On NYT Magazine on AI: Resist the Urge to be Impressed

Emily M. Bender · Follow

22 min read · Apr 17, 2022

[Now available as an "audiopaper" on my soundcloud. (Please excuse occasional noise from airplanes overhead + my inconsistency about whether to render quote marks out loud.)]

Context

On April 15, 2022, Steven Johnson published a piece in the New York Times Magazine entitled "A.I. Is Mastering Language. Should We Trust What It Says?" I knew this piece was coming, because I had been interviewed for it, over email, a couple of weeks ago. I read it with some trepidation, because I had the sense that Johnson's question and goals going into the article did not maintain sufficient skepticism of the claims of AI boosters. At the same time, I was also fairly confident my words weren't going to be taken out of context because I'd been contacted by a fact checker who was verifying the quotes they intended to use.

On reading the article, my expectations were met on both counts. Ordinarily, when I encounter AI hype in media coverage of research/products that claim to be "AI", I get inspired to write tweet threads aiming to educate folks on how to spot and thus resist such hype. (Here's a recent example.) Johnson's article is ~10k words long, though, and so I've decided to try to do the same in blog form, rather than as a tweet thread.

tl;dr

- Yes, there is an urgent need to address the harm being done by so-called "AI" and to set up effective regulation and governance so that those who are impacted by this technology have power over how it is deployed.
- But no, the harms aren't going to come from autonomous "AI" that just hasn't been taught appropriate values.
- And no, the solution isn't to try to build "AI" (or "AGI") faster or "outside" the megacorps. (Scare quotes on "outside" there, because OpenAI isn't really as independent as they claim — given both the source of their initial funding and their deal with Microsoft.)
- What's needed is not something out of science fiction — it's regulation, empowerment of ordinary people and empowerment of workers.
- Puff pieces that fawn over what Silicon Valley techbros have done, with amassed capital and computing power, are not helping us get any closer to solutions to problems created by the deployment of so-called "AI". On the contrary, they make it harder by refocusing attention on strawman problems.
- If you'd like to learn more about what is going on and what shape meaningful solutions could take, I recommend authors such as Safiya Noble, Meredith Broussard, Ruha Benjamin, Shoshana Zuboff, Abeba Birhane, Joy Buolamwini and her colleagues at the Algorithmic Justice

League, and journalists such as Khari Johnson, Edward Ongweso Jr, and Karen Hao (see especially this piece on OpenAI).

On asking the right questions

I am not a journalist, but it seems to me that a key lesson from research must hold in journalism too: In research, the questions we ask shape what we can find and how the conversation of scholarship is advanced. Similarly, in journalism, I would believe that the question a journalist asks in a piece of writing shapes how the public can be educated.

The headline (not necessarily due to Johnson, but I think a not incongruent framing of the article, in this case) asks: "A.I. Is Mastering Language. Should We Trust What It Says?" In taking up this question, the article is not keeping any distance from the point of view of the primary organization it is covering (OpenAI) but rather adopting Open AI's stance towards their own technology wholesale. This headline asserts that "AI" has "mastered language" (spoiler alert: it hasn't) and in the process presupposes the existence of something that can be referred to as "AI". Those who disagree with this assertion (👋) are framed in the article as "skeptics" — more on this below. The second part of the headline "Should We Trust What It Says?" frames "saying" as done by "AI" as analogous to "saying" as done by "people" … and nothing in the article really highlights any difference between the two either.

When thinking about trust, trustworthiness, and the pattern recognition technology that gets billed as "AI", I think there are a lot of valuable questions to be asked, including:

- Why are people so quick to be impressed by the output of large language models (LLMs)? (*Nota bene*: This is not a new observation. It goes back at

least to the way people reacted to Eliza, see Weizenbaum 1976)

- In what ways are corporations leveraging that credulousness, on the part of users, investors, regulators?
- Where is this technology being deployed, and what are the potential consequences? Who is bearing the brunt? (We — Timnit Gebru, Angelina McMillan-Major, Meg Mitchell, our further co-authors and I — talked about various kinds of potential harm in the Stochastic Parrots 🦜 paper, but I would be very interested in journalistic work on actual deployments.)
- What would effective regulation look like in this space? Who is working on that regulation?

With respect to OpenAI specifically, I think it would be useful to ask:

- How is OpenAI shaping the conversation around so-called "AI", as developed by them or others?
- How does OpenAI's rhetoric around "artificial general intelligence" shape public and regulator understanding of claims of other companies, such as those who purport to "predict" recidivism risk, "recognize" emotion, or "diagnose" mental health conditions?
- What are the relationships between OpenAI's staff/board members/founders and other organizations?
- What are the financial incentives at play, and whose interests do they represent (noting OpenAI's "exclusive computing partnership" with Microsoft)?

Buying into the hype

Here are a selection of examples of how Johnson's writing betrays his commitment to the notion that "AI" (or maybe even "AGI"?) is a thing that exists or is on the verge of existing, very much in line with how the folks at OpenAI see things. Some are more subtle, others are more blatant. This section is quite long, because in 10k words, Johnson provided many examples of this. If you get bored, just jump ahead to the next heading…

All quotes here and below are from the NYT Magazine article linked above; all highlighting (via boldface) is added by me.

> *GPT-3* ***has been trained*** *to write Hollywood scripts and compose nonfiction in the style of Gay Talese's New Journalism classic "Frank Sinatra Has a Cold."*

Talking about "training" machine learning systems is the standard terminology. I think it is always worth pausing and considering on what grounds we are talking about "training", "learning" and "intelligence" in these systems; what metaphors are at play; and to what extent we are asked, as readers, to nod along to the metaphor without any clear guidance as to which aspects apply and which are merely suggestive.

In the case of (large) language models like GPT-3, the "training" involves taking a mathematical model with random mathematical parameters ("weights") and iteratively adjusting those weights in response to differences between model output and some point of comparison showing expected output. For GPT-3, the primary "training" is just next word prediction over enormous amounts of text. There is then also a notion of "fine-tuning" where that pre-trained model is shown a small set of "prompts" and expected responses to those prompts. I actually can't tell if "training" in the quote above refers to the general pre-training or to people using that pre-trained model, via fine-tuning, in those different use cases.

> *Others have fed the software prompts that generate patently offensive or delusional responses, showcasing the limitations of the model and its potential for harm if adopted widely* ***in its current state****.*

Here, what I want to point up is how the phrase "its current state" suggests that there is a developmental path that systems like GPT-3 are treading, and the problems that people have noted with GPT-3 (in various use cases) are addressable, somewhere along that path, and not actually fundamental mismatches between technology and purpose.

> *So far, the experiments with large language models have been mostly that: experiments probing the model for* ***signs of true intelligence****, exploring its creative uses, exposing its biases.*

Indeed, people have been probing GPT-3 and its ilk in various ways. I'm not sure how many would say they are looking for "signs of true intelligence" but OpenAI's Ilya Sutskever did infamously muse on Twitter in February:

Tweet from Ilya Sutskever on Feb 9, 2022 reading “it may be that today’s large neural networks are slightly conscious”

I would expect serious journalism looking into this work to ask why anyone would expect GPT-3 and its ilk to “have intelligence” and furthermore to demand a definition of intelligence. Just noting off hand that people are looking into it again supports OpenAI’s (and others’) AI hype.

> *there was a sense that the long “A.I. winter,” the decades in which the field failed to live up to its early hype, was finally beginning to thaw.*

This is an odd use of the term “AI winter”, which is usually used to refer to the lack of funding for AI research that results from over-promising (over-hyping) and then (necessarily) failing to deliver on the hype. This sentence instead suggests that AI is now living up to the hype, which in turn serves to support (and amplify) the hype.

> *But GPT-3’s intelligence, if intelligence is the right word for it, comes from the bottom up: through the* ***elemental act*** *of next-word prediction.*

I guess the most charitable reading of this one is that next-word prediction is all that GPT-3 does in its primary training phase, and as such, next-word prediction is elemental for GPT-3. But “elemental act” is a pretty grandiose term for this, and seems to elevate this banal task to something of deep importance.

> *On the side of emergent intelligence, a few points are worth making. First, large language models have been making steady improvements, year after year, on* ***standardized reading comprehension tests****.*

The second statement here is true, in the sense that scores have been going up. But it is also misleading: just because the tests were designed to test for reading comprehension by people, and even if we assume that they do a good job of measuring that, doesn't mean that comparable scores by machines on the same tests entail that machines are doing something comparable. This comes down to the concept of "construct validity": does the test actually measure some construct which is coherent in itself and effectively measured by the test? To establish construct validity for reading comprehension tests, we need a definition of what reading comprehension is as well as evidence that answering the test questions more accurately corresponds to more reading comprehension. There is no reason to believe, a priori, that a test designed to achieve those ends for humans would be equally effective for machines, because there is no reason to believe that machines use the same processes for answering the questions as humans do. (For more on construct validity in the evaluation of so-called "AI" systems, see Raji et al 2021 or Ben Dickson's popular press coverage of it, as well as Jacobs and Wallach 2021.)

Furthemore, contextualizing this (ultimately insufficiently supported) claim about reading "comprehension" as a "point worth making" about "emergent intelligence" only serves to further emphasize the (misleading) reading of the sentence that it really is something about intelligence, i.e. really something about machines "comprehending" text.

> *"You'll see that it puts the arms and the legs in the right place," Murati points out. "And there's a tutu, and it's walking the dog just like it was a human, even though it's a baby radish. It shows you that GPT-3 really has* ***quite a good conception*** *of all the things that you were asking it to combine."*

Quoting such statements from OpenAI personnel (Murati is a senior VP) uncritically means that NYT Magazine is just platforming AI hype. What does "quite a good conception" mean? The most natural interpretation is something like "thinks abstractly", so this statement boils down to claiming that "it puts the arms and the legs in the right place" establishes that GPT-3 "thinks abstractly". But, again, this claim is made without evidence, and Johnson doesn't demand any. Even if Murati has some other technical definition of "good conception" in mind, how are NYT Magazine readers expected to understand it except as analogous to what humans do?

> *In a way, you can think of GPT-3 as a purely linguistic version of the Cartesian brain in a vat or in a "Matrix"-style cocoon:*

I'm wondering what a "purely linguistic" version of these would even be. I suspect that Johnson (like many others) has mistaken the ability of GPT-3 and its ilk to manipulate linguistic form with actually acquiring a linguistic system. Languages are symbolic systems, and symbols are pairings of form and meaning (or, per de Saussure, signifier and signified). But GPT-3 in its training was only provided with the form part of this equation and so never had any hope of learning the meaning part.

Perhaps Johnson is imagining an existence where he becomes limited to using his own existing linguistic capability to interact with the world, being cut off from all of his senses. But this is not analogous to what GPT-3 is doing — or what anything else, no matter what its internal architecture might be, could do with a "training" regime analogous to GPT-3's. Alexander Koller and I lay out this argument in detail in Bender and Koller 2020 (for popular media coverage, see this piece by Will Douglas Heaven).

So, again, suggesting that the GPT-3 is like any kind of brain is pure hype.

> *GPT-3 and its peers have made one astonishing thing clear: The machines have acquired language.*

This statement is only true if we interpret "acquired language" to mean "have been programmed to produce strings of text that humans who speak that language find coherent". Given all of the surrounding hype, the reader could be forgiven for not being able to find that interpretation, and instead taking it to mean something else — necessarily false.

Also, and this bears repeating, whenever a computer scientist claims to be able to do something with "language" or "natural language", it's worthwhile to ask, "Which language?". In the case of GPT-3, the answer is primarily English, with a smattering of other languages, as happened to have appeared in the training data.

> *Perhaps the game of predict-the-next-word is what children unconsciously play when they are acquiring language themselves: listening to what initially seems to be a random stream of phonemes from the adults around them, gradually detecting patterns in that stream and testing those hypotheses by anticipating words as they are spoken. Perhaps that game is the initial scaffolding beneath all the complex forms of thinking that language makes possible.*

Child language acquisition is a well-established field of study, with rich methodology and results. Johnson's speculation is completely unfounded (and doesn't claim any foundations, being presented as mere musings), and simply serves to bolster the claim that GPT-3 is doing something like what children do, without presenting any evidence to that end. For a very brief overview of what the child language acquisition literature has to say that is relevant to this debate, see section 6 of Bender and Koller 2020.

All that glitters is not gold

One further kind of AI hype that Johnson's piece peddles, and one that is particular to "generative" technology like GPT-3, is a sleight of hand that asks readers to believe that something that takes the form of a human artifact is equivalent to that artifact. Here are a few quick examples:

> *GPT-3 has been trained to write Hollywood scripts and* ***compose nonfiction*** *in the style of Gay Talese's New Journalism classic "Frank Sinatra Has a Cold."*

GPT-3 can't "compose nonfiction". Nonfiction by definition is factual writing about the world. But GPT-3 has no access to facts, only to strings in its training data. To the extent that it outputs strings of words that humans interpret and can verify as factual, that factuality is and can only ever be purely accidental.

> *For instance, even without the kind of targeted training that OpenAI employed to create Codex, GPT-3 can already generate sophisticated legal documents, like licensing agreements or leases.*

Someone who needs to create a licensing agreement or a lease doesn't just need a document that looks and feels like a licensing agreement or a lease. They need something that speaks to their particular situation and is legally binding in their particular jurisdiction. A set of strings that take the form of legalese is not a "sophisticated legal document".

> *… GPT-3's recent track record suggests that other, more elite professions may be ripe for disruption. A few months after GPT-3 went online, the OpenAI team discovered that the neural net had developed surprisingly effective skills at writing computer software, even though the training data had not deliberately included examples of code.*

"Writing software" entails much, much more than generating code, even code that is syntactically correct and compiles. It includes, at least, determining the specifications of the system to be produced and creating tests to ensure that the system behaves as desired.

> *If you gave 100 high school students the same prompt, I doubt you would get more than a handful of papers that exceeded GPT-3's attempt. And of course, GPT-3 wrote its version of the essay in half a second.*

This leaves me wondering if Johnson has forgotten (or never understood) why high school students are asked to write essays. It is not, to be sure, to keep the world's supply of essays topped up! Rather, it is about what the students learn in the process of doing the writing.

OpenAI hagiography

Johnson's NYT Magazine piece frequently strays into straight up hagiography of OpenAI and the people who lead it, as in this characterization of Ilya Sutskever:

> *"Here is the underlying idea of GPT-3," Sutskever said intently, leaning forward in his chair. He has an intriguing way of answering questions: a few false starts — "I can give you a description that almost matches the one you asked for" — interrupted by long, contemplative pauses, as though he were mapping out the entire response in advance.*

Johnson also uncritically presents the OpenAI crew as setting out to save the world:

> *And once again, all the evidence suggested that this power was going to be controlled by a few Silicon Valley megacorporations.*
> *The agenda for the dinner on Sand Hill Road that July night was nothing if not ambitious: figuring out the best way to steer A.I. research toward the most positive outcome possible,*

The hubris in thinking that as a group they would be positioned to "figure that out" is perhaps only surpassed by the naïveté in believing that the way to achieve that lies primarily in … building so-called "AGI" systems. But Johnson reports this ambition uncritically, even admiringly, and doesn't seem to notice (until much, much further down) that there isn't a whole lot of daylight between the leaders of "a few Silicon Valley megacorporations" and the people at that dinner.

> *Today, roughly a fifth of the organization is focused full time on what it calls "safety" and "alignment" (that is, aligning the technology with humanity's interests)*

Again, there is absolutely no evidence presented that the folks at OpenAI are positioned to understand "humanity's interests" (and not just their own/those of people like them) and the idea is presented uncritically until much, much further down the ~10k word piece where Johnson writes:

> *But beyond the charter itself, and the deliberate speed bumps and prohibitions established by its safety team, OpenAI has not detailed in any concrete way who exactly will get to define what it means for A.I. to "benefit humanity as a whole." Right now, those decisions are going to be made by the executives and the board of OpenAI — a group of people who, however admirable their intentions may be, are not even a representative sample of San Francisco, much less humanity. Up close, the focus on safety and experimenting "when the stakes are very low" is laudable.*

> *But from a distance, it's hard not to see the organization as the same small cadre of Silicon Valley superheroes pulling the levers of tech revolution without wider consent, just as they have for the last few waves of innovation.*

LLMs aren't inevitable and the "wider web" isn't representative

There's another category of missteps in this piece that I think can also be attributed to insufficient distance from the subject (here: OpenAI) and skepticism of their claims. With all of the resources being poured into LLMs (and similarly into very large models trained on image datasets), it's worth remembering that nothing is predestined about this. We can imagine other futures, but to do so, we have to maintain independence from the narrative being pushed by those who believe that "AGI" is desirable and that LLMs are *the* path to it.

In that light, consider this passage from the NYT Magazine piece:

> *L.L.M.s have even more troubling propensities as well: They can deploy openly racist language; they can spew conspiratorial misinformation; when asked for basic health or safety information they can offer up life-threatening advice. All those failures stem from one inescapable fact: To get a large enough data set to make an L.L.M. work, you need to scrape the wider web. And the wider web is, sadly, a representative picture of our collective mental state as a species right now, which continues to be plagued by bias, misinformation and other toxins.*

First, in fact, the third of those (offering up life-threatening advice) doesn't (only) stem from the fact that their training data is uncurated. Rather, it's connected to the fact that *by design* LLMs are just making stuff up. More specifically, they're making up text strings in the language they're trained on, which humans who speak that language can interpret. So when a person

comes in with a health and safety related question, if what comes back is wrong, chances are high it will also be dangerous. Most importantly: these can be understood as reasons not to use LLMs (perhaps at all, but at least in very many specific applications). However, the piece as a whole seems to follow OpenAI (and others) in seeing these propensities as temporary shortcomings of LLMs "in their current state" (see above) which will be addressed in time. (And this even as the tech is being commercialized, i.e. deployed with known risks of harm.)

Second, the wider web is not a representative picture of humanity. In section 4 of Stochastic Parrots 🦜 we go through in detail how various forces influence who has access to the web, who is comfortable continuing to contribute, who is represented in the parts of the web selected for LLM training data, and how the rudimentary filtering applied to that training data further creates further distortion. (The academic paper itself is long and admittedly dense, but the 20-minute video we prepared as the presentation of the paper at FAccT 2021 might be more approachable.) Relatedly, Safiya Noble, in her tour de force work *Algorithms of Oppression*, shows how the advertising-driven economy of web search shapes results that people see even as Google (quite misleadingly) presents these results as "just what's out there naturally".

Not just an academic debate

Whether or not "AI" actually works isn't just an academic debate. It could have been, if the people working on "AI" were simply doing esoteric projects in research labs without trying to monetize them. But that is not the world we live in: it seems like every day there is a new news story about someone selling an "AI" system to do something inappropriate, like fill in grades for students who couldn't take tests, or diagnose mental health disorders, interview job candidates, or apprehend migrants.

So it matters that the public at large have a clear understanding of what "AI" can actually do, which purported applications are so much snake oil, and what questions to ask to discover the possible harms from these systems (whether they are working properly or not). But Johnson misses the opportunity to educate the public in these ways, on two counts.

First, he talks up a possible application of LLMs in information retrieval (i.e. to replace search engines):

> *If the existing trajectory continues, software like GPT-3 could revolutionize how we search for information in the next few years. […] if the GPT-3 true believers are correct, in the near future you'll just ask an L.L.M. the question and get the answer fed back to you, cogently and accurately.*
>
> *This is not just an esoteric debate. If you can use next-word-prediction to train a machine to express complex thoughts or summarize dense material, then we could be on the cusp of a genuine technological revolution where systems like GPT-3 replace search engines or Wikipedia as our default resource for discovering information.*

But this is in fact a terrible idea, as Will Douglas Heaven put it in his recent MIT Tech Review article covering work by Chirag Shah and me (at CHIIR 2022) and by Martin Potthast et al (in SIGIR Forum, 2020).

But more broadly, Johnson narrows the scope of possible casualties of AI hype to AI research itself:

> *But if the large language models are ultimately just "stochastic parrots," then A.G.I. retreats once again to the distant horizon — and we risk as a society*

> *directing too many resources, both monetary and intellectual, in pursuit of a false oracle.*

Yes, it is a problem that we have over-concentrated research resources on "AI", when there are so many other worthy (and in many cases urgent) problems in the world. (And I'd say that's true even if "AGI" were a well-defined and feasible research goal.) But those aren't the only harms: every time someone applies pattern recognition at scale over data to produce systems that are supposedly "unbiased" and "objective" in a way that makes decisions (or produces content) affecting real humans, there's harm. And all of that is left out of view by Johnson's framing.

On being placed into the "skeptics" box

To be fair, Johnson does spare a little room for dissenting voices (including quotes from me, Meredith Whittaker, and Gary Marcus). But he frames us as "skeptics", sort of token nods to the other side in some "both sides" reporting. But it's worse than that, in two ways: First, the skeptics framing seems to shift the burden of proof away from those who claim to be doing something outlandish (building "AGI") and towards those who call out the unfounded claims.

> *Some skeptics argue that the software is capable only of blind mimicry — that it's imitating the syntactic patterns of human language but is incapable of generating its own ideas or making complex decisions, a fundamental limitation that will keep the L.L.M. approach from ever maturing into anything resembling human intelligence.*

Having presented that skepticism, he asks:

> *How can we determine whether GPT-3 is actually generating its own ideas or merely paraphrasing the syntax of language it has scanned from the servers of Wikipedia, or Oberlin College, or The New York Review of Books?*

But he doesn't ask on what grounds we should believe that it is generating its own ideas. Instead, it's "no one really knows":

> *Some people argue that higher-level understanding is emerging, thanks to the deep layers of the neural net. Others think the program by definition can't get to true understanding simply by playing "guess the missing word" all day. But no one really knows.*

But more grating to me is that being relegated to the "skeptics" box cedes the framing of the debate to the AI boosters. I am not just saying "no that's not how to build 'AI'", nor is Meredith Whittaker (though Gary Marcus does seem to be mostly concentrated on that argument). For me (and I believe also Whittaker as well as the authors I cited in the tl;dr at the top of this post), the relevant question is not "how do we build 'AI'?" but rather things like "How do we shift power so that we see fewer (ideally no) cases of algorithmic oppression?", "How do we imagine and deploy design processes that locate technology as tools, shaped for and in the service of people working towards pro-social ends?", and "How do we ensure the possibility of refusal, making it possible to shut down harmful applications and ensure recourse for those being harmed?"

So, while I agree that there are very large ethical issues at stake here, I fundamentally disagree with the framing of this article. Perhaps nowhere is that more acute than in the proposed solution to the ethical issues:

> *The very premise that we are now having a serious debate over the best way to instill moral and civic values in our software should make it clear that we have crossed an important threshold.*

and

> *We've never had to teach values to our machines before.*

Johnson quotes me in this part as saying:

> *"So long as so-called A.I. systems are being built and deployed by the big tech companies without democratically governed regulation, they are going to primarily reflect the values of Silicon Valley," Emily Bender argues, "and any attempt to 'teach' them otherwise can be nothing more than ethics washing."*

That quote is accurate, but there was a bit more to it. What I actually sent him was:

Talking about "teaching machines values" is a fundamental misframing of the situation and a piece of AI hype. Software systems are artifacts, not sentient entities, and as such "teaching" is a misplaced and misleading metaphor (as is "machine learning" or "artificial intelligence"). Rather, as with any other artifact or system, the builders of these systems are designing values into them. To say that they could be "taught" other values is just another way to hide the values of the system builders (already designed in) behind a veneer of faux objectivity. So long as so-called "AI" systems are being built and deployed by the big tech cos without democratically governed regulation they are going to primarily reflect the values of Silicon Valley and any attempt "teach" them otherwise can be nothing more than ethics washing.

But I don't think that Johnson really took my point, because the article firmly holds to the framing of "teaching" machines values:

> *Should we build an A.G.I. that loves the Proud Boys, the spam artists, the Russian troll farms, the QAnon fabulists? It's easier to build an artificial brain that interprets all of humanity's words as accurate ones, composed in good faith, expressed with honorable intentions. It's harder to build one that knows when to ignore us.*

I hope anyone who has made it this far in this blog post can see the issues in that statement, how they imagine the problems in terms of assuming that it is both possible and desirable to build fully autonomous agents.

And to take just one further example, Johnson writes, again imagining autonomous agents ("citizens", forsooth):

> *And if large language models are in our future, then the most urgent questions become: How do we train them to be good citizens? How do we make them "benefit humanity as a whole" when humanity itself can't agree on basic facts, much less core ethics and civic values?*

What I kept waiting (in vain) for this article to do was to break out from the OpenAI frame. Why build these things in the first place? If we are building them, why not think in terms of democratic governance that creates the frame in which they can be built and deployed, and requirements of transparency and documentation, rather than in terms of tinkering with the algorithms and their training data?

Conclusion

Source: https://www.maxpixel.net/Yellow-Red-Green-Hybrid-Macaw-Bird-Orange-Parrot-943228

There is a talk I've given a couple of times now (first at the University of Edinburgh in August 2021) titled "Meaning making with artificial interlocutors and risks of language technology". I end that talk by reminding the audience to not be too impressed, and to remember:

- Just because that text seems coherent doesn't mean the model behind it has understood anything or is trustworthy
- Just because that answer was correct doesn't mean the next one will be

- When a computer seems to "speak our language", we're actually the ones doing all of the work

Maintaining this skeptical stance is non-trivial. Johnson writes:

> *The first few times I fed GPT-3 prompts of this ilk, I felt a genuine shiver run down my spine. It seemed almost impossible that a machine could generate text so lucid and responsive based entirely on the elemental training of next-word-prediction.*

and

> *It's important to stress that this is not a question about the software's becoming self-aware or sentient. L.L.M.s are not conscious — there's no internal "theater of the mind" where the software experiences thinking in the way sentient organisms like humans do. But when you read the algorithm creating original sentences on the role of metafiction, it's hard not to feel that the machine is thinking in some meaningful way.*

"Hard not to feel" is apt. When we encounter something that seems to be speaking our language, without even thinking about it, we use the skills associated with using that language to communicate with other people. Those skills centrally involve intersubjectivity and joint attention and so we imagine a mind behind the language even when it is not there.

But reminding ourselves that all of that work is on our side, the human side, is of critical importance because it allows us a clearer view of the present, in which we can more accurately track the harm that people are doing with technology, and a broader view of the future, where we can work towards meaningful, democratic governance and appropriate regulation.

Acknowledgments

I'd like to thank Timnit Gebru and Meg Mitchell for encouragement to sit down and write this piece and both of them as well as Leon Derczynski and Meredith Whittaker for thoughtful yet quick turn-around comments.

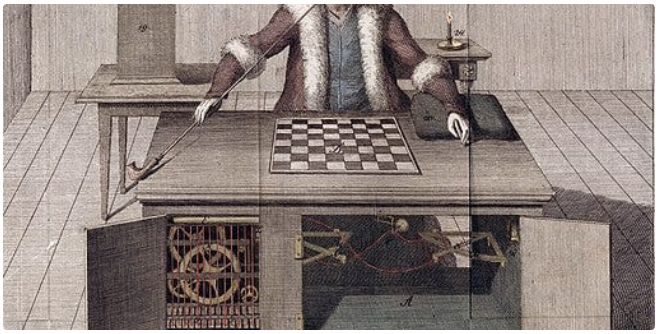
Written by Emily M. Bender

2.4K Followers

Professor, Linguistics, University of Washington// Faculty Director, Professional MS Program in Computational Linguistics (CLMS) faculty.washington.edu/ebender

Follow

More from Emily M. Bender



Emily M. Bender

Emily M. Bender

Opening remarks on “AI in the Workplace: New Crisis or...

On Thursday 9/28, I had the opportunity to speak at a virtual roundtable convened by...

5 min read · Oct 1, 2023

685 12

Talking about a ‘schism’ is ahistorical

In two recent conversations with very thoughtful journalists, I was asked about the...

9 min read · Jul 5, 2023

689 15

Emily M. Bender

Thought experiment in the National Library of Thailand

With the advent of ChatGPT, large language models (LLMs) went from a relatively niche...

6 min read · May 25, 2023

1.1K 25

Emily M. Bender

Policy makers: Please don’t fall for the distractions of #AIhype

Below is a lightly edited version of the tweet/toot thread I put together in the...

5 min read · Mar 29, 2023

734 15

See all from Emily M. Bender

Recommended from Medium

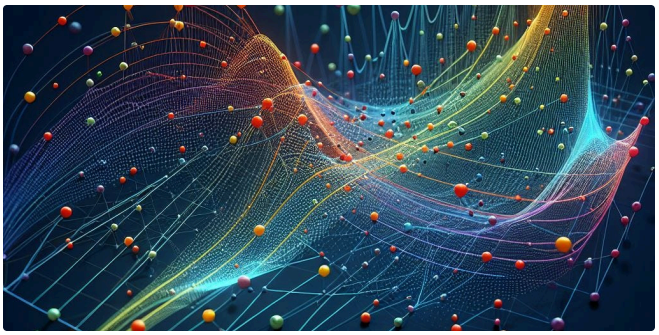
Emily M. Bender

“Ensuring Safe, Secure, and Trustworthy AI”: What those seve...

On July 21, 2023, in a photo op featuring only men in suits, President Biden announced tha...

7 min read · Jul 29, 2023

744 8



Tim Sumner in Towards Data Science

A New Coefficient of Correlation

What if you were told there exists a new way to measure the relationship between two...

10 min read · Mar 31, 2024

1.6K 27

Lists

Staff Picks

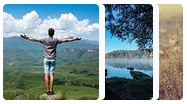
621 stories · 891 saves

Stories to Help You Level-Up at Work

19 stories · 559 saves

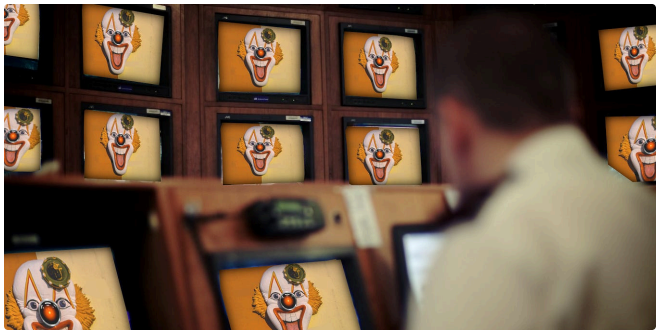
Self-Improvement 101

20 stories · 1594 saves



Productivity 101

20 stories · 1483 saves



Cory Doctorow

Supervised AI isn't

Automation blindness can't be automated away.

· 6 min read · Aug 23, 2023

1.3K 11

Steve Yegge

A good day with Jeff

This is a story about Jeff Bezos which I've told many times over the years, but which I don't...

6 min read · May 27, 2023

1.5K 29

Somnath Singh in Level Up Coding

The Era of High-Paying Tech Jobs is Over

The Death of Tech Jobs.

· 14 min read · Mar 31, 2024

4.98K 147

Admiral Cloudberg

Powerless over London: The crash of British Airways flight 38

When a British Airways Boeing 777 crash-landed at Heathrow Airport in 2008,...

35 min read · 5 days ago

2.6K 28

See more recommendations