Deloitte.

aiStudio: Benchmarking Model Guardian

15.5.2021 | David Thogmartin | aiStudio

Model Guardian – Designed to For In-Depth Analysis of Model Bias

An award-winning paper by Lee and Singh (2021) finds gaps in open source fairness toolkits resolved by Deloitte's Model Guardian.

Open source fairness toolkits	Model Guardian (quantitative) + Scorecard (qualitative)
Steep learning curve: all six toolkits require the user to "learn a lot of things before [he/she] could get going." One interviewee said it would take "at least weeks of reading" to get up to speed on relevant literature	**No prior experience required**: Model Guardian walks the user step-by-step through each stage of the analysis without assuming prior knowledge on algorithmic fairness
Awkward user interface: one-size-fits-all design that is either deceptively simple or too complex to navigate	**Intuitive design**: Model Guardian highlights the key takeaways while offering the option to drill down into details
Limited coverage of model pipeline: the tools only address biases in model build and evaluation phases, overlooking others, e.g. data collection	**End-to-end model lifecycle coverage**: the scorecard questionnaire addresses the biases introduced through people/process in model pipeline
Incomplete mitigation strategy: the "de-biasing" pre-/in-/post-processing methods are controversial for only solving for a narrow version of fairness. These toolkits only offer technical mitigation, some of which may be inconsistent with anti-discrimination laws	Targeted mitigation: the best mitigation strategy may not be technical. If biases are introduced through data collection and labelling, they should be addressed through changes in marketing strategy and training of labellers. The scorecard questionnaire helps address the bias at its source
Limited customisation: a lot of the tool is hard-coded to their data, and "a lot of extra work is needed" to adapt the toolkits to a particular use case	**Bespoke, custom set-up**: Model Guardian is set up as an accelerator to be easily adapted to contextual and regulatory considerations
Complicated integration: practitioners found the toolkits difficult to "plug and play" with existing workflows, partially due to limited data security	**Built for plug-and-play**: Model Guardian contains building blocks in its code library to easily integrate into any organisational workflow
Lack of consensus: each toolkit espouses a different vision of what "fairness" means with limited guidance on which approach is appropriate. Ex) see the framing difference between Fairness 360 and Fairlearn.	**Bespoke expertise**: with each Model Guardian implementation, we draw from Deloitte domain and regulatory experts and work with each organisation to provide a holistic understanding of the ethical issues

Deloitte.