A Regulation Dilemma in Artificial Intelligence Development

The Anh Han[1], Francisco C. Santos[3,4], Luís Moniz Pereira [2], Tom Lenaerts[4,5]

[1] School of Computing and Digital Technologies, Teesside University
[2] NOVA Laboratory for Computer Science and Informatics (NOVA LINCS), Universidade Nova de Lisboa
[3]INESC-ID and Instituto Superior Técnico, Universidade de Lisboa
[4] Machine Learning Group, Université Libre de Bruxelles
[5] Artificial Intelligence Lab, Vrije Universiteit Brussel
Email: t.han@tees.ac.uk

Abstract

We examine a social dilemma that arises with the advancement of technologies such as AI, where technologists can choose a safe (SAFE) vs risk-taking (UNSAFE) course of development. SAFE is costlier and takes more time to implement than UNSAFE, allowing UNSAFE strategists to further claim significant benefits from reaching supremacy in a certain technology. Collectively, SAFE is the preferred choice when the risk is sufficiently high, while risk-taking is preferred otherwise. Given the advantage of risk-taking behaviour in terms of cost and speed, a social dilemma arises when the risk is not high enough to make SAFE the preferred individual choice, enabling UNSAFE to prevail when it is not collectively preferred (leading to a smaller population/social welfare). We show that the range of risk probabilities where the social dilemma arises depends on many factors, the most important among them are the time-scale to reach supremacy in a given domain (i.e. short-term vs long-term AI) and the speed gain by ignoring safety measures. Moreover, given the more complex nature of this scenario, we show that incentives such as reward and punishment (for example, for the purpose of technology regulation) are much more challenging to supply correctly than in case of cooperation dilemmas such as the Prisoner's Dilemma and the Public Good Games.

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, genetics, are generating an anxiety which makes companies, nations and regions think they should respond competitively (Armstrong et al., 2016; Baum, 2017; Bostrom, 2017; Cave and ÓhÉigeartaigh, 2018; Lee, 2018). AI appears for instance to have instigated a race among chip builders, simply because of the requirements it imposes on the 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 (AI-Roadmap-Institute, 2017; Cave and ÓhÉigeartaigh, 2018; Apps, 2019).

Races for supremacy in a domain through AI may however have detrimental consequences since participants to the race may well ignore ethical and safety checks in order to speed up the development and reach the market first. AI researchers and governance bodies, such as the EU, are urging to consider together both the normative and the social impact of major technological advancements concerned (Declaration, 2018; Russell et al., 2015; Jobin et al., 2019; European Commission, 2020; Future of Life Institute, 2019). However, given the breadth and depth of AI and its advances, it is not an easy task to assess when and which AI technology in a concrete domain needs to be regulated. This issue was, among others, highlighted in the recent EU White Paper on AI (European Commission, 2020). Data to estimate the risk of a technology is usually limited, especially at an early stage of its development or deployment. Its potential adverse effects may however begin to be ascertained by the expert reviewers of scientific publications (Hutson, 2021).

Here, we summarise our previous works (Han et al., 2020, 2021b) examining this problem theoretically, resorting to a novel innovation dilemma where technologists can choose a safe (SAFE) vs risk-taking (UNSAFE) course of development. Companies race towards the deployment of some AI -based product in some domain X. They can either carefully consider all data and AI pitfalls along the way (SAFE) or else take undue risks by skipping some tests just to speed up the process (UNSAFE). Overall, SAFE are costlier strategies and take more time to implement than UNSAFE strategies, allowing UNSAFE strategists to further claim significant benefits from reaching technological supremacy.

In more detail, we posit that it requires time to reach domain supremacy through AI (DSAI), modelling this by a number of development steps or technological advancement rounds (Han et al., 2020). In each round the development teams (or players) need to choose between one of two strategic options: to follow safety precautions (the SAFE action) or ignore safety precautions (the UNSAFE action). Because it takes more time and more effort to comply with precautionary requirements, playing SAFE is not just costlier, but implies slower development speed too, compared to playing UNSAFE. We consequently assume that to play SAFE involves paying a cost $c > 0$, while playing UNSAFE costs

nothing ($c = 0$). Moreover, the development speed of playing UNSAFE is $s > 1$ whilst the speed of playing SAFE is normalised to $s = 1$. The interaction is iterated until one or more teams establish DSAI, which occurs probabilistically, i.e. the model assumes, upon completion of each round, that there is a probability ω that another development round is required to reach DSAI—which results in an average number $W = (1 - \omega)^{-1}$ of rounds per competition/race. We thus do not make any assumption about the time required to reach DSAI in a given domain. Yet once the race ends, a large benefit or prize B is acquired that is shared amongst those reaching the target simultaneously.

We pitch all development teams (e.g. AI companies, nations) together in the time evolution of the adoption of SAFE and UNSAFE policies resorting to the tools of evolutionary game theory (Sigmund, 2010). As the ALife community has shown repeatedly, this sort of modelling and simulation has been applied successfully to understand the evolution of social and biological behaviours at all scales, from the social organization level to the human agent and even gene ones.

As a result, we identify conditions under which safe or risk-taking behaviour emerges, and when they are collectively preferred, leading to a greater population social welfare. We next explored ways to influence it towards safe and beneficial outcomes, namely when and how to sanction unsafe decisions made by stake-holders or reward compliant ones. Finally, we identify when regulations need to be put in place to favour outcomes most beneficial for all, but at the same time taking care to avoid strict regulations that would be introduced too early and thereby stifle innovation (Han et al., 2020, 2021b, 2019).

Lessons for AI governance policies

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 regulatory actions are required (Han et al., 2020). For instance, it would probably take very long until we have an AI capable of doing anything that humans do (one usually termed Artificial General Intelligence). Still, in many domains, such as chess playing, AI already outperforms humans. It would not take very long until self-driving cars become safer than average human drivers. 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 requirements of regulations depend on finding a balance between the desirable innovation speed and the risk of its 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 compliance with safety measures, should be regulated for society's benefit. It goes without saying that, in both time-scales, only when individual benefits conflict with the overall societal interests, will explicit regulation of unsafe actions become paramount.

These findings indicate that, when defining codes of conduct and regulatory policies for AI, first of all, a clear understanding about the timescale of the race is required for effective AI governance. Regulation might not always be necessary and could even have detrimental effects if not timely applied in the right circumstances.

Indeed, we explicitly tested in our simulations what would happen if always companies that take risks are sanctioned (Han et al., 2021b), reducing their speed but at the cost of speed reduction by the sanctioning party. As anticipated, over-regulation, conducive to beneficial innovation being stifled, occurred whenever the gain from speeding up outbenefited the taking of risk.

Yet an issue remains to be solved for proper regulation: Even if we can assess the game's timescale, we still need to estimate the measures of risk and gain associated with risk-taking behaviours. We need data to do so, but it is usually not yet available at an early stage of development.

Our latest finding though suggests a way out based on the idea of voluntary safety agreements. That is, desirable outcomes are achievable without any over-regulation whatsoever if companies have the freedom of choice between independently pursuing their course of actions or else establishing instead binding agreements to act safely. Sanctioning can then be applied only against those that do not abide by their commitment pledges (Han et al., 2021a). Thus, our analysis indicates the need to facilitate this option, enabling AI companies to voluntarily commit to safety agreements without repercussion should they choose to opt out.

Concluding Remarks

We have described how evolutionary modelling and simulations using game theory can be powerful to generate useful insights into the behavioural dynamics and the impact of interventions, in the context of AI development and governance. This approach has been widely adopted to study biological and artificial life systems (Nowak, 2006; Andras et al., 2018; Perc et al., 2017; Smaldino and Lubell, 2014; Han et al., 2021c), which once again has here shown its usefulness to study a complex issue of significant importance.

Acknowledgements T.A.H., L.M.P., T.L. are supported by Future of Life Institute grant RFP2-154. FCT-Portugal supported with grants UIDB/04516/2020 (L.M.P.), and UIDB/50021/2020, PTDC/MAT-APL/6804/2020, PTDC/CCI-INF/7366/2020 (F.C.S). Fonds de la Recherche Scientifique (F.R.S)-F.N.R.S. supported through the projects PDR31257234 and FuturICT2.0 (www.futurict2.eu) withing the call FLAG-ERA JCT 2016. T.A.H. is also supported by Leverhulme Research Fellowship (RF-2020-603/9).

References

AI-Roadmap-Institute (2017). Report from the ai race avoidance workshop, tokyo.

Andras, P., Esterle, L., Guckert, M., Han, T. A., Lewis, P. R., Milanovic, K., Payne, T., Perret, C., Pitt, J., Powers, S. T., et al. (2018). Trusting intelligent machines: Deepening trust within socio-technical systems. *IEEE Technology and Society Magazine*, 37(4):76–83.

Apps, P. (2019). Are China, Russia winning the AI arms race? [Reuters; Online posted 15-January-2019].

Armstrong, S., Bostrom, N., and Shulman, C. (2016). Racing to the precipice: a model of artificial intelligence development. *AI & society*, 31(2):201–206.

Baum, S. D. (2017). On the promotion of safe and socially beneficial artificial intelligence. *AI & Society*, 32(4):543–551.

Bostrom, N. (2017). Strategic implications of openness in AI development. *Global Policy*, 8(2):135–148.

Cave, S. and ÓhÉigeartaigh, S. (2018). An AI Race for Strategic Advantage: Rhetoric and Risks. In *AAAI/ACM Conference on Artificial Intelligence, Ethics and Society*, pages 36–40.

Declaration, M. (2018). The montreal declaration for the responsible development of artificial intelligence launched. https://www.canasean.com/the-montreal-declaration-for-the-responsible-development-of-artificial-intelligence-launched/.

European Commission (2020). White paper on Artificial Intelligence – An European approach to excellence and trust. Technical report, European Commission.

Future of Life Institute (2019). Lethal autonomous weapons pledge. https://futureoflife.org/lethal-autonomous-weapons-pledge/.

Han, T. A., Lenaerts, T., Santos, F. C., and Pereira, L. M. (2021a). Voluntary safety commitments provide an escape from over-regulation in AI development. Preprint: arxiv.org/abs/2104.03741.

Han, T. A., Pereira, L. M., and Lenaerts, T. (2019). Modelling and Influencing the AI Bidding War: A Research Agenda. In *Proceedings of the AAAI/ACM conference AI, Ethics and Society*, pages 5–11.

Han, T. A., Pereira, L. M., Lenaerts, T., and Santos, F. C. (2021b). Mediating Artificial Intelligence Developments through Negative and Positive Incentives. *PLOS ONE*, 16(1):e0244592.

Han, T. A., Pereira, L. M., Santos, F. C., and Lenaerts, T. (2020). To Regulate or Not: A Social Dynamics Analysis of an Idealised AI Race. *Journal of Artificial Intelligence Research*, 69:881–921.

Han, T. A., Perret, C., and Powers, S. T. (2021c). When to (or not to) trust intelligent machines: Insights from an evolutionary game theory analysis of trust in repeated games. *Cognitive Systems Research*, 68:111–124.

Hutson, M. (2021). Who Should Stop Unethical A.I.? https://www.newyorker.com/tech/annals-of-technology/who-should-stop-unethical-ai.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, pages 1–11.

Lee, K.-F. (2018). *AI superpowers: China, Silicon Valley, and the new world order*. Houghton Mifflin Harcourt.

Nowak, M. A. (2006). *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard University Press, Cambridge, MA.

Perc, M., Jordan, J. J., Rand, D. G., Wang, Z., Boccaletti, S., and Szolnoki, A. (2017). Statistical physics of human cooperation. *Phys Rep*, 687:1–51.

Russell, S., Hauert, S., Altman, R., and Veloso, M. (2015). Ethics of artificial intelligence. *Nature*, 521(7553):415–416.

Sigmund, K. (2010). *The Calculus of Selfishness*. Princeton University Press.

Smaldino, P. E. and Lubell, M. (2014). Institutions and cooperation in an ecology of games. *Artificial life*, 20(2):207–221.