Time-scale Differences will Influence the Regulation Required in an Idealised AI Race Game

The Anh Han[1], Luis Moniz Pereira [2], Francisco C. Santos[3,4], Tom Lenaerts[4,5]
1:Teesside University, 2: Universidade Nova de Lisboa, 3:IST, Universidade de Lisboa,
4: Université Libre de Bruxelles, 5: Vrije Universiteit Brussel
Email: T.Han@tees.ac.uk

Abstract

Rapid technological advancements in Artificial Intelligence (AI), as well as the growing deployment of intelligent technologies in new application domains, have generated serious anxiety and a fear of missing out among different stakeholders, fostering a racing narrative. Whether real or not, the belief in such a race for domain supremacy through AI can make it real simply from its consequences. These consequences may be negative, as racing for technological supremacy creates a complex ecology of choices that could push stakeholders to underestimate or even ignore ethical and safety procedures. Given the breadth and depth of AI and its advances, it is difficult to assess which technology needs regulation and when. As there is no easy access to data describing this alleged AI race, theoretical models are necessary to understand its potential dynamics, allowing for the identification of when procedures need to be put in place to favour outcomes beneficial for all. We show in [Han et al. 2020], that next to the risks of setbacks and being reprimanded for unsafe behaviour, the time-scale in which domain supremacy can be achieved plays a crucial role. When this can be achieved in the short term, those who completely ignore the safety precautions are bound to win the race but at a cost to society, apparently requiring regulatory actions [Han et al. 2021]. For a long-term situation, conditions can be identified that require the promotion of risk-taking as opposed to compliance with safety regulations in order to improve social welfare. These results remain robust both when two or several actors are involved in the development process and when the negative outcomes affect either the unsafe actor or the entire group. Thus, when defining codes of conduct and regulatory policies for AI applications, a clear understanding of the time-scale of the race is thus required, as this may induce important non-trivial effects.

The current work is based on the publication in [Han et al. 2020], which has not been presented at a major AI conference with archival proceedings before.

Acknowledgement

T.A.H., L.M.P. and T.L. have been supported by Future of Life Institute grant RFP2-154. T.A.H. is also supported by a Leverhulme Research Fellowship (RF-2020-603/9). L.M.P. is also supported by NOVA LINCS (UIDB/04516/2020) with the financial support of FCT-Fundação para a Ciência e a Tecnologia, Portugal, through national funds. F.C.S. acknowledges support from FCT Portugal (grants UIDB/50021/2020, PTDC/MAT-APL/6804/2020, and PTDC/CCI-INF/7366/2020). T.L. and F.C.S. acknowledge the support by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215. T.L. acknowledges support by the FuturICT2.0 (www.futurict2.eu) project funded by the FLAG-ERA JTC 2016.

References

[Han *et al.*, 2020] The Anh Han, Luis Moniz Pereira, Francisco C. Santos, and Tom Lenaerts. To Regulate or Not: A Social Dynamics Analysis of an Idealised AI Race. *Journal of Artificial Intelligence Research*, 69:881–921, 2020.

[Han *et al.*, 2021] The Anh Han, Luis Moniz Pereira, Tom Lenaerts, and Francisco C. Santos. Mediating Artificial Intelligence Developments through Negative and Positive Incentives. *PLOS ONE*, 16(1):e0244592, 2021.