A Model-driven and Formal Approach to Operationalizing Notions of Fairness

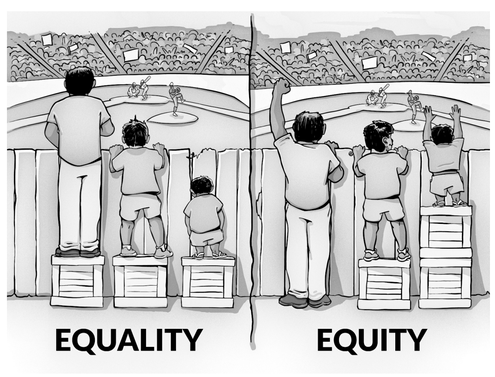
Making Fairness Actionable

Julian Alfredo Mendez — julian.mendez@cs.umu.se — umu.se/en/staff/julian-mendez
Umeå University

Introduction

- Can we design a **readable language** to express **fairness** to monitor AI systems?
- Can the specifications be efficiently **prototyped**?

Problem



How to formalize a fairness scenario?

Method

Social Fairness Desiderata
(human interpretation)
Graphical Specification of Fairness
(compiles to)
Declarative Language-based Specification
(code for) *(code for)*
Proof Checker | Mainstream Development

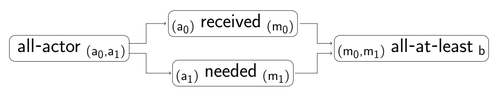
Tiles

A software framework to model and agree on **definitions of fairness**.

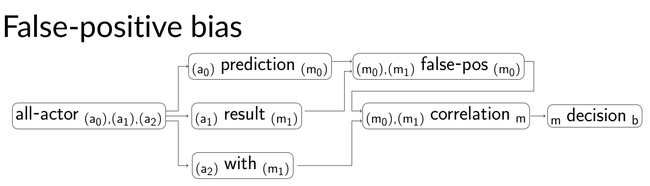
Representation with Tiles (julianmendez.github.io/tiles):

- Equality

- Equity



- False-positive bias



– all-actor: tuples of actors;
– prediction: the original prediction on an actor;
– result: the actual result of an actor;
– with: if the actor has a protected attribute;
– false-pos: the false positives;
– correlation: correlation between the sources;
– decision: whether there is a significant bias.

Soda

Tiles is implemented in Soda (julianmendez.github.io/soda), a **purely-functional language** with **object-oriented notation**. The descriptions are translated to Scala version 3, and then to Java Virtual Machine (JVM) bytecode. The proofs are translated to and verified in the proof assistant Lean version 4.

- Functions in Soda
  $f(x : A) : B = g(x)$
  `lambda` $x \longrightarrow f(x)$
  `if` b `then` e_1 `else` e_2
  `match` x `case` $A\_(y) \implies f(x)(y)$
- Classes and Packages in Soda
  `class` A `extends` B `...` `end`
  `abstract` $x : A$
  `this`, `package`, `import`
- Formal Verification (External)
  `directive`

Conclusion and Future Work

Tiles is a framework and Soda is a formal language to model **definitions of fairness**. We plan to show how **formal verification** can be applied in more general contexts, like for **multi-agent systems**.

Bibliography

- **Julian Alfredo Mendez**, *Making Fairness Actionable*, Licentiate thesis, 2024. URN:urn:nbn:se:umu:diva-232384
- **Andrea Aler Tubella, Flavia Barsotti, Rüya Gökhan Koçer, and Julian Alfredo Mendez.** *Ethical implications of fairness interventions: what might be hidden behind engineering choices? Ethics and Information Technology*, volume 24, issue 1, article 12, Springer 2022. DOI:10.1007/s10676-022-09636-z.
- **Andrea Aler Tubella, Dimitri Coelho Mollo, Adam Dahlgren Lindström, Hannah Devinney, Virginia Dignum, Petter Ericson, Anna Jonsson, Timotheus Kampik, Tom Lenaerts, Julian Alfredo Mendez, and Juan Carlos Nieves.** *ACROCPoLis: A Descriptive Framework for Making Sense of Fairness*. *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1014–1025, 2023. DOI:10.1145/3593013.3594059.
- **Julian Alfredo Mendez, Timotheus Kampik, Andrea Aler Tubella, and Virginia Dignum.** *A Clearer View on Fairness: Visual and Formal Representation for Comparative Analysis*. In Florian Westphal, Einav Peretz-Andersson, Maria Riveiro, Kerstin Bach, and Fredrik Heintz, editors, *14th Scandinavian Conference on Artificial Intelligence, SCAI 2024*, pages 112–120. Swedish Artificial Intelligence Society, June 2024. DOI:10.3384/ecp208013.
- **Julian Alfredo Mendez.** *Soda: An Object-Oriented Functional Language for Specifying Human-Centered Problems*. arXiv. DOI:10.48550/arXiv.2310.01961.
- **Julian Alfredo Mendez, Timotheus Kampik** *Can Proof Assistants Verify Multi-Agent Systems?*. In *Proceedings of the 21st European Conference on Multi-Agent Systems, EUMAS*, 2024. DiVA urn:nbn:se:umu:diva-232383.

s.cs.umu.se/1m536z

UMEÅ UNIVERSITY

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM