Mediating Artificial Intelligence Developments through Negative and Positive Incentives

T.A. Han[1]*, L.M. Pereira*[2]*, T. Lenaerts*[3] *, F.C. Santos*[4] *(1) Teesside University, Middlesbrough, UK, Email: T.Han@tees.ac.uk; (2) Universidade Nova de Lisboa, Caparica, Portugal; (3) Université Libre de Bruxelles, Brussels, Belgium; (4) Universidade de Lisboa, Lisbon, Portugal*

The field of Artificial Intelligence (AI) is going through a period of great expectations, introducing a certain level of anxiety in research, business and also policy. This anxiety is further energised by an AI race narrative that makes people believe they might be missing out. Whether real or not, a belief in this narrative may be detrimental as some stake-holders will feel obliged to cut corners on safety precautions, or ignore societal consequences just to "win". Starting from a baseline model [1] that describes a broad class of technology races where winners draw a significant benefit compared to others (such as AI advances, patent race, pharmaceutical technologies), we investigate here [2] how positive (rewards) and negative (punishments) incentives may beneficially influence the outcomes. We uncover conditions in which punishment is either capable of reducing the development speed of unsafe participants or has the capacity to reduce innovation through over-regulation. Alternatively, we show that, in several scenarios, rewarding those that follow safety measures may increase the development speed while still ensuring safe choices. Moreover, in the latter regimes, rewards do not suffer from the issue of over-regulation as is the case for punishment. Overall, our findings provide valuable insights into the nature and kinds of regulatory actions most suitable to improve safety compliance in the contexts of both smooth and sudden technological shifts.

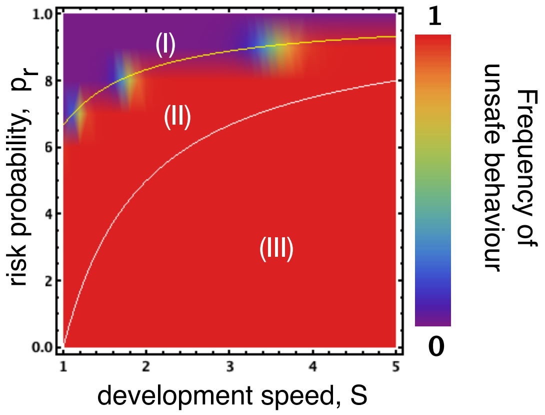

Figure 1: Frequency of unsafe behaviour as a function of development speed and the disaster risk, in absence of incentives (see Ref [1]). In regions (**I**) and (**III**), safe and unsafe/innovation, respectively, are the preferred collective outcome and are selected by natural selection, thus no regulation being required. Region (**II**) requires regulation as safe behaviour is preferred but not selected. This talk is meant to explore how to promote safe behaviour in this dilemma region using incentives (peer reward vs peer punishment) - see preprint in Reference [2].

References

[1] T.A. Han, L.M. Pereira, F.C. Santos and T. Lenaerts. 2020. *To Regulate or Not: A Social Dynamics Analysis of an Idealised AI Race*. Journal AI Research (in Press). arXiv:2010.00403

[2] T.A. Han, L.M. Pereira, F.C. Santos and T. Lenaerts (2020). *Mediating Artificial Intelligence Developments through Negative and Positive Incentives*. arXiv:2010.00403