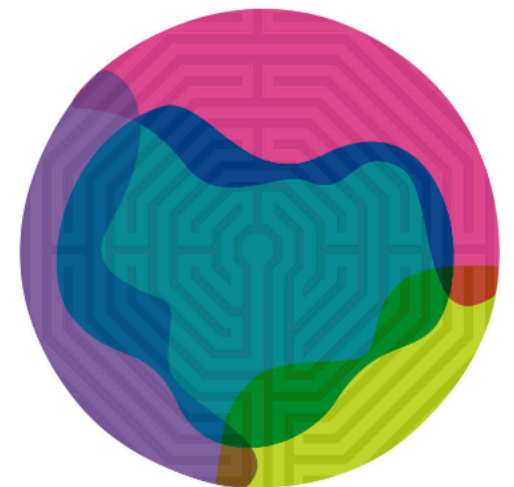


Resisting Dehumanization in the Age of “AI”

Emily M. Bender
CogSci 2022

29 July 2022
Toronto, Canada

@emilymbender



**COGNITIVE
DIVERSITY**
CogSci 2022

“AI” research, development and sales involves dehumanization on many levels

- Computational metaphor
- Digital physiognomy
- “Ground lies”
- Irrelationality
- Ghost work
- Reinforcing the white racial frame

Cognitive scientists are well positioned to resist this, and we have many roles to play

- Problematize simplified tasks
- Critically analyze claims of “AI” capabilities
- Decenter whiteness/WEIRDs/English
- Advocate for broader distribution of research funds
- Envision alternative pro-human development paths
- Engage in & support public scholarship

Roadmap

- Researcher stance
- Dehumanization: working definition
- Dehumanization in “AI” research, development & sales
- What cognitive scientists can do

Researcher stance/Who am I?

- PhD training in syntax and sociolinguistics
- Long experience with multilingual grammar engineering: building grammars in software, across (mostly spoken) languages
- Since 2016: methodologies for supporting consideration of societal impacts of language technology—in NLP research, development, and education.
- Broader conversation about identifying and mitigating harms done in the name of “AI”

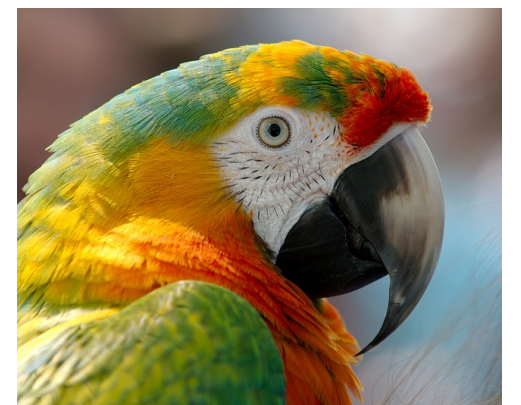
Climbing towards NLU: On Meaning, Form and Understanding in the Age of Data (Bender & Koller 2020)

- Written in reaction to widespread claims that language models “understand” language
- But language is a system of signs (pairing of form & meaning; de Saussure)
- Language models (GPT-3 et al) are trained on the task of string prediction: their only input is form
- Comparisons to child language acquisition are misleading: child learn language in socially rich, socially situated interactions
- Octopus thought experiment: posit an intelligent learner, given access only to form; all that is learned is patterns in form



On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 (Bender, Gebru et al 2021)

- Observed a trend towards ever larger language models, and asked:
- What are the possible risks associated with this technology and what paths are available for mitigating those risks?
- Environmental costs & environmental racism
- Financial costs & impact on research participation
- Datasets filled with hegemonic viewpoints & worse; no/minimal documentation and accountability
- Synthetic text generating machines can reproduce systems of oppression
- Synthetic text generating machines can mislead humans



But how do I know that
you're not just a
stochastic parrot?

Dehumanization: Definitions

- “**Dehumanization happens when people are depicted, regarded, or treated as not human or less human.** [...] I start with such a thin notion since not much agreement exists beyond it in the scholarship on dehumanization, not even with respect to the above examples. Most scholars will count them as dehumanizing, while others will not.” (Kronfeldner 2021:xvii)
- “If racialization is understood not as a biological or cultural descriptor but as a conglomerate of sociopolitical relations that discipline humanity into full humans, not-quite-humans, and nonhumans, then blackness designates a changing system of **unequal power structures that apportion and delimit which humans can lay claim to full human status and which humans cannot.**” (Weheliye 2014:3)

Dehumanization: Working definition

1. Cognitive state of failing to perceive another human as fully human
2. Acts that express that cognitive state or otherwise entail the assertion that another human is not fully human
3. Experience of being subjected to acts that express lack of perception of one's humanity and/or deny human experience or human rights

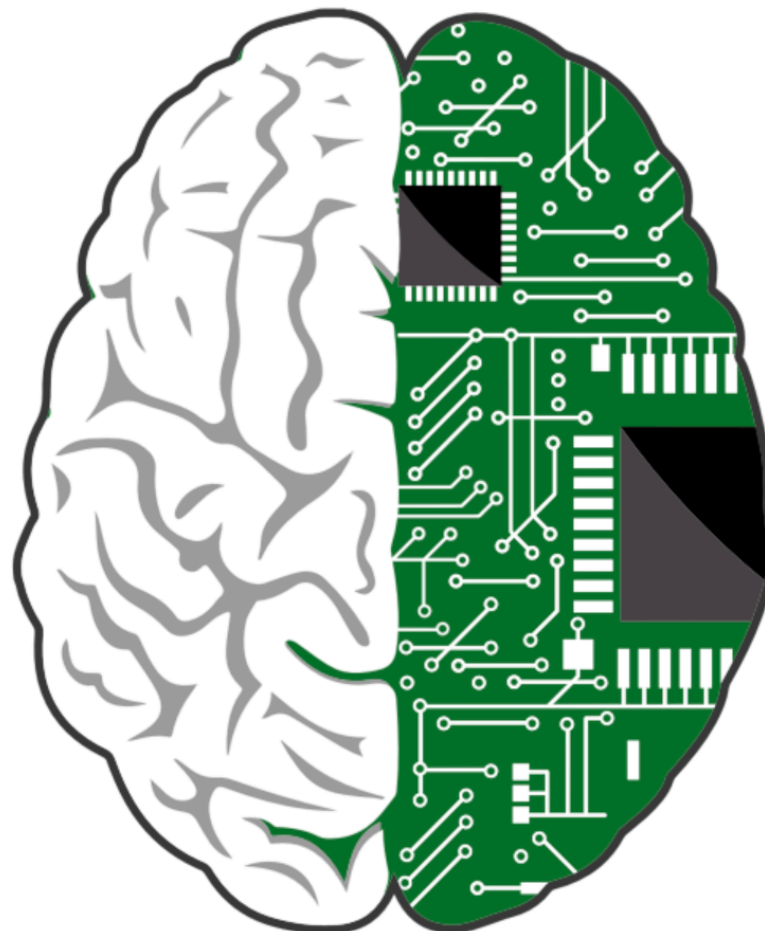
Fully human

- Entitled to all rights recognized as human rights
- Equally in possession of an internal life and point of view
- Welcomed and known as one's full self

Dehumanization in the research,
development and sales of “AI”

The Computational Metaphor (Baria & Cross 2021)

- Scientific metaphor used and debated in neuroscience: THE BRAIN IS A COMPUTER
- PR metaphor used by technologists: THE COMPUTER IS A BRAIN



The Computational Metaphor (Baria & Cross 2021)

- “afford[s] the human mind less complexity than is owed, and the computer more wisdom than is due.” (p.2)
- “the Computational Metaphor rests on other well-ingrained ideologies in which a hierarchy of human value is **tied to a particular notion of intelligence** such that the quality of **being emotional is considered inferior to being rational.**” (p.6)
- “This notion of intelligence extends to the justified subjugation of beings considered less rational to those considered (or propagandized as) more rational, whether animals to humans, women to men, or one race of humans to another. According to this logic, **in its fake-ness as a human intelligence, AI paradoxically succeeds in being a more trustworthy form of intelligence**, by being the epitome of rational thought.” (p.6)

On anthropomorphism in science (Dijkstra 1985)

- “A more serious byproduct of the tendency to talk about machines in anthropomorphic terms is **the companion phenomenon of talking about people in mechanistic terminology**. The critical reading of articles about computer-assisted learning [...] leaves you no option: in the eyes of their authors, the educational process is simply reduced to a caricature, something like the building up of conditional reflexes. For those educationists, Pavlov’s dog adequately **captures the essence of Mankind** —while I can assure you, from intimate observations, that **it only captures a minute fraction of what is involved in being a dog—.**”



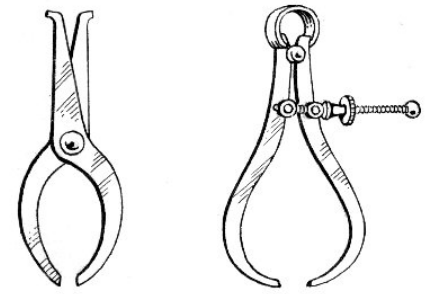
Appropriating experiences of disabled people to assert humanity of “AI”

- Agüera y Arcas (blog, 12/2021) asserts that LLMs are like Deafblind people
- Under the heading “modality chauvinism” calls on the writings of Daniel Kish and Helen Keller to argue that no one sensory system is required for humans to develop concepts
- But he can’t show that large language models are like people, with internal lives, relationships, and full personhood
- The analogy ends up dehumanizing Blind and Deafblind people, by saying they are like something that is patently not human, specifically because of their disability.

<https://bit.ly/EMB-blog1>

Digital Physiognomy

(see Agüera y Arcas, Mitchell & Todorov 2017)



- Classification of people into identity categories or personality characteristics based on computer processing of photos, voice signals, or other
 - Claims of predicting: criminality, sexual orientation, employability, political leaning, psychopathy, etc (see Stark & Hutson 2022)
 - Gender, race, etc classification similarly problematic
- Flattens human identities and emotional experiences into intrinsic, externally observable categories of classification
- Pseudoscience of physiognomy made apparently “objective” through the application of computers

“Ground lies”

- Data sets used in training “AI” systems are mythologized as representative, due to size or lack of curation (Paullada et al 2021, Raji et al 2021, Scheuerman et al 2021)
- Decisions at every point: where to collect from, how to collect, how to filter, what labels to apply, who should apply them, how to verify labels
- If we don’t actively work to curate the datasets we want, we *will* be collecting datasets representative of dehumanizing ideologies like white supremacy
- “Data sets so specifically built in and for white spaces represent the constructed reality, not the natural one. **To have accuracy calculated in the absence of my lived experience** not only offends me, but also **puts me in real danger.**” (Raji 2020)

Humans are not just social but thoroughly relational

- “The self thus never just *is* but rather emerges continuously and jointly relying on behavior and action and on doing and being together with others.” (Kyselo 2014:8)
- “Humans are inherently historical, social, cultural, gendered, politicized, and contextualized organisms. Accordingly, their knowing and understanding of the world around them necessarily takes place through their respective lenses.” (Birhane, 2021:5)

AI “knowing” is irrelational

- “Data science and data practices reincarnate rationalism in many forms, including [...] the manner in which the dominant view is taken as the “God’s eye view” (Birhane 2021:3)
- Machines aren’t designed to be in relationship, to jointly make meaning, or to apply “*metis*” (Scott 1998)

Irrelationality:

Devaluing humanity while leaving no space for it

- “But there is also an underlying presupposition almost always at play that suggests, tacitly and otherwise, that the **dehumanized and anonymous decision-making** done by computers in a way that mimics—but replaces—that of human actors is somehow **more just or fair.**” (Roberts 2021:52)
- “Damage manifests most profoundly not only when errors get made, but when people are compelled to endure those errors. [...] absurdity follows when algorithmic systems deny the people they mistreat the status to lodge complaints, let alone the power to repair, resist, or escape the world that these systems create. In other words, **when algorithmic systems try to insist that they live in their utopias.**” (Alkhatib 2021:3)

Ghost work:

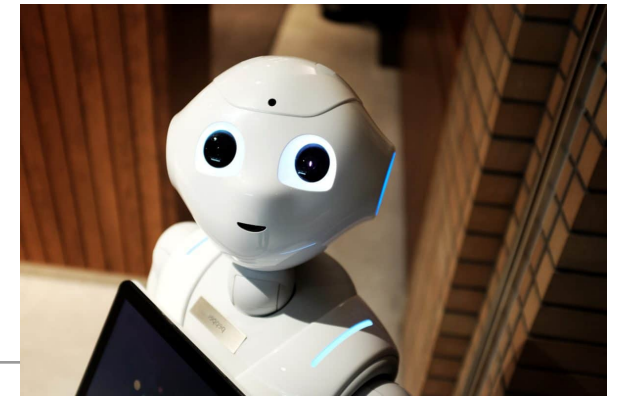
Humans as hidden software components

- Human effort is everywhere in so-called “AI” systems: data labeling, system design, evaluation, and backstop for when the task is too difficult for the machine
- Tech firms hide the labor and humanity of microworkers in systems designed to produce the illusion of AI
- Crowdwork platforms hide humanity of microworkers from requestors by representing workers only through their worker IDs and selling them as interchangeable

(Gray and Suri 2019, Roberts 2021)

Reinforcing the white racial frame

(Cave & Dihal 2020)

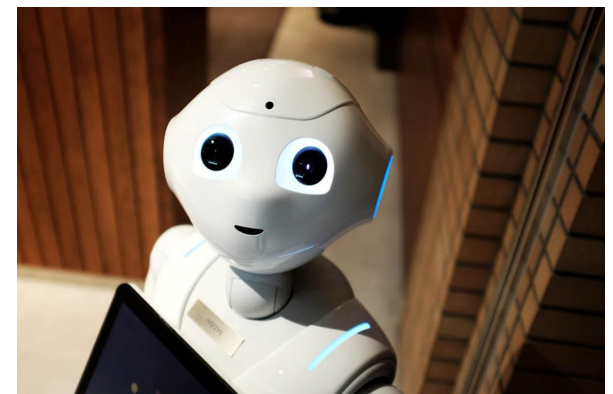


- Within Anglo Western culture at least, AI is racialized as white
 - Robots are frequently literally made with white (color) exteriors
 - More humanoid robots are designed to be perceived as racially white
 - Even voice assistants & text-based chatbots mostly “talk white”
 - Weizenbaum’s ELIZA used white language features (Marino 2014)
 - Siri released in 2011, African American voices for it only in 2021

Reinforcing the white racial frame

(Cave & Dihal 2020)

- Cave & Dihal's hypothesized causes for this:
 - Disproportionately white workforce in AI
 - The traits associated with AI (intelligence, professionalism, power) are those that the white racial frame ascribes to white people
 - “The Whiteness of the machines allows the White utopian imagination to fully exclude people of colour.” (p.698)



What can cognitive scientists do about this?

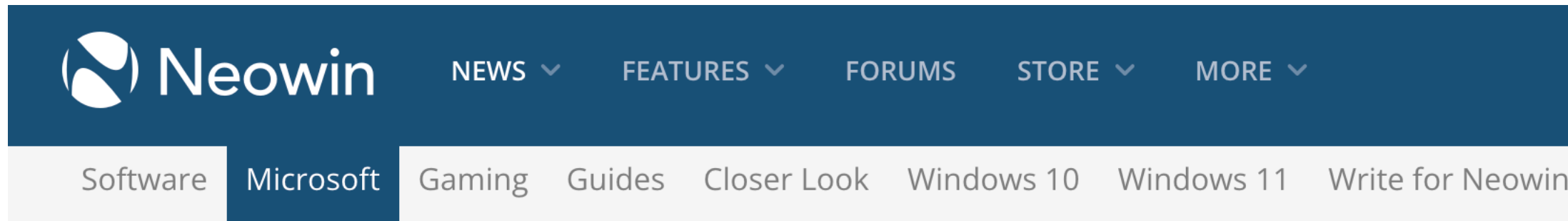
Problematize simplified tasks

- Much work in ML is driven by *tasks* which are meant to illustrate *capabilities*
- A *task* can be defined intensionally or extensionally (Schlangen 2021)
 - Intensionally: informal description of what the algorithm does (translate from Thai to Navajo; transcribe spoken Kinyarwanda to standard orthography)
 - Extensionally: through a dataset of paired inputs and outputs
- *Capability*: something more general hypothesized to underlie the possibility of doing the task
 - The tasks in SuperGLUE illustrate “understanding” (Wang et al 2019)

Problematize simplified tasks

- Culture of computer science: Emphasis on problem solving, frequently motivated by ‘expense’ of human experts
- Culture of machine learning: Solutions should be general; it’s bad to look at the data
- (Partial) Division of labor between dataset producers and algorithm builders
- Result: Wild overclaims (repeated and amplified in the media) of computers “surpassing humans” at tasks like “natural language understanding”

AI hype



Microsoft's AI model has outperformed humans in natural language understanding

Usama Jawad  · Jan 7, 2021 09:48 EST · **HOT!**

 0

[Microsoft](#) is heavily invested in artificial intelligence models with expertise in natural language understanding (NLU). To that end, the company has [acquired startups studying natural language processing \(NLP\)](#) and also has an [exclusive license to OpenAI's GPT-3 language model](#). Now, the [Redmond tech giant has announced that its AI model has outperformed humans](#) in SuperGLUE benchmarks.

AI hype



NEWS ▾

FEATURES ▾

FORUMS

STORE ▾

MORE ▾

Microsoft Research Blog

Microsoft DeBERTa surpasses human performance on the SuperGLUE benchmark

Published January 6, 2021

By [Pengcheng He](#), Principal SDE; [Xiaodong Liu](#), Principal Researcher; [Jianfeng Gao](#), Distinguished Scientist & Vice President; [Weizhu Chen](#), Partner Science Manager

Research Area

 [Artificial intelligence](#)

AI hype

[NEWS](#) ▾[FEATURES](#) ▾[FORUMS](#)[STORE](#) ▾[MORE](#) ▾

Microsoft Research Blog

Microsoft De SuperGLUE

Published January 6, 2021

By [Pengcheng He](#), Principal SI
Distinguished Scientist & Vice

Natural language understanding (NLU) is one of the longest running goals in AI, and SuperGLUE is currently among the most challenging benchmarks for evaluating NLU models. The benchmark consists of a wide range of NLU tasks, including question answering, natural language inference, co-reference resolution, word sense disambiguation, and others. Take the causal reasoning task (COPA in Figure 1) as an example. Given the premise “the child became immune to the disease” and the question “what’s the cause for this?” the model is asked to choose an answer from two plausible candidates: 1) “he avoided exposure to the disease” and 2) “he received the vaccine for the disease.” While it is easy for a human to choose the right answer, it is challenging for an AI model. To get the right answer, the model needs to understand the causal relationship between the premise and those plausible options.

the

Problematize simplified tasks

- Cognitive scientists are domain experts in many of the tasks of “AI”
- We (unfortunately) have a job to do here
- Bender & Koller 2020: No, language models are not “understanding”
- Raji et al 2021: Claims of “general” language understanding/“general” visual understanding are unsupported and problematic

Critically analyze claims of “AI” capabilities

- Unending supply of over-hyped stories about “AI” in the news media
- The skills we hone critically examining simplified tasks in our domain transfer!



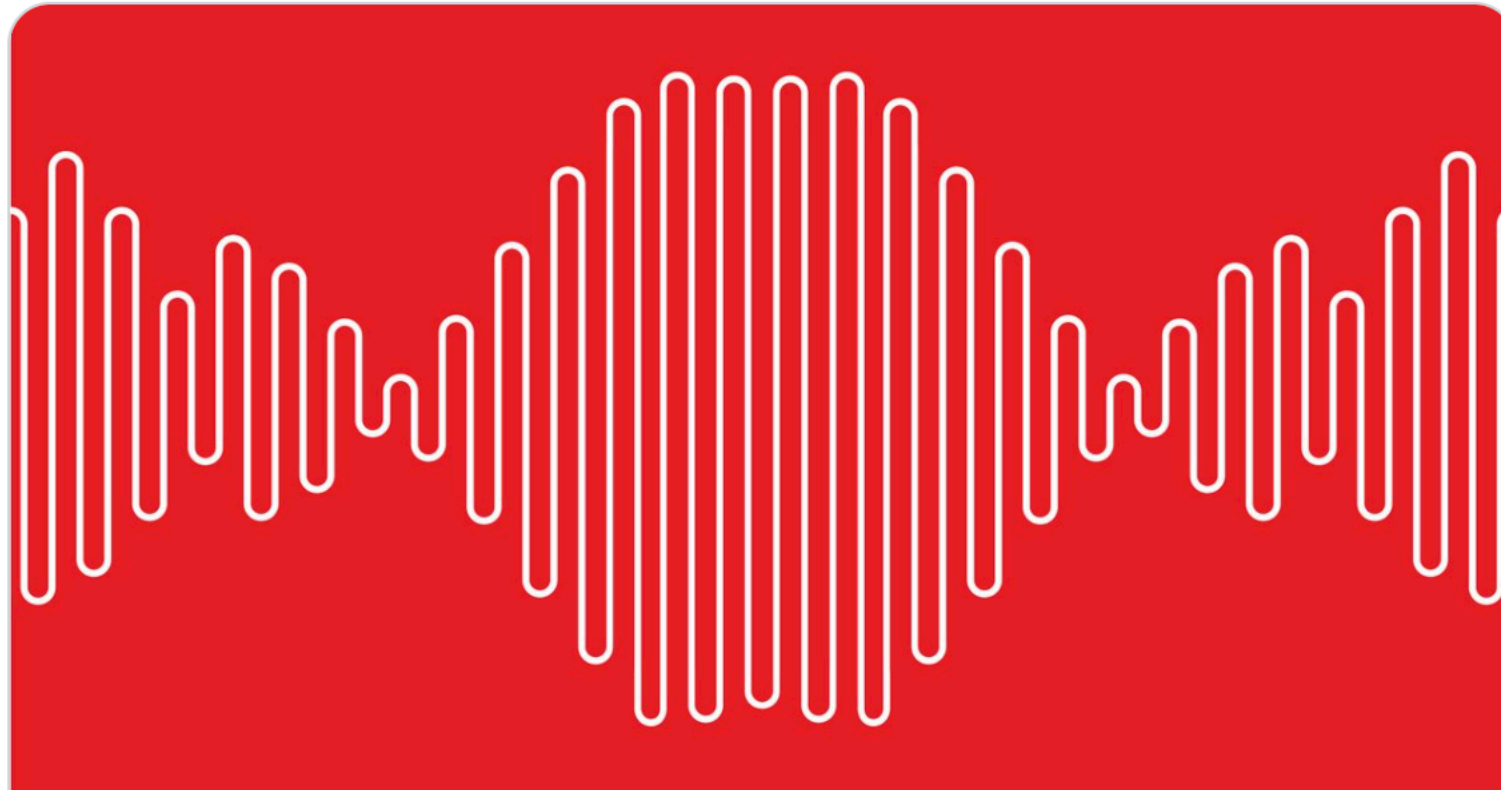
Emily M. Bender

@emilymbender



I find this reporting infuriating, so I'm going to use it to create a mini-lesson in detecting [#AIhype](#).

If you're interested in following this lesson, please read the article, making note of what you think sounds exciting and what makes you skeptical.



nytimes.com

Can A.I.-Driven Voice Analysis Help Identify Mental Disorders?

Early tests have been promising, but issues involving bias, privacy and mistrust of “black box” algorithms are possible pitfalls.



Emily M. Bender @emilymbender · Jul 3



Not it **effing** can't. This headline is breathtakingly irresponsible.

h/t [@hypervisible](#)



bloomberg.com

Algorithm Claims to Predict Crime in US Cities Before It Happens

A new computer algorithm can now forecast crime in a big city near you — apparently.

85

799

3,332



Key questions to ask: Ideally out loud

- How is the task defined? What's the input / what's the output?
- Is there any reason to believe that the input provides sufficient information to produce accurate output?
- Where did the training data come from and how was it validated?
- Who benefits from assuming the output is accurate?
- Can this technology be used for surveillance, harassment, or otherwise denying people their rights?



Emily M. Bender

@emilymbender



And now we are squarely in the ML techno-solutionism danger zone: It's established that it would be beneficial to have something that can do X with only Y input, but not that it's actually possible to do X with only Y input.

11:56 AM · Apr 6, 2022 · Twitter Web App

28 Retweets **9** Quote Tweets **163** Likes



Emily M. Bender @emilymbender · Apr 6



Replying to @emilymbender

On the other hand, you can always train an ML system that takes y_s (elements of Y) as input and gives x_s (elements of X) as output and thus LOOKS LIKE it's doing X with only Y input.



1



17



140



Critically analyze claims of “AI” capabilities: Sometimes (often) it’s actually people

- "A clear alternative to “AI” is to focus on the people present in the system. If a program is able to distinguish cats from dogs, don’t talk about how a machine is learning to see. Instead talk about how people contributed examples in order to define the visual qualities distinguishing “cats” from “dogs” in a rigorous way for the first time. There's always a second way to conceive of any situation in which AI is purported. **This matters, because the AI way of thinking can distract from the responsibility of humans.**"

Lanier & Weyl 2020

<https://www.wired.com/story/opinion-ai-is-an-ideology-not-a-technology/>

Decenter whiteness/WEIRDs/English

- Insist on specificity:
 - e.g. “natural language” is not a synonym for “English” (#BenderRule, Bender 2019)
- Insist on success criteria that don’t leave concerns of minoritized people as an afterthought or nice-to-have (Raji 2020, Birhane 2021)
- Question the entire metaphor of “artificial intelligence”
 - for how it aligns “AI” with whiteness
 - for how it devalues cognitive capabilities outside those prized by rationality

Advocate for broader distribution of research funds

- Computer science in general and ML/AI in particular is suffering from too much funding
 - Both industry funds & national research support
- Impossible “pace” of research within ML/AI
- Neglect of alternative research paths
- Power imbalance between computer science and domain areas computer science should be partnering with



Advocate for broader distribution of research funds

- Large language models are especially problematic, providing the illusion of very general systems
 - Can generate seemingly coherent text (in English, and some other languages) on a very wide variety of topics
 - Reach exceeds their grasp: the generated text is untethered from any model of the domain
 - Reach exceeds our grasp: people inventing tasks without clear use cases
=> unable to evaluate or determine safety parameters
 - cf Talat et al 2022 “unsafe at any accuracy”

Envision alternative pro-human research paths

- Present “AI” research is throwing tremendous resources at essentially made up problems
 - Including carbon budget and other natural resources (Strubell et al 2019, Schwartz et al 2019, Borning et al 2020)
- Hundred billion parameter models trained on enormous datasets aren’t naturally occurring phenomena that we should seek to understand
- Identify practical problems that could benefit from computational solutions (rather than ML solutions that could benefit from problems)
- Identify scientific questions in the cognitive sciences that could be elucidated with ML or other computational techniques

Envision alternative pro-human research paths

- Engage in the on-going discussion of ethical considerations/broader impacts
- As cognitive scientists, we are positioned to see the people involved
 - As research participants
 - As others impacted by technology as developed & deployed

Engage in public scholarship

- “AI” has captured the public imagination
- Tech firms selling “AI” are pushing to shape the regulatory landscape
 - Asserting claims to data and the digital world (Zuboff 2019)
 - Selling surveillance technology and other deeply problematic applications of “AI” (predictive policing, recidivism prediction)
- We can’t get to sensible regulation without an informed public and informed policy makers

Public scholarship on Twitter

- Cultivate a set of accounts to follow, especially people who experience different forms of oppression
- Build a network of people speaking out about similar things
 - Backchannel support is critical
- Be prepared to say the same thing over and over in many ways
- (It's also fine to not do Twitter; it's not for everyone)

@emilymbender

Public scholarship in the media

- You don't have to take all media requests, but it can be very valuable to take some
- See if your institution has media training
- Check the work of the journalist contacting you (due diligence)
- Ask to fact check direct quotes
- Speak from your expertise only
- Be prepared to say the same thing over and over in many ways

Hold space for public scholarship by others

- The academy is still struggling to recognize public scholarship as scholarship (Hale 2008, Matias 2021)
- Including broad engagement with the media
- As well as specific engagement with the communities our research ostensibly serves

“AI” research, development and sales involves dehumanization on many levels

- Computational metaphor
- Digital physiognomy
- “Ground lies”
- Irrelationality
- Ghost work
- Reinforcing the white racial frame

Cognitive scientists are well positioned to resist this, and we have many roles to play

- Problematize simplified tasks
- Critically analyze claims of “AI” capabilities
- Decenter whiteness/WEIRDs/English
- Advocate for broader distribution of research funds
- Envision alternative pro-human development paths
- Engage in & support public scholarship


Cognitive scientists are well positioned to resist this, and we have many roles to play

- Problematize simplified tasks
- Critically analyze claims of “AI” capabilities
- Decenter whiteness/WEIRDs/English
- Advocate for broader distribution of research funds
- Envision alternative pro-human development paths
- Engage in & support public scholarship

Thank you!

<https://bit.ly/EMB-COGSCI22>

References

- Agüera y Arcas, B. (2021). Do large language models understand us? Blog post on Medium.com, <https://medium.com/@blaisea/do-large-language-models-understand-us-6f881d6d8e75>.
- Agüera y Arcas, B., Mitchell, M., and Todorov, A. (2017). Physiognomy’s new clothes. Blog post on Medium.com, <https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a>.
- Alkhatib, A. (2021). To live in their utopia: Why algorithmic systems create absurd outcomes. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–9.
- Baria, A. T. and Cross, K. (2021). The brain is a computer is a brain: Neuroscience’s internal debate and the social significance of the computational metaphor. <https://arxiv.org/abs/2107.14042>.
- Bender, E. M. (2019). The #benderrule: On naming the languages we study and why it matters. *The Gradient*. <https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/>.
- Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S., and et al (2021). On the dangers of stochastic parrots: Can language models be too big?  In *Proceedings of FAccT 2021*.
- Bender, E. M. and Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.
- Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, 2(2):100205.
- Borning, A., Friedman, B., and Logler, N. (2020). The ‘invisible’ materiality of information technology. *Communications of the ACM*, 63(6):5764.
- Cave, S. and Dihal, K. (2020). The whiteness of AI. *Philosophy & Technology*, 33(4):685–703.
- Dijkstra, E. W. (1985, September 25). On anthropomorphism in science. Philosophers’ Lunch, Austin, TX, transcript available at <https://www.cs.utexas.edu/users/EWD/ewd09xx/EWD936.PDF>.
- Hale, C. R., editor (2008). *Engaging Contradictions: Theory, Politics, and Methods of Activist Scholarship*. University of California Press, Berkeley and Los Angeles CA.
- Kronfeldner, M., editor (2021). *The Routledge Handbook of Dehumanization*. Routledge, New York.
- Kyselo, M. (2014). The body social: An enactive approach to the self. *Frontiers in Psychology*, 5:1–16.
- Lanier, J. and Weyl, E. G. (2020). Ai is an ideology, not a technology. *WIRED*.
- Marino, M. C. (2014). The racial formation of chatbots. *CLCWeb: Comparative Literature and Culture*, 16(5):13.
- Matias, J. N. (2021). Tenure lessons for engaged scholars. Blog post on Medium.com, <https://natematias.medium.com/tenure-lessons-for-engaged-scholars-6fe0b1a6745>.
- Paullada, A., Raji, I. D., Bender, E. M., Denton, E., and Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2.
- Raji, D. (2020). How our data encodes systematic racism. *MIT Technology Review*.
- Raji, I. D., Bender, E. M., Paullada, A., Denton, E., and Hanna, A. (2021). Ai and the everything in the whole wide world benchmark. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*.
- Roberts, S. T. (2021). Your AI is a human. In Mullaney, T. S., Peters, B., Hicks, M., and Philip, K., editors, *Your Computer is on Fire*, pages 51–70. MIT Press.
- Scheuerman, M. K., Hanna, A., and Denton, E. (2021). Do datasets have politics? Disciplinary values in computer vision dataset development. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).
- Schlangen, D. (2021). Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, Online. Association for Computational Linguistics.
- Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. (2020). Green AI. *Communications of the ACM*, 63(12):5463.
- Scott, J. C. (1998). *Seeing Like A State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, New Haven, CT.

- Stark, L. and Hutson, J. (2022). Physiognomic artificial intelligence. *Fordham Intellectual Property, Media and Entertainment Law Journal*, 32(4).
- Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.
- Talat, Z., Blix, H., Valvoda, J., Ganesh, M. I., Cotterell, R., and Williams, A. (2022). On the machine learning of ethical judgments from natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 769–779, Seattle, United States. Association for Computational Linguistics.
- Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3266–3280.
- Weheliye, A. G. (2014). *Habeas Viscus: Racializing Assemblages, Biopolitics, and Black Feminist Theories of the Human*. Duke University Press, Durham, NC.
- Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Public Affairs Books, New York.