Karlsruher Institut für Technologie

Institut für Informationssicherheit und Verlässlichkeit (KASTEL)
Anwendungsorientierte Formale Verifikation
Prof. Dr. Bernhard Beckert

Dynamisches Prüfen von Fairness-Eigenschaften mit Hilfe von Laufzeitmonitoren

1 Hintergrund

Immer häufiger werden relevante Entscheidungen über Resourcenverteilungen nicht mehr von Menschen, sondern von Algorithmen getroffen. Diese ziehen dabei eine Menge von numerischen Merkmalen zu Rate, um eine Entscheidungsempfehlung zu treffen. Solche Algorithmen werden bereits heute in Bereichen wie der Vergabe von Krediten, Risikobeurteilung von Straftätern und weiteren sozialkritischen Aufgabenfeldern eingesetzt.[1]
Wichtig ist dabei, dass die getroffenen Entscheidungen fair sind und keine Minderheiten diskriminieren. Bspw. sollten Merkmale wie Hautfarbe oder Geschlecht keinen Einfluss auf diese Entscheidungen haben. Außerdem müssen hierbei auch Merkmale Beachtung finden, die eine direkte Korrelation mit solchen kritischen Merkmalen haben.

2 Aufgabe

Ziel dieses Projektes ist es, mit Hilfe von Laufzeit-Verfahren programmierte Entscheidungsalgorithmen auf ausgewählte Fairness-Eigenschaften zur Laufzeit oder zur Testzeit zu prüfen.
Ihre Aufgabe besteht darin, zu untersuchen wie Fairnesseigenschaften formalisiert und zur Laufzeit verifiziert werden können. Insbesondere sollen dabei probabilistische Spezifikationsmethoden in Erwägung gezogen werden. Zudem sollen für die somit formalisierten Spezifikationen Methoden entwickelt werden, die die spezifizierten Eigenschaften zur Laufzeit prüfen [1]. Es besteht auch die Möglichkeit im Rahmen des Projekts Methoden zur a priori Verifikation (bspw. Methoden zur Verifikation von neuronalen Netzwerken) mit Laufzeitmonitoren zu verknüpfen.

3 Kontakt / Betreuung

Jonas Klamroth, klamroth@fzi.de, Raum 1.1.27 (FZI)
Dr. Michael Kirsten, kirsten@kit.edu, Raum 228 (Geb. 50.34)
Samuel Teuber, teuber@kit.edu, Raum 203 (Geb. 50.34)

Literatur

[1] Lars Grunske. „An effective sequential statistical test for probabilistic monitoring“. In: *Information and Software Technology* 53.3 (2011), S. 190–199. DOI: 10.1016/j.infsof.2010.10.003.

[1] https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses