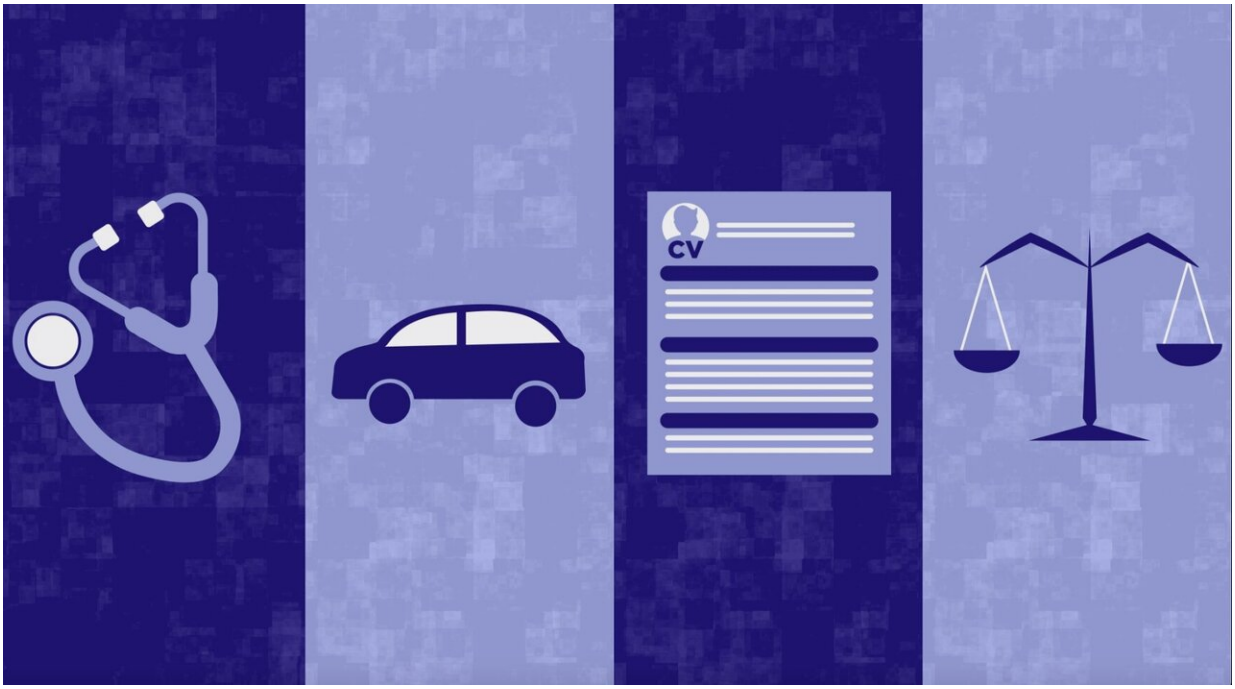
New algorithms train AI to avoid specific bad behaviors

November 21 2019



As robots, self-driving cars and other intelligent machines weave AI into everyday life, a new way of designing algorithms can help machine-learning developers build in safeguards against specific, undesirable outcomes like racial and gender bias. Credit: Deboki Chakravarti

Artificial intelligence has moved into the commercial mainstream thanks to the growing prowess of machine learning algorithms that enable computers to train themselves to do things like drive cars, control robots

or automate decision-making.

But as AI starts handling sensitive tasks, such as helping pick which prisoners get bail, policy makers are insisting that computer scientists offer assurances that automated systems have been designed to minimize, if not completely avoid, unwanted outcomes such as excessive risk or racial and gender bias.

A team led by researchers at Stanford and the University of Massachusetts Amherst published a paper Nov. 22 in *Science* suggesting how to provide such assurances. The paper outlines a new technique that translates a fuzzy goal, such as avoiding gender bias, into the precise mathematical criteria that would allow a machine-learning algorithm to train an AI application to avoid that behavior.

"We want to advance AI that respects the values of its human users and justifies the trust we place in autonomous systems," said Emma Brunskill, an assistant professor of computer science at Stanford and senior author of the paper.

Avoiding misbehavior

The work is premised on the notion that if "unsafe" or "unfair" outcomes or behaviors can be defined mathematically, then it should be possible to create algorithms that can learn from data on how to avoid these unwanted results with high confidence. The researchers also wanted to develop a set of techniques that would make it easy for users to specify what sorts of unwanted behavior they want to constrain and enable machine learning designers to predict with confidence that a system trained using past data can be relied upon when it is applied in real-world circumstances.

"We show how the designers of machine learning algorithms can make it

easier for people who want to build AI into their products and services to describe unwanted outcomes or behaviors that the AI system will avoid with high-probability," said Philip Thomas, an assistant professor of computer science at the University of Massachusetts Amherst and first author of the paper.

Fairness and safety

The researchers tested their approach by trying to improve the fairness of algorithms that predict GPAs of college students based on exam results, a common practice that can result in gender bias. Using an experimental dataset, they gave their algorithm mathematical instructions to avoid developing a predictive method that systematically overestimated or underestimated GPAs for one gender. With these instructions, the algorithm identified a better way to predict student GPAs with much less systematic gender bias than existing methods. Prior methods struggled in this regard either because they had no fairness filter built-in or because algorithms developed to achieve fairness were too limited in scope.

The group developed another algorithm and used it to balance safety and performance in an automated insulin pump. Such pumps must decide how big or small a dose of insulin to give a patient at mealtimes. Ideally, the pump delivers just enough insulin to keep blood sugar levels steady. Too little insulin allows blood sugar levels to rise, leading to short term discomforts such as nausea, and elevated risk of long-term complications including cardiovascular disease. Too much and blood sugar crashes—a potentially deadly outcome.

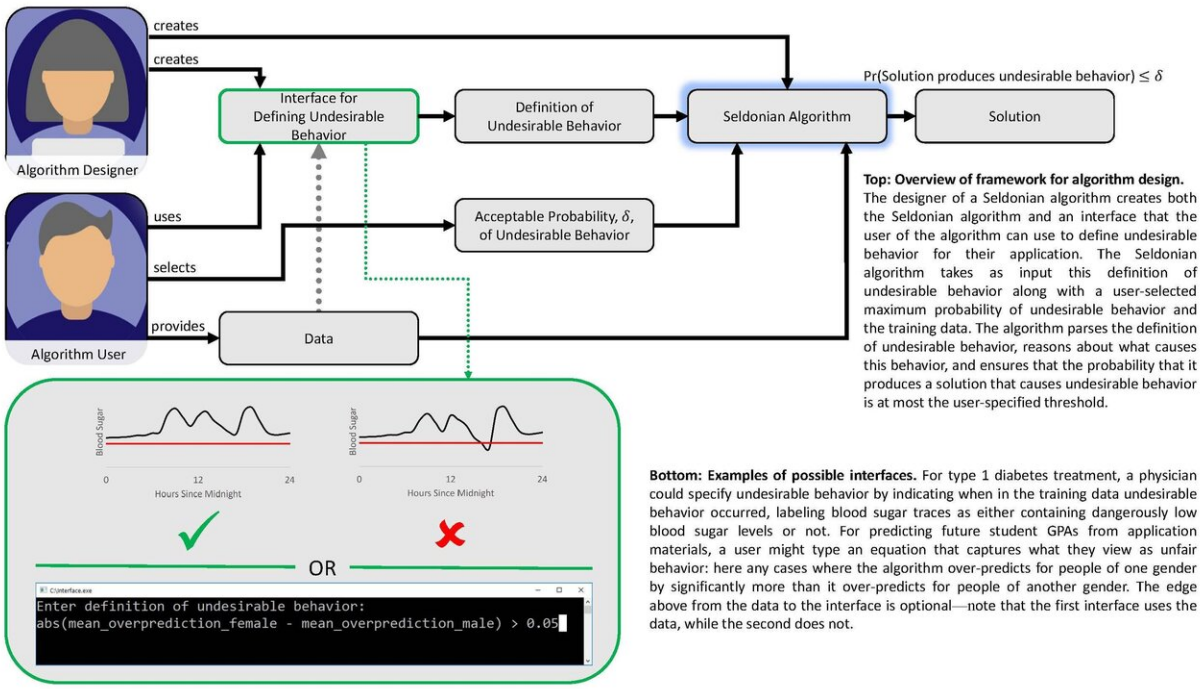


Diagram depicting paper's framework. Credit: Philip Thomas

Machine learning can help by identifying subtle patterns in an individual's blood sugar responses to doses, but existing methods don't make it easy for doctors to specify outcomes that automated dosing algorithms should avoid, like low blood sugar crashes. Using a blood glucose simulator, Brunskill and Thomas showed how pumps could be trained to identify dosing tailored for that person—avoiding complications from over- or under-dosing. Though the group isn't ready to test this algorithm on real people, it points to an AI approach that might eventually improve quality of life for diabetics.

In their *Science* paper, Brunskill and Thomas use the term "Seldonian algorithm" to define their approach, a reference to Hari Seldon, a character invented by science fiction author Isaac Asimov, who once proclaimed three laws of robotics beginning with the injunction that "A

robot may not injure a human being or, through inaction, allow a human being to come to harm."

While acknowledging that the field is still far from guaranteeing the three laws, Thomas said this Seldonian framework will make it easier for machine learning designers to build behavior-avoidance instructions into all sorts of algorithms, in a way that can enable them to assess the probability that trained systems will function properly in the real world.

Brunskill said this proposed framework builds on the efforts that many computer scientists are making to strike a balance between creating powerful algorithms and developing methods to ensure that their trustworthiness.

"Thinking about how we can create algorithms that best respect values like safety and fairness is essential as society increasingly relies on AI," Brunskill said.

More information: P.S. Thomas el al., "Preventing undesirable behavior of intelligent machines," *Science* (2019). science.sciencemag.org/lookup/ … 1126/science.aag3311

Provided by Stanford University

Citation: New algorithms train AI to avoid specific bad behaviors (2019, November 21) retrieved 18 April 2024 from https://techxplore.com/news/2019-11-machine-algorithms-safety-fairness.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.