<AI & Equality> A Human Rights Toolbox

Sofia Kypraiou[1,2], Caitlin Kraft-Buchman[2], and Daniel Gatica-Perez[1]

[1]École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[2]Women at the Table, Geneva, Switzerland

Extended Abstract

A new generation of researchers responsible for creating algorithmic systems have solid technical backgrounds but often lack substantial human rights knowledge or frameworks for using this technical knowledge as "AI for Social Good". As the data science/machine learning field evolves quickly, many universities do not have time to adjust the material and adapt to the changes. There is a deficit of bandwidth and knowledge regarding the relationship between 'technique' and human rights, which may leave another generation of scientists disempowered to leverage their education to impact the world needs.

We present *<AI & Equality> A Human Rights Toolbox* [1]. Our methodology includes a workshop consisting of a Human Rights module and code, outreach and community plan incorporating human rights concepts with data science and integrating International Human Rights frameworks with current concepts of fairness for designing an educational tool for computer scientists.

The Toolbox is based on an idea of the NGO Women at the Table [2] its support for this project, and its collaboration with the Office of the United Nations High Commissioner for Human Rights (OHCHR). Piloted at EPFL through workshops, *<AI & Equality> A Human Rights Toolbox* addresses the problem head-on in language computer science students can both understand and use. This Toolbox is the first and to date only foray of OHCHR into the world of university computer science students in order to jumpstart a conversation about a human rights-based approach being the baseline from which we should create new algorithms and new models.

We focus on human rights instead of ethics, which are often better defined and measurable, as most are defined under international or national law. Human rights are rights we have because we exist as human beings. These universal rights are inherent to us all, regardless of nationality, sex, national or ethnic origin or any other status. The United Nations (1948). They are based on international law and provide an ethical lens to transcend national and cultural boundaries. They put people at the centre of decision-making and can be used to assess and address any unintended harm.

[1] https://aiequalitytoolbox.com/
[2] https://www.womenatthetable.net/

In its current state, our methodology consists of OHCHR / EPFL driven workshops that include a Jupyter notebook that takes the Human Rights module and marries it with how human rights interplay with decisions made at various points of the data and model lifecycle.

The workshop consists of 2 parts: I. Human Rights Module, and an applied research conversation, II. applied coding Toolbox. A different expert in the specific domain presents each part. At the end of each session, we strongly encourage discussion, questions, and knowledge sharing with and between the students. We employ a broad to narrow approach, meaning we first introduce basic information about human rights principles.

The human rights expert presents digestible definitions of human rights and human rights principles, equality and non-discrimination, and poses a critical analysis on what a human rights-based approach might look like in machine learning. Group (2003), The United Nations (1948)

Then we explore how this plays out in current research and present practical ways to translate human rights principles through code.

The Jupyter notebook investigates how code can be de-biased and improved to support the Human Rights principlesGroup (2003). Participants experiment with data to see how different mathematical and data concepts of fairness interrelate, begin a critical analysis checklist of the data process and apply some of the concepts and debiasing literature to hands-on exercise.

This clear methodology provokes critical analysis on where and when to intervene in the fairness pipeline: pre-processing (training data), in-processing (model design) and post-processing (predictions).

For the practical sessions, we have curated a list of questions taken from the work Datasheets for Datasets Gebru et al. (2018) and Model Cards for Model reporting Mitchell et al. (2018).

For example, the human rights principle of participation and inclusion is brought to the fore through the following question Gebru et al. (2018): "Does the dataset identify any subpopulations (e.g., by age, gender)?" This is translated through the exploratory data analysis by presenting demographic plots of our use-case dataset.

We evaluated our methodology in an iterative process of three workshops at two different Universities, EPFL and University College Dublin (UCD). The participants of the majority were master-level computer and data science students. We found that the students responded well, with a marked increase in their awareness of human rights principles. They improved their ability to identify and analyze how gender, racial and other bias occurred or can occur in the research, design, and development of AI, in addition to their ability to identify and use tools and techniques to mitigate bias in AI.

We intend to scale the methodology and workshops by blending workshops at new participating universities with a guest Professor or Doctoral Candidate from that university sharing their applied research and how it interplays with a Human Rights framework. It is anticipated over time that participating universities will bring two professors or doctoral candidates from a social science discipline such as Law/Philosophy/Ethics, and lecturer/Professor from Computer Science (with these presentations being added to the Notebook and AIEqualityToolbox site) so that the learnings and examples evolve and community is strengthened.

We wish to bring an international university generation to understand the scientist's unique potential of social impact in the real world, bridging science and human rights policy to foster systemic resilience and more equal, just, robust democracies.

Going forward, we want to create a space and content for young policymakers and young scientists to gather and find resources and one another to create the technology we need and the technology we deserve, in line with the human rights values we all embrace.

References

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, and Kate Crawford. Datasheets for datasets. 3 2018. URL http://arxiv.org/abs/1803.09010.

UN Sustainable Development Group. The human rights based approach to development cooperation towards a common understanding among un agencies, 9 2003. URL https://unsdg.un.org/resources/human-rights-based-approach-development-cooperation-towards-common-understanding-among-un.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pages 220–229, 10 2018. doi: 10.1145/3287560.3287596. URL http://arxiv.org/abs/1810.03993http://dx.doi.org/10.1145/3287560.3287596.

The United Nations. *Universal Declaration of Human Rights*. December 1948.