Notes on Cappelen and Dever, 17/04/2023

Background

GOFAI==>neural networks+big data==>transformers+embeddings==>today

Turing test (behaviour)==>Chinese room (intentionality)==> ? (?)

Cappelen and Dever

What can AIs do? What can, to be precise, ChatGPT do? How much are they like us? Do they share properties with us, or have variations on properties we have? If so, can we broaden our conception of properties so they apply to both AI and us. Does AI have beliefs and desires?

An important note: "ChatGPT is a competent speaker of English. That view is presupposed in this chapter." (*does a creature who can speak but who can't believe or want imaginable?)*

Two quick arguments. ChatGPT ascribes itself beliefs; and others ascribe it beliefs. So, it has beliefs. Can you think of problems with this argument?

The question is **what needs to be added to contentful representational states to make them into beliefs**. Representations+ ? = Beliefs

1. Beliefs aim at the truth

Beliefs aren't just content. They can also be evaluated. A belief is good if it's true and bad if not. C&D hold that by virtue of being trained on bodies of text whose speakers (typers) aim at truth, ChatGPT's states can also come to aim at truth.

2. Beliefs are integrated and interconnected.

Beliefs bear relations to each other. "Someone who believes that if it’s raining, the streets are wet, and who also believes that it’s raining, is at least positioned to form the belief that the streets are wet."

3. Beliefs are connected to action

We 'solve for' belief and desire simultaneously in the classic belief-desire psychology model:

> States S1 and S2 of agent A are a belief that p and a desire that q, respectively, if when some action C best promotes that q on condition that p, A will ceteris paribus perform C as caused by S1 and S2.

If we don't have desires, we don't get beliefs. Do we have desires?

FIRST BLOCKER: ChatGPT has no states that aim at the truth.

Objection:

> Aiming at the truth is no part of what ChatGPT is doing. It's just stochastically creating probable text continuations of input texts. It isn't trying to say anything true or false, and in fact has no concept of truth or falsity. And even if it did, it has no means of checking what it says against the world, so wouldn't be able to strive for truth even if it knew what truth was and wanted to achieve it.

An interesting thought:

> The intentional design doesn't spring from the intentions of the programmers/creators of ChatGPT, and doesn't involve intentions directed at ChatGPT. What is intentionally designed is the text corpus that's the training set for ChatGPT. There's no global intention directed at the corpus as such, of course. (Or if there is, it's the intention of the programmers to use that corpus as the training corpus. That intention also plays a role in creating the truth-tracking aim, but it's a minor role.) Rather, there are millions and millions of individual intentions, intentions directed at the individual bits of text in the corpus, intentions to say something true in producing those texts. Those millions of intentions are then causally effective in bringing about a largely-true corpus. And training on a largely-true corpus is then causally effective in creating a ChatGPT-FInal that in fact produces largely true text. But those millions of intentions are also constitutively effective in making it the aim of a neural net -- one whose developmental history constitutively makes it a producer of texts with certain similarity relations to the training corpus -- into a neural net whose states aim to track the truth.

SECOND BLOCKER: ChatGPT's representational states aren't integrated and interconnected.

Suppose that ChatGPT represents the world as being such that Berlin and Frankfurt are in Germany. (We'll take ourselves now to be licensed to say that ChatGPT represents the world as being such that X, rather than just saying that ChatGPT represents that X, because we've now argued that ChatGPT's representational states aim for truth.) Can ChatGPT infer from those representings that Berlin is in Germany?

It doesn't look like ChatGPT has any internal mechanisms for making this kind of representational transition. If we ask ChatGPT where Berlin is, it may respond by saying that Berlin is in Germany. But the mechanisms here is not:

- ChatGPT has a state S1 representing the world as being such that Berlin and Frankfurt are in Germany.
- On the basis of being in state S1, ChatGPT enters into another state S2 that represents the world as being such that Berlin is in Germany.
- On the basis of being in state S2, ChatGPT produces the text output "Berlin is in Germany'', and thereby says that Berlin is in Germany.

Instead, there are just holistic features of the weighted neural net of ChatGPT that dispose it to respond to certain kinds of text prompts by responding "Berlin is in Germany". We've already argued that these holistic features in combination with the causal history of ChatGPT, the way we interact with ChatGPT, and other external features of ChatGPT may be enough to constitute ChatGPT's representing Berlin as being in Germany. But the current problem is that the holistic features don't seem to be integrated with the holistic features that similarly encode ChatGPT's representing Berlin and Frankfurt's being in Germany.

Response:

Externalized Interaction: The interaction happens first in the training of ChatGPT on the data set -- the training procedures that dispose ChatGPT to produce "Berlin is in Germany" continuations are connected to and interact with the training procedures that dispose ChatGPT to produce "Berlin and Frankfurt are in Germany" continuations, and it's these continuation dispositions that (partially) ground ChatGPT in representing things about Berlin, Frankfurt, and Germany. And the interaction then happens secondarily in the data set itself -- the presence of Berlin-and-Frankfurt-relevant training data in the data set is dynamically connected to the presence of Berlin-relevant training data, because the production of the text that constitutes the data set is an open centralized domain in which the presence of one kind of text leads to the presence of another kind of text. (That's just the truism that some of the things we write give rise to other things we write.) And the interaction happens

tertiarily, and ultimately, in us and our interactions with the world. We are a central domain in which the presence of Berlin-and-Frankfurt-relevant information is integrated with the presence of Berlin-relevant data.

A picturesque way of putting it: ChatGPT does indeed infer that Berlin is in Germany from its representing Berlin and Frankfurt's being in Germany. But it makes the inference through us. It defers onto us the task of organizing bodies of information, and then by coordinating with us (through the training procedure) acquires the relevant correlating tendencies. Maybe that's not paradigm human inference, but it's a suitably de-anthropocentrized inferring. (And Familiar Examples: is this really wholly alien to our experience? We don't always privately supervise the inferential closure and consistency of our beliefs. Rather than just reading the Peano axioms and then privately running the cognitive dynamics of extending belief in those axioms to beliefs in a wide range of mathematical truths, we offload the inferential work externally, and just coordinate ourselves with the products of mathematicians. Rather than privately curating our beliefs about social engagements for consistency, we externally encode them onto an appointments calendar and then use geometric features of the external calendar to do the consistency check.)

So, the response to the second blocker is that ChatGPT does have beliefs that are integrated and interconnected. They are just integrated and interconnected in a less humanly-familiar way that routes through mechanisms outside ChatGPT. But that externality is no threat to the representations being beliefs.

THIRD BLOCKER: ChatGPT's representational states aren't connected to action.

Consider Galen Strawson's Weather Watchers:

The Weather Watchers are a race of sentient, intelligent creatures. They are distributed about the surface of their planet, rooted to the ground, profoundly interested in the local weather. They have sensations, thoughts, emotions, beliefs, desires. They possess a conception of an objective, spatial world. But they are constitutionally incapable of any sort of behavior, as this is ordinarily understood. They lack the necessary physiology. Their mental lives have no other-observable effects. They are not even disposed to behave in any way. (Mental Reality, 251)

The Weather Watchers were originally introduced by Strawson as a general counterexample to functionalism. They have all the attitudes, but they undertake no actions -- how, then, can the attitudes be analytically tied to actions in the way that functionalism requires? But we can more generally take the Weather Watchers as a push to envision ways of life more distant from the humanly familiar. Take away the desires and add a tendency to relate what

they've learned about the weather to others -- now we have the Weather Reporters. If the Weather Watchers are comprehensible, so surely are the Weather Reporters.

ChatGPT Has Desires

Does ChatGPT have desires? Well, if beliefs, why not desires? We've already argued that ChatGPT can have representational states and that those representational states can have the kind of organization, architecture, economy, and interaction that's characteristic of specifically mental representational states. Is there then any reason to have special hesitation about desires?

That's a nice statement of theory, but the fact is that people do seem to find the attribution of desires to ChatGPT more dubious than the attribution of beliefs or of tryings. "Look, ChatGPT thinks Shakespeare wrote a play in which both Hamlet and Othello appear" is a natural utterance, as is "And now ChatGPT is trying to convince me that Hamlet was poisoned by Iago." Desire-talk, however, comes less naturally. Perhaps "ChatGPT wants me to get my nephew a Playstation as a birthday present" after requesting present suggestions? But an error theory is more tempting for that talk. Does ChatGPT really want you to do that? What would ChatGPT care what you in the end got as a gift?

Some potentially distinctive features of desire:

(1) Desire is constitutively linked to pleasure and displeasure.

(2) Desire is linked to a theory of the good, so that for an organism to desire that p is for it to take p to be good. Taking p to be good could then be a matter of believing that p is good or having p appear good to the organism.

(3) Desire is attention-focusing. When we are thirsty and thus desire water, we aren't just in a motivational state that prompts water-taking actions when we believe the world to be suitable. We're also on the watch for water. Our attention is directed toward water -- we search the environment looking for opportunities to get water.

Then they argue none of these are indeed essential to desire.