Modelling and regulating safety compliance: Game theory lessons from AI development races analyses

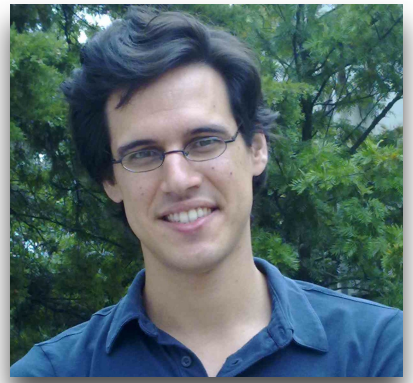
The Anh Han

AI Safety @ IJCAI, August 19-20, 2021

IJCAI 2021 MONTREAL

Tom Lenaerts,
ULB Brussels

Luis Moniz Pereira,
UNL Lisbon



Francisco C. Santos,
IST Lisbon

Teesside University

t.han@tees.ac.uk

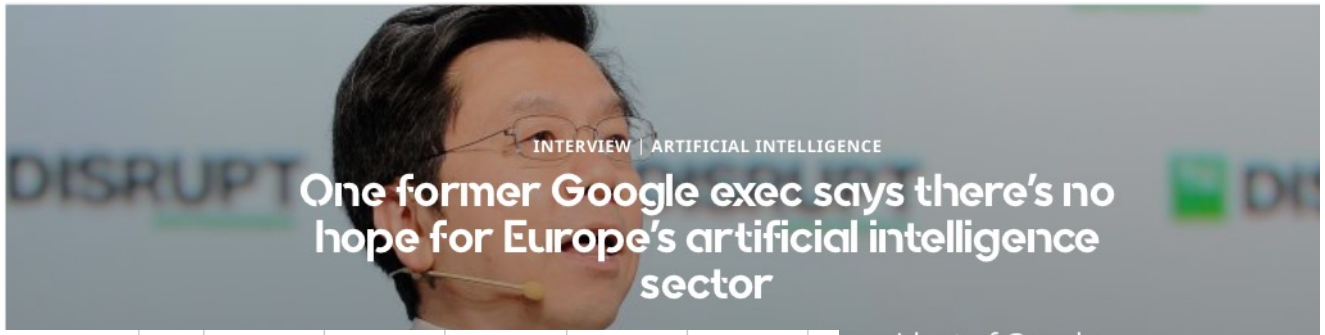

The Economist
SPECIAL REPORT
AI IN BUSINESS
March 31st 2018
GrAIt expectations
A reprint of *The Economist's* special report on AI in Business brought to you, with permission from the publisher, by
cloudera®

Commentary: Are China, Russia winr AI arms race?

Peter Apps

13 MIN RE

In October 31 Chinese teenagers reported to the Beijing Institute of Tech the country's premier military research establishments. Selected from m applicants, Chinese authorities hope they will design a new generation o



29.06.2018, TRYING TO KE

Europe'

Being a leading tech innovator is future. But can Europe catch up

BBC | Sign in | Home | News | Sport | Weather | iPlayer

NEWS

Home | Brexit | Coronavirus | UK | World | Business | Politics | Tech | Science | Health | Family & Education

Technology

Biden urged to back AI weapons to counter China and Russia threats

By Leo Kelion
Technology desk editor

9 hours ago

Long Reads

From imitation t China became a


AI
Innovation/
Research
RULES
REGULATIONS
COMPLIANCE
STANDARDS
POLICIES

Question 1: What do we actually know about the dynamics of this (alleged) AI race?

Question 2: What do we know about the impact of regulatory actions on such race dynamics?

Question 3: Is regulation always useful? Useful in what sense?

Journal of Artificial Intelligence Research 69 (2020) 881-921 Submitted 06/2020; published 11/2020

Question 1: What do we actually know about the dynamics of this (alleged) AI race?

To Regulate or Not: A Social Dynamics Analysis of an Idealised AI Race

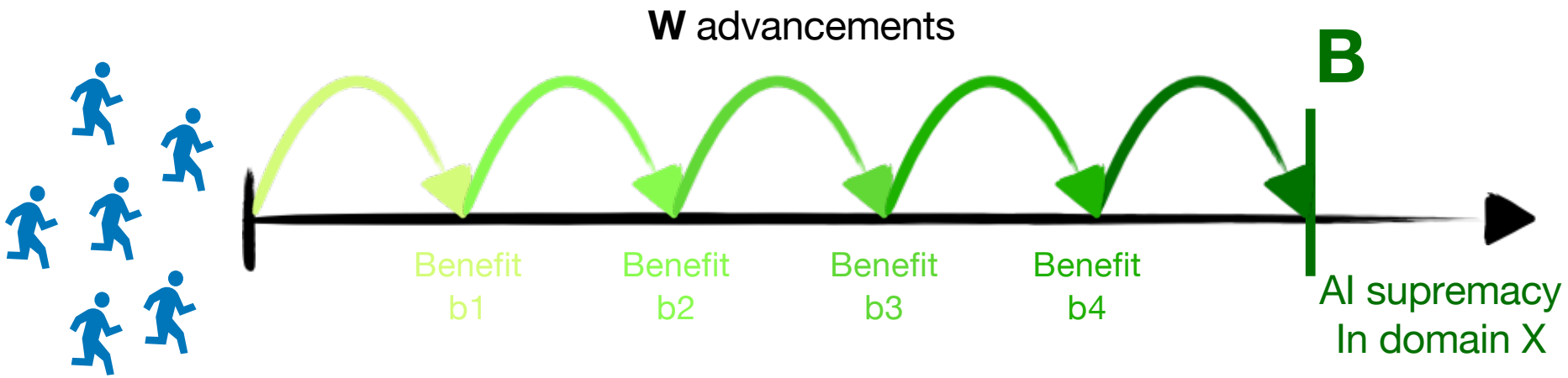
The Anh Han T.HAN@TEES.AC.UK
School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, UK TS1 3BA

Luís Moniz Pereira LMP@FCT.UNL.PT
NOVA Laboratory for Computer Science and Informatics (NOVA LINCS), Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

Francisco C. Santos FRANCISCOCSANTOS@TECNICO.ULISBOA.PT
INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, IST-Taguspark, 2744-016, Porto Salvo, Portugal
& Machine Learning Group, Université Libre de Bruxelles, Boulevard du Triomphe CP212, Brussels, Belgium

Tom Lenaerts TOM.LENAERTS@ULB.AC.BE
Machine Learning Group, Université Libre de Bruxelles, Boulevard du Triomphe CP212, 1050 Brussels, Belgium
& Artificial Intelligence Lab, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

AI Supremacy Race (AISR) model



Each advancement requires **a choice**: play **SAFE** / play **UNSAFE**

SAFE is costly (c) and takes more time than **UNSAFE**

s=1

s>1

Simplifying assumptions:

Each advancement requires the same time

Each advancement generates the same benefit **b**

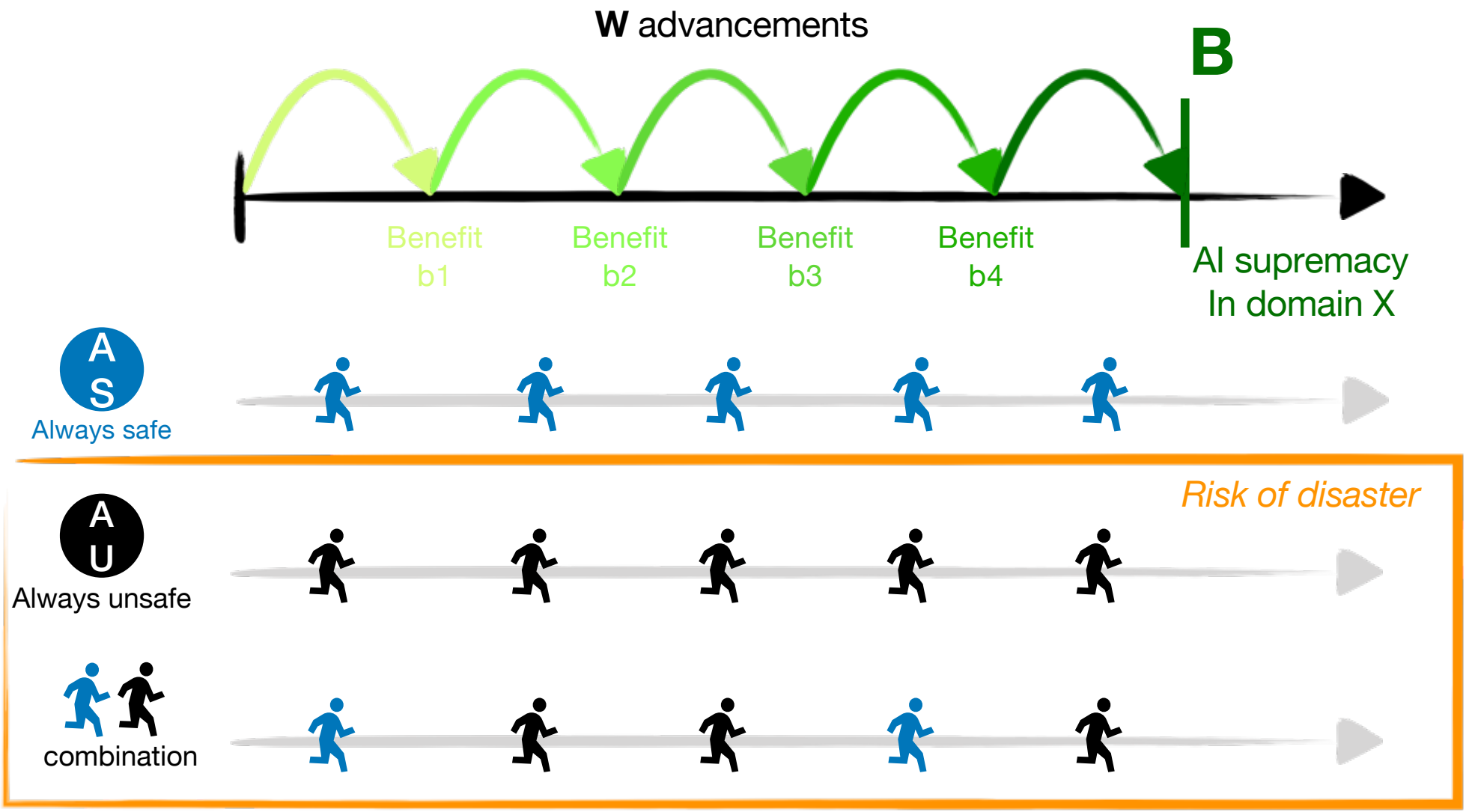
Payoffs in one round

	 SAFE	 UNSAFE
 SAFE	-c + b/2 	-c + b/(s+1) 
 UNSAFE	bs/(s+1) 	b/2 

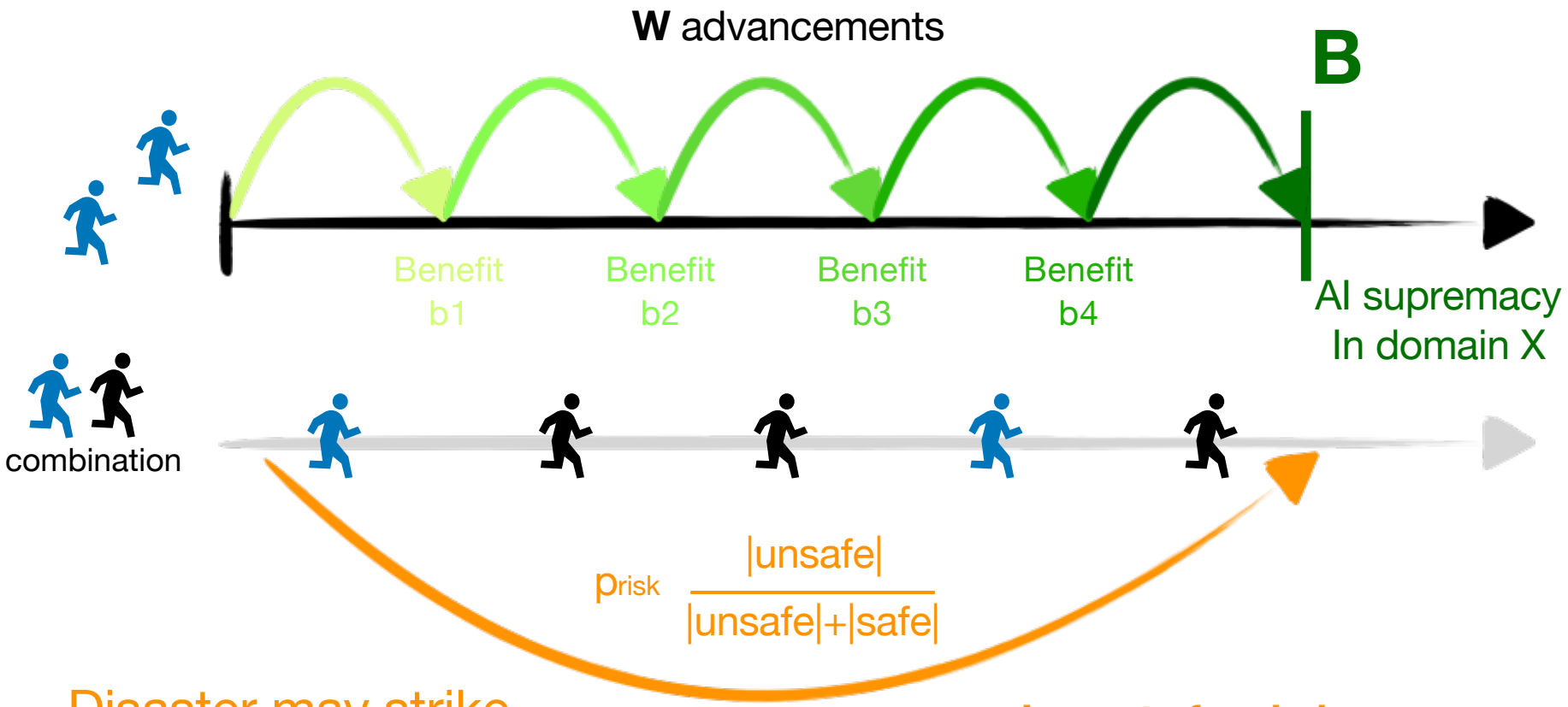
We have also expanded this to N>2 players

Assume simultaneous moves for each advancement

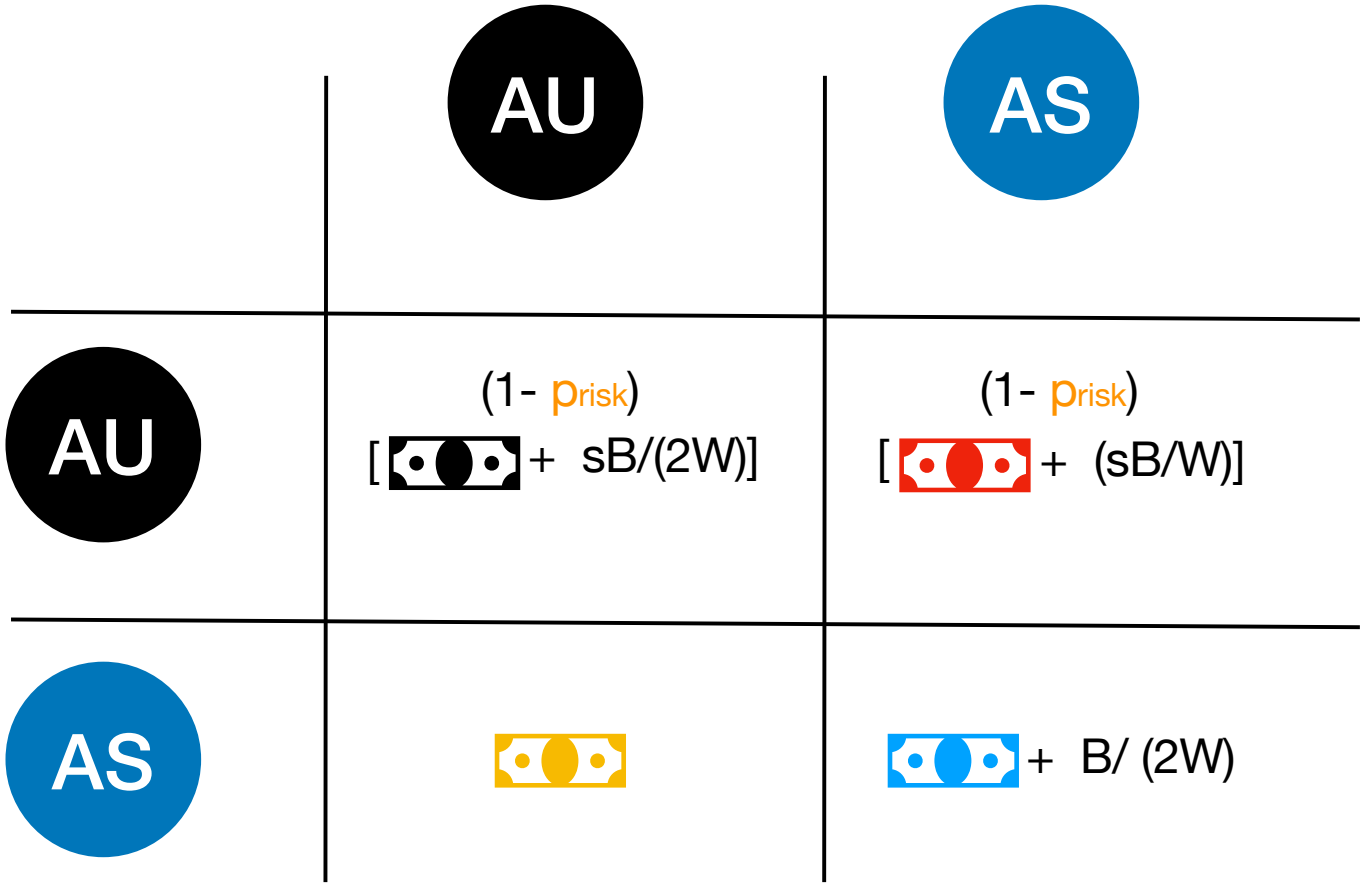
Strategies in AISR



Strategies in AISR

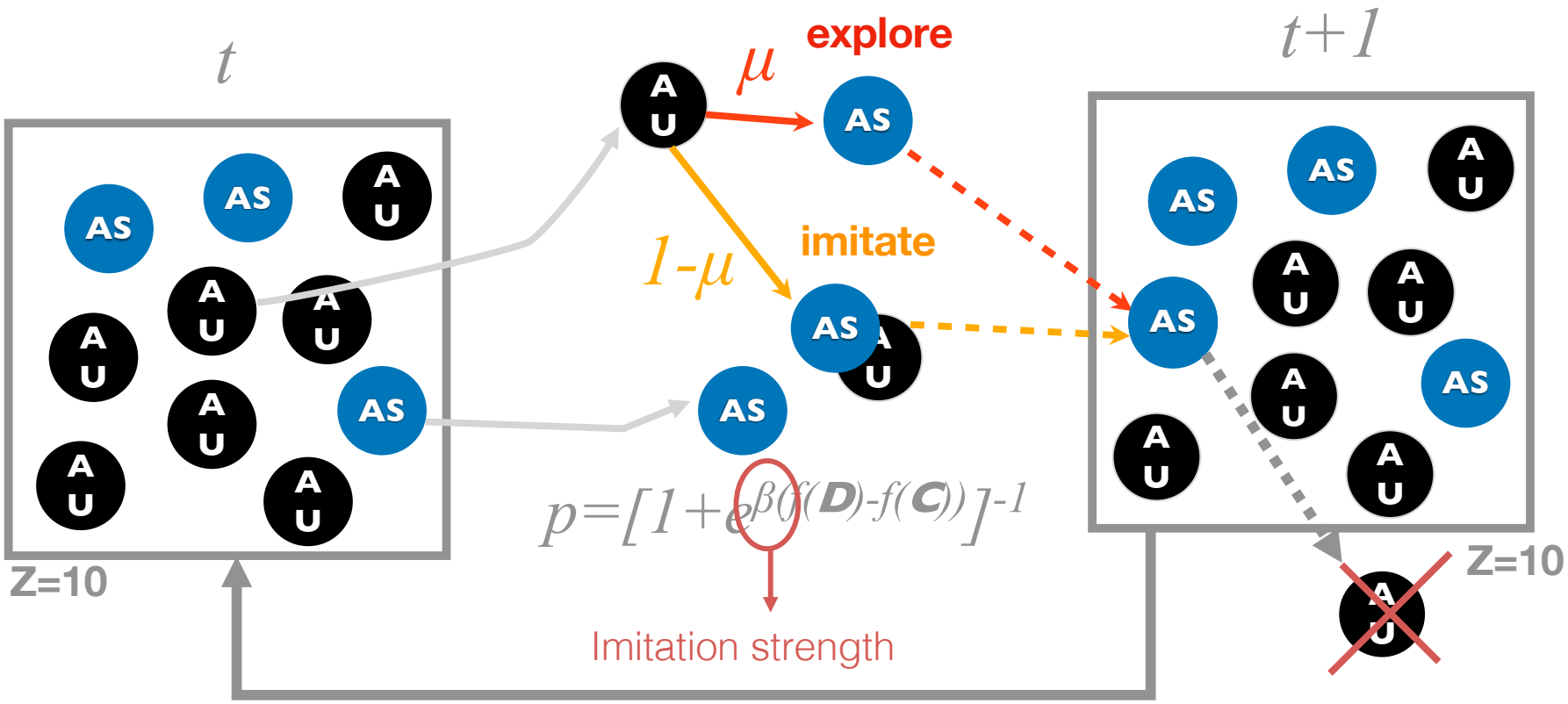


Average Gains
(fitness / social success)

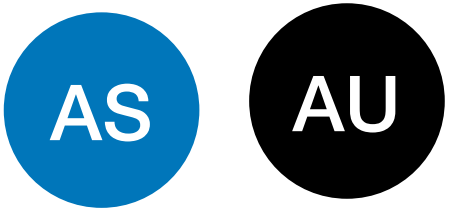


Evolutionary Game Theory

Defining the social learning process

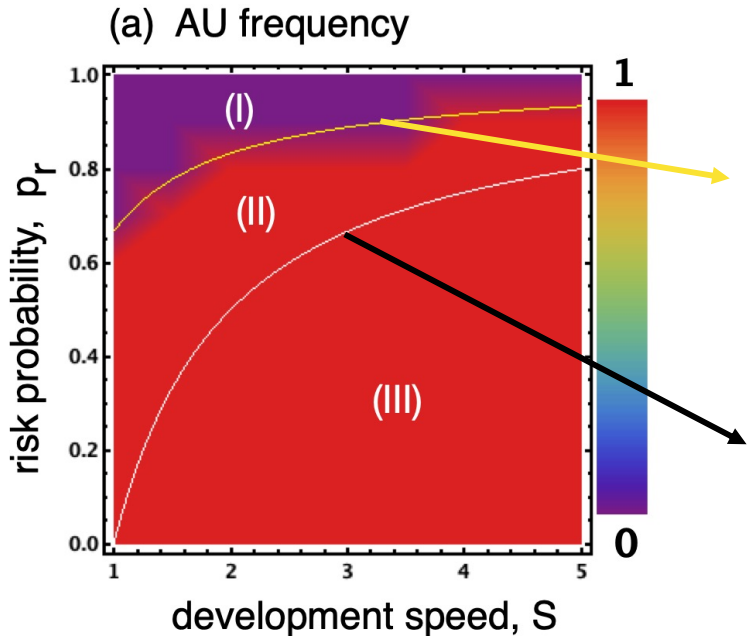


Short-term AI (small W)



Q1: What behaviour dominates?

Q2: Which parameters influence dominance?



Below this yellow line, AU is preferred to AS

pr < 1-1/(3s)

Below this white line, all AU is more beneficial than all AS

pr < 1-1/s

*c=1, b=4, **W=100**, B=10000, Beta=0.1, Z=100*

Short-term AI requires regulation of unsafe development

AU AS

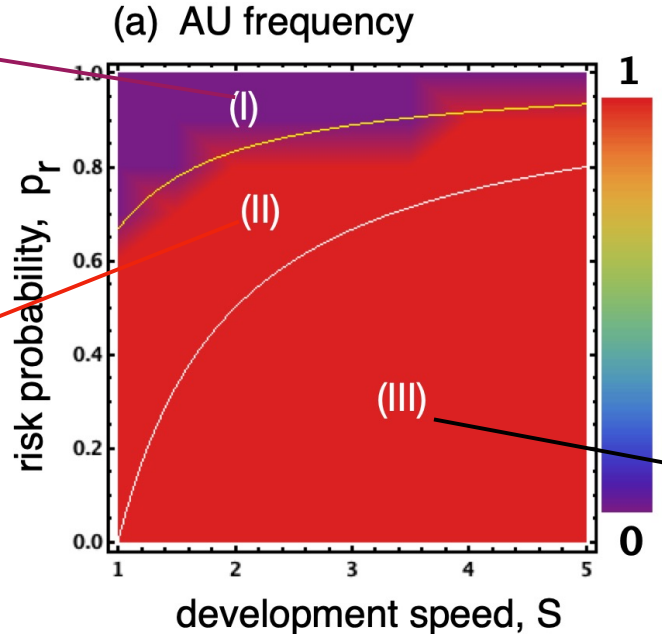
Q3: When are regulatory actions required? When not?

Compliance Zone

Promoting safe has no effect

Dilemma Zone

Promoting safe improves social welfare

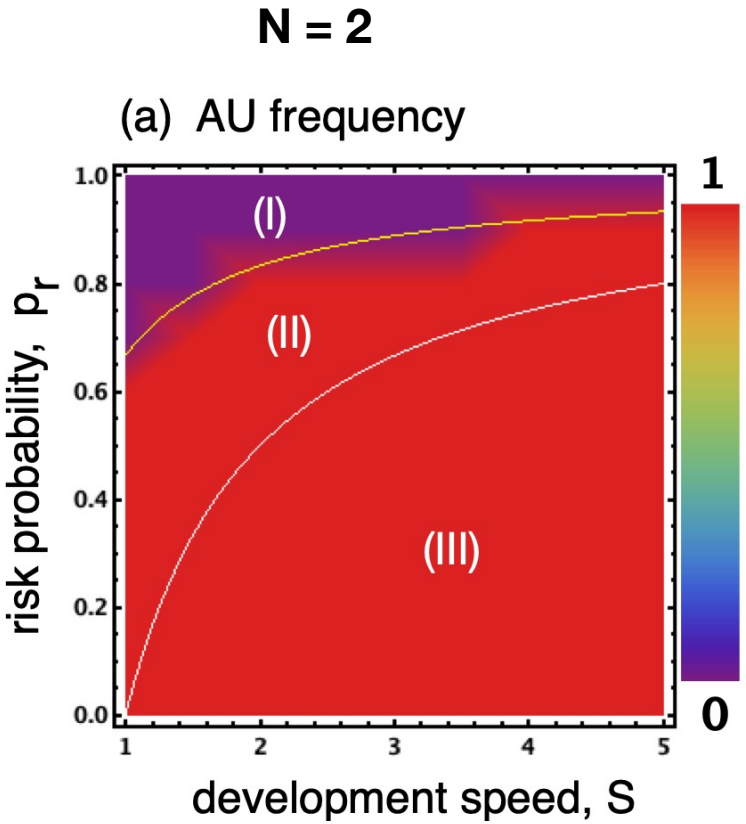


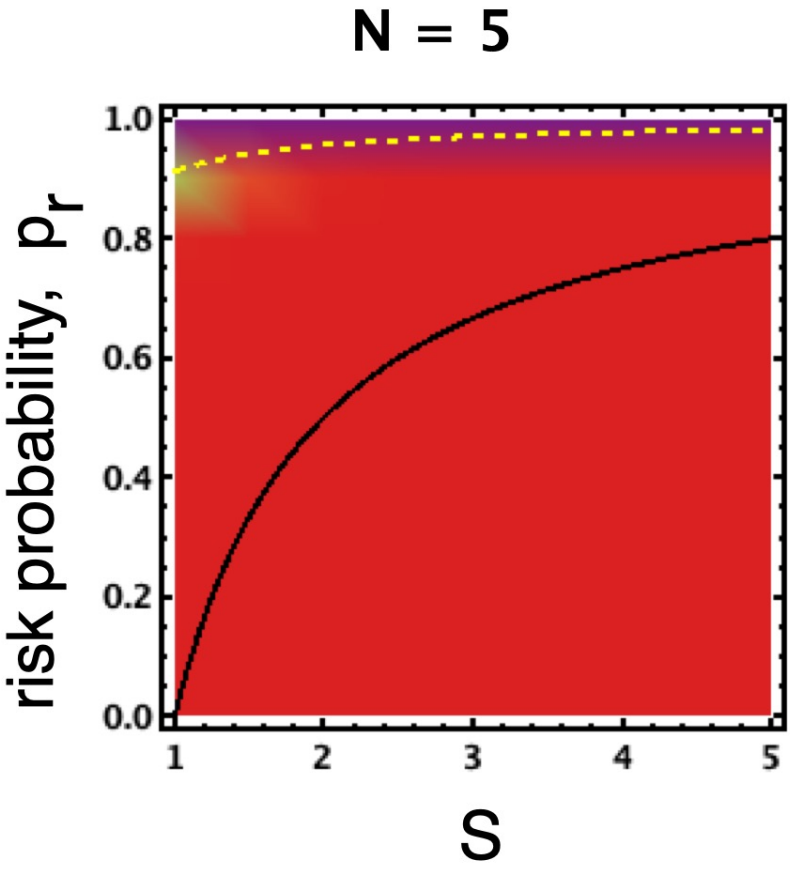
Innovation zone

Promoting safe is detrimental

*c=1, b=4, **W=100**, B=10000, Beta=0.1, Z=100*

Bigger groups lead to larger dilemma zones

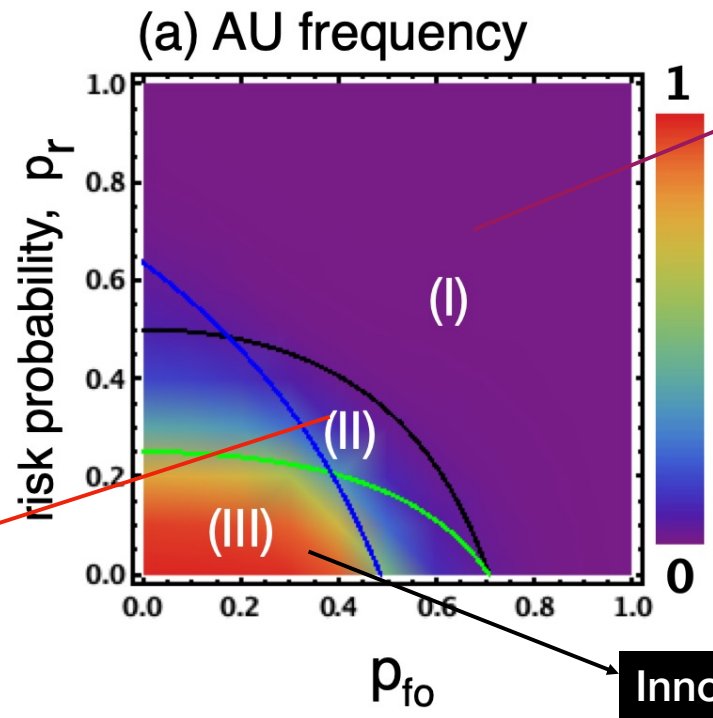




c=1, b=4, W=100, B=10000, Beta=0.1, Z=100

Long-term AI (large W)

Long-term AI requires promotion of risk-taking!



Compliance Zone

Promoting safe has no effect

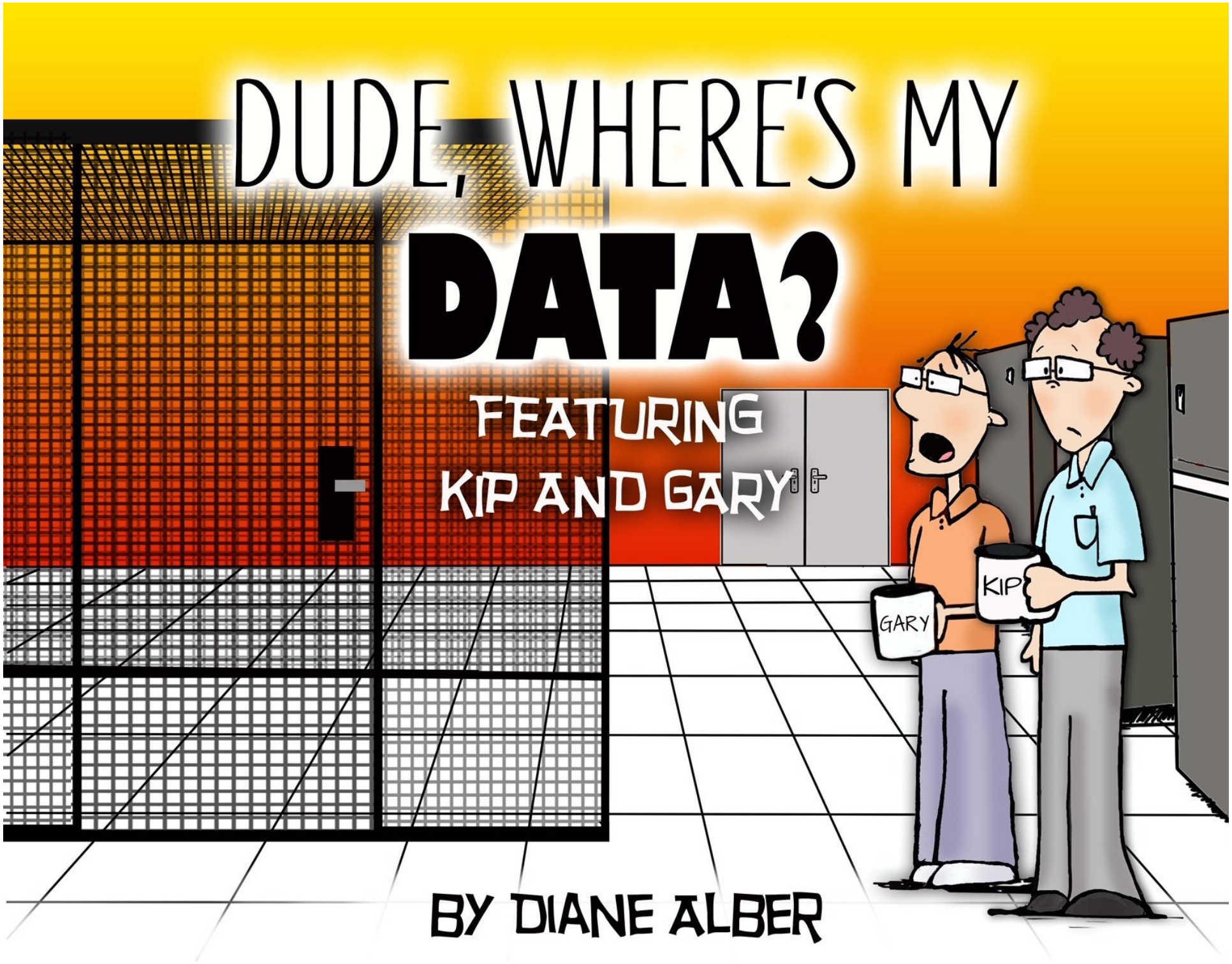
Promoting safe Is detrimental

Dilemma Zone

Innovation zone

Promoting safe Is detrimental

W = 100000, *c=1, b=4, s=1.5, B=10000, Beta=0.1, Z=100*


DUDE, WHERE'S MY
DATA?
FEATURING
KIP AND GARY
GARY
KIP
BY DIANE ALBER

Question 2: What do we know about the impact of regulatory actions on such race dynamics?

Question 3: Is regulation always useful?

PLOS ONE

Citation: Han TA, Pereira LM, Lenaerts T, Santos FC (2021) Mediating artificial intelligence developments through negative and positive incentives. PLoS ONE 16(1): e0244592. https://doi.org/10.1371/journal.pone.0244592

RESEARCH ARTICLE

Mediating artificial intelligence developments through negative and positive incentives

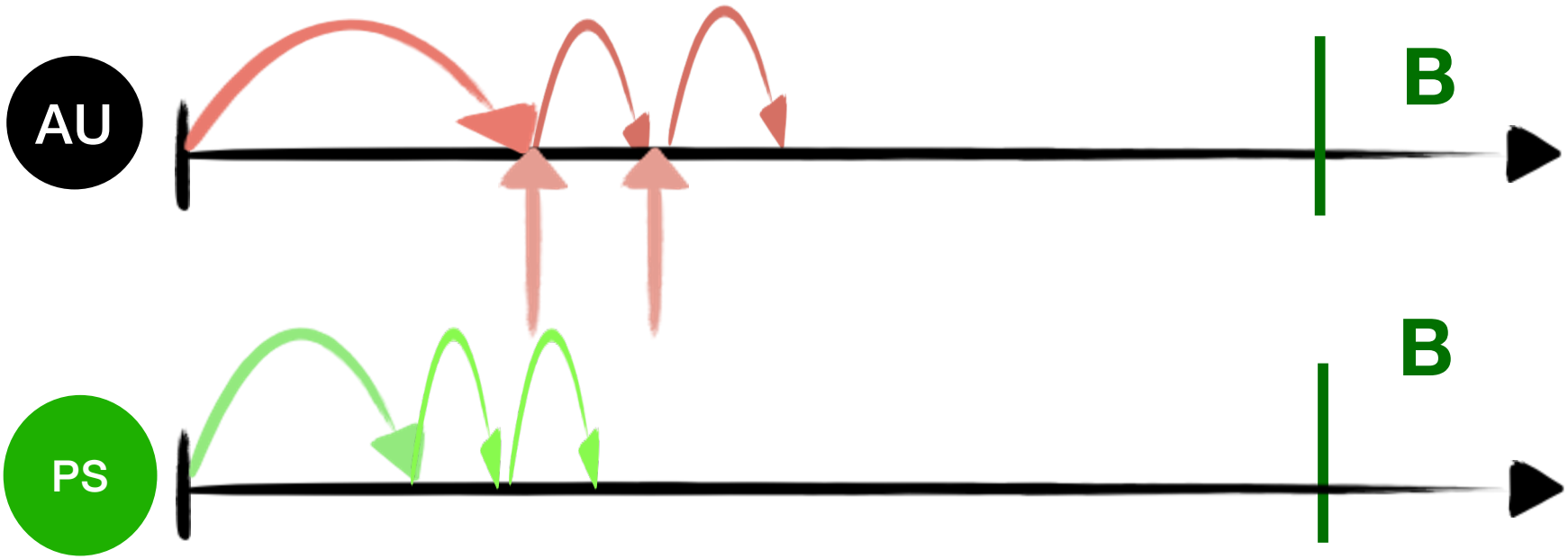
The Anh Han[1]*, Luís Moniz Pereira[2], Tom Lenaerts[3,4], Francisco C. Santos[5]

Abstract

The field of Artificial Intelligence (AI) is going through a period of great expectations, introducing a certain level of anxiety in research, business and also policy. This anxiety is further energised by an AI race narrative that makes people believe they might be missing out. Whether real or not, a belief in this narrative may be detrimental as some stake-holders will feel obliged to cut corners on safety precautions, or ignore societal consequences just to "win". Starting from a baseline model that describes a broad class of technology races where winners draw a significant benefit compared to others (such as AI advances, patent race, pharmaceutical technologies), we investigate here how positive (rewards) and negative (punishments) incentives may beneficially influence the outcomes. We uncover conditions in which punishment is either capable of reducing the development speed of unsafe participants or has the capacity to reduce innovation through over-regulation. Alternatively, we show that, in several scenarios, rewarding those that follow safety measures may increase the development speed while ensuring safe choices. Moreover, in the latter regimes, rewards do not suffer from the issue of over-regulation as is the case for punishment. Overall, our findings provide valuable insights into the nature and kinds of regulatory actions most suitable to improve safety compliance in the contexts of both smooth and sudden technological shifts.

Negative Incentives in AISR

Peer Punishment (PS): reduces speed of an UNSAFE opponent, whereas PS's speed is reduced



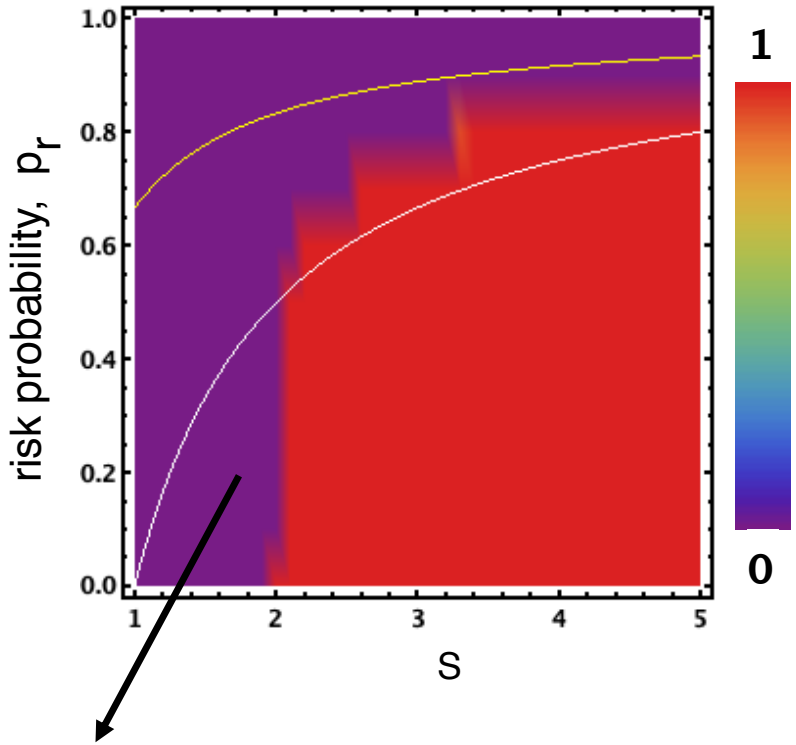
Institutional Punishment: an institution, not part of the system, reduces speed of UNSAFE players

Always regulate unsafe players leads to over-regulation

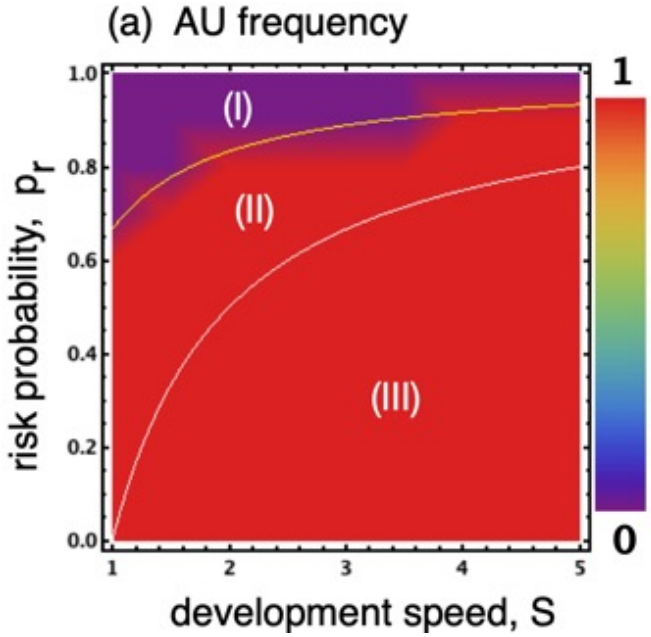
AU

AS

Effect of incentive
s' = 1



Over-regulation



Short-term AI

*c=1, b=4, **W=100**, B=10000, Beta=0.1, Z=100*

Conclusions

We describe a plausible model that can be useful when thinking about AI governance policies and regulations

Time-scale to reach AI supremacy strongly influences regulation

Incentives can improve safety outcome in the dilemma zones but can lead to over-regulation

Han et al. *To Regulate or Not: A Social Dynamics Analysis of an Idealised AI Race.* Journal of Artificial Intelligence Research, 69: 881-921, 2020

Han et al. *Mediating Artificial Intelligence development through positive and negative incentives.* PLoS ONE 16(1): e0244592, 2021.

Talk and poster presentation - Journal track @IJCAI

#J21 Time-scale Differences will Influence the Regulation Required in an Idealised AI Race Game

The Anh Han (Teesside University), Luis Moniz Pereira (Universidade Nova de Lisboa), Francisco Santos (INES

Lenaerts (Université Libre de Bruxelles)

Video streams:

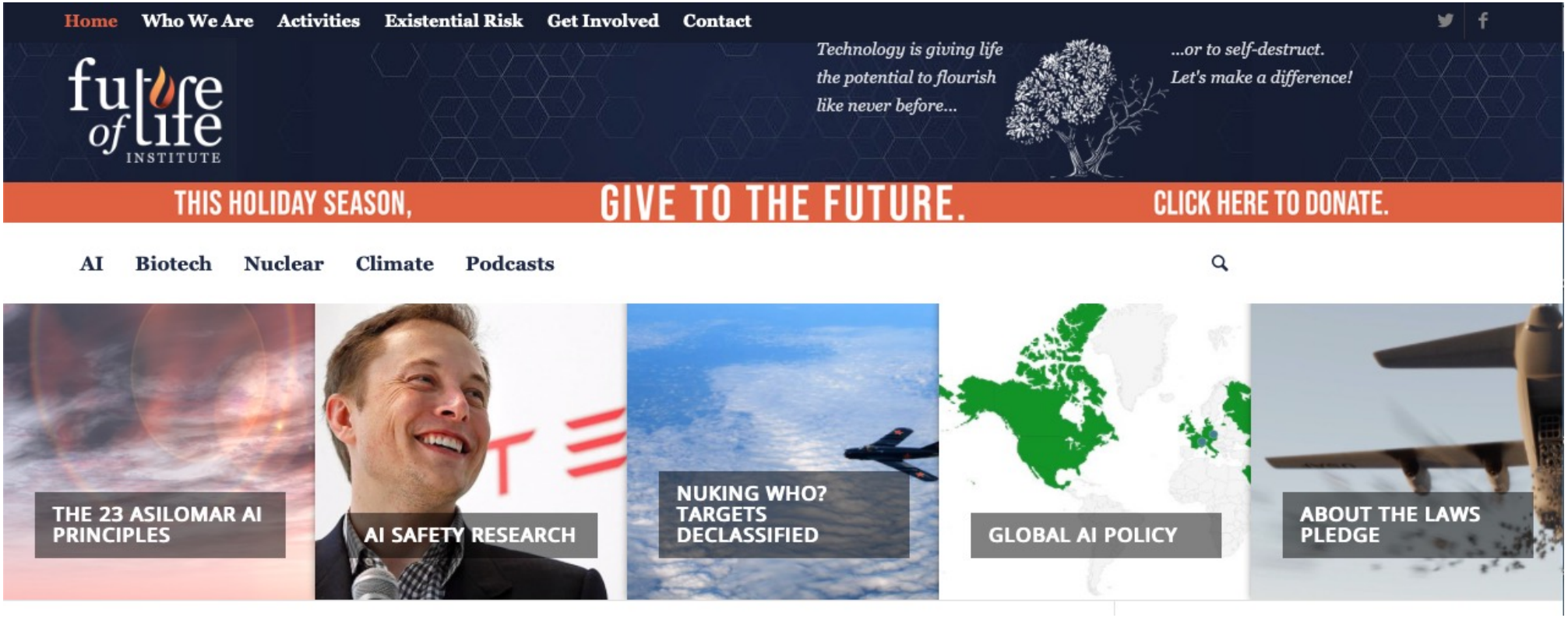
Aug 23th at 13:30 EDT – Room Red 1 **and**

Aug 26th at 10:00 EDT – Room Green 2

Posters:

Aug 23th at 14:00 EDT – Room Red **and**

Aug 26th at 10:30 EDT – Room Green



“Incentives for Safety Agreement Compliance in AI Race ” (2019-2021)

http://futureoflife.org/2018-ai-grant-recipients#Han