Network Diversity Promotes Safety Adoption in Swift Artificial Intelligence Development

Theodor Cimpeanu[1], Francisco C. Santos[3], Luís Moniz Pereira[2],
Tom Lenaerts[4,5], and The Anh Han[1,⋆]
[1] School of Computing, Engineering 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

Abstract

Regulating the development of advanced technology such as Artificial Intelligence (AI) has become a principal topic, given the potential threat they pose to humanity's long term future. First deploying such technology promises innumerable benefits, which might lead to the disregard of safety precautions or societal consequences in favour of speedy development, engendering a race narrative among firms and stakeholders due to value erosion. Building upon a previously proposed game-theoretical model describing an idealised technology race, we investigated how various structures of interaction among race participants can alter collective choices and requirements for regulatory actions. Our findings indicate that strong diversity among race participants, both in terms of connections and peer-influence, can reduce the conflicts which arise in purely homogeneous settings, thereby lessening the need for regulation.

Introduction

Researchers and stakeholders alike have urged for due diligence in regard to AI development on the basis of several concerns. The desire to be at the foreground of the state-of-the-art, or the pressures imposed by upper management, might tempt developers to ignore safety procedures or apprehensions about ethical consequences (Armstrong et al., 2016; Cave and ÓhÉigeartaigh, 2018). Regulation and governance of advanced technologies such as Artificial Intelligence (AI) have become increasingly more important given their potential implications, such as for associated risks and ethical issues (European Commission, 2020; Declaration, 2018; Russell et al., 2015; Future of Life Institute, 2015, 2019). With the tremendous benefits promised from being first able to supply such technologies, stake-holders might cut corners on safety precautions in order to ensure rapid deployment, in a race towards AI market supremacy (AIS) (Armstrong et al., 2016; Cave and ÓhÉigeartaigh, 2018).

With this aim in mind, a baseline model of an innovation race has been recently proposed (Han et al., 2020), in which innovation dynamics are pictured through the lens of Evolutionary Game Theory (EGT) and where all race participants are equally well-connected in the system. The baseline results have showed the importance of accounting for different time-scales of development, and also exposed the dilemmas that arise when what is individually preferred by developers differs from what is globally beneficial. However, real-world stakeholders and their interactions are far from homogeneous (Schilling and Phelps, 2007; Newman, 2004; Barabasi, 2014). Some individuals are more influential than others, or play different roles in the unfolding of new technologies. Technology races are shaped by complex networks of exchange, influence, and competition where diversity abounds. Here we summarise a recent work (Cimpeanu et al., 2022) studying impacts of network topology on the adoption of safety measures in innovation dynamics.

Models and Methods

Assuming that winning the race towards supremacy is the goal of the development teams and that a number of development steps are required, the players have two strategic options at each step: to follow safety precautions (denoted by SAFE) or to ignore them (denoted by UNSAFE) (Han et al., 2020). As it takes more time and effort to comply with the precautionary requirements, playing SAFE is not only costlier but also implies a slower development speed, compared to playing UNSAFE. Let us also assume that to play SAFE players need to pay additional costs. The interactions are iterated until one or more teams achieve a designated objective, after having completed W development steps. As a result, the players obtain a large benefit B, shared among those who reach the target objective at the same time. However, a setback or disaster can happen with some probability, which is assumed to increase with the number of times the safety requirements have been omitted by the winning team(s). Although many potential AI disaster scenarios have been sketched (Armstrong et al., 2016; Pamlin and Arm-

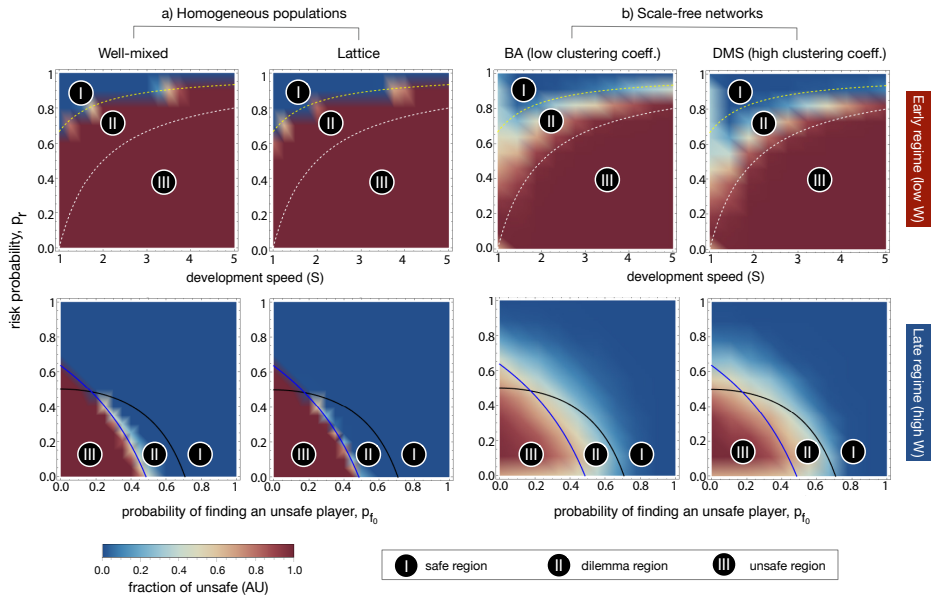

Figure 1: Colour gradients indicating the average fraction of AU (unsafe strategy) for (a) homogeneous (well-mixed and lattices) populations and (b) scale-free networks (BA and DMS models). In the early regime, region II indicates the parameters in which safe AI development is the preferred collective outcome, but unsafe development is expected to emerge and regulation may be needed. In regions I and III, safe and unsafe AI development, respectively, are both preferred collective outcomes and expected to emerge from self-organization. In the late regime, the solid black line marks the boundary above which safety is the preferred outcome, whereas the blue line indicates the boundary above which safety becomes risk dominant against unsafe development.

strong, 2015; Han et al., 2019, 2022), the uncertainties in accurately predicting these outcomes are high. When such a disaster occurs, risk-taking participants lose all their benefits. We denote by p_r the risk probability of such a disaster occurring when no safety precaution is followed at all.

To study the effect of network structures on the safety outcome, we have analysed the following types of networks, from simple to more complex (Cimpeanu et al., 2022): well-mixed populations (complete graph), where each agent interacts with all other agents in a population; square lattice of size with periodic boundary conditions; and scale-free (SF) networks (Barabási and Albert, 1999; Dorogovtsev, 2010; Newman, 2003), generated by means of two growing network models — the widely-adopted Barabási-Albert (BA) model (Barabási and Albert, 1999; Albert and Barabási, 2002) and the Dorogovtsev-Mendes-Samukhin (DMS) model (Dorogovtsev, 2010), the latter of which allowed us to assess the role of a large number of triangular motifs (i.e. high clustering coefficient).

Results and Conclusions

We initially considered the roles of degree-homogeneous graphs in the evolution of safety in the AI race game. They simulated the AI race game in well-mixed populations (Figure 1, first column), and then explored the same game on a square lattice, where each agent can interact with its four edge neighbours (Figure 1, second column). They show that the trends remain the same when compared with well-mixed populations, with very slight differences in numerical values between the two. That is, homogeneous spatial variation is not enough to influence safe technological development.

Investigating beyond homogeneous structures, we make use of two SF network models. Contrary to the findings on homogeneous networks, SF structures produce marked improvements in almost all parameter regions of the AI race game (Figure 1). Given that innovation in the field of AI (more broadly, technological advancement), should be profitable (and robust) to developers, shareholders and society altogether, it is important to discuss the analytical loci where these objectives can be fulfilled. Assuredly, it is observed that diversity in players introduces two marked improvements in both early and late safety regimes. Firstly, very little regulation is required in the case of a late AI race, principally concerning the existing observations in homogeneous settings. Intuitively, this suggests that there is little encouragement needed to promote risk-taking in late AIS regimes: diversity enables beneficial innovation. Secondly, the region for early AIS regimes in which regulation must be enforced is diminished, but not completely eliminated. Consequently, governance should still be prescribed when developers are racing towards an early or otherwise uncertain timeline to reaching transformative AI. It stands to reason that insight into what regime type the AI race operates in is therefore paramount to the success of any potential regulatory actions.

References

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47.

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

Barabasi, A.-L. (2014). *Linked-how Everything is Connected to Everything Else and what it Means F*. Perseus Books Group.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.

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

Cimpeanu, T., Santos, F. C., Pereira, L. M., Lenaerts, T., and Han, T. A. (2022). Artificial intelligence development races in heterogeneous settings. *Scientific Reports*, 12(1):1–12.

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/.

Dorogovtsev, S. (2010). *Complex networks*. Oxford: Oxford University Press.

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

Future of Life Institute (2015). Autonomous Weapons: An Open Letter from AI \& Robotics Researchers. Technical report, Future of Life Institute, Cambridge, MA.

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. (2022). Voluntary safety commitments provide an escape from over-regulation in ai development. *Technology in Society*, 68:101843.

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., 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.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences*, 101(suppl 1):5200–5205.

Pamlin, D. and Armstrong, S. (2015). Global challenges: 12 risks that threaten human civilization. *Global Challenges Foundation, Stockholm*.

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

Schilling, M. A. and Phelps, C. C. (2007). Interfirm collaboration networks: The impact of large-scale network structure on firm innovation. *Management science*, 53(7):1113–1126.