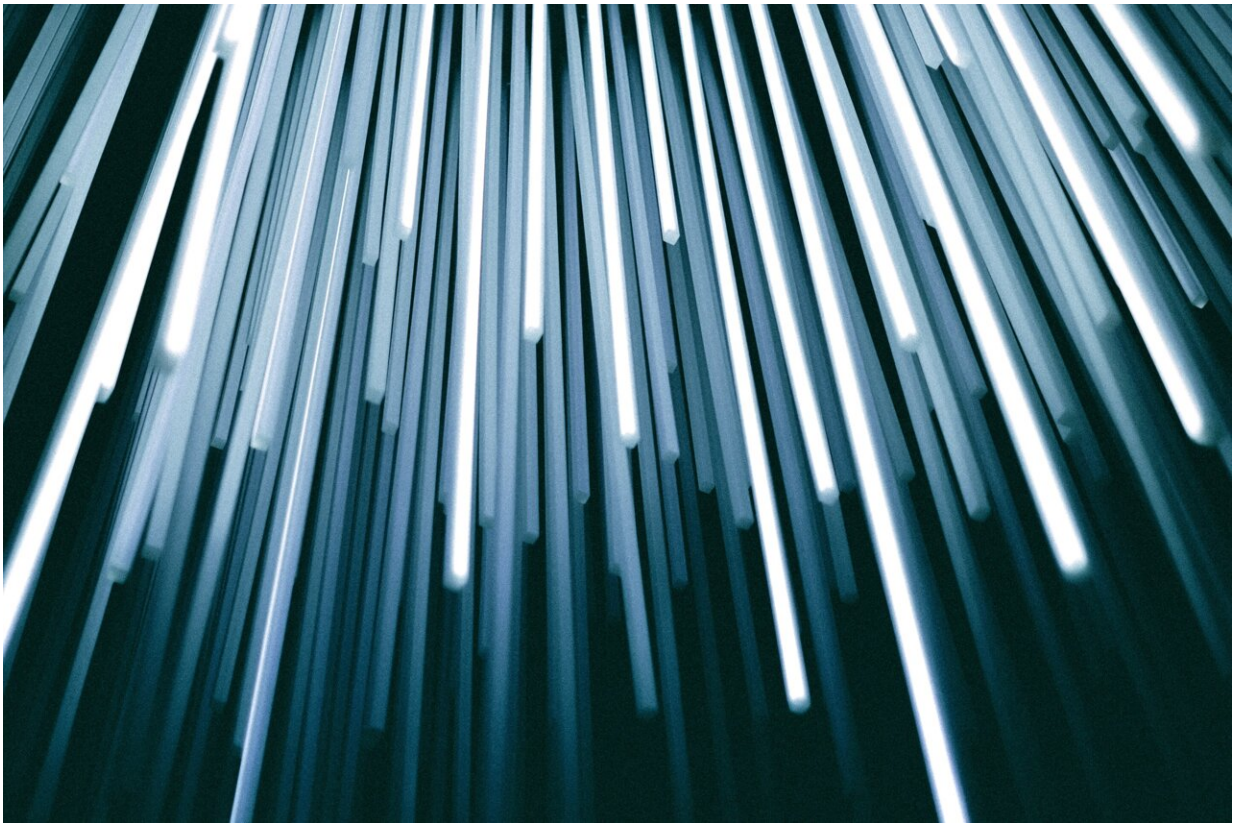
Why marking essays by algorithm risks rewarding the writing of 'bullshit'

October 20 2017, by Kai Riemer



Credit: CC0 Public Domain

Picture this: you have written an essay. You researched the topic and carefully constructed your argument. You submit your essay online and receive your grade within seconds. But how can anyone read,

comprehend and judge your essay that quickly?

Well, the answer is no one can. Your essay was marked by a computer. Would you trust the mark you received? Would you approach your next essay with the same effort and care?

These are the questions that parents, teachers and unions are asking about automated essay scoring (AES). The Australian Curriculum, Assessment and Reporting Authority (ACARA) proposes to use this program to grade essays, like persuasive writing questions, in its NAPLAN standardised testing scheme for primary and secondary schools.

ACARA has defended its decision and suggested that computer-based marking can match or even surpass the consistency of human markers.

In my view, this misses the point. Computers are unable to genuinely read and understand what a text is about. A good argument has little worth when marks are awarded by a structural comparison with other texts and not by judging its ideas.

More importantly though, we risk encouraging the writing of text that follows "the script" but essentially says nothing of worth. In other words, the writing of "bullshit".

How does algorithmic marking work?

It's not entirely clear how AES functions, but let's assume, in line with previous announcements, that it employs a form of machine-learning.

Here's how that could work: a machine-learning algorithm "learns" from a pool of training data – in this case, it may be "trained using more than 1,000 NAPLAN writing tests scored by human markers".

But it generally does not learn the criteria by which humans mark essays. Rather, machine learning consists of multiple layers of so-called "artificial neurons". These are statistical values that are gradually adjusted during the training period to associate certain inputs (structural text patterns, vocabulary, key words, semantic structure, paragraphing and sentence length) with certain outputs (high grades or low grades).

When marking a new essay, the algorithm makes a statistical inference by comparing the text with learned patterns and eventually matches it with a grade. Yet the algorithm cannot explain why this inference was reached.

Importantly, high grades are awarded to papers that show the structural features of highly persuasive writing – papers that follow the "persuasion rulebook", so to speak.

Rewarding bullshit

Are the claims by ACARA that algorithmic marking can match the consistency of human markers wrong? Probably not, but that's not the issue.

It's possible that machine-learning could reliably award higher grades for those papers that follow the structural script for persuasive writing. And it might indeed do this with higher consistency than human markers. Examples from other fields show this – for instance, in the classification of images in medical diagnosis. It will certainly be quicker and cheaper.

But it will not matter what a text is *about*: whether the argument is ethical, offensive or outright nonsensical, whether it conveys any coherent ideas or whether it speaks effectively to the intended audience.

The only thing that matters is that the text has the right structural

patterns. In essence, algorithmic marking might reward the writing of "bullshit" – text written with little regard for the subject matter and solely to fulfil the algorithm's criteria.

Not simply lying, analysts use "bullshit" to describe empty talk or meaningless jargon. Princeton philosopher Harry Frankfurt argues that talking bullshit may actually be worse than lying, because the lie at least reaffirms the truth:

"It is impossible for someone to lie unless he thinks he knows the truth. Producing bullshit requires no such conviction. A person who lies is thereby responding to the truth, and he is to that extent respectful of it … For the bullshitter, however, all these bets are off: he is neither on the side of the true nor on the side of the false. His eye is not on the facts at all, as the eyes of the honest man and of the liar are, except insofar as they may be pertinent to his interest in getting away with what he says."

Unlike humans, algorithms are incapable of truly understanding when something is nonsense rather than genuine ideas and argumentation. It doesn't know whether a text has any worth or relationship to our world at all.

That's why algorithmic marking, whether in NAPLAN or otherwise, risks rewarding the writing of bullshit.

Encouraging the wrong thing

Our politics, businesses and media are already flooded with empty arguments and jargon. Let's not reward the skill of writing it.

Any application of algorithmic decision-making creates feedback loops. It influences future behaviour by rewarding and foregrounding some aspects of human practice and backgrounding others.

This is particularly the case when incentives are tied to the outcomes of algorithmic decision-making. In the case of NAPLAN, we know that the government rewards schools that score highly. As a result, there is already an entire industry geared towards "cracking the script" of NAPLAN in order to secure high marks.

Imagine what happens when students realise that genuine ideas and valid arguments are not rewarded by the algorithm.

This article was originally published on The Conversation. Read the original article.

Provided by The Conversation

Citation: Why marking essays by algorithm risks rewarding the writing of 'bullshit' (2017, October 20) retrieved 19 April 2024 from https://phys.org/news/2017-10-essays-algorithm-rewarding-bullshit.html