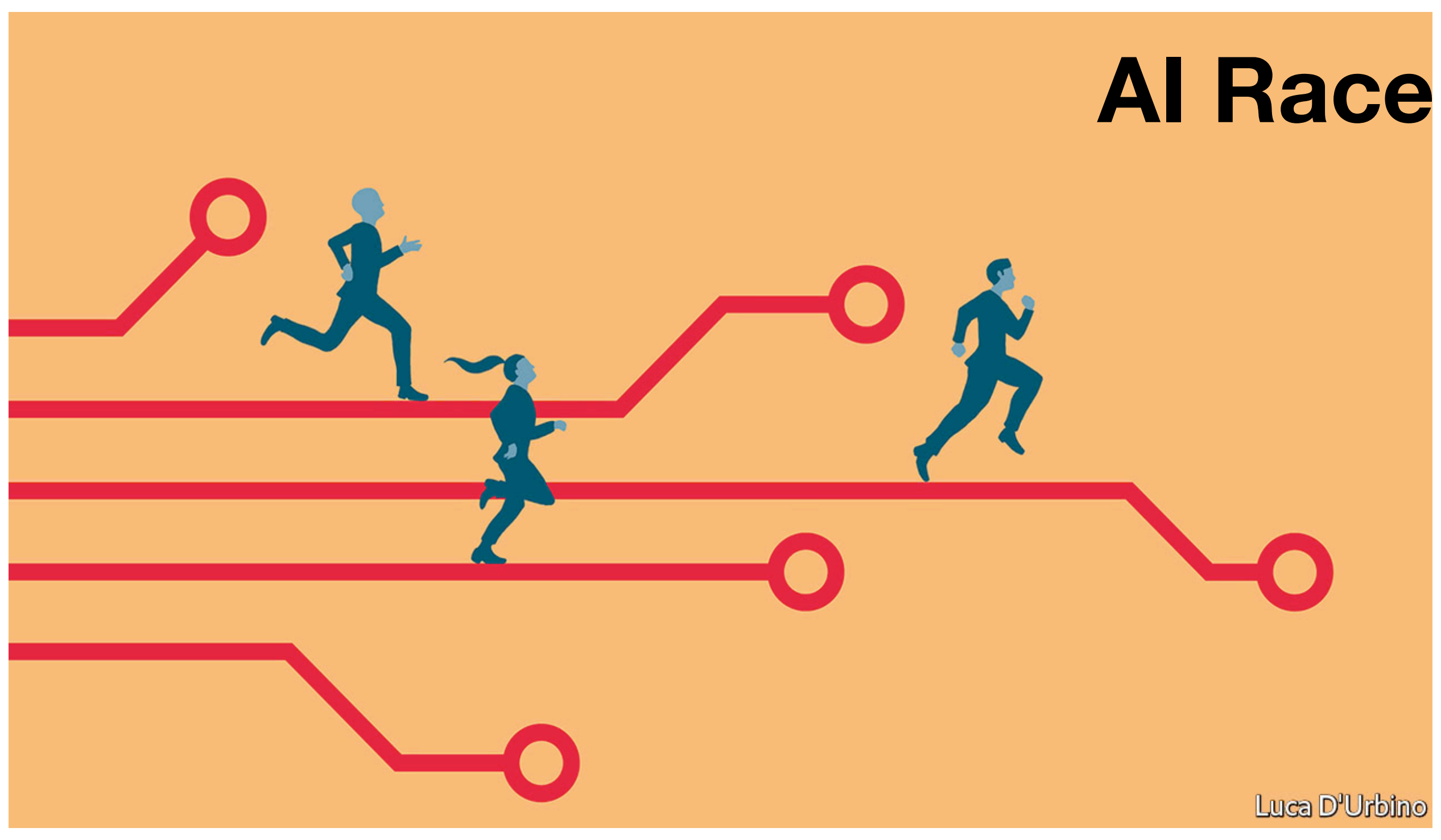
Modelling and Influencing the AI Bidding War: A Research Agenda

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

1)Computer Science department, Teessides University 2) NOVA-LINCS, Universidade Nova de Lisboa 3) MLG group, Université Libre de Bruxelles, 4) AI Lab, Vrije Universiteit Brussel

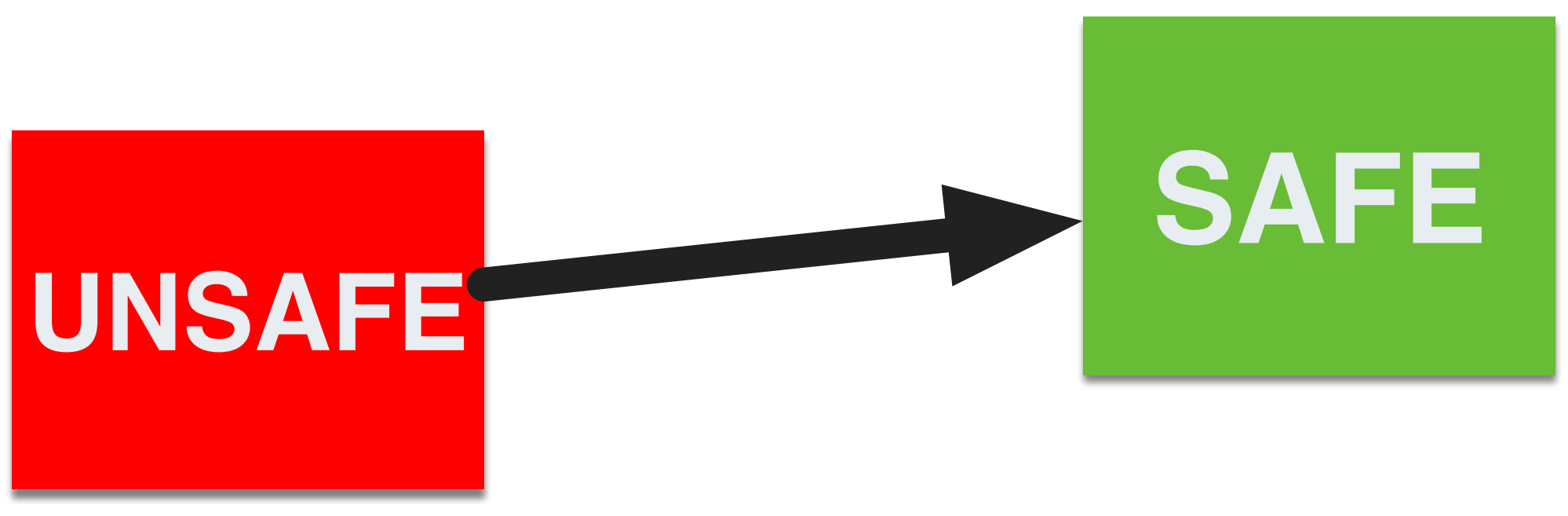
- ◆ A race for technological supremacy could lead to serious negative consequences (e.g. unsafe extra speedy development).
- ◆ Little attention has been given to understanding the dynamics and emergent behaviours arising from an AI race.
- ◆ We use Evolutionary Game Theory (**EGT**) to build models of competition and cooperation among AI development teams.
- ◆ Propose research agenda for **modelling** the AI race to understand its dynamics and how to influence it in a beneficial way.

AI Race Modelling Research Agenda



What are the key factors influencing the AI Race?
1) Openness
2) Risk perception
3) Inequalities (resources, capabilities, etc)
…..

How incentives, viz. peer vs institutional, negative vs positive, and their combinations, can be used to ensure safety compliance?



AI Safety Agreement

EGT dynamical modelling of **agreements** & **incentives**

Two-team AI Race Models (preliminary)

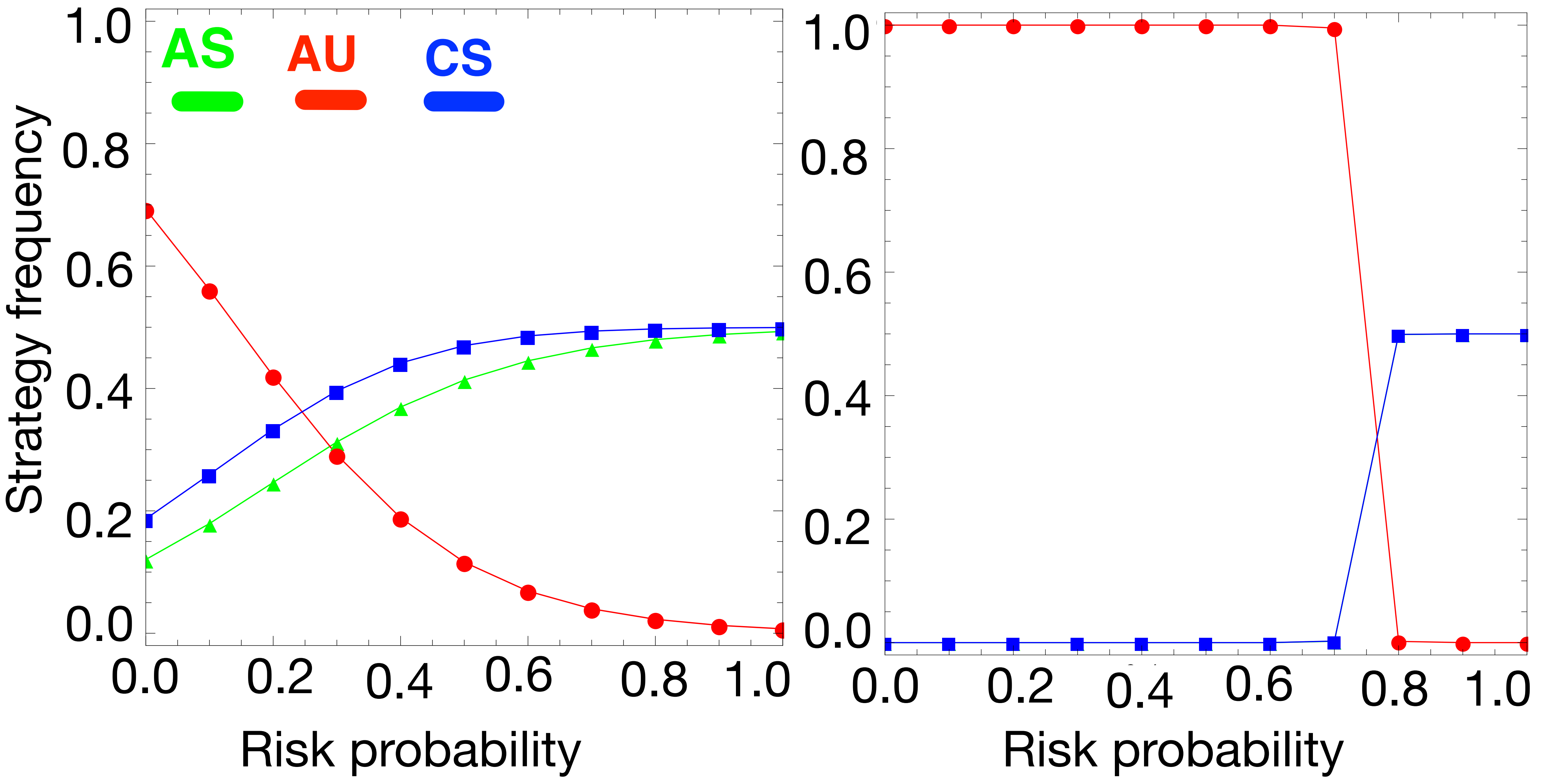
AI Race as a repeated game with two options SAFE and UNSAFE in each round.

Playing SAFE is more costly and takes more time than playing UNSAFE.

We consider a well-mixed population of players adopting one of three strategies
1) AS: always plays SAFE
2) AU: always plays UNSAFE
3) CS: conditionally playing SAFE

When benefit from winning the race is large, Always Unsafe (AU) dominates population

Race Winning benefit: SMALL **Race Winning benefit: LARGE**



NEXT STEPS

◆ Incorporate key factors into the models (group size, openness, inequalities, etc)
Incentives for promoting safety behaviour and agreement compliance (peer vs institutional, rewards vs punishment)

*t.han@tees.ac.uk