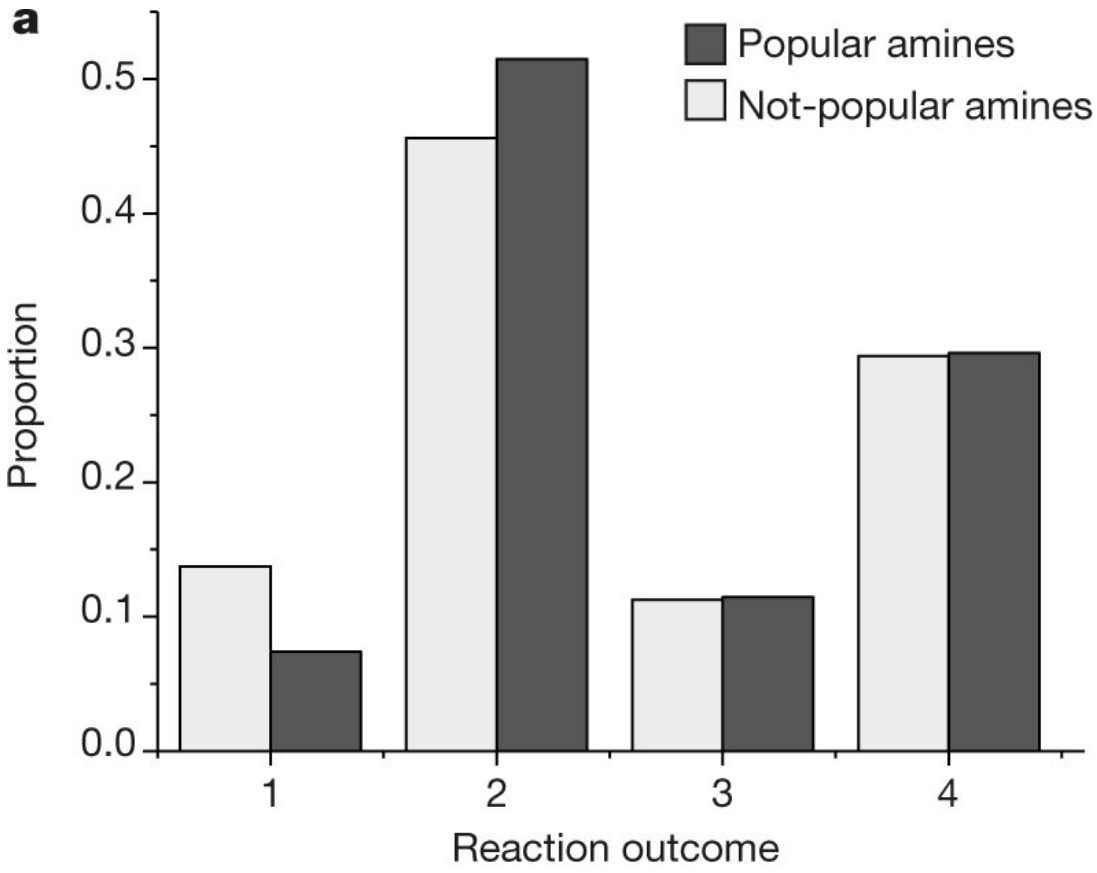
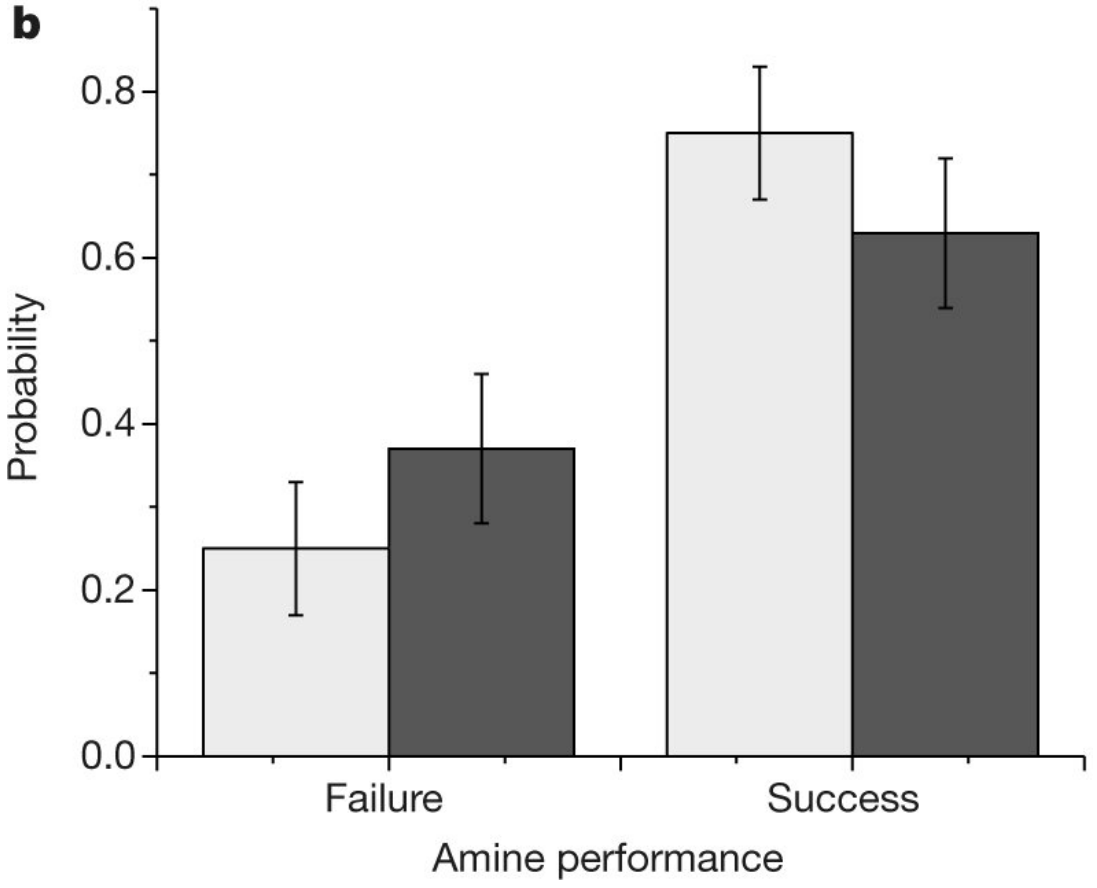
Chemists show how bias can crop up in machine learning algorithm results

September 12 2019, by Bob Yirka


a
Popular amines
Not-popular amines
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
Reaction outcome


b
Probability
0.0
0.2
0.4
0.6
0.8
Failure
Success
Amine performance

a, The proportion by outcome for each reaction, using the outcome scale described in Methods, for the popular and not-popular amines in the human-selected dataset. b, Estimated probability of observing at least one successful reaction (outcome 4) or failure (outcomes 1, 2 and 3) for a given amine, for the N = 27 popular and N = 28 not-popular amines among the human-selected dataset. Centre values indicate observed proportion of outcomes. Error bars indicate a bootstrap estimate of the standard deviation. Credit: *Nature* (2019). DOI: 10.1038/s41586-019-1540-5

A team of material scientists at Haverford College has shown how human bias in data can impact the results of machine-learning algorithms used to predict new reagents for use in making desired products. In their paper published in the journal *Nature*, the group describes testing a machine-learning algorithm with different types of datasets and what they found.

One of the more well-known applications of machine-learning algorithms is in facial recognition. But there are possible problems with such algorithms. One such problem occurs when a facial algorithm intended to look for an individual among many faces has been trained using people of just one race. In this new effort, the researchers wondered if bias, unintentional or otherwise, might be cropping up in machine learning algorithm results used in chemistry applications designed to look for new products.

Such algorithms use data describing the ingredients of reactions that result in the creation of a new product. But the data the system is trained on could have a major impact on the results. The researchers note that currently, such data is obtained from published research efforts, which means they are typically generated by humans. They note that the data

from such efforts could have been generated by the researchers themselves, or by other researchers working on separate efforts. Data could even come from a single person simply relating from memory, or from a professor's suggestion, or a graduate student with a bright idea. The point is, the data could be biased in terms of the background of the resource.

In this new effort, the researchers wanted to know if such biases might have an impact on the results of machine-learning algorithms used for chemistry applications. To find out, they looked at a specific set of materials called amine-templated vanadium borates. When they are synthesized successfully, crystals form—an easy way to determine if a reaction was successful.

The experiment consisted of training a machine-learning algorithm on data surrounding the synthesis of vanadium borates, and then programming the system to create its own. Some of the data collected by the researchers was human-generated, and some of it was collected randomly. They report that the algorithm trained on the random data did better at finding ways to synthesize the vanadium borates than when it used data generated from humans. They claim this shows a clear bias in the data that was created by humans.

More information: Xiwen Jia et al. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis, *Nature* (2019). DOI: 10.1038/s41586-019-1540-5

© 2019 Science X Network

Citation: Chemists show how bias can crop up in machine learning algorithm results (2019, September 12) retrieved 19 April 2024 from https://phys.org/news/2019-09-chemists-bias-crop-machine-algorithm.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.