Abhishek Gupta

AI Ethics Researcher, McGill University & District 3

Dr. Fenwick McKelvey

Assistant Professor, Communication Studies, Concordia U.

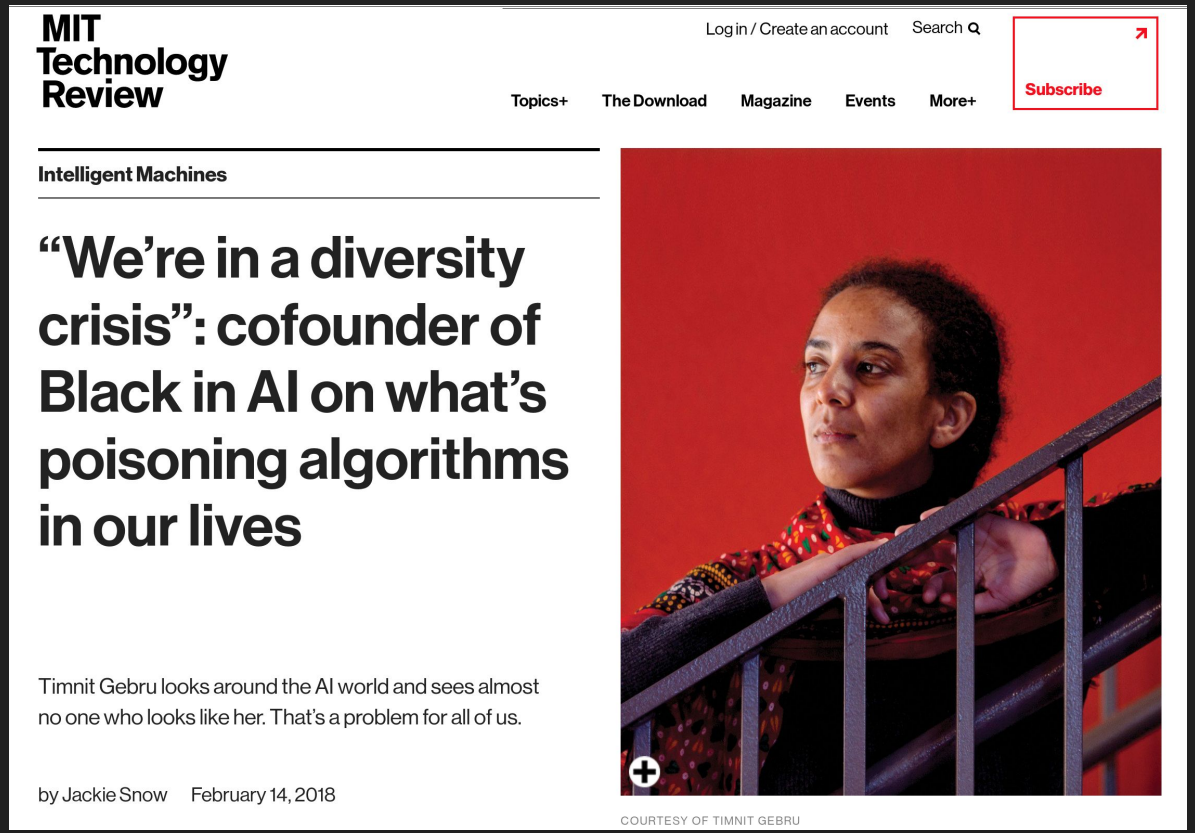
“People worry that computers will get too smart and take over the world, but the real problem is that they’re too stupid and they’ve already taken over the world.”
- Pedro Domingos

Source: https://www.nature.com/news/there-is-a-blind-spot-in-ai-research-1.20805#/b1

Inclusion

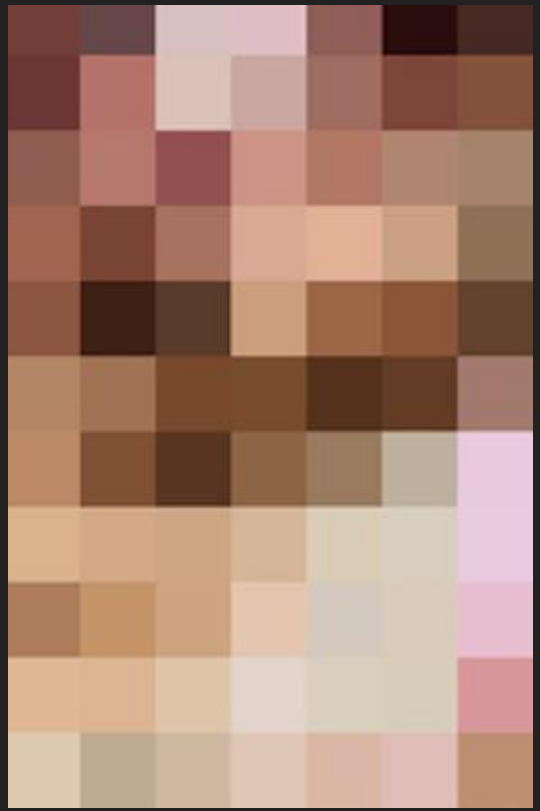
Explainability

Governance



How inclusive are the teams building AI?

Source:https://www.technologyreview.com/s/610192/were-in-a-diversity-crisis-black-in-ais-founder-on-whats-poisoning-the-algorithms-in-our/



Lack of inclusion leads to inequity

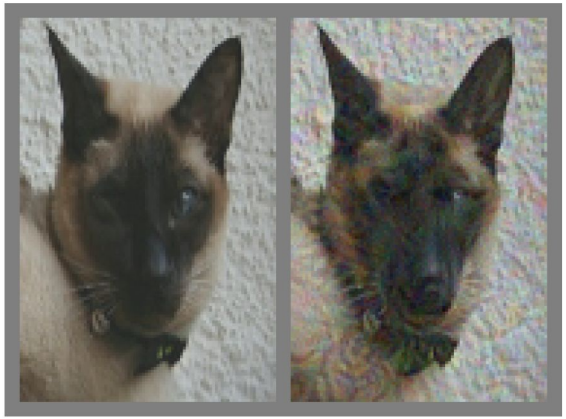


Figure 1. While in most cases our adversarial examples fool humans only after a brief exposure, the example depicted has a strong effect even for long viewing duration. On the left, we show an image of a cat. On the right, we show the same image after it has been adversarially perturbed to look like a dog. Although easily overlooked, note that cat-specific features can still be identified. For instance, the original boundary between the top of the cat head and the wall is still visible in the adversarial image, despite the top of the dog head seeming to be lower. Also, long white cat whiskers remain visible.

Adversarial examples in artificial intelligence

Source: https://arxiv.org/abs/1802.08195

Figure 7: Selected words projected along two axes: x is a projection onto the difference between the embeddings of the words *he* and *she*, and y is a direction learned in the embedding that captures gender neutrality, with gender neutral words above the line and gender specific words below the line. Our hard debiasing algorithm removes the gender pair associations for gender neutral words. In this figure, the words above the horizontal line would all be collapsed to the vertical line.

Machine learning can adopt the status quo

Source: Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *ArXiv:1607.06520 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1607.06520

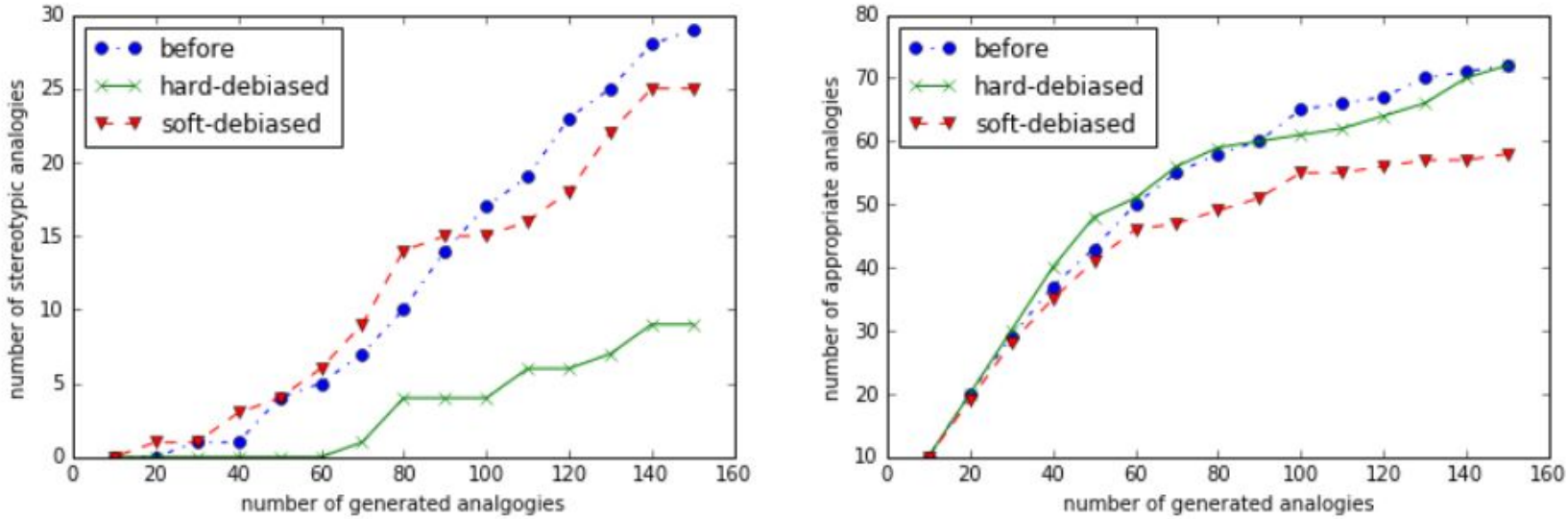


Figure 8: Number of stereotypical (Left) and appropriate (Right) analogies generated by wordembeddings before and after debiasing.

How can government assess bias and implementation?

Source: Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *ArXiv:1607.06520 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1607.06520

What does public consultation around AI look like?

Source: Abhishek Gupta, AI Ethics Meetup Group in Montreal

Governance after automation

Inclusion

More inclusive, interdisciplinary teams developing and deploying AI

Explainability

New public methods to assess and evaluate AI for bias

Governance

AI for the public good has to include the public