Modelling and regulating safety compliance: Game theory lessons from AI development races analyses

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: Universidade de Lisboa,
4: Université Libre de Bruxelles, 5: Vrije Universiteit Brussel

Introduction

Rapid technological advancements in Artificial Intelligence (AI), together with the growing deployment of AI in new application domains such as robotics, face recognition, self-driving cars and genetics, are generating an anxiety which makes companies, nations and regions think they should respond competitively. AI appears, for instance, to have instigated a race among the chip builders, just because of the requisites it imposes on that technology. Governments are furthermore stimulating economic investments in AI research and development as they fear of missing out, resulting in a racing narrative that increases further the anxiety among stake-holders.

Innovation races for supremacy in a domain involving AI may, however, trigger detrimental consequences. Participants may well ignore ethical and safety checks so as to speed up development and thereby reach the market first. AI researchers and its governance bodies, such as the EU, are urging to consider together both the normative aspects and the social impacts of all major technological advancements concerned. However, given the breadth and depth of AI and its advances, it is no easy task to assess when and which AI technologies in a concrete domain need regulation. Data to estimate the risk of a technology is usually limited, especially at an early stage of its development and deployment.

In our recent works [Han *et al.*, 2020; Han *et al.*, 2021], we examine this problem theoretically, resorting to a novel innovation dilemma where technologists can choose a safe (SAFE) vs risk-taking (UNSAFE) course of development. Companies are held to race towards the deployment of some AI-based product in a domain X. They can either carefully consider all data and AI pitfalls along the way (the SAFE ones) or else take undue risks by skipping recommendable testing so as to speed up the processing involved (the UNSAFE ones). Overall, SAFE are costlier strategies and take more time to implement than UNSAFE ones, therefore permitting UNSAFE strategists to claim significant further benefits from reaching technological supremacy first.

Lessons for AI Safety governance policy

We find that the time-scale in which domination or supremacy in an AI domain can be achieved plays a crucial role in determining when exactly regulatory actions are required [Han *et al.*, 2020]. For instance, it would probably take very long until we have an AI capable of achieving anything that done by humans (one usually dubbed Artificial General Intelligence). Still, in many domains, such as chess playing, AI already outperforms humans. Arguably, it would not take very long until self-driving cars become safer than average human drivers. And other examples abound.

We find that, in short-term result scenarios, companies that ignore safety precautions are bound to win in our simulations, and hence they should be regulated. Nonetheless, in this case, the exact regulation requirements depend on finding a balance between a desirable innovation speed and its risk of engendering negative externalities.

Differently, in a long-term result scenario, screening for unsafe actions ensures that only when the risk is low will winning companies act in an unsafe manner. Such risk-taking, as opposed to full compliance with safety measures, should be regulated with society's benefit in mind. It goes without saying that, in either time-scale, only when individual benefits conflict with the overall societal interests, explicit regulation of unsafe actions becomes paramount.

These findings imply that, when defining codes of conduct and regulatory policies for AI, then first of all a clear understanding about the timescale of the race is a desirable prerequisite for effective AI governance. Regulation might not always be necessary and could even have detrimental effects if not timely applied in the right circumstances. We explicitly tested in our simulations what would happen if one always sanctioned companies that take risks [Han *et al.*, 2021]. As anticipated, over-regulation is conducive to beneficial innovation being stifled, and occurred whenever the gain from speed up out-benefited that of risk taking.

Main speaker's bio: The Anh Han is an associate professor at Teesside University. His research interest includes behavioral modelling, evolutionary game theory, agent-based simulations. He has published over 80 peer-reviewed articles in top-tier AI conferences and high-ranking scientific journal. His research has been funded by Future of Life Institute, Leverhulme Trust Foundation, and FWO Belgium. He regularly serves in programme committees of top tier conferences (e.g., AAAI, IJCAI, AAMAS) and on editorial boards of international journals (e.g., PLoS One, Adaptive Behavior).

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. *J. 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.