Explaining Algorithmic Decisions with respect to Fairness

Qusai Ramadan[1] Amir Shayan Ahmadian[1] Jan Jürjens[1,2] Steffen Staab[1,3] Daniel Strüber[4]

Keywords: Software Fairness; Explainable Software; Model-Based Analysis; UML

1 Abstract

Decision-Making Software (D-MS) may exhibit biases against people on grounds of *protected characteristics* such as *gender* and *ethnicity*. Such undesirable behavior should not only be detected but also explained. To avoid complicated explanations and expensive fixes, fairness awareness has to be proactively embedded in the design phase of the system development. With *fairness by design*, system developers have to be supported with tools that detect and explain discriminations during the system architecture design [Ra18a].

Only avoiding protected characteristics in a D-MS does not prevent discrimination. Due to data correlations, other data may act as proxies for protected characteristics, thereby causing the so-called *discrimination by proxy* [GBM17]. There are two possible explanations for the correlations: **(1) Societal fact.** For instance, if females are more likely to have *long hair* more than males, then *long hair* can act as a proxy for the *gender*. **(2) Information flow.** The actual input of a D-MS may contain data that resulted from processing protected characteristics. For example, in an insurance company, it might be authorized to use the *gender* for identifying an *insurance tariff* but it might be not allowed to use *gender* for deciding about the *reimbursement factor*. However, if the *insurance tariff* is used as input to the reimbursement D-MS, a discrimination against *gender* have to be reported because the *insurance tariff* indirectly leaks a signal about the *gender* to the reimbursement D-MS.

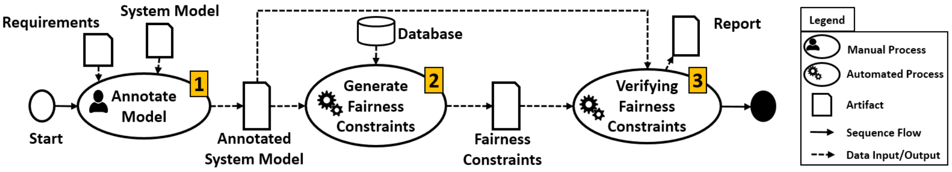
Existing works only consider two approaches: *white-* (e.g., [Da17]) and *black-box* approaches (e.g., [GBM17]). While white box approaches can uncover discrimination, they cannot uncover discrimination in the above mentioned sense, as they did not consider possible information flow between system components. While black box approaches may solve this, they do not produce a witness to describe where and how a data flow can happen. Moreover, both approaches cannot be used in the early phase of the system design [Ra18a].

We aim to support developers with tools to reason about hidden flows for protected characteristics to a D-MS during the modeling of the system architecture. Detecting hidden information is a key challenge in security engineering [De76]. However, a model-based information flow analysis approach that supports fairness analysis is lacking [Ra18a]. We propose to develop a model-based discrimination detection framework (see Fig. 1):

Input: (i) a *requirements document* containing fairness requirements. (ii) a model (UML) that is automatically generated from a system implementation or manually created based on a system specifications. The model have to describe the structural and behavioral aspects of

[1] Universität Koblenz-Landau, Germany. Email: qramadan,ahmadian,juerjens,staab@uni-koblenz.de
[2] Fraunhofer Institute for Software and System Engineering, Germany
[3] University of Southampton, UK
[4] University of Gothenburg, Gothenburg, Sweden. Email:danstru@chalmers.se

Fig. 1: High-level overview of a model-based discrimination analysis framework.



the system. (ii) a *database* of historical data. **Process:** (i)*Annotating a system model* with fairness requirements. For this, we plan to extend the privacy UML profile in [Ah17b]. (iii) *Generating fairness constraints* in term of formal specifications (e.g., Computation Tree Logic). In this step, proxies are also identified by analyzing the system database and encoded in the constraints. (iii) *Verifying the fairness constraints* against the system model (using a model checker). **Output:** A witness reporting a fairness violation in the form of a sequence of actions that can explain where and how discrimination can happen. Our framework can be realized by extending CARiSMA, a model-based security analysis tool support [Ah17a].

Other open challenges. Information flow is not the only source for discriminations. Other sources are: First, a nonalignment between the organizational needs and the system models due to misunderstandings between expert stakeholders about fairness terminologies. This challenge can benefit from the model transformation technology, as proposed in our work in [Ra17]. Second, trade-offs between fairness and privacy requirements. A privacy requirement may disallow a D-MS from accessing protected characteristics. However, this requirement will prevent the D-MS from being able to uncover up-to-date proxies, as a proxy identification requires accessing to protected characteristics. For this, we plan to extend our work on conflicts detection [Ra18b].

Literatur

[Ah17a] Ahmadian, A. S.; Peldszus, S.; Ramadan, Q.; Jürjens, J.: Model-based privacy and security analysis with CARiSMA. In: Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering. ACM, S. 989–993, 2017.

[Ah17b] Ahmadian, A. S.; Strüber, D.; Riediger, V.; Jürjens, J.: Model-based Privacy Analysis in Industrial Ecosystems. ECMFA, Springer/, 2017.

[Da17] Datta, A.; Fredrikson, M.; Ko, G.; Mardziel, P.; Sen, S.: Use Privacy in Data-Driven Systems. In: Proceedings of the ACM CCS Conference. 2017.

[De76] Denning, D. E.: A lattice model of secure information flow. Communications of the ACM 19/5, S. 236–243, 1976.

[GBM17] Galhotra, S.; Brun, Y.; Meliou, A.: Fairness testing: testing software for discrimination. In: Proceedings of the 2017 11th Joint Meeting on ESEC/FSE. ACM, S. 498–510, 2017.

[Ra17] Ramadan, Q.; Salnitri, M.; Strüber, D.; Jürjens, J.; Giorgini, P.: From Secure Business Process Modeling to Design-Level Security Verification. In: 20th ACM/IEEE MODELS 2017 International Conference. S. 123–133, 2017.

[Ra18a] Ramadan, Q.; Ahmadian, A. S.; Strüber, D.; Jürjens, J.; Staab, S.: Model-based discrimination analysis: a position paper. In: Proceedings of the International Workshop FairWare@ICSE 2018, Gothenburg, Sweden. 2018.

[Ra18b] Ramadan, Q.; Strüber, D.; Salnitri, M.; Riediger, V.; Jürjens, J.: Detecting Conflicts Between Data-Minimization and Security Requirements in Business Process Models. In: ECMFA 2018, Held as Part of STAF 2018, 2018, Proceedings. S. 179–198, 2018.