International Conference on Artificial Intelligence and Cybersecurity (ICAIC 2025)

Conference Proceedings   November 27-28, 2025 (Virtual)

A Method to Test the Ethics of Some AI Classifiers - The Example of School Dropouts Problem

Antonio Ballarin

University Canada West, Vancouver, Canada

Abstract
If an AI artifact is an emulation of human behavior in relation to the performance of some activity and if the human being, in carrying out that activity, is required to respect a behavioral framework defined by laws, rules, regulations, procedures, best practices, etc., then the AI that emulates that human behavior is also required to respect the same behavioral framework. The idea of ethics tests is developed on this principle and, precisely on the basis of this principle, a pragmatic methodology can be developed that can test the correspondence in the observance of the artefact to the behavioral framework within which it will necessarily be placed in its operation. The approach proposed in this work allows us to offer an extremely pragmatic solution to the search for an ?ethical behavior? for AI artifacts, bypassing the difficult applicability of the complex and abstract legislation currently in force on this topic. In order to explain the applicability of this methodology to a concrete problem, this work considers the problem of school dropout as an example and describes the construction of two classifiers, one based on a neural network and one on a decision tree, able to predict the phenomenon. The application of the methodology clearly shows how the explainability offered by a symbolic system, such as a decision tree, is not applicable as an element of explainability in the behavior of a neural classifier.

Keywords: Ethics Tests, Explainability, Classifications System, Bias, Categorical Polarization