Machines also make mistakes

Research group develops algorithm that enables user to prevent unwanted behaviors of Artificial Intelligence

First published - February 27, 2020

By Thauane Silva

Foto: Gustavo Diehl/UFRGS

In our daily life, we make constant use of technologies controlled by Artificial Intelligence (AI). It is present in our mobile phones, in computer programs, in diagnostics and medical treatments, in job recruitment and, more commonly, in digital platforms such as Netflix and Spotify. AI processes data provided directly or indirectly by the user, identifies patterns, and uses them to achieve objectives and perform specific tasks. However, it is possible that, on the path chosen to achieve the a given result, there may be discrimination of gender or race, for example. Issues like these frame cases of unwanted behaviors to be avoided.

A study conducted by researchers from the Federal University of Rio Grande do Sul (UFRGS) and the American universities of Massachusetts and Stanford, published in the journal Science, analyzed the possible mistakes made by AIs and ways to avoid them. From there, the algorithms named by scientists as "Seldonians" were created – in honor of the character Hari Seldon from the science fiction books of the writer Isaac Asimov. Unlike existing algorithms, these – the Seldonians – work for any type of situation and provide greater security for users, since they allow the restriction of certain machine behaviors.

Algorithms would be the step by step to achieve a goal, such as the ones we follow in a recipe. They are usually planned from different situations that lead to the same result initially defined. The novelty brought by Seldonian algorithms is the possibility of customizing the available paths to reach the necessary solution in each case. "Our article then proposes another way to develop these algorithms, in order to make it easier to regulate their behavior, specifying what they should not do," explains professor Bruno Castro da Silva, research member of the study and faculty member of the Institute of Informatics at UFRGS.

According to the researcher, the idea is that the user can restrict certain behaviors of the programmed machine in order to reduce errors in the predictions of artificial intelligence. This will be possible through two simple commands described in percentage: the stipulation of the maximum degree of discrimination that will be accepted and the probability of guaranteeing the given solution, that is, how much certainty is expected so that the algorithm does not miss the result. After that, the machine presents the best path, considering the restrictions stipulated by the user, or warns that there is no possible solution within these limitations.

To prove the existence of the errors, a test was conducted to try to predict how 43,000 undergraduate students from UFRGS would do in their first three semesters of the course. During the AI training, the scores that each one achieved in the vestibular tests were provided to serve as the basis of the analysis. The obtained results with the existing algorithms showed evidence of discrimination according to the student's gender: they predicted that men would do much better than women. Meanwhile, Seldonian algorithms showed the same percentage of error for both genders.

A second experiment was carried out by the group to verify the feasibility of the algorithm created. In this, a human metabolism simulator from the Food and Drug Administration – the U.S. federal agency responsible for food and drug control – was used to observe how different people with type 1 diabetes react to certain amounts of insulin. It was asked from the AI to determine how much insulin each patient would need before a carbohydrate meal, without risk of hypoglycemia. Existing algorithms have indicated a new treatment regardless of the consequences; the Seldonian algorithm did not indicate a dose of insulin until it was proven that the patient's sugar levels would remain stable.

The researchers continue to conduct tests to confirm the effectiveness of the Seldonian algorithm before it can be released for public use. Current research focuses on the AI that runs autonomous cars, which are already circulating in the United States, and can make mistakes according to changes in road signs, for example.

https://youtu.be/vccLrBRnTa4

The work is part of the activities of the Artificial Intelligence Research Group of the Institute of Informatics, which completes 30 years in 2020. Check out the program "Getting to know UFRGS," from UFRGS TV, about artificial intelligence and the group's work (in Portuguese): https://youtu.be/zZnlpm3Vb2A

Scientific Article

THOMAS, Philip S. et al. Preventing undesirable behavior of intelligent machines. Science Journals, v. 366, issue 6468, p. 999-1004, 22 Nov. 2019

Translated into English by Caroline Cristiane Vargas de Souza, under the supervision and translation revision of Elizamari R. Becker (P.h.D.) – IL/UFRGS.