Published in Aleph Alpha Blog

Aleph Alpha
Aug 17 · 8 min read · Listen

Alignment and emotions for humans and AI: If we want to trust AI, we may need to hurt it

AI becomes meaningfully more capable almost every month. With this new fidelity and transformative use-cases, society is rethinking the human-machine-collaboration and ethical alignment. These conversations need to involve a wider audience and holistic perspective, which sometimes brings along a misunderstanding about how these AI systems work: Modern large AI models show impressive results, but they don't have any agency, nor do they hold any views. The lack of agency becomes obvious, for example, when trying to find consistency in the model output in facts or even opinions. Prompting Aleph Alpha's luminous model for the most beautiful German cities, even with two very similar prompts, we get different results. Of course, this is not a question that can even be answered objectively — as is clear from a human perspective. A language model cannot think in categories of subjectively-learned patterns and objective fact: A question about grammar, style, taste, or factual knowledge is no different — for the language model, every text it writes is learned probability distributions.

The most incredibly beautiful German towns are: the city of Heidelberg and its castle, Rothenburg ob der Tauber with its medieval streets. In Bavaria you will find Munich, capital of state in Germany, as well at other cities like Regensburg or Passau that offer a lot of attractions for tourists.

Of all the German cities of varying sizes the most beautiful are: Berlin and Hamburg. They both have a lot of history, but they also offer modern attractions.

Example of large language model (here: Aleph Alpha's Luminous) reacting to the question about the most beautiful German cities

These deep learning models show all their capabilities: correctly reacting in logical situations, answering questions, and writing summaries simply by internally solving the question — what is the most likely new text to follow in any given context. The answer to this question is built on the structure and interdependencies the model has seen during their training. The training observations include all kinds of taste, outdated misinformation, and offensive speech and language. The model can easily pull things out of context — it is not impossible that documents in the training data set describing the horrors of World War II and the Nazi regime may lead the model to using the word 'Nazi' in a context where we would prefer it would rather not.

Alignment research is trying to prevent bad outcomes and making sure human preferences are met One approach for alignment and safety is trying to change the world that these world models know of by filtering training data or fine-tuning the model so that it behaves more according to our goals. I believe this can help in some cases. But eventually, this is a wrong approach that will not only fail to achieve the goal and introduce new risks: A model that has never heard about Nazis won't know how to behave if it encounters that concept. In addition to introducing new risks, this approach will force the ideology of whoever builds the model onto all users — without transparency about how the creator's world view and values have influenced the models. Maybe because the native European perspective is more pluralistic, I believe responsible AI should do the opposite: transparent and non-ideological (as neutral and adaptable as possible).

Strong convictions — loosely held

Large language or multimodal models are limited world models, modeling the world of human-created language. Humans are creating and using world models. These are not unlike the huge AI models; understanding patterns and structure. There is evidence that our whole human perception and consciousness is just the output of our individual world models influenced by sensory input and learned interdependencies. We know of different preferences for German cities, and we wouldn't be surprised if a stranger would express certainty about any of the towns above being the "most beautiful".

For us, however, there is a significant difference between things we would consider "likely", "probable", or "possible" and our actions and convictions. In any given situation, we judge our actions and the actions of others not by their likelihood but by the correctness and (deontological) ethics.

It seems possible that humans act with a direct output of their world model. I like the analogy of "System 1" thinking here. Maybe when drunk or unconcentrated, we might just "instinctively" act with learned behavior and patterns. Of course, these world model outputs are not just built by observation but by our actions and thoughts. A human world model is a lot more than all inputs/ observations of the past. Every human is an essential part of their own world, and a crucial aspect of that world model is to learn the interdependencies between it's creator and the environment. Our world model contains the results of past actions and forces, making them habitual and internalized. This world knowledge includes learned and hard-coded emotional links to states and activities that can be retrieved and shaped similar to perceptual and factual patterns.

There are forces beyond likelihood-estimation within humans that are always active (and vital) influencing our actions in any given situation. Let's look at how we could build an AI system that shows similar behavior and agency — what are the puzzle pieces we might need?

From the map to the expedition

Machine learning researchers have successfully built agents acting in a moderately complex environment based on world models. Those agents have a functional component that is using the world model to plan, evaluate and predict. The value functions driving this have to come at least partly from outside the world model that contains information about values, norms, and emotions but does not weigh them.

For example, I am prompting luminous with an unpleasant situation, and one of the possible completions is rather confrontative.

Example of a language model (here: Aleph Alpha's Luminous) completing a prompt about human conflict — here with a more confrontative solution

We would probably all agree that this outcome is possible, but is it desirable? Asking luminous, we can retrieve some information about the link to emotions, norms, and potential consequences from the world model:

The AI's generated solution above can be judged from a human perspective by the model itself: What might be a human reaction that follows? Why?

The learned link to human values can provide vital input to choose the right course of action (weighing pros and cons), but there is a crucial step missing: This is a variant of the classical is/ ought problem — the world model tells us what is (and might be) while giving us little help determining what we should strive for.

In reality, the lines are blurry, the best model of human behavior is debatable, and there are many great perspectives on this. Thinking about how I would replicate human capabilities and forces in AI, this architecture is perhaps a useful first draft:

rough functional idea for a system with human-like capabilities

Comparing these boxes with models like Aleph Alpha's luminous or GPT-3, only a small subset can reasonably claim to be within reach of these models (here colored in yellow).

Building blocks of Trust

The capabilities of large (language) models are intensely debated currently — however mostly with the focus on "intelligence" and cognitive abilities. These conversations relate to the thoughts presented here in some essential ways. The design displayed may provide a perspective on which functional blocks are missing, and how to build something with comparable effect. There is an interesting discussion about emergent properties, hypothesizing that AGI may eventually emerge just by scaling systems (without any feature engineering). Given the experience with deep learning in the last years this seems not impossible. This article is not making the case for the need for feature-engineered functionality as the only way to achieve the desired effect. These components may turn out to be just a way we understand functionality.

Feelings

- Feelings help us to come up with a value judgment for current or possible future states, our actions, and the actions of others
- This is arguably even the case for rationally driven values and goals or charitable work. Ancient Greece had an interesting concept for a positive emotion driving correct actions as eudaimonia
- Emotional links seem to come from
  a) Biology/ evolution
  b) Culture/ epigenetics
  c) Personal experience
  d) Personal development
- Emotional evaluation is strongly driven by potential future outcomes. The current models contain some knowledge about this, so it might be possible to add an emotional module that has some hard-coded human-like values
- Feelings not only help to come up with the right action but are also driving the construction and update of our world model as curiosity and love for art or dreams that replay scenarios we should (emotionally) be paying attention to

Logic & conceptual thinking

- Humans seem to have developed potential for symbolic reasoning. We invented math and logic to solve problems that cannot be approached by similarities and observation alone. The rational mind generates a signal for factual correctness
- Even small children seem to be able to map observations and thoughts into a conceptual space that allows for logical processing
- This is a recently hotly debated topic. At Aleph Alpha, we are currently working on something (hopefully) unique here that I don't want to spoil.

Thinking loop (init and return)

- There seems to be a trigger that can start a system-2 "thinking loop". Once the process is active, the world model incrementally builds new context by combining knowledge and logic. With the context growing in complexity, operations are possible that would not be possible in a single forward pass/ thought. For thoughts that surpass mental capacity, tools like a piece of paper or computer can be useful. An additional criterion can tell us when a problem is solved, returning a result and exiting the process.
- For reinforcement learning agents, typically, this is not optional. They do some planning based on learned knowledge all the time and for every single action. This reduces complexity but does not allow for some powerful human features:
  a) fast and effortless action in a calm and harmless situation
  b) dynamic effort and time allocated based on the severity of the decision

"You're entering a world of pain"

Current large (world) models are "untrustworthy". These models are neither consistent truth machines nor acting based on any (deontological) values. Giving them a sense for what's desirable in the world, and in action may teach them many of the behaviors that for humans are an essential part of intelligence. Combined with ways to iteratively "think" (build context recursively with the help of learned patterns) learned values would allow for reliable navigation in a complex environment guided by structure in world models. Human guiding emotions are connected to the perceived and imagined world and self, existing on a different level of abstraction — allowing us to handle even radically new situations mostly consistent. These tools for navigating possibilities are missing in today's large AI-models. No level of data distibution shift will eventually solve this: if we want to trust AI we may need to hurt it.

. . .

Author: Jonas Andrulis (Founder & CEO of Aleph Alpha)

This article was originally published on April 9, 2022 on http://andrulis.tech/index.html

5

5

More from Aleph Alpha Blog
Follow

This is the official Aleph Alpha Blog to keep you up to date to all topics related to the research and development of AGI.

Aleph Alpha · Jul 28
Ethics and bias in generalizable AI — Hate the player, don't hate the game
Since 2008, the second wave of AI applications has brought forth surprising functionality and raised expectations to the point where eve…
Research 10 min read

Share your ideas with millions of readers. Write on Medium

Aleph Alpha · Jun 13
AGI and knowledge: we have ways of making him talk
Our conception of information and data in IT Systems is heavily shaped by the capabilities and design of these systems. In a traditional system, it is easy to differentiate between the data, application logic and user…
Artificial Intelligence 6 min read

Aleph Alpha · Jun 9
German startup Aleph Alpha raises $27M Series A round to build 'Europe's OpenAI'
With Microsoft now being an investor in OpenAI the field is more open for new insurgents into the open-source AI arena. Now a German…
Press 3 min read

Aleph Alpha · Jun 8
AI hardware: data was never the new oil
Artificial Intelligence (AI) technology has long been predicted to have a massive impact on the economy, international relations and society at large. In recent years, these promises have started to be fulfilled, and…
Artificial Intelligence 7 min read

Aleph Alpha · Feb 14
The end of the era ImageNet
World-models are overtaking specialized and supervised models in performance and generalizability. With our new research we lift this development from the language space (with models like GPT-3) into a…
AI 4 min read

Read more from Aleph Alpha Blog

Get unlimited access

Search

Aleph Alpha
35 Followers
We are an independent European company researching, developing and operationalizing a new foundational AI technology for the public and private sector
Follow

More from Medium

Jan Marcel Kezmann in MLearning.ai
Google Colab Pro Vs MacBook Pro M1 Max 24 Core

Jim Clyde Monge in MLearning.ai
How To Turn Yourself Into Anything With MidJourney's V4 Model

Alberto Romero
GPT-4 Rumors From Silicon Valley

Jesus Rodriguez in Towards AI
DeepMind's Idea to Build Neural Networks that can Replay Past Experiences Just Like Humans Do

Help Status Writers Blog Careers Privacy Terms About Text to speech