Voluntary safety pledges overcome over-regulation dilemma in AI development: an evolutionary game analysis

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

[1] School of Computing, Engineering and Digital Technologies, Teesside University (Email: t.han@tees.ac.uk)

[2] INESC-ID and Instituto Superior Técnico, Universidade de Lisboa

[3] NOVA Laboratory for Computer Science and Informatics (NOVA LINCS), Universidade Nova de Lisboa

[4] Machine Learning Group, Université Libre de Bruxelles

[5] Artificial Intelligence Lab, Vrije Universiteit Brussel

Abstract

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, we show that socially unwanted outcomes may be produced when sanctioning is applied unconditionally to risk-taking, i.e., potentially unsafe behaviours. 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. Overall, our 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.

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; 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) an the UK National AI strategy.

Several proposals for mechanisms on how to avoid, mediate, or regulate the development and deployment of AI, have been made (Baum, 2017; Cave and ÓhÉigeartaigh, 2018; Geist, 2016; Shulman and Armstrong, 2009; Han et al., 2019; Vinuesa et al., 2020; Nemitz, 2018; Taddeo and Floridi, 2018; Askell et al., 2019; O'Keefe et al., 2020; Cimpeanu et al., 2022). Essentially, regulatory measures such as restrictions and incentives are proposed to limit harmful and risky practices in order to promote beneficial designs (Baum, 2017). Examples include financially supporting the research into beneficial AI (McGinnis, 2010) and making AI companies pay fines when found liable for the consequences of harmful AI (Gurney, 2013).

Although such regulatory measures may provide solutions for particular scenarios, one needs to ensure that they do not overshoot their targets, leading to a stifling of novel innovations, hindering investments into the development into novel directions as they may be perceived to be too risky (Hadfield, 2017; Lee, 2018). Worries have been expressed by different organisations and academic societies that too strict policies may unnecessarily affect the benefits and societal advances that novel AI technologies may have to offer (EDRI, 2021). Regulations affect moreover big and small tech companies differently: A highly regulated domain makes it more difficult for small new start-ups, introducing an inequality and dominance of the market by a few big players (Lee, 2018). It has been emphasised that neither over-regulation nor a laissez-faire approach suffices

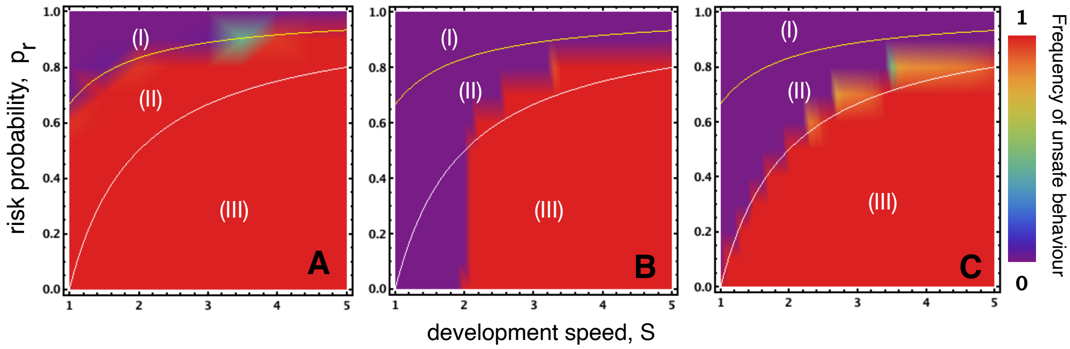

Figure 1: Frequency of unsafe behaviour as a function of development speed (s) and the disaster risk (p_r). **Panel A, in absence of incentives** (Han et al., 2020), the parameter space can be split into three regions. In regions (I) and (III), safe and unsafe/innovation, respectively, are the preferred 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. **Panel B, when unsafe behaviour is sanctioned unconditionally** (Han et al., 2021), while unsafe behaviour is reduced in region II, over-regulation occurs in region III, reducing beneficial innovation. **Panel C, unsafe behaviour is sanctioned only in presence of a voluntary commitment** (Han et al., 2022), unsafe behaviour is significantly reduced in region II while avoiding over-regulation.

when aiming to regulate AI technologies (Dawson et al., 2019). In order to find a balanced answer, one clearly needs to have first an understanding of how a competitive development dynamic actually could work and how governance choices impact this dynamic, a task well-suited for dynamic systems or agent-based models.

Here, we highlight main results from our recent work (Han et al., 2022) examining this problem theoretically, using methods from Evolutionary Game Theory (Sigmund, 2010), see Figure 1. It resorts to a baseline model describing a development competition where technologists can choose a safe (SAFE) vs risk-taking (UNSAFE) course of development (Han et al., 2020). Namely, it considers that to reach domain supremacy through AI in a certain domain, a number of development steps or technological advancement rounds are required (Han et al., 2020). In each round the technologists (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. Moreover, there is a probability that a disaster occurs if UNSAFE developments take place during this competition (see (Han et al., 2020) for a full description).

We first demonstrate that unconditional sanctioning will negatively influence social welfare in certain conditions of a short-term race towards domain supremacy through AI technology (Han et al., 2021), leading to over-regulation of beneficial innovation (see Figure 1B). Since data to estimate the risk of a technology is usually limited (especially at an early stage of its development or deployment), simple sanctioning of unsafe behaviour (or reward of safe behaviour) could not fully address the issue.

To solve this critical over-regulation dilemma in AI development, we propose an alternative approach (Han et al., 2022), which is to allow technologists or race participants to voluntarily commit themselves to safe innovation procedures, signaling to others their intentions (Han et al., 2015; Nesse, 2001; Han, 2022). Specifically, this bottom-up, binding agreement (or commitment) is established for those who want to take a safe choice, with sanctioning applied to violators of such an agreement. It is shown that, by allowing race participants to freely pledge their intentions and enter (or not) in bilateral commitments to act safely and avoid risks, accepting thus to be sanctioned in case of misbehavior, high levels of the most beneficial behaviour, for the whole, are achieved in all regions of the parameter space, see Figure 1C. These results are directly relevant for the design of self-organized AI governance mechanisms and regulatory policies that aim to ensure an ethical and responsible AI technology development process.

Acknowledgements

This work was supported by a Future of Life Institute AI grant (RFP2-154).

References

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

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.

Askell, A., Brundage, M., and Hadfield, G. (2019). The Role of Cooperation in Responsible AI Development. *arXiv preprint arXiv:1907.04534*.

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.

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.

Dawson, D., Schleiger, E., Horton, J., McLaughlin, J., Robinson, C., Quezada, G., Scowcroft, J., and S, H. (2019). Artificial Intelligence: Australia's Ethics Framework. Technical report, Data61 CSIRO, Australia.

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/.

EDRI (2021). Civil society calls for AI red lines in the European Union's Artificial Intelligence proposal. Technical report, European Commission. Accessed January-29-2021.

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/.

Geist, E. M. (2016). It's already too late to stop the ai arms race: We must manage it instead. *Bulletin of the Atomic Scientists*, 72(5):318–321.

Gurney, J. K. (2013). Sue my car not me: Products liability and accidents involving autonomous vehicles. *U. Ill. JL Tech. & Pol'y*, page 247.

Hadfield, G. K. (2017). *Rules for a flat world: why humans invented law and how to reinvent it for a complex global economy*. Oxford University Press.

Han, T. A. (2022). Institutional incentives for the evolution of committed cooperation: ensuring participation is as important as enhancing compliance. *Journal of The Royal Society Interface*, 19(188):20220036.

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., Lenaerts, T., and Santos, F. C. (2021). 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., Santos, F. C., Lenaerts, T., and Pereira, L. M. (2015). Synergy between intention recognition and commitments in cooperation dilemmas. *Scientific reports*, 5(9312).

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.

McGinnis, J. O. (2010). Accelerating AI. *Nw. UL Rev.*, 104:1253.

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180089.

Nesse, R. M. (2001). *Evolution and the capacity for commitment*. Foundation series on trust. Russell Sage.

O'Keefe, C., Cihon, P., Garfinkel, B., Flynn, C., Leung, J., and Dafoe, A. (2020). The windfall clause: Distributing the benefits of ai for the common good. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 327–331.

Shulman, C. and Armstrong, S. (2009). Arms control and intelligence explosions. In *7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, July*, pages 2–4.

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

Taddeo, M. and Floridi, L. (2018). Regulate artificial intelligence to avert cyber arms race. *Nature*, 556(7701):296–298.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S., Tegmark, M., and Nerini, F. F. (2020). The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, 11(233).