

Resisting Dehumanization in the Age of ‘AI’

Emily M. Bender

Department of Linguistics

University of Washington

Abstract

The production and promotion of so-called ‘AI’ technology involves dehumanization on many fronts. I explore these processes of dehumanization and the role that cognitive science can play by bringing a richer picture of human cognition to the discourse.

Keywords Artificial intelligence, cognitive science, dehumanization, interdisciplinarity

1 Introduction

The ways in which so-called ‘artificial intelligence’ (‘AI’) is described in the research literature, the popular press, and blogs or other advertizing copy from tech companies involves dehumanization in many ways.¹ Fortunately, cognitive scientists are well-positioned to resist this trend, based on our research practice and expertise. Figure 1 summarizes both the kinds of dehumanization and the possibilities of resistance. In this paper, I will detail each of these ways in which the practice of AI enacts dehumanization (Section 2) and the ways in which cognitive scientists can push back (Section 3). To set the stage, I describe my path to this topic and provide a working definition of dehumanization.

Researcher’s Path My path to this topic paper runs through two papers I co-authored. The first is Bender & Koller, 2020, written in reaction to widespread claims that language models actually understand language. Language models are systems trained to output plausible sequences of words

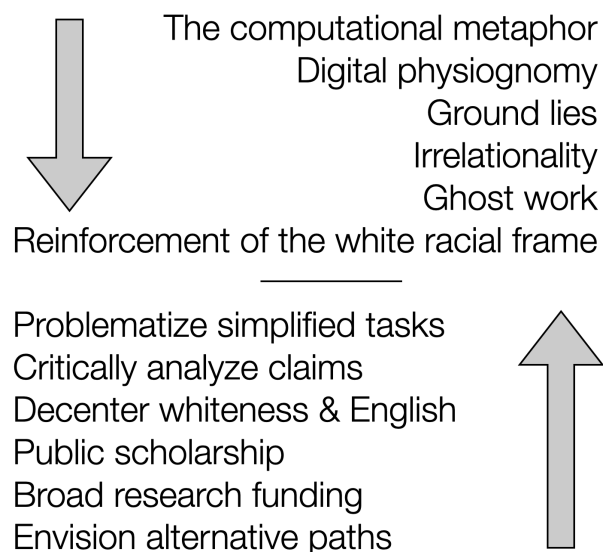


Figure 1: Six ways in which ‘AI’ practice enacts dehumanization (top) and six ways cognitive scientists can resist (bottom)

or letters based on a context of sequences of words or letters. Since about 2019, language models for English (e.g. GPT-3; Brown et al., 2020) model distribution of word forms closely enough to produce seemingly coherent text. A linguistic perspective, however, shows that this coherence is in the eye of the beholder: language models have no communicative intent nor understanding of what the word sequences mean. Language is a system of signs, i.e., pairings of form and meaning or signifier and signified (de Saussure, 1959), but language models are trained only on form and have no access to the meaning side of it. What’s not in the training data can’t be ‘learned’ by the model.

The second paper is Bender, Gebru, McMillan-Major, & Shmitchell, 2021, in which we observed the trend, already very evident in 2020, towards ever larger language models and asked: What could go wrong here? We surveyed the literature on risks associated with language models and mitigation strategies. We considered environmental costs through the lens of environmental racism; financial costs and their impact on research participation; how the training data sets come to be filled with hegemonic viewpoints and worse, without documentation or accountability for their content; and finally how synthetic text generating machines can reproduce the systems of oppression from their data sets, while also misleading humans who can’t help but make sense of text that we encounter.

We coined the term *stochastic parrots* to evoke the way in which large language models, run as text synthesis machines, “haphazardly [stitch] together sequences of linguistic forms [...] according to probabilistic information about how they combine, but without any reference to meaning” (p. 617).

I now often get asked “How do I know that you’re not just a stochastic parrot?” I have decided I am not going to have conversations with people who will not posit my humanity as a basic axiom of the conversation. Acknowledging my privilege (living as a white cis person in the US) to not have noticed before, that phrasing led me to think about dehumanization across AI.

Dehumanization: Working Definition From the broad literature on dehumanization, I draw on two sources as touch points in making a working definition:

“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” (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)

Kronfeldner’s words are from the preface of an edited volume on dehumanization and reflect the difficulty of operationalizing a definition. Weheliye’s remarks, not directly a definition of dehumanization, are informative for their clarity about the experience of racialization.

From those starting points I come to a three-part working definition. *Dehumanization* is any of:

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

Here, I am using the phrase *fully human* to mean entitled to all rights recognized as human rights, equally in possession of internal life and point of view, and welcomed as one’s full self. This working definition acknowledges the cognitive process of the person doing the dehumanization, the acts that express it, and the experience of being the target of it.

2 Dehumanization in the Research, Development, and Sales of ‘AI’

I outline six ways the development and sales of ‘artificial intelligence’ contributes to dehumanization.

The Computational Metaphor Baria and Cross (2021) analyze the computational metaphor in neuroscience: THE BRAIN IS A COMPUTER.² They note it is a bi-directional metaphor, where the other half, THE COMPUTER IS A BRAIN, is used pervasively by technologists. About the metaphor as a whole, Baria and Cross (2021, 2) write it “afford[s] the human mind less complexity than is owed, and the computer more wisdom than is due.” In short, this metaphor builds up computers at the expense of how we understand humans. Furthermore Baria and Cross identify in the computational metaphor a hierarchy of human value defined in terms of ideologies around intelligence, where ‘rationality’ is valued above ‘emotionality’, affording more power to those who display more machine-like qualities. Ultimately, “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.” (*Ibid.*, 6)

Especially pernicious is the appropriation of the experiences of disabled people to assert the humanity of AI. This rhetorical turn is elaborated by Aguëra y Arcas (2021), who asserts that large language models are like Deafblind people. Under the heading of ‘modality chauvinism’ he calls on the writings of Daniel Kish, who’s blind, and Helen Keller, who’s Deafblind, to argue that no one sensory system is required for humans to develop concepts, even sensory concepts. But his purpose

in doing so is to argue that large language models might therefore also be developing concepts. He can't show that large language models are like people with internal lives and relationships and full personhood and so he ends up dehumanizing blind and Deafblind people by saying that they are like something that is patently not human, specifically because of their disability.

Digital Physiognomy Researchers pursuing *digital physiognomy* claim to predict such things as criminality, sexual orientation, employability, political leanings, and psychopathy based on photos, videos, voice samples, etc. (see Stark & Hutson, 2022). Thus the long-discredited pseudoscience of physiognomy has come back, using computers for a veneer of objectivity (Agüera y Arcas, Mitchell, & Todorov, 2017). Classifying people by gender or race based on how they look is equally problematic: it flattens human identities and experiences into fabricated categories which are falsely imagined to be intrinsic, immutable, and externally observable. Not only are such classifications fundamentally not possible (the information simply isn't in the input signal) but attempting them is harmful, entailing the objectification of the people being subjected to such systems, the rigidification of categories, and the misattribution of characteristics and identities.

'Ground Lies' Next consider the way training data sets for 'AI' are mythologized as being representative (Paullada, Raji, Bender, Denton, & Hanna, 2021; Raji, Bender, Paullada, Denton, & Hanna, 2021; Scheuerman, Hanna, & Denton, 2021). This viewpoint holds that data collected without care is 'naturally occurring' and therefore a true representation — despite biasing decisions of: where to collect data from, how to collect it, how to filter it, what labels to apply, who should apply the labels, how the labels are verified, and more. If we don't actively work to curate the data sets that we want, then we will be collecting data sets that are representative of dehumanizing ideologies like white supremacy and calling lies 'ground truth' (Raji, 2020).³

Irrelationality Humans are thoroughly relational in our experience of ourselves, our lives, and our world. As Birhane (2021, 5) and others argue, our understanding of our world is inherently bound up in our culture, history, and lived experience, all of which are primarily built with and through our relationships to other people. Kyselo (2014, 8) makes the case that even our very selves

emerge in interaction. The ‘knowing’ that we program into ‘AI’ is, in contrast, *irrelational*, that is, ostensibly abstracted from the web of relations within which we have all of these experiences. Birhane (2021, 3) observes and elucidates the deep commitment to rationalism and the supposed potential of a ‘God’s eye view’ built into these systems — where the ‘God’s eye view’ aligns with the perspective of those with power in society.

Machines aren’t designed to apply what Scott (1998) termed *metis*: the way people working with rules creatively navigate through them based on the facts at hand. Computers can work with hard, hand-coded rules or statistical processing (either so-called deep learning or more traditional statistical methods) based on historical data, but never in relationship to the full situation at hand and thus never with wisdom (Weizenbaum, 1976). This *irrelationality* ends up devaluing humanity while also leaving no space for it. We must recognize that attempting to make decision-making more fair by replacing computers with humans is wishful thinking: pushing off difficult decisions to supposedly impartial machines that encode hegemonic values and lack flexibility is an exercise in shirking accountability, while again devaluing the human web of relationships (Roberts, 2021; Alkhatib, 2021).

Ghost Work Human effort is everywhere in these systems: labeling data, design and evaluation, and as a backstop for when the computer fails on some input. Tech firms ship those tasks off to microworkers on crowdsourcing platforms (Gray & Suri, 2019; Roberts, 2021), hiding often grueling labor (e.g., content moderation) and the humanity of the microworkers behind the illusion of ‘AI’. Furthermore, crowdwork platforms encapsulate microworkers in worker IDs so requestors are encouraged to treat workers as interchangeable software components accessible through an application programming interface (Gray & Suri, 2019).

Reinforcement of the White Racial Frame The (Anglophone) discourse around ‘AI’ reinforces the white racial frame: the implicit and assumed understanding of racial categories that supports systemic racism and white supremacy, shaping decision-making wherever it is not actively resisted. Cave and Dihal (2020) document how (within Anglo Western culture) ‘AI’ is racialized as white: depictions of robots and actual robots are frequently made with white exteriors

and even robots without physicality (voice assistants or text-based chatbots) adopt white-coded speaking styles (Marino, 2014).⁴ Cave and Dihal (2020) hypothesize that this shows the influence of the white racial frame: the traits that are associated with ‘AI’—intelligence, professionalism, power—are those that the white racial frame ascribes to white people. Further and damningly, they hypothesize that white people overrepresented in the ‘AI’ workforce are designing a set of servants who would let them avoid interacting with people who aren’t white. This is problematic on its face, but also in more subtle ways: The whiteness of ‘AI’ is dehumanizing because the white racial frame itself is dehumanizing to anybody who is not ascribed whiteness—recall Weheliye’s (2014) description of blackness within the white racial frame.

3 What Can Cognitive Scientists Do about This?

My purpose is not just to surface these many facets of dehumanization in ‘AI’, but also to talk about what we can do about it. Here are six suggestions.

Problematize Simplified Tasks Machine learning (ML) research is driven by *tasks*, defined either through informal descriptions of what the algorithm is supposed to do (e.g., transcription of spoken Kinyarwanda from audio recordings) or through data sets pairing inputs with expected outputs (Schlangen, 2021). Tasks are supposed to represent vague *capabilities*, hypothesized to underlie the possibility of doing the task. However, many ML tasks lack construct validity (see Raji et al., 2021 and works cited there): we don’t know that an algorithm’s score on test data means that it has the capability that a human would use to do the task ostensibly represented by the examples.

How does such misinterpretation of the results of ML tasks come about? I think it follows from the idea that computer science (CS) should provide general solutions. To achieve this, many researchers don’t look at specific data so as to avoid creating systems overly tailored to that data. But this also prevents understanding the shape of the problem. The division of labor between those who construct datasets and those who build the algorithms sets the conditions for wild over-claims about system capabilities. The ‘marketing’ parts of dataset papers get repeated as

serious, well-founded descriptions of the capabilities the tasks supposedly represent. For example, the SuperGLUE paper frames their benchmark as a “rigorous test of language understanding” composed of tasks that “test a system’s ability to understand and reason about texts in English” (Wang et al., 2019, 2,4). This leads to a 2021 blogpost⁵ from Microsoft research titled “Microsoft DeBERTa surpasses human performance on the SuperGLUE benchmark,” and making such claims as “To get the right answer, the model needs to understand the causal relationship between the premise and those plausible options.” The media then reports it as “Microsoft’s AI model has outperformed humans in natural language understanding.”⁶

Cognitive scientists are the domain experts in the capabilities that these tasks supposedly test for. We must bring our expertise to bear in contextualizing how the tasks relate to the capabilities we study. Bender and Koller (2020) respond to the claim that language models ‘understand’ (supported by scores on benchmarks like SuperGLUE) by laying out a linguist’s perspective on meaning and understanding. Raji et al. (2021) similarly problematize claims of ‘generality’ in ML. Everyone who is researching things that humans do with our cognition can probably find something where folks doing ML are making spurious claims—and then be in a position to say: That’s not how that works!

Critically Analyze Claims of ‘AI’ Capabilities Problematizing simplified tasks hones skills for critically examining claims of ‘AI’ capabilities. Faced with hype-driven headlines like ‘Can A.I.-Driven Voice Analysis Help Identify Mental Disorders?’⁷ and ‘Algorithm Predicts Crime in US Cities Before It Happens’,⁸ one can ask:

- How is this task defined?
- What’s the input and what’s the output?
- Does the input provide sufficient information to produce accurate output?
- Where did the training data come from and how was it validated?
- Can this technology be used for surveillance, harassment, or otherwise denying people their rights?

Exploring these questions publicly, even just raising them, helps deflate ‘AI’ hype.

The hype is especially dangerous in cases where it would be beneficial to have something that can do X with only Y input, but we haven’t established that it’s possible. For example, it would be useful to give people accurate actionable mental health diagnoses based only on their voice (assuming we could prevent the negative surveillance use cases). This is a danger zone for ML tech solutionism because we can always construct ML systems that look like they are doing the job, taking Y and giving X. But if we can’t validate system outputs, they’re useless. Until and unless we have comprehensive and robustly enforced regulation of this kind of misuse of ‘AI’ technologies, it is up to us to expose it for what it is.

Another angle for critical analysis is looking for the people in the ‘AI’ system. Lanier and Weyl (2020) argue that ‘AI’ is an ideology, not a technology, reminding us that “the AI way of thinking can distract from the responsibility of humans.” Here, we are concerned with the people who designed the system and decided how to use it, whom we should ask: why is this a safe thing to do, why did you frame the task this way, and whose interests does it serve?

Decenter Whiteness/English A third act of resistance is the decentering of identities or characteristics that get accorded the status of ‘default’ or ‘unmarked’ (e.g., whiteness, speaking English). If we don’t name English when we’re working on it (Bender, 2019), or white people, Western society, or middle and upper class people when we’re working on them, then we misinterpret results about those ‘unmarked’ groups as general and results about other specific groups as parochial. This both weakens our research and contributes to maintaining (in the US context, at least) the white racial frame. Similarly we should insist on success criteria that don’t leave the concerns of minoritized people as an afterthought (Raji, 2020; Birhane, 2021). A system is not accurate if it’s not accurate for everybody: if it is failing on non-white people it is failing.

We should also question the entire metaphor of artificial intelligence for how it aligns intelligence with whiteness and for how it devalues cognitive capabilities that are outside of those prized by rationality. Here, I believe that the field of psychology has some work to do, taking accountability for how ‘intelligence’ is discussed and how it relates to the white racial frame. Doing so will be a powerful force for resisting dehumanization from ‘AI’.

Engage in Public Scholarship ‘AI’ has captured the public imagination, helped by decades of science fiction. Tech firms selling ‘AI’ are shaping the regulatory landscape, asserting claims to data in the digital world (Zuboff, 2019) and selling surveillance technology and other deeply problematic applications of ‘AI’. Sensible regulation requires an informed public and informed policy makers — which in turn requires public scholarship.

One way to do this is on social media. As a first step, cultivate a set of accounts to follow and learn from, especially people who experience different forms of oppression. Then build a network of people who are speaking out about similar things: spaces to offer mutual support are key to doing this work without burning out. Similarly, connecting with traditional media can be both valuable and time consuming. Prepare for this by doing institutional media training (if available), learning to vet journalists before engaging with them, and speaking within one’s expertise. For both social and traditional media, be prepared to say the same things over and over, to educate new audiences.

Public scholarship also includes engagement with policymakers and policy advocacy. When I have been invited to talk with policymakers, I have taken the opportunity to advocate for policy goals on the basis of my understanding of how technology like large language models works and how human language processing capability leaves us vulnerable to being misled by the technology. These goals include transparency (of the fact of the use of automation, of the data used for training), accountability (held by people for system output), application of existing regulations (not assuming that new technology means that previous protections of rights become moot), and sufficient funding of social science and humanities research to be able to understand the impact of technology on society.

Finally, in addition to engaging in public scholarship, we must also hold space for others who are doing so. We can support public scholarship whenever we review tenure packets, allocate grant funding, or otherwise wield power.

Advocate for Broader Distribution of Research Funds I believe that CS, and especially ‘AI’/ML, is overfunded, creating a power imbalance between CS and the domain areas it should be partnering with. A stark example is the nominally interdisciplinary US NSF Program on Fairness in Artificial Intelligence in Collaboration with Amazon (FAI),⁹ where proposals required principal

investigators from CS departments. When questions are about fairness of technology, the core scholarship area required to answer those questions isn't necessarily CS. With more equitably distributed research funding, computer scientists would have to enter into interdisciplinary collaborations with others as equals.

Large language models are especially problematic here. With systems that can generate what looks like legal contracts, medical advice, scientific papers, etc., we might think we are about to have systems that can actually do those things. Without significant funding of non-CS research in the content areas implicated and the way technology in those content areas would affect society, we risk finding ourselves in a situation where the computer scientists proclaim 'solved it!' on the strength of their stochastic parrots and regulators and other decision makers believe them.

Envision Alternative Pro-Human Research Paths 'AI' research is currently throwing resources at made-up problems such as automating morality judgments, including not just financial or research time resources, but also things like carbon budget and other natural resources (Strubell, Ganesh, & McCallum, 2019). We should be identifying practical problems that could benefit from computational solutions, not ML solutions that could benefit from problems. One step is to engage in the ongoing discussion of ethical considerations in ML conferences. Cognitive scientists are particularly well-positioned to focus on the people involved in this kind of research, so we have a role to play in helping to shape that conversation. The more work in 'AI' and ML is held accountable for its impacts in the world, the more we can shift the research focus to humanistic concerns.

4 Conclusion

I have reviewed the ways that 'AI' research development and sales involves dehumanization, from the computational metaphor, to digital physiognomy, to our 'ground lies', to 'AI's insistence on irrelationality, to the way we hide ghost work, and finally the way in which 'AI' reinforces the white racial frame. I have also explored the ways that cognitive scientists are well positioned to resist this dehumanization. We have many roles to play: we can problematize simplified tasks, critically analyze claims of 'AI' capabilities, work to decenter whiteness and WEIRDs and English, engage

in and support public scholarship, advocate for broader distribution of research funds, and envision alternative pro-human development paths.

Recommended Reading

- ‘How our data encodes systematic racism’ (Raji, 2020) succinctly and powerfully summarizes the dangers posed to Black people by datasets that encode systemic racism.
- *Artificial Unintelligence: How Computers Misunderstand the World* (Broussard, 2019) exposes the phenomenon of *techochauvanism* or the belief that technology can solve all problems and traces how attempted ‘solutions’ cause more harm.
- ‘To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes’ (Alkhatib, 2021) captures the tragedy and absurdity that follows when we structure our world around algorithmic rather than human systems.
- ‘The Brain Is a Computer Is a Brain: Neuroscience’s Internal Debate and the Social Significance of the Computational Metaphor’ (Baria & Cross, 2021) explores the origins of the computational metaphor in neuroscience, its uptake in computer science, and its effects.

Acknowledgments

This paper has benefitted greatly from discussions with Leon Derczynski, whom I also thank for his encouragement. Likewise, I am grateful to Andy Perfors and the other organizers of the COGSCI 2022 conference for their invitation to present the keynote talk that is the basis of this paper. I am also indebted to Deb Raji, Amandalynne Paullada, Alex Hanna, Remi Denton, Timnit Gebru and Margaret Mitchell for the wide-ranging discussions (largely over group chat) which helped shape these ideas. Finally, I thank the reviewers and editor for CDPS for their input.

Notes

¹The term ‘artificial intelligence’ is poorly defined and primarily has a marketing function. In this paper, I use scare quotes to emphasize this fact.

²All caps follows the tradition of metaphor theory (Lakoff & Johnson, 1980).

³There is also interesting work to be done in exploring the bias of datasets to better understand it (e.g., Garg, Schiebinger, Jurafsky, & Zou, 2018). Here, too, the data should be curated, but in this case to be representative of some population of interest, rather than to minimize bias.

⁴Apple’s Siri, released in 2011, got African-American voices in 2021.

⁵<https://www.microsoft.com/en-us/research/blog/microsoft-deberta-surpasses-human-performance-on-the-superglue-benchmark/>, accessed March 14, 2023

⁶<https://www.neowin.net/news/microsofts-ai-model-has-outperformed-humans-in-natural-language-understanding/>, accessed March 14, 2023


⁷<https://www.nytimes.com/2022/04/05/technology/ai-voice-analysis-mental-health.html>, accessed March 15, 2023

⁸Later corrected to ‘Algorithm Claims to Predict Crime in US Cities Before It Happens’, <https://www.bloomberg.com/news/articles/2022-06-30/new-algorithm-can-predict-crime-in-us-cities-a-week-before-it-happens>, accessed March 15, 2023

⁹<https://www.nsf.gov/pubs/2021/nsf21585/nsf21585.htm>

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., & 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* (pp. 1–9).
- Baria, A. T., & Cross, K. (2021). *The brain is a computer is a brain: Neuroscience’s internal debate and the social significance of the computational metaphor*. arXiv. Retrieved from <https://arxiv.org/abs/2107.14042> (<https://arxiv.org/abs/2107.14042>) doi: 10.48550/ARXIV.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. (2021). On the dangers of stochastic parrots: Can language models be too big?  In *Proceedings of facct 2021*.
- Bender, E. M., & 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* (pp. 5185–5198). Online: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/2020.acl-main.463> doi: 10.18653/v1/2020.acl-main.463
- Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, 2(2), 100205.
- Broussard, M. (2019). *Artificial unintelligence: How computers misunderstand the world*. Cambridge, MA: MIT Press.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hassel, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020, neurips 2020, december 6-12, 2020, virtual*. Retrieved from <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>
- Cave, S., & Dihal, K. (2020). The whiteness of AI. *Philosophy & Technology*, 33(4), 685–703.
- de Saussure, F. (1959). *Course in general linguistics*. New York: The Philosophical Society. (Translated by Wade Baskin)
- Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16), E3635–E3644. Retrieved from <https://www.pnas.org/doi/abs/10.1073/pnas.1720347115> doi: 10.1073/pnas.1720347115
- Gray, M. L., & Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass*. New York: Eamon Dolan Books.

- Kronfeldner, M. (Ed.). (2021). *The routledge handbook of dehumanization*. New York: Routledge.
- Kyselo, M. (2014). The body social: An enactive approach to the self. *Frontiers in Psychology*, 5, 1–16. Retrieved from <https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00986/full> doi: <https://doi.org/10.3389/fpsyg.2014.00986>
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.
- Lanier, J., & Weyl, E. G. (2020). Ai is an ideology, not a technology. *WIRED*. Retrieved from <https://www.wired.com/story/opinion-ai-is-an-ideology-not-a-technology/>
- Marino, M. C. (2014). The racial formation of chatbots. *CLCWeb: Comparative Literature and Culture*, 16(5), 13.
- Paullada, A., Raji, D., Bender, E. M., Denton, E., & 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*. Retrieved from <https://www.technologyreview.com/2020/12/10/1013617/racism-data-science-artificial-intelligence-ai-opinion/>
- Raji, D., Bender, E. M., Paullada, A., Denton, E., & 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 T. S. Mullaney, B. Peters, M. Hicks, & K. Philip (Eds.), *Your computer is on fire* (pp. 51–70). MIT Press.
- Scheuerman, M. K., Hanna, A., & Denton, E. (2021). Do datasets have politics? Disciplinary values in computer vision dataset development. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2). Retrieved from <https://doi.org/10.1145/3476058> doi: 10.1145/3476058
- 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)* (pp. 670–674). Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.acl-short.85> doi: 10.18653/v1/2021.acl-short.85
- Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have*

- failed*. New Haven, CT: Yale University Press.
- Stark, L., & Hutson, J. (2022). Physiognomic artificial intelligence. *Fordham Intellectual Property, Media and Entertainment Law Journal*, 32(4). Retrieved from <https://ir.lawnet.fordham.edu/iplj/vol32/iss4/2>
- Strubell, E., Ganesh, A., & 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* (pp. 3645–3650). Florence, Italy: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/P19-1355> doi: 10.18653/v1/P19-1355
- Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... Bowman, S. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in neural information processing systems* (pp. 3266–3280).
- Weheliye, A. G. (2014). *Habeas viscus: Racializing assemblages, biopolitics, and black feminist theories of the human*. Durham, NC: Duke University Press.
- Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. San Francisco CA: Freeman.
- Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. New York: Public Affairs Books.