Twitter Thread by ■■❄■Emily M. Bender❄■■■

■■❄■Emily M. Bender❄■■■
@emilymbender

They sure will. Here's a quick analysis of how, from my perspective as someone who studies societal impacts of natural language technology:

> I have no idea how, but these will end up being racist https://t.co/pjZN0WXnnE
>
> — Michael Hobbes (@RottenInDenmark) December 16, 2020

First, some guesses about system components, based on current tech: it will include a very large language model (akin to GPT-3) trained on huge amounts of web text, including Reddit and the like.

It will also likely be trained on sample input/output pairs, where they asked crowdworkers to create the bulleted summaries for news articles.

The system will be some sort of encoder-decoder that "reads" the news article and then uses its resulting internal state to output bullet points. Likely no controls to make sure the bullet points are each grounded in specific statements in the article.

(This is sometimes called "abstractive" summarization, as opposed to "extractive", which has to use substrings of the article. Maybe they're doing the latter, but based on what the research world is all excited about right now, I'm guessing the former.)

So, in what ways will these end up racist?

1) The generator, giving the outputs, will be heavily guided by its large language model. When the input points it at topics it has racist training data for, it may spit out racist statement, even if they aren't supported by the article.

2) If sample input/output pairs will likely have been created by people who haven't done much reflecting on their own internalized racism, the kinds of things they choose to highlight will probably reflect a white gaze which the system will replicate.

[Ex: A Black person is murdered by the police. The article includes details about their family, education, hobbies, etc as well as statements by the police about possibly planted evidence. Which ends up in the summary?]

1&2 are about being racist in the sense of saying racist things. It will also likely be racist in the sense of disparate performance:

3) The system, trained mostly on mainstream/white-gaze texts, when asked to provide a summary for an article written from a BIPOC point of view, won't perform as well.

4) When the system encounters names that are infrequent in American news media, they may not be recognized as names of people, sending the generator down weird paths.

1-4 are all about system performance, but what about its impact in the world?

5) System output probably won't come with any indication of its degree of uncertainty/general level of accuracy/accuracy with the type of text being fed in. So people will pick up 'factoids' from these summaries that are wrong (and racist).

6) System output won't indicate what kinds of things usually make it into the summary, so the already racist patterns of e.g. how Black victims of police violence are described in the press as consumed by readers will get worse (see pt. 2).

I'm sure there's more, but that's my quick analysis for tonight.