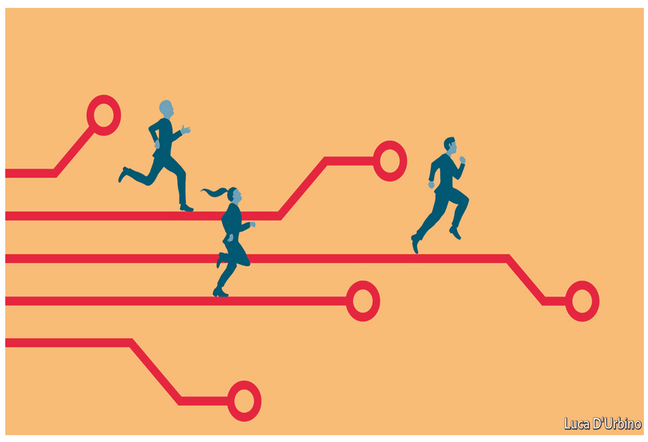
Time-scale Differences Will Influence the Regulation Required in an Idealised AI Race Game

The Anh Han[1,*], L.M. Pereira[2] , F.C. Santos[3], T. Lenaerts[4]

1)Computing&Games depart., Teessides University 2) NOVA-LINCS, Universidade Nova de Lisboa
3) IST Lisbon, 4) MLG group, Université Libre de Bruxelle

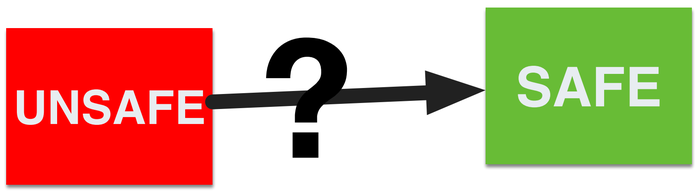
- A race for technological supremacy could lead to serious negative consequences (e.g. unsafe extra speedy development).
- Little attention has been given to understanding the dynamics and emergence of safety behaviours arising from an AI race.
- We use Evolutionary Game Theory (**EGT**) to build models of competition and cooperation among AI development teams.
- Besides the level of risk, the timescale to reach supremacy in an AI domain decides the regulatory action required for maximizing societal benefit.

An Evolutionary Game Model of AI Racing

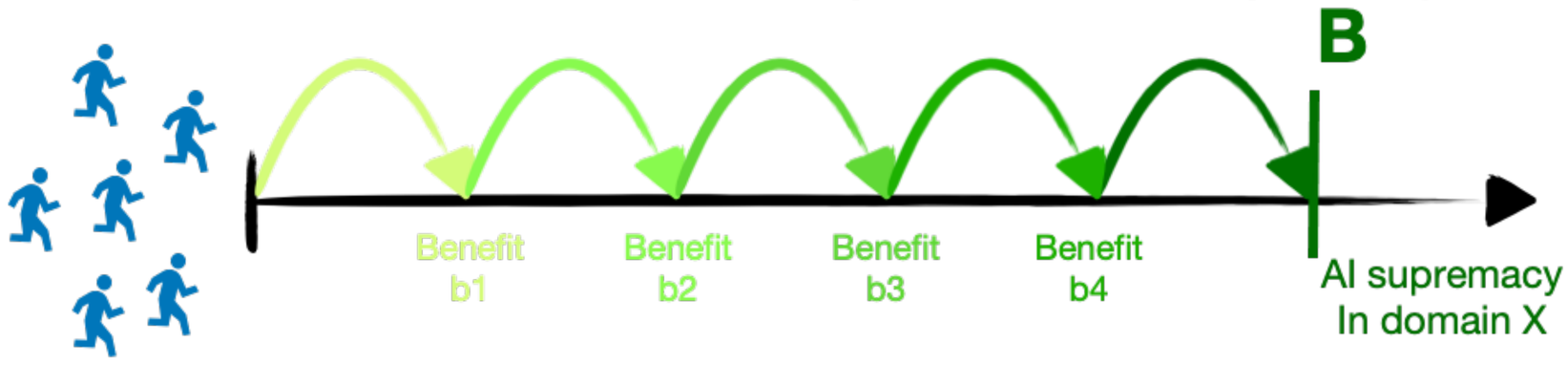


What are the key factors influencing the AI Race?

- Time scale of the AI race
- Risk perception
- Inequalities, networks
- Incentives, regulation



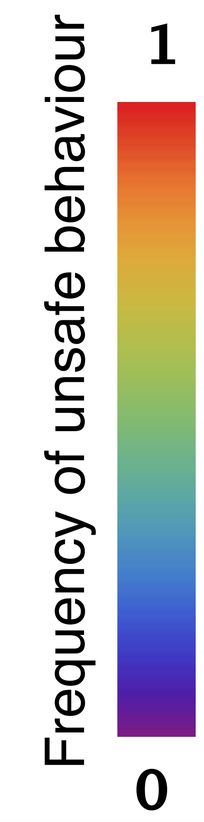
Several advancements are required to reach supremacy



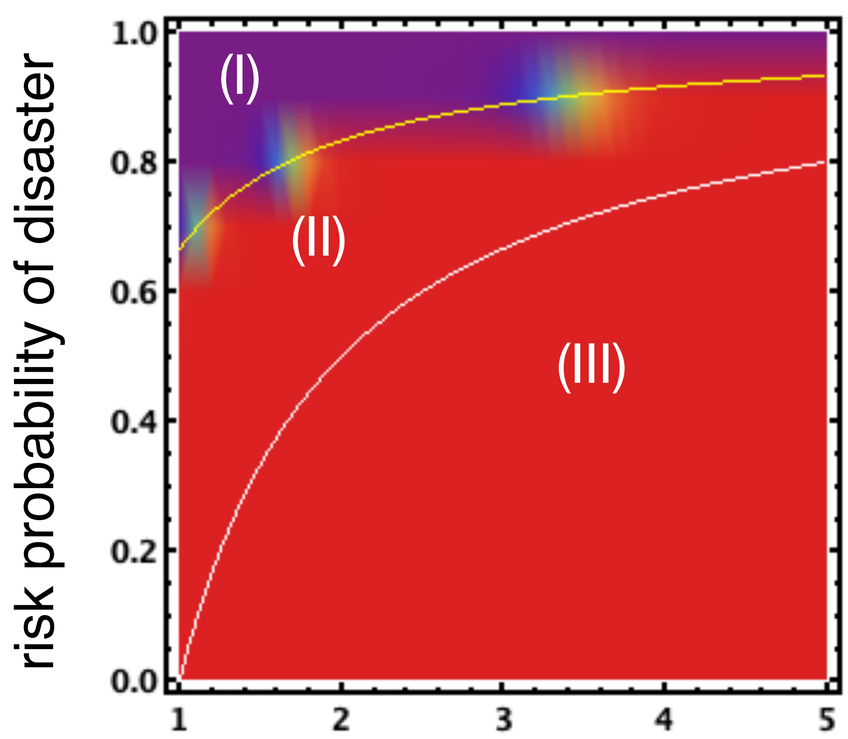
✓ AI Race is modeled as a repeated game with two options SAFE and UNSAFE in each round.
✓ Playing SAFE is more costly and takes more time than playing UNSAFE.
✓ We study a well-mixed population of AI teams
- AS: always plays SAFE
- AU: always plays UNSAFE
- CS: conditionally playing SAFE

When regulation is required?

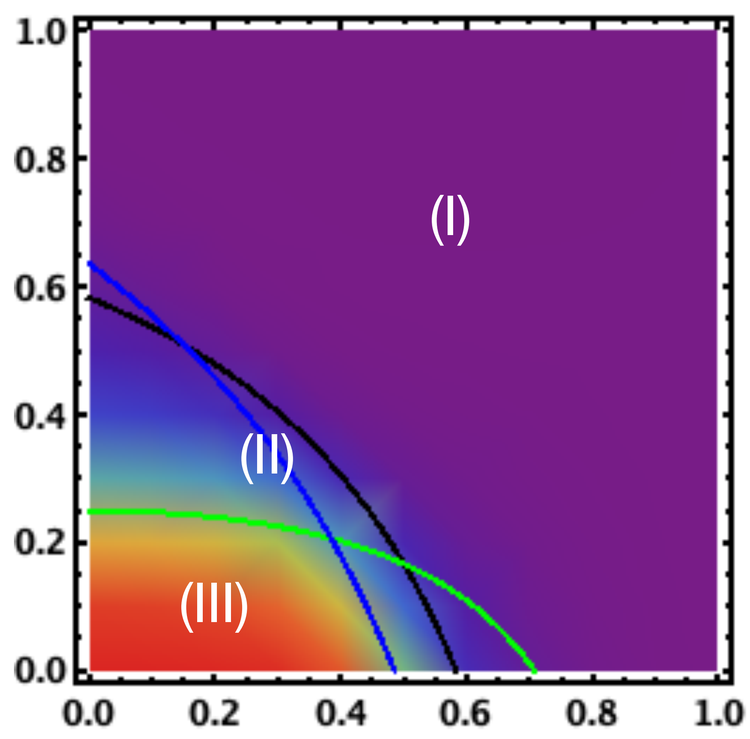
(I) Compliance Zone
(II) Dilemma Zone
(III) Innovation Zone



Short-term AI requires regulation of unsafe



Long-term AI requires promotion of risk-taking



REFERENCES

1) Han et al. *To Regulate or Not: A Social Dynamics Analysis of an Idealised AI Race.* JAIR, 69: 881-921, 2020
2) Han et al. *Mediating Artificial Intelligence development through positive and negative incentives.* PLoS ONE16(1):e0244592, 2021

*t.han@tees.ac.uk