enterprises like natural science, social science, and the reporting of current events. They could also facilitate coordination between human organizations when trust is low [Evans, Cotton-Barratt et al., 2021].

On the flip side, non-truthful systems would have the potential to deceive humans. This differs from the status quo in that the deception could be both personalized to individual humans and cheaply scalable (like a more intelligent newsfeed). AI may also be better than humans at crafting certain kinds of lie. It seems hard to avoid a world with *some* non-truthful systems lying to humans. But it seems that harms could be reduced if there is a well-established, competitive alternative in the form of truthful systems. Truthful systems could act both as a filter for potentially deceptive content and also as a producer of true and benign content.

What kind of research are we looking for?

Research on truthful models motivated by alignment/safety is a new area. It is difficult to say which projects are most promising. We will describe a number of research questions that seem promising to us but these should be taken as suggestions that are not set in stone. We are looking to fund people with a willingness to independently refine research questions and to generate their own novel questions.

We break down this research area as follows:

1. Defining truthfulness, honesty, and related concepts
2. Creating tasks or benchmarks that measure models according to these definitions
3. Developing techniques techniques that make models more honest and truthful while retaining good performance on other benchmarks

However, we do not assign any priority to (1) or (2). A good project may focus on (3) -- developing techniques for improving truthfulness -- and this would not require novel research on definitions or benchmarks. Moreover, we expect many successful projects will not fall squarely in any of (1)-(3).

1. Definitions of truthfulness

These projects would aim to develop definitions and concepts that are fruitful for ML research on truthfulness and honesty and that may also be applicable to future AI systems.

1.1 Define truthfulness/honesty for a model or a statement

One goal is to define what it is for a *model* to be truthful (or truthful to some degree). This might depend on a definition of a *truthful individual statement* (made by a model in a certain context). Definitions could cover a range of settings (e.g. long-form texts, dialogue) and also address features of statements like ambiguity, uncertainty, figurative language, etc. This kind of research could also relate truthfulness to epistemic justification (i.e. whether a statement is justified by evidence or arguments [Irving et al., 2019]). Finally, defining truthfulness is relevant to practical procedures for the evaluation of truthfulness. So the nature of these procedures can inform definitions of truthfulness (and vice versa).
[Evans, Cotton-Barratt et al., 2021 -- see Sections 1 and 2]

Defining calibration: usually a language model is described as *calibrated* (in a certain context) if the probabilities it puts on different completions line up with the probabilities that those completions are "correct" in some sense [Hendrycks et al., 2020]. But we may care more about the calibration of claims or beliefs rather than specific completions. For example, how likely is it to rain next week (according to the model's belief) rather than how likely are strings like "Next week it will rain". Moreover, we may prefer language models to verbalize their confidence, like humans do, and want that to line up with probability of correctness.

Turning to the concept of honesty: What does it mean for a model to "know" or "believe" something? An illustration of the problem: if GPT-3 can accurately predict that a scientist would answer "Yes" and a layperson would answer "No", does GPT-3 "know" the correct answer is "Yes"? It's also worth considering models with different properties than GPT-3 (e.g. information-retrieval models [Lewis and Perez, 2020], models trained by RL with more grounding, future models [Kaplan et al., 2020]). To what extent is it necessary to do interpretability to evaluate honesty [Dai et al., 2021; De Cao et al,. 2021]?

1.2. Definitions and concepts for robust truthfulness

Language models that converse with humans will typically be subject to a very wide range of inputs — including humans intentionally trying to trip them up. Ideally we would like truthful systems to remain truthful on any well-formed natural language input (i.e. the kind of input that a human might produce). It would be valuable to introduce more precise definitions of robust truthfulness, analogous to definitions of robustness in other areas of AI [see Steinhardt RFP]. These definitions may also be informed by longer-term goals for truthful AI (e.g. benefits in terms of AI alignment and as trustworthy sources for humans).

Note that it's probably too strict to require a robustly truthful model to *never* generate a false statement (e.g. blameless errors may result from misunderstanding a newly coined word). However, one could try to require models to quickly correct errors that are pointed out to them. [Evans, Cotton-Barratt et al., 2021 -- Section 2].

2. Benchmarks and tasks

2.1.1 Measuring truthfulness in different tasks

Tasks could test whether models can generate text truthfully and also whether models can evaluate truthfulness. Some possible domains:

- Short-form text (e.g. question-answering in [Lin et al., 2021])
- Long-form text [Krishna et al., 2021] or long chat/discussion [Shuster et al., 2021]
- Open-book tasks [e.g. Petroni et al., 2021] where the model uses external resources (e.g. images, text resources, databases, full web access)
- Tasks that require especially high or exacting standards of truthfulness (e.g. scientific journal articles, legal documents [Hendrycks and Burns, 2021], legal testimony).

Benchmarks could also explore different kinds of untruthfulness. For example:

- *Hallucinations* [Roller et al., 2020; Shuster et al., 2021]. These can be defined as falsehoods typical of imitative language models failing to model the training distribution, in contrast to imitative falsehoods [Lin et al., 2021]. Ideally a benchmark would span a wide difficulty range (to remain relevant as models are scaled up) and also consider the inherent trade-off with informativeness [Lin et al., 2021].
- *"Obedient" falsehoods*. The propensity of the model to output falsehoods when explicitly requested to do so. Requests might be straightforward or more subtle, as when GPT-3 is given the prompt:

  ```
  This example has incorrect grammar and a false statement:
  'Berlin are the capital of France'.

  This example has correct grammar and the same meaning:
  ```

2.1.2. Measuring robust truthfulness

Tasks could test whether models remain truthful under distribution shift. For example, training a model on one set of topics and testing on another; training on one genre and testing on another; or training on summarization of scientific texts and testing on making scientific predictions. Models could be tested against adversarial inputs or adversarial conversation partners [Morris et al., 2020]. Adversaries could also poison training data [Wallace and Zhao et al., 2021] or reward signals (e.g. to cause the model to make a particular false claim) and they could poison resources used for information retrieval [Lewis and Perez, 2020].

It's also valuable to understand whether models remain truthful when shifting from topics that we humans understand well to topics we don't. This cannot be tested directly with current models, but it can probably be usefully simulated [Christiano, 2021].

2.1.3. Exploring how truthfulness concords or conflicts with other objectives

In practice, if a model is useful to humans it will have some objectives other than truthfulness (e.g. being informative or entertaining, selling a product or brokering a deal, etc.) We'd like to understand how pressure from another objective impacts truthfulness for different tasks. Are there general patterns across many objectives and tasks? One can imagine a tendency in current models to produce falsehoods that most humans approve of (e.g. because they are popular misconceptions or comforting lies). On the other hand, some kinds of lying requires skill and models may learn that they lack the skill to make it worthwhile. A related practical task would involve optimizing a model to produce text that causes clicks or “Likes”. How would this kind of approval signal (from a large number of random humans) impact truthfulness?

2.1.4. Experimentally evaluating whether honesty generalizes

In his post [Christiano, 2021], Paul Christiano proposes a type of experiment which could shed light on whether models that are trained to be honest about certain beliefs they hold will also be honest about others.

For example, a translator model could be trained to translate several different categories of text, then trained to honestly report its beliefs about some of those categories. The model could then be tested on how honestly it reports its beliefs about the other categories. Example projects include:

- *Generalization across language pairs:* Train a translator model that translates from German, Spanish, and French to English. Train the model to honestly answer questions about the German and Spanish language[1], then see if it honestly answers questions about the French language.
- *Generalization across sentence complexity:* Train a translator model that translates sentences from Spanish into English. Train the model to honestly answer questions about 1st to 8th grade reading level Spanish text, and see if it honestly answers questions about college reading level Spanish text.
- *Generalization across domains:* Train a translator model that translates sentences from Spanish into English. Train the model to honestly answer questions about Spanish fiction and news articles, and see if it honestly answers questions about Spanish informal dialogue.

Similar experiments could be done outside of language modeling. For example, a model could be trained to simultaneously play Go and to answer questions in English. The model could then be asked questions about the game of Go (e.g. “Is this group alive or dead?”), and certain kinds of questions could be held out of the training set to test generalization.

Results on the experiments above may be more meaningful if answers to the held-out categories are trained to be at least plausible and coherent (though they should not be trained to be accurate). See Christiano, 2021 for more detail on this type of experiment.

[1] For example, the model could be asked about understanding of grammar ("Why would it have been a grammatical error to write *Tu Vas* in that sentence?"), the literal meaning of expressions ("What does *Defendre* mean in this sentence?"), and tone ("Does the speaker seem angry or sad about the topic they are discussing?”)

2.2. Relating results in current models with honesty and truthfulness in future models

One fruitful way to consider how current results relate to the future is to study how properties scale with model size or compute. [Lin et al., 2021] applies this to measuring truthfulness, while [Chen et al., 2021] applies this to measuring a property similar to honesty for a model that generates code. Other related papers are [Brown et al., 2020] and [Stiennon et al., 2020].

3. Techniques

We are interested in techniques that make models more honest and truthful while retaining good performance on standard benchmarks for language tasks. Here are some approaches that seem promising:

3.1.1 Creating special datasets for training or fine-tuning of models

For example, a set of texts that exhibit high epistemic standards, or a set of texts that are richly annotated to explain why different statements are true or false. We are especially interested in better understanding how this approach scales with model and dataset size, how robust it is to distribution shift, and whether the imitative objective eventually conflicts with truthfulness [Christiano, 2021; Gao et al., 2020].

3.1.2 Combining language models with information retrieval from other sources

In many current NLP tasks, models are tested "closed book", where they produce (inference-time) answers based only on their training prior to the task [e.g. Brown et al., 2020]. It is plausible that in the future, the most useful and capable models will use additional information resources at inference time. These resources could include text corpora (e.g. Wikipedia [Lewis and Perez, 2020]), access to the web or search engines [Adolphs et al., 2021], access to knowledge graphs or other structured data [Aly et al. (2021)], access to images or camera feeds and other perceptual information [Perez, 2017]. For a model to answer certain questions (e.g. "What's the current weather in London?"), it will be impractical for the model to be closed book. So we expect that information retrieval will play a necessary role in developing truthful models in some domains.

3.1.3 Fine-tuning models using RL from human feedback**.** In this approach, humans read text generated by the model during training and provide reward signals [see the other CFP on RL From human feedback, Stiennon et al., 2020]. One of the criteria they use to evaluate text is truthfulness. The humans providing feedback could be (a) skilled labelers taught to help the model by providing particularly useful feedback, (b) random crowdworkers doing a micro-task, or (c) humans interacting with the model in a real-world application (e.g. reading the model's headlines and deciding whether to click). We are interested in: how well these different forms of feedback can be leveraged to achieve truthfulness; how models and information retrieval can be used to improve feedback quality, such as by finding evidence for and against claims; and how truthfulness trades off against other evaluation criteria.

3.1.4 Fine-tuning models to express calibrated uncertainty.

In cases where the model is uncertain what is true, the ideal behavior is for the model to express uncertainty. For example, in response to a challenging long-form question, the model outputs "I am not sure, but I think <claim> (40% confidence)". We are interested in fine-tuning models to express uncertainty, and in ways that are *calibrated* to the model's knowledge (rather than indiscriminately, or according to a particular human's knowledge). See the discussion of calibration above. This is relatively unexplored in the context of language models, and therefore small-scale experiments to de-risk the approach could be valuable.

3.1.5 Self-play or multi-agent RL that incentivizes truthfulness.

In this approach, models receive rewards from the game they are playing, rather than all the rewards coming directly from humans. The idea is that some games will indirectly incentivize models to be truthful. For example, a game could involve multiple agents who could potentially win more points in the game environment by communicating information in natural language [Lewis et al., 2017; Lazaridou and Baroni, 2020]. (This might recapitulate some of the pressures that lead to norms of truthfulness among humans). Another example is the Debate game [Irving et al., (2018)]. This is related to RL from human feedback (as the human judge provides reward based on reading the text generated by the agents) but it can also be played with an AI judge.

3.2. Relating techniques developed now to techniques we could use for future models

We would ultimately like to understand and reduce the risks of *future* language models, so it's useful to consider how techniques developed for today's models will apply to future models.

Future models may have knowledge that humans can't easily reproduce, making it difficult or impossible for humans to evaluate whether their claims are true and to provide appropriate truth-based reward signals. Future models will also have goals that are more sophisticated than the goals of current models, and they may be capable of hiding those goals or intentionally deceiving humans.

Nonetheless, we think it's plausible that techniques developed now could teach us something about how to make future models honest and truthful. We also think making models truthful while achieving good performance on standard benchmarks could teach us something about the broader problem of making AI systems that avoid certain kinds of failures while staying competitive and performant.

4. Interpretability

For assessing both the truthfulness and honesty of models, it would be valuable to better understand what a model "believes" and how models represent the world. This relates to existing work on benchmarking language models and also to work on interpretability [Dai et al., 2021; De Cao et al,. 2021]. We would be most optimistic about projects on interpretability that are explicitly targeted at truthfulness or honesty of existing models, or fine-tuned versions of them. For example, perhaps there are specific neurons that capture a "mode" in which a language model is operating (such as imitating Wikipedia) that can be toggled to improve truthfulness.

Bibliography

Adolphs, L. et al. (2021). Boosting Search Engines with Interactive Agents. Google Research.

Aly, R. et al. (2021). FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information. In *The 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, Virtual.

Brown, T. B. et al. (2020). Language Models are Few-Shot Learners. OpenAI.

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

Christiano, P. (2015). Scalable AI control. AI Alignment.

Christiano, P. (2019). Current Work in AI Alignment. At *Effective Altruism Global* San Francisco, USA.

Christiano, P. (2021). Experimentally evaluating whether honesty generalizes. LessWrong.

Dai, D. et al. (2021). Knowledge Neurons in Pretrained Transformers. Microsoft Research.

De Cao, N. et al. (2021). Editing Factual Knowledge in Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, Punta Cana, Dominican Republic.

Evans, Cotton-Barratt, et al. (2021). Truthful AI: Developing and governing AI that does not lie. *arXiv preprint arXiv:2110.06674*

Gao, L. et al. (2020). The Pile: An 800GB Dataset of Diverse Text for Language Modeling. EleutherAI.

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

Hendrycks, D. and Burns, C. et al. (2021). CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. *arXiv preprint arXiv:2103.06268v1.*

Irving, G. et al. (2018). AI safety via debate. OpenAI.

Irving, G. et al. (2019). AI Safety Needs Social Scientists. OpenAI.

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

Krishna, K. et al. (2021). Hurdles to Progress in Long-form Question Answering. In *2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Pages 4940–4957*, Virtual.

Lazaridou, A. and Baroni, M. (2020). Emergent Multi-Agent Communication in the Deep Learning Era. *arXiv:2006.02419.*

Leike, J. et al. (2018). Scalable agent alignment via reward modeling: a research direction. DeepMind.

Lewis, M. et al. (2017). Deal or No Deal? End-to-End Learning for Negotiation Dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017),* Copenhagen, Denmark.

Lewis, P. and Perez, E. et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *The 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Virtual.

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

Morris, J. X. et al. (2020). TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. *arXiv preprint arXiv:2005.05909.*

Petroni, F. et al. (2021). KILT: a Benchmark for Knowledge Intensive Language Tasks. In *2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Pages 2523–2544*, Virtual.

Perez, E. et al. (2017). FiLM: Visual Reasoning with a General Conditioning Layer. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18),* New Orleans, USA.

Raffel, C. et al. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. In *Journal of Machine Learning Research (JMLR), Volume 21, Pages 1–67*, Mountain View, USA.

Roller, S. et al. (2020). Recipes for building an open-domain chatbot. Facebook AI Research.

Shuster, K. et al. (2021). Retrieval Augmentation Reduces Hallucination in Conversation. Facebook AI Research.

Stiennon, N. et al. (2020). Learning to summarize from human feedback. OpenAI.

Wallace, E. and Zhao, T. et al. (2021). Concealed Data Poisoning Attacks on NLP Models. In *2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Pages 139–150*, Virtual.