Issue 3

PARROTS

Mapping the Latent Spaces of Culture

Ted Underwood

August 2022
doi:10.5281/zenodo.6567481

The technology I need to discuss in this paper doesn't yet have a consensus name. Some observers point to an architecture, the Transformer.[1] "On the Dangers of Stochastic Parrots" focuses on size and discusses "large language models."[2] A paper from Stanford emphasizes applications: "foundation models" are those that can adapt "to a wide range of downstream tasks."[3] Each name identifies a different feature of recent research as the one that matters. To keep that question open, I'll refer here to "deep neural models of language," a looser category.

However we define them, neural models of language are already changing the way we search the web, write code, and even play games. Academics outside computer science urgently need to reflect on them. "On the Dangers of Stochastic Parrots" deserves credit for starting that discussion — especially since its publication required tenacity and courage. I am honored to be part of a forum exploring its significance for the humanities.

The argument that Bender et al. advance has two parts: first, that large language models pose social risks, and second, that they will turn out to be a "misdirected research effort" anyway, since they pretend to perform "natural language understanding" but "do not have access to meaning." (615).

In historical disciplines, it is far from obvious that all meaning boils down to intentional communication between individuals.

I agree that the trajectory of recent research has risks. But to understand the risks language models pose, I think we will need to understand how they produce meaning. The premise that they simply "do not have access to meaning" tends to prevent us from seeing the models' social role. I hope humanists can help illuminate that role by offering a wider range of ways to think about the work language does.

Language models as models of culture

It is true that language models don't yet represent their own communicative goals or an interlocutor's state of mind. These are important aspects of language, and for "Stochastic Parrots," they are the whole story: the article defines *meaning* as "meaning conveyed between individuals" and "grounded in communicative intent" (616).

But in historical disciplines, it is far from obvious that all meaning boils down to intentional communication between individuals. Historians often use *meaning* to describe something more collective, because the meaning of a literary work, for example, is not circumscribed by intent. It

is common for debates about the meaning of a text to depend more on connections to books published a century earlier (or later) than on reconstructing the author's conscious plan.[4]

I understand why researchers in a field named "artificial intelligence" would associate meaning with mental activity and see writing as a dubious proxy for it. But historical disciplines rarely have access to minds, or even living subjects. We work mostly with texts and other traces. For this reason, I'm not troubled by the part of "Stochastic Parrots" that warns about "the human tendency to attribute meaning to text" even when the text "is not grounded in communicative intent" (618, 616). Historians are already in the habit of finding meaning in genres, nursery rhymes, folktale motifs, ruins, political trends, and other patterns that never had a single author with a clear purpose.[5] If we could find meaning only in intentional communication, we wouldn't find much meaning in the past at all. So not all historical researchers will be scandalized when we hear that a model is merely "stitching together sequences of linguistic forms it has observed in its vast training data" (617). That's what we often do too, and we could use help.

A willingness to find meaning in collective patterns may be especially necessary for disciplines that study the past. But this flexibility is not limited to scholars. The writers and artists who borrow language models for creative work likewise appreciate that their instructions to the model acquire meaning from a training corpus. The phrase "Unreal Engine," for instance, encourages a neural network called CLIP to select pictures with a consistent, cartoonified style. But this has nothing to do with the dictionary definition of "unreal." It's just a helpful side-effect of the fact that many video game screenshots are captioned with the name of the game engine (Unreal Engine) that produced them.

> *Intelligence only starts to interest us after it mixes with time to become a biased, limited pattern of collective life.*

In short, I think people who use neural models of language typically use them for a different purpose than "Stochastic Parrots" assumes. The immediate value of these models is often not to mimic individual language understanding, but to represent specific cultural practices (like styles or expository templates) so they can be studied and creatively remixed. This may be disappointing for disciplines that aspire to model general intelligence. But for historians and artists, cultural specificity is not disappointing. Intelligence only starts to interest us after it mixes with time to become a biased, limited pattern of collective life. Models of culture are exactly what we need.

Language models in historical research

While I'm skeptical that language models are devoid of meaning, I do share other concerns raised by the authors of "Stochastic Parrots." For instance, I agree that researchers will need a way to understand the subset of texts that shape a model's response to a given prompt. Culture is historically specific, so models will never be free of omission and bias. But by the same token, we need to know which practices they represent.

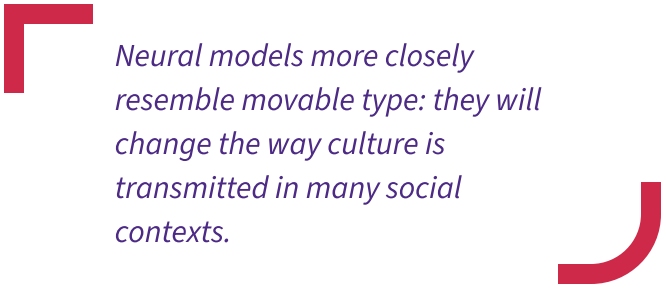
If companies want to offer language models as a service to the public — say, in web search — they will need to do even more than know what their models represent. Somehow, a single model will need to produce a picture of the world that is acceptable to a wide range of audiences, without amplifying harmful biases or filtering out minority discourses (Bender et al., "Stochastic Parrots," 614). That's a delicate balancing act.

doi:10.5281/zenodo.6567481

Historians don't have to compress their material as severely as that. Since history is notoriously a story of conflict, and our sources were interested participants, few people expect historians to represent all aspects of the past with one correctly balanced model. On the contrary, historical inquiry is usually about comparing perspectives. Machine learning is not the only way to do this, but it can help. For instance, researchers can measure differences of perspective by training multiple models on different publication venues or slices of the timeline.[6]

When research is organized by this sort of comparative purpose, the biases in data are not usually a reason to refrain from modeling — but a reason to create more corpora and train models that reflect a wider range of biases. On the other hand, training a variety of models becomes challenging when each job requires thousands of GPUs. Tech companies might have the resources to train many models at that scale. But will universities?

One way around this impasse is to train a single model that can explicitly distinguish multiple perspectives. At present, researchers create this flexibility in a rough and ready way by "fine-tuning" BERT on different samples. A more principled approach might design models to recognize the social structure in their original training data. One recent paper associates each text with a date stamp, for instance, to train models that respond differently to questions about different years.[7] Similar approaches might produce models explicitly conditioned on variables like venue or nationality — models that could associate each statement or prediction they make with a social vantage point. Producing models that can represent hundreds of vantage points may become easier if this technology evolves — as increasingly seems likely — to separate the language model proper from a larger database that explicitly encodes world knowledge.[8] Training a new language model is computationally expensive. But editing the knowledge base to reflect a particular period or set of sources might be relatively cheap.



If neural language models are to play a constructive role in research, universities will also need alternatives to material dependence on tech giants. In 2020, it seemed that only the largest corporations could deploy enough silicon to move this field forward. In October 2021, things are starting to look less dire. Coalitions like EleutherAI are reverse-engineering language models.[9] Smaller corporations like Hugging Face are helping to cover underrepresented languages. NSF is proposing new computing resources.[10] The danger of oligopoly is by no means behind us, but we can at least begin to see how scholars might train models that represent a wider range of perspectives.

The effects of modeling culture

Of course, scholars are not the only people who matter. What will language models (and models of culture) mean for people outside universities?

I agree with the authors of "Stochastic Parrots" that neural language models are dangerous. But I am not sure that critical discourse has alerted us to the most important dangers yet. Critics often prefer to say that these models are dangerous only because they don't work and are devoid of meaning. Perhaps that seems to be the strongest rhetorical position, since it concedes no value to the models. But I suspect this hard line also prevents critics from envisioning what the models might be good for and how they're likely to be (mis)used.

doi:10.5281/zenodo.6567481

Consider the surprising art scene that sprang up when CLIP was released. OpenAI still hasn't released the DALL-E model that uses the numbers CLIP assigns to text to find a corresponding point in a latent space of hypothetical images.[11] But that didn't stop graduate students and interested amateurs from duct-taping CLIP to various generative image models and using the contraption to explore visual culture in dizzying ways.[12]

The angel of air. Unreal Engine. VQGAN + CLIP, Aran Komatsukaki, May 31, 2021.

Does the emergence of this subculture make any sense if we assume that CLIP is just a failed attempt to reproduce individual language use? In practice, the people tinkering with CLIP don't expect it to respond like a human reader. More to the point, they don't want it to. They're fascinated because CLIP uses language *differently* than a human individual would — mashing together the senses and overtones of words and refracting them into the potential space of internet images like a new kind of synesthesia.[13] The pictures produced are fascinating, but (at least for now) too glitchy to impress most people as art. They're better understood as postcards from an unmapped latent space.[14] The point of a postcard, after all, is not to be itself impressive, but to evoke features of a larger region that looks fun to explore. Here the "region" is a particular visual culture; artists use CLIP to find combinations of themes and styles that could have occurred within it (although they never quite did).

doi:10.5281/zenodo.6567481

The clockwork angel of air flying over a rocky coast, Kodak Portra film. Ted Underwood, using Katherine Crowson's cc12m_1 diffusion model, December 28, 2021.

Will models of this kind also have negative effects? Absolutely. The common observation that they could reinforce existing biases is the mildest possible example. If we approach neural models as machines for mapping and rewiring collective behavior, we will quickly see that they could do much worse than reinforce existing biases: for instance, deepfakes could create new hermetically sealed subcultures and beliefs that are difficult to contest.

My goal in this essay is not to decide whether neural language models are good or bad — just to clarify what's being modeled, why people care, and what kinds of (good or bad) effects we might expect. Reaching a comprehensive judgment is likely to take decades. After all, models are easy to distribute. So this was never a problem, like gene splicing, that could stay bottled up as an ethical dilemma for one profession that controlled the tools. Neural models more closely resemble movable type: they will change the way culture is transmitted in many social contexts. Since the consequences of movable type included centuries of religious war in Europe, my analogy is not meant to reassure. I just mean that questions on this scale don't get resolved quickly or by experts. We are headed rather for a broadly political debate about antitrust, renewable energy, and the shape of human culture itself — a debate where everyone will have some claim to expertise.[15]

doi:10.5281/zenodo.6567481

Let me end, however, on a positive note. I have suggested that approaching neural models as models of culture rather than intelligence gives us even more reason to worry about them. But it also gives us more reason to hope. It is not entirely clear what we plan to gain by modeling intelligence, since there are already more than seven billion intelligences on the planet. By contrast, it is easy to see how exploring spaces of possibility implied by the past of human culture could support a more reflective and more adventurous approach to our future. I can imagine a world where generative models of culture are used grotesquely or locked down as IP for Netflix. But I can also imagine a world where fan communities use them to remix plot tropes and gender norms, making "mass culture" a more self-conscious, various, and participatory phenomenon than the twentieth century usually allowed it to become.

I don't know which of those worlds we will build. But either way, I suspect we will need to reframe our conversation about artificial intelligence as a conversation about models of culture and the latent spaces they imply. Philosophers and science fiction writers may enjoy debating whether software can have mental attributes like intention. But that old argument does little to illuminate the social questions new technologies are really raising. Neural language models are dangerous and fascinating because they can illuminate and transform shared patterns of behavior — in other words, cultural practices. When the problem is redescribed this way, the concerns about equity foregrounded by "Stochastic Parrots" still matter deeply. But the imagined contrast between mimicry and meaning in the article's title no longer connects with any satirical target. Culture clearly has meaning. But I'm not sure that anyone cares whether a culture has autonomous intent, or whether it is merely parroting human action.

doi:10.5281/zenodo.6567481

1. Ashish Vaswani et al., "Attention Is All You Need," 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, 2017. https://arxiv.org/abs/1706.03762. ↩

2. Emily M. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ," *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2021), 610–623, https://doi.org/10.1145/3442188.3445922. Further citations of this work are given in the text. ↩

3. Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models," CoRR Aug 2021, 3–4, https://arxiv.org/abs/2108.07258. ↩

4. "It is language which speaks, not the author." Roland Barthes, "The Death of the Author," in *Image / Music / Text*, trans. Stephen Heath (New York: Hill and Wang, 1977), 143. ↩

5. To this list one might also add the material and social aspects of book production. In commenting on "Stochastic Parrots," Katherine Bode notes that book history prefers to paint a picture where "meaning is dispersed across . . . human and non-human agents." Katherine Bode, qtd. in Lauren M. E. Goodlad, "Data-Centrism and Its Discontents," *Critical AI* (blog), Oct. 15, 2021, https://criticalai.org/2021/10/14/blog-recap-stochastic-parrots-ethics-of-data-curation/. ↩

6. Sandeep Soni, Lauren F. Klein, and Jacob Eisenstein, "Abolitionist Networks: Modeling Language Change in Nineteenth-Century Activist Newspapers," *Journal of Cultural Analytics* 6, no. 1 (Jan. 18, 2021), https://culturalanalytics.org/article/18841-abolitionist-networks-modeling-language-change-in-nineteenth-century-activist-newspapers; Ted Underwood, "Machine Learning and Human Perspective," *PMLA* 135, no. 1 (Jan. 2020): 92–109, http://hdl.handle.net/2142/109140. ↩

7. Bhuwan Dhingra et al., "Time-Aware Language Models as Temporal Knowledge Bases," *Transactions of the Association for Computational Linguistics* 10 (Mar. 18, 2022): 257–73, https://doi.org/10.1162/tacl_a_00459. ↩

8. Sebastian Borgeaud et al., "Improving Language Models by Retrieving from Trillions of Tokens," arXiv, Dec. 8, 2021, https://arxiv.org/abs/2112.04426v1. ↩

9. See for instance, Sid Black et al., "GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow," version 1.0, Mar. 21, 2021, https://doi.org/10.5281/zenodo.5297715. ↩

10. White House Briefing Room, "The White House Announces the National Artificial Intelligence Research Resource Task Force," June 10, 2021, https://www.whitehouse.gov/ostp/news-updates/2021/06/10/the-biden-administration-launches-the-national-artificial-intelligence-research-resource-task-force/. ↩

11. Aditya Ramesh et al., "Zero-Shot Text-to-Image Generation," arXiv, updated Feb. 26, 2021, https://arxiv.org/abs/2102.12092. ↩

12. One early technique was eventually published as Katherine Crowson et al., "VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance," arXiv, Apr. 18, 2022, https://arxiv.org/abs/2204.08583. ↩

13. One good history of this scene is titled "Alien Dreams"—a title that concisely indicates how little interest artists have in using CLIP to reproduce human behavior. Charlie Snell, "Alien Dreams: An Emerging Art Scene," *Machine Learning at Berkeley* (blog), June 30, 2021, https://ml.berkeley.edu/blog/posts/clip-art/. ↩

14. For a skeptical history of this spatial metaphor, see Nick Seaver, "Everything Lies in a Space: Cultural Data and Spatial Reality," in "Towards an Anthropology of Data," ed. Rachel Douglas-Jones, Antonia Walford, and Nick Seaver, special issue, *Journal of the Royal Anthropological Institute* 27, no. 1 (Apr. 2021): 43–61, https://doi.org/10.1111/1467-9655.13479. We also skeptically probe the limits of spatial metaphors for culture (but end up confirming their value) in Ted Underwood and Richard Jean So, "Can We Map Culture?" *Journal of Cultural Analytics* 6, no. 3 (June 17, 2021), https://doi.org/10.22148/001c.24911. ↩

15. I haven't said much about the energy cost of training models. For one thing, I'm not fully informed about contemporary efforts to keep that cost low. More importantly, I think the cause of carbon reduction is actively harmed by pitting different end users against each other. If we weigh the "carbon footprint" of your research agenda against my conference travel, the winner will almost certainly be British Petroleum. Renewable energy is a wiser thing to argue about if carbon reduction is actually our goal. For another version of this idea, see Mark Kaufman, "The Carbon Footprint Sham: A 'Successful, Deceptive' PR Campaign," *Mashable,* July 9, 2021, https://mashable.com/feature/carbon-footprint-pr-campaign-sham. ↩