Voluntary safety commitments provide an escape from over-regulation in AI development

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

With the introduction of Artificial Intelligence (AI) and related technologies in our daily lives, fear and anxiety about their misuse, as well as the hidden biases in their creation, have led to a demand for regulation to address such issues. Yet, blindly regulating an innovation process that is not well understood may stifle this process and reduce benefits that society might gain from the generated technology, even under the best of intentions. Starting from a baseline game-theoretical model that captures the complex ecology of choices associated with a race for domain supremacy using AI technology [1], we show how socially unwanted outcomes may be produced when sanctioning is applied unconditionally to risk-taking, i.e., potentially unsafe behaviours [2]. As an alternative to resolve the detrimental effect of over-regulation, we propose a voluntary commitment approach, wherein technologists have the freedom of choice between independently pursuing their course of actions or else establishing binding agreements to act safely, with sanctioning of those that do not abide to what they have pledged [3]. Overall, this work reveals for the first time how voluntary commitments, with sanctions either by peers or by an institution, leads to socially beneficial outcomes in all scenarios that can be envisaged in the short-term race towards domain supremacy through AI technology. These results are directly relevant for the design of governance and regulatory policies that aim to ensure an ethical and responsible AI technology development process.

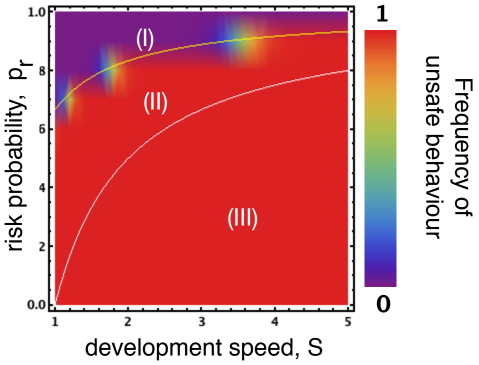

Figure 1: Frequency of unsafe behaviour as a function of development speed and the disaster risk, in absence of incentives [1]. In regions **(I)** and **(III)**, safe and unsafe/innovation, respectively, are the preferred collective outcome also selected by natural selection, thus no regulation being required. Region **(II)** requires regulation as safe behaviour is preferred but not the one selected. This talk explores how to promote safe behaviour in this dilemma region by using voluntary commitments, without engendering over-regulation in other regions [3].

References

[1] Han, T. A., Pereira, L. M., Santos, F. C., & Lenaerts, T. (2020). To Regulate or Not: A Social Dynamics Analysis of an Idealised AI Race. *J. Artificial Intell. Research*, *69*, 881-921.

[2] Han, T. A., Pereira, L. M., Lenaerts, T., & Santos, F. C. (2021). Mediating artificial intelligence developments through negative and positive incentives. *PloS one*, *16*(1).

[3] Han, T. A., Lenaerts, T., Santos, F. C., & Pereira, L. M. (2021). Voluntary safety commitments provide an escape from over-regulation in AI development. arXiv preprint arXiv:2104.03741.