Negative Human Rights as a Basis for Long-term AI Safety and Regulation (Abstract Reprint)

Ondrej Bajgar[1] and **Jan Horenovsky**[2]
[1]Department of Engineering Science Future of Humanity Institute, University of Oxford, United Kingdom
[2]Department of Politology and Sociology, Faculty of Law, Charles University, Czech Republic
ondrej@bajgar.org, horenovj@prf.cuni.cz

Abstract Reprint. This is an abstract reprint of a journal by [Bajgar and Horenovsky, 2023].

Abstract

If autonomous AI systems are to be reliably safe in novel situations, they will need to incorporate general principles guiding them to recognize and avoid harmful behaviours. Such principles may need to be supported by a binding system of regulation, which would need the underlying principles to be widely accepted. They should also be specific enough for technical implementation. Drawing inspiration from law, this article explains how negative human rights could fulfil the role of such principles and serve as a foundation both for an international regulatory system and for building technical safety constraints for future AI systems.

References

[Bajgar and Horenovsky, 2023] Ondrej Bajgar and Jan Horenovsky. Negative human rights as a basis for long-term AI safety and regulation. *J. Artif. Intell. Res.*, 76:1043–1075, 2023.