# A Human Rights-Based Approach to Responsible AI

Vinodkumar Prabhakaran, Margaret Mitchell, Timnit Gebru, Iason Gabriel

Research on fairness, accountability, transparency and ethics in AI assumes an implicit consensus on the values and principles that guide these interventions, which is at odds with the pluralistic world we live in.

We explore the potential of **human rights as globally salient and cross-culturally recognized set of values,** as a potential grounding framework for explicit value alignment in responsible AI, and as a framework for global civil society partnership and participation.

## Three aspects of Human Rights

Human rights are ***a set of moral claims*** anchored in the notion that human life has value and that there are aspects of personhood that must be protected.

Human rights are ***part of a legal regime and set of instruments*** that aim (in part) to make these values a reality

Human rights are ***part of a cultural practice and global social movement*** that focuses on advocacy, empowerment and critique of existing institutions.

## Illustrations

**The right to be free from discrimination** is a negative right not to be harmed in certain ways, and it is heavily impacted by prevalent forms of algorithmic bias, and is directly relevant to most of FATE work.

**This right to health** enshrines everyone to the enjoyment of the highest attainable standard of physical and mental health. Responsible AI research has the potential to enable this right at scale, through AI-enabled services and diagnostic tools, for instance in reducing the cost of healthcare, or increasing access to health.

**The right to science** enshrines everyone the right to share in scientific advancement and its benefits. This right asserts the importance of Responsible AI to ultimately benefit all sections of humanity, including those historically excluded from the benefits of scientific advances.

## Discussion

We propose future research along three pathways:

**Translational** research aimed at building a shared vocabulary of concerns, values, and expected outcomes to bridge the gap between technology researchers and civil society activists working in responsible AI

**Functional** aspects of a human rights based inquiry that studies what a human rights based inquiry into responsible AI reveal that existing methodologies do not.

Enabling potential of AI in order to strengthen human rights fulfilment around the world, including also enabling equitable global access to the advancements in AI.

## Limitations

- efficacy as a framework across the globe: universality vs. cultural relativism
- enforceability as a legal framework across the globe