Technology in Society 68 (2022) 101843

Contents lists available at ScienceDirect

Technology in Society

journal homepage: www.elsevier.com/locate/techsoc

Voluntary safety commitments provide an escape from over-regulation in AI development

The Anh Han[a,*], Tom Lenaerts[b,c,d,e], Francisco C. Santos[f], Luís Moniz Pereira[g]

[a] School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, TS1 3BA, UK
[b] FARI Institute, Université Libre de Bruxelles-Vrije Universiteit Brussel, 1050, Brussels, Belgium
[c] Machine Learning Group, Université Libre de Bruxelles, Boulevard du Triomphe CP212, 1050, Brussels, Belgium
[d] Artificial Intelligence Lab, Vrije Universiteit Brussel, Pleinlaan 2, 1050, Brussels, Belgium
[e] Center for Human-Compatible AI, UC Berkeley, 2121 Berkeley Way, 94702 Berkeley, CA, USA
[f] INESC-ID and Instituto Superior Tecnico, Universidade de Lisboa, Portugal
[g] NOVA Laboratory for Computer Science and Informatics (NOVA LINCS), Universidade Nova de Lisboa, 2829- 516, Caparica, Portugal

ARTICLE INFO

Keywords:
Evolutionary game theory
AI development Race
Commitments
Incentives
Safety

ABSTRACT

With the introduction of Artificial Intelligence (AI) and related technologies in our daily lives, fear and anxiety about their misuse as well as their inherent biases, incorporated during their creation, have led to a demand for governance and associated regulation. Yet regulating an innovation process that is not well understood may stifle this process and reduce benefits that society may gain from the generated technology, even under the best intentions. Instruments to shed light on such processes are thus needed as they can ensure that imposed policies achieve the ambitions for which they were designed. Starting from a game-theoretical model that captures the fundamental dynamics of a race for domain supremacy using AI technology, we show how socially unwanted outcomes may be produced when sanctioning is applied unconditionally to risk-taking, i.e. potentially unsafe, behaviours. We demonstrate here the potential of a regulatory approach that combines a voluntary commitment approach reminiscent of soft law, wherein technologists have the freedom of choice between independently pursuing their course of actions or establishing binding agreements to act safely, with either a peer or governmental sanctioning system of those that do not abide by what they pledged. As commitments are binding and sanctioned, they go beyond the classic view of soft law, akin more closely to actual law-enforced regulation. Overall, this work reveals how voluntary but sanctionable commitments generate socially beneficial outcomes in all scenarios envisageable in a short-term race towards domain supremacy through AI technology. These results provide an original dynamic systems perspective of the governance potential of enforceable soft law techniques or co-regulatory mechanisms, showing how they may impact the ambitions of developers in the context of the AI-based applications.

1. Introduction

With the rapid advancement of AI and related technologies, there has been significant fear and anxiety about their potential misuse as well as the social and ethical consequences that may result from biases within the design of such systems [1–4]. While expectations associated with these advanced technologies increase and monetary profits stimulate rapid deployment, there is a serious risk for taking unethical or risky short cuts to enter a market first with the next innovation, ignoring safety checks and ethical development procedures. As different disagreeable examples have emerged [5], governments and regulating bodies have been catching up by debating new forms and frameworks for regulating this technology [6–9]; notably, the recent EU draft proposal for AI regulation based on identifiable risks [10,11]. Such debates have produced proposals for mechanisms on how to avoid, mediate, or regulate the development and deployment of AI [6,7,9,12–18]. Essentially, regulatory measures such as restrictions and incentives are proposed to limit harmful and risky practices in order to promote beneficial designs [6]. Examples of such approaches [6] include financially supporting the research into beneficial AI [19] and making AI companies pay fines when found liable for the consequences of harmful AI [20].

Although such regulatory measures may provide solutions for

[*] Corresponding author.
E-mail address: t.han@tees.ac.uk (T.A. Han).

https://doi.org/10.1016/j.techsoc.2021.101843
Received 14 September 2021; Received in revised form 28 November 2021; Accepted 21 December 2021
Available online 28 December 2021

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 [21,22]. Worries have been expressed by different organisations/societies that too strict policies may unnecessarily affect the benefits and societal advances that novel AI technologies may have to offer [23]. 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 [22]. It has been emphasised that neither over-regulation nor a laissez-faire approach suffices when aiming to regulate AI technologies [24]. 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.

As an intermediate step in regulating novel technologies, less formal and more flexible tools, referred to as soft law, are frequently introduced [25,26], most often by the stake-holders themselves. Soft law involves a series of instruments that are often voluntary in nature and not directly enforceable by governments (which contrasts it from hard or traditional law) [27,28]. Examples are professional guidelines, private standards, codes of conduct, and best practices. Yet, to ensure that these soft law measures are followed, supporting mechanisms such as indirect enforcement by entities like insurance companies, journal publishers, grant funding agencies, and governmental enforcement programs against unfair or deceptive business practices, need to be put in place [28]. For technology with a global impact, it remains to be seen if such simple incentives to follow a set of rules of conduct are sufficient. A globally enforceable system [27] involving also the needs of society, may be required here: just those defined in case of well-defined international/national waters or borders; specific treaties on fishing; war rules of engagement; types of weapon development and deployment; airplanes' flying aptness, etc.

Starting from a baseline game theoretical model, referred to as the DSAIR model [29] (a model to examine the dynamics of a domain supremacy race), which defines the process through which multiple stake-holders aim for market supremacy, we demonstrate first that unconditional sanctioning will negatively influence social welfare in certain conditions of a short-term race towards domain supremacy through AI technology. Afterwards, we examine an alternative mechanism for resolving the issue, leading to less detrimental effects than unconditional sanctioning. Our approach is to allow technologists or race participants to voluntarily commit themselves to safe innovation procedures, signaling to others their intentions. 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. The latter differentiates our approach from soft-law [26,28]; voluntary participating in the safety commitment also implies accepting the possibility to face hard repercussions when violating the commitment. As will be shown in the results, a soft-law, voluntary approach, needs to be combined with a form of enforcement, either peer or institutional, to be effective. Enforcement against involuntary actions may be detrimental to an innovation process like the AI revolution we are experiencing at the moment.

Previous Developments. As was shown in Ref. [29] participants either follow safety precautions (the SAFE option) or ignore them (the UNSAFE option) in each step of the development process of the DSAIR model. The main assumption in the model was that it requires more time and more effort to comply with the precautionary requirements, making the SAFE option not only costlier, but also slower compared to the UNSAFE option. Accordingly, it was assumed that in playing SAFE, participants must pay a cost $c > 0$, whereas playing UNSAFE costs nothing. Furthermore, whenever playing UNSAFE, the development speed differs and is $s > 1$ whereas in playing SAFE the speed is simply normalized to 1. Decisions to act SAFE or UNSAFE in AI development are repeated until one or more teams attain the designated objective, which can be translated into having completed W development steps, on average [29]. As a result, they earn a large benefit or prize B (e.g. windfall profits [16]), equally shared among those reaching the target at the same time. A development disaster or a setback may however come to occur with some probability, which is presumed to increase with the number of times that safety requirements were ignored by the winning team(s) at each step. Whenever a disaster of this kind occurs, all the benefits of a risk-taking participant are lost. This risk probability is denoted by p_r (see the Models and Methods section for more details).

It was observed in the DSAIR model that, in case the time-scale to reach the target is short, so that over the whole of the development process the average of the accumulated benefit is much smaller than the final benefit B, only for a certain window of parameter settings societal interest conflicts with the individual ones: In that region, individual unsafe behaviour dominates, despite that safe development would lead to a larger collective outcome or social welfare (cf. region **II** in Fig. 1). From the regulatory perspective, it is only region **II** that thus requires governance in order to promote or enforce safe actions in order to avoid any disaster that may occur during the technology development race.

A peer punishment mechanism against unsafe behaviour was proposed in order to mediate the behavior in that region (without affecting the desirable safe outcome in region **I**) [30]. It however may lead to a reduction of the societal welfare that can be obtained in region **III** (see Fig. 1) where the desired unsafe (risk-taking/innovative) behaviour becomes significantly reduced whenever punishment is not very costly to the punisher while strongly affecting the punished. The problem with applying system-wide sanctioning for any risk-taking behavior is that it does not take into account the region wherein the AI innovation is taking place, as was visualised by Fig. 1. That is, if it is the case that a development race falls in region **III** (low risk and innovation is collectively beneficial and preferred), unsafe behaviour should not be punished as it is beneficial for the overall social welfare. To enact such a region-dependent targeted punishment scheme, one would require the ability to estimate exactly the risk level associated with each AI development scenario, i.e. knowing beforehand the risk as well as the speed of development. Clearly, that may not be easy due to the lack of data on how such an AI innovation dynamics works. This is especially true in the

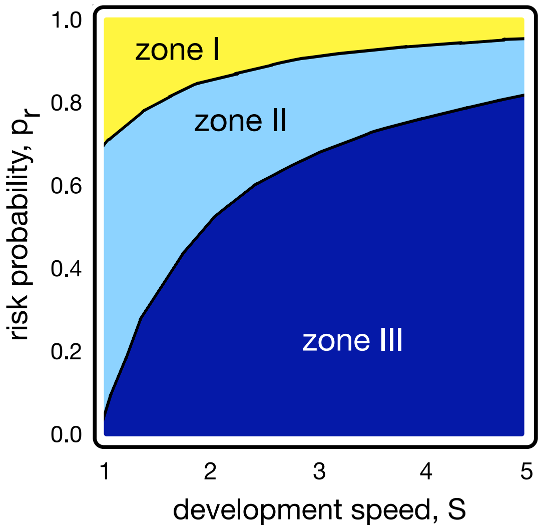

Fig. 1. Behavioural regions (zones) as identified in 29 when the time to reach domain supremacy is short. Region **II**: Inside the plots, the two solid lines delimit the parameters' boundaries wherein the collective prefers safety compliant behavior, yet unsafe development is individually preferred. Regions **III** and **I** exhibit where unsafe (respectively, safe) development is both the individually and collectively preferred outcome. Parameters: $c = 1$, $b = 4$, $W = 100$, $p_{f_0} = 0.5$, $B = 10^4$, $\beta = 0.1$, and $Z = 100$.

early stages of the development and adoption of many new technologies, which has become known as the so-called double-bind problem [31]: the impact of technologies, including AI ones cannot be predicted until it becomes a reality, and controlling or changing it is difficult or no longer possible at that point. Additionally, for all its benefits, the large-scale adoption of AI technologies can lead to enormous and unforeseen threats, as pointed out in Ref. [2].

In the remainder of this manuscript, we will focus on a DSAIR model with two participants. In Ref. [29] the results for more than two participants were also provided, showing that the conclusions remained stable, with the exception that the region **I** disappears with an increasing number of race participants.

Goals. In this article an alternative solution is proposed that circumvents the need of being able to estimate correctly the speed of development and risk in order to appropriately regulate UNSAFE behavior: race participants can choose whether or not to establish a bilateral commitment to act safely, which also exposes them to a sanction in case they do not uphold their commitment. While, on the one had, the development teams are allowed to work in an UNSAFE manner without repercussions if they do not commit; a prior agreement allows, on the other hand, the safety compliant participants to identify easily unsafe ones while having also the capacity to punish the dishonest ones (who might also be sanctioned by an external party such as a regulating institution).

Our results reveal that this freedom of choice, if enabled through a prior bilateral commitment to SAFE actions, can, on the one hand, avoid over-regulating unsafe behaviour and, on the other, improve significantly the desired safety outcome in the dilemma zone when compared to punishment alone. Interestingly, this type of binding pledges has been argued to be of relevance within other types of global conflicts and dilemmas, such as environmental governance [32–34].

2. Models and Methods

We first recall the DSAIR model with pairwise interactions, previously introduced in Ref. [29], whose understanding is required before one can fully grasp the novel concepts and solutions that are proposed here. Concretely, we extend the prior work with the option to bilaterally, but voluntarily, commit to acting safely, plus the associated punishment of violations of such commitments when one is found out. As will be shown, this novelty leads to surprisingly beneficial outcomes in the potential race dynamics that are being studied here. The Evolutionary Game Theory (EGT) methods being used to analyse the novel models are also described in this section.

2.1. Summary of the DSAIR model and prior results

The DSAIR model [29] was originally defined as a two-player game repeated with a certain probability, consisting thus on average of W rounds.[1] At each round of development, players gather benefits arising from their intermediate AI developments, subject to whether or not they chose to act UNSAFE or SAFE. Presuming some fixed benefit, b, resulting from the AI market, the teams will share this gain proportionately to their development speed. Accordingly, at every round of the race one can write a payoff matrix denoted by Π with respect to row players i, whose entries are denoted by Π_{ij} (j corresponding to some column), as shown

$$\Pi = \begin{array}{c} \\ SAFE \\ UNSAFE \end{array} \left(\begin{array}{cc} SAFE & UNSAFE \\ -c+\dfrac{b}{2} & -c+\dfrac{b}{s+1} \\ \dfrac{sb}{s+1} & \dfrac{b}{2} \end{array} \right) \tag{1}$$

The payoff matrix can be explained as follows. Firstly, wherever there is an interaction between two players selecting the SAFE action, each shall pay a cost c and the resulting benefit b is shared. Differently, whenever interaction is between two players selecting the UNSAFE action, they shall share benefit b without having had to pay cost c. Whenever an UNSAFE choice is matched with a SAFE one, the SAFE choice necessitates a cost c and receives a (smaller) part $b/(s+1)$ of b, whereas the UNSAFE choice collects a larger $sb/(s+1)$ whilst not ever having had to pay c. Note that Π is a simplification of the matrix defined in Ref. [29] for, in the current time-scale, it was shown that the parameters as defined here sufficiently explain the obtained results.

We analyse the evolutionary outcomes of this game in a well-mixed and finite population consisting of Z players. Given the choices each player can make and the fact that these choices need to be repeated for W round, each adopts player one of the two following strategies [29]:

- **AS**: complies every time with safety precautions, acting thus SAFE in every round.
- **AU**: complies not once with safety precautions, acting thus UNSAFE in every round.

The averaged payoffs for AS vs AU are expressed by payoff matrix

$$\begin{array}{c} \\ AS \\ AU \end{array} \left(\begin{array}{cc} AS & AU \\ \dfrac{B}{2W}+\Pi_{11} & \Pi_{12} \\ p\left(\dfrac{sB}{W}+\Pi_{21}\right) & p\left(\dfrac{sB}{2W}+\Pi_{22}\right) \end{array} \right), \tag{2}$$

wherein, for presentation purposes alone, let us denote $p = 1 - p_r$ (note that p_r was explained in the Introduction section).

As has been shown in Ref. [29], by contemplating where AU is risk-dominant against AS (cf. Methods below), then three distinct regions are identifiable within the parameter space s-p_r (cf. Fig. 1): (**I**) if $p_r > 1 - \frac{1}{3s}$, AU is risk-dominated by AS: safety compliance affords both the collectively preferred outcome and the one evolution selects; (**II**) if $1 - \frac{1}{3s} > p_r > 1 - \frac{1}{s}$: even though safety compliance is the more desirable strategy for ensuring the highest collective outcome, the social learning dynamics leads the population to the state within which safety precautions have been mostly ignored; (**III**) if $p_r < 1 - \frac{1}{s}$ (AU is risk-dominant against AS), then unsafe development is both collectively preferred and selected by the social learning dynamics.

So it is important to remember for the rest of the paper that UNSAFE actions are preferred and established in zone **III**, SAFE actions are preferred and established in zone **I** and a conflict exists between the individual and the collective in zone **II** as the former prefers UNSAFE actions and the latter SAFE actions.

2.2. Bilateral commitment strategies to ensure beneficial outcomes

We now extend the DSAIR model with strategies that can bilaterally commit to safety compliant behavior: Before an interaction, participants can commit to play SAFE in each round. The commitment stands when all parties agree. The players can refuse to commit, preferring to proceed without being pushed into the safe direction and being able to take risks. Those that committed but later select the UNSAFE action are potentially subject to sanctioning. Two sanctioning scenarios are considered here: (a) peer punishment (PP), which is performed by the co-player who kept

[1] An N-player version of this game was discussed also in Ref. [29]. Yet in order to keep things easy to access, we focus here on the two-player scenario.

her side of the deal, and (b) institutional punishment (IP), which is performed by a third-party that is not actively participating in the race for supremacy in some domain through AI (e.g. the European Union or United Nations). Each player has the freedom of behavioural choice whereby those who do not commit will not be punished when playing UNSAFE in the DSAIR model. We call the latter behavior "honest" unsafe behavior whereas those that act unsafely after committing are referred to as "dishonest".

Sanctioning an opponent who played UNSAFE in a previous round consists in imposing a reduction s_β on the opponent's speed [30]. In case of PP, the punishing player also incurs a reduction s_α on her own speed. Committing may also be costly for all that do as they give up other choices. A commitment cost ε (per round) is thus introduced for those performing this pre-play action.

With the possibility of joining or not a commitment to behave safely and sanctioning dishonest unsafe behaviour, one can now define the possible strategies. AS and AU (as defined above) can either commit to safe actions, and furthermore, when involved in a commitment, decide whether to punish a dishonest co-player. If no commitment can be made, the player will select the UNSAFE action.[2] The choices listed before lead to five strategies:

1. **AS-in**: willing to commit, plays SAFE when commitment is in place and UNSAFE otherwise, but does not use (costly) punishment. This strategy can be considered a second-order free-rider on the punishment effort of others;
2. **AS-out**: does not commit, but always selects the SAFE action;
3. **AU-in**: claiming to commit to the SAFE action, but always plays UNSAFE in the interaction. This strategy makes a commitment, trying to exploit safe players who only want to interact under an agreement;
4. **AU-out**: does not commit and plays UNSAFE in the interaction. This strategy wants to freely take risk or innovate without any repercussions otherwise imposed by a commitment to follow SAFE actions;
5. **PS**: willing to commit and plays SAFE when the other player also commits; plays UNSAFE otherwise, and also punishes an UNSAFE action of a co-player that committed with her. This strategy is only present in the case of PP, and is not present in the IP sanctioning model.

Given the current setting, there are other possible strategies, such as a committed unsafe strategist (AU-in) that punishes other committed unsafe players. These different strategies were omitted for the sake of brevity of exposition since they would be disadvantageous in the presence of the current set of strategies. For example, the aforementioned punishing AU-in would perform worse than the non-punishing AU-in included in the model, since, whereas they would behave equivalently when playing against other strategies, however, within a homogeneous population of their own, the former has a lower payoff than the latter and will be gainsaid by it.

2.3. *Evolutionary dynamics in finite populations*

Herein are adopted the methods of EGT for finite populations [35–37], whether in the analytical or numerical results obtained here. In such settings, the payoffs of players stand for their social *success* or *fitness*, and the evolutionary dynamics is shaped by social learning [38], in accordance to which the players that are most successful will tend more often to be copied by other players. The so-called pairwise rule of comparison is utilised to model social learning [37], which ensures a player A with fitness f_A resorts to adopt the strategy of player B with fitness f_B with a probability established by the Fermi function, $P_{A,B} = \left(1 + e^{-\beta(f_B - f_A)}\right)^{-1}$, where the intensity of selection is conveniently described by β. In a population wherein several strategies are in co-presence, their long-term frequency can be computed simply by calculating the stationary distribution of a Markov chain the states of which represent each strategy. Absent behavioural exploration or mutations, the end states of evolution are inevitably monomorphic. Meaning whenever such a state is reached, escape by imitation is impossible. Hence, we presume further that, given some mutation probability, each agent may freely explore its behavioural space (which consists of the two actions, UNSAFE and SAFE, in our case), by randomly adopting an action as a result of mutation. At the limit of a small probability of mutating, the population is comprised of at most one of two strategies at any time. Accordingly, the social dynamics is describable utilising a Markov Chain, in which each state represents a monomorphic population whose transition probabilities to another state are expressed by the fixation probability of one single mutant [39,40]. The Markov Chain's stationary distribution depicts the average time the whole population spends at each monomorphic end state (see the examples in Fig. 3 for illustration).

Let $\pi_{X,Y}$ denote the payoff some strategist X gathers from a pairwise interaction with some strategist Y (as defined in the payoff matrix). Assume there exist two strategies at most in the population, for example k agents using strategy A ($0 \le k \le Z$) and ($Z - k$) agents using instead strategy B. Hence, the (average) payoff of agents using A and B can be formulated, respectively, as

$$\begin{aligned} \Pi_A(k) &= \frac{(k-1)\pi_{A,A} + (Z-k)\pi_{A,B}}{Z-1}, \\ \Pi_B(k) &= \frac{k\pi_{B,A} + (Z-k-1)\pi_{B,B}}{Z-1}. \end{aligned} \tag{3}$$

As a result, at each step in time, the probability of changing by ±1 of the number of k agents using strategy A is specified as [37]

$$T^{\pm}(k) = \frac{Z-k}{Z}\frac{k}{Z}\left[1 + e^{\mp\beta[\Pi_A(k) - \Pi_B(k)]}\right]^{-1}. \tag{4}$$

The fixation probability of a single mutant adopting strategy A, in a population ($Z - 1$) of agents adopting B, is defined by [37,40]

$$\rho_{B,A} = \left(1 + \sum\nolimits_{i=1}^{Z-1} \prod\nolimits_{j=1}^{i} \frac{T^-(j)}{T^+(j)}\right)^{-1}. \tag{5}$$

In the limit of neutral selection (i.e. $\beta = 0$), $\rho_{B,A}$ equals the inverse of the population size, $\rho_N = 1/N$. When considering a set $\{1, \ldots, s\}$ of different strategies, such probabilities of fixation define the Markov Chain transition matrix $M = \{T_{ij}\}_{i,j=1}^{s}$, with $T_{ij,j\neq i} = \rho_{ji}/(s-1)$ and $T_{ii} = 1 - \sum_{j=1,j\neq i}^{s} T_{ij}$. The normalized eigenvector of the transposed matrix of M associated with eigenvalue 1 produces the above defined stationary distribution [39], depicting the relative time the population stays adopting each of the strategies.

A major standpoint of comparison of two strategies A and B is in which direction the transition is more probable or stronger, the one of some B mutant fixating in a population of agents that employ A, $\rho_{A,B}$, or that of a mutant A fixating in the population of agents that employ B, $\rho_{B,A}$. At the limit, for a large enough population size (i.e. a large Z), the condition simplifies to [36]

$$\pi_{A,A} + \pi_{A,B} > \pi_{B,A} + \pi_{B,B}. \tag{6}$$

3. Results

We first focus on analysing the system of self-enforcing bilateral commitment (PP) then showing the results of bilateral commitment complemented with institutional enforcement (IP).

[2] We also consider the version where these players unconditionally play SAFE regardless of any commitment. The safety outcomes in all regions are similar, just the strategies' dominance is slightly different. Results are provided in Appendix (see Figures A3 and A4).

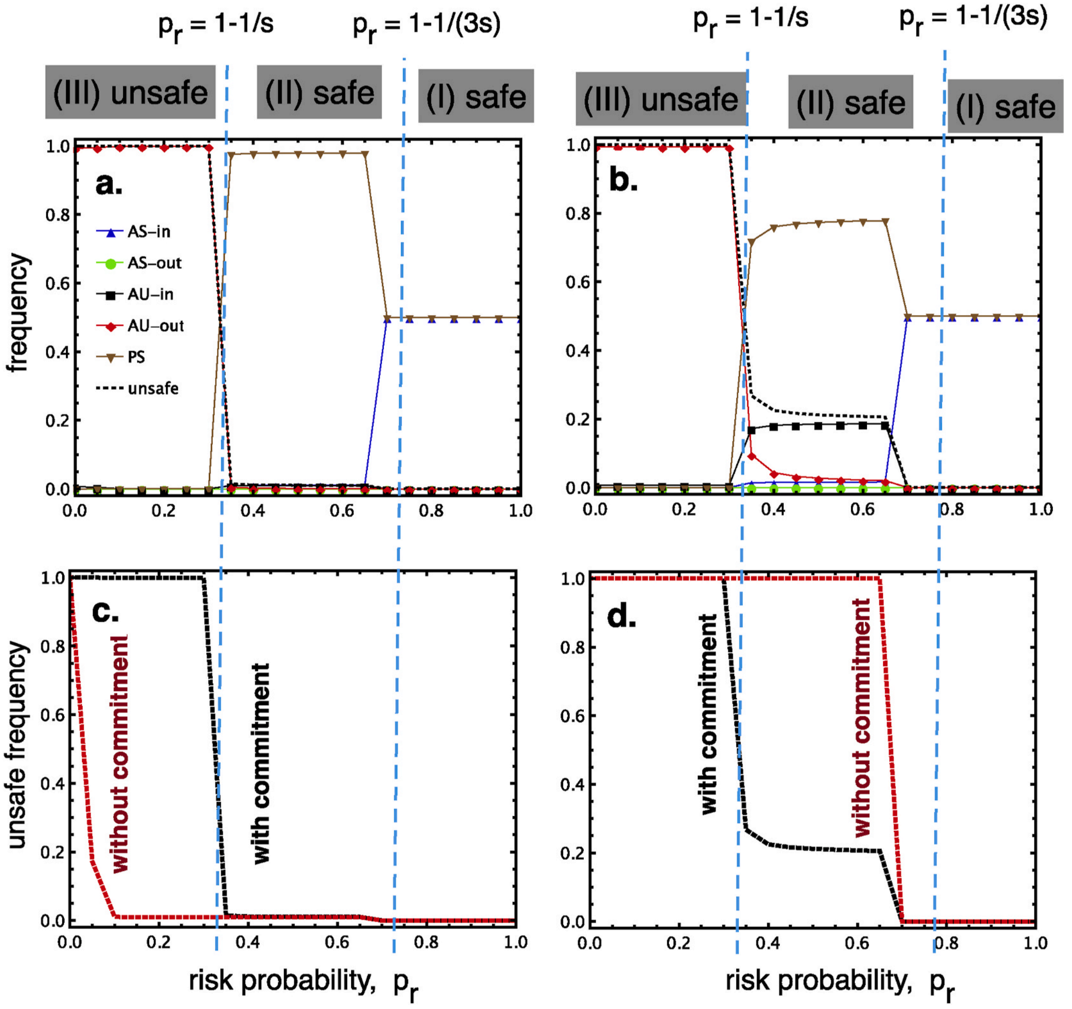

Fig. 2. Behavioural dominance in different zones for varying p_r, in presence of prior commitments (top row: panels a, b) and comparison of its overall unsafe behaviour against when there are no commitments (bottom row: panels c, d). The black dotted lines in panels a and b indicate the total unsafe frequency (i.e. the sum of AU-in and AU-out frequencies). The desired collective behaviour is indicated for each zone (i.e. unsafe in zone **III** and safe in zones **I** and **III**). We show results for two important scenarios: when efficient punishment can be made for a small cost (*left column:* $s_\alpha = 0.3$, $s_\beta = 1$) and when punishment is not highly efficient (*right column:*, $s_\alpha = 1$, $s_\beta = 1$). Parameters: $b = 4$, $c = 1$, $s = 1.5$, $W = 100$, $B = 10^4$, $\beta = 1$, $Z = 100$.

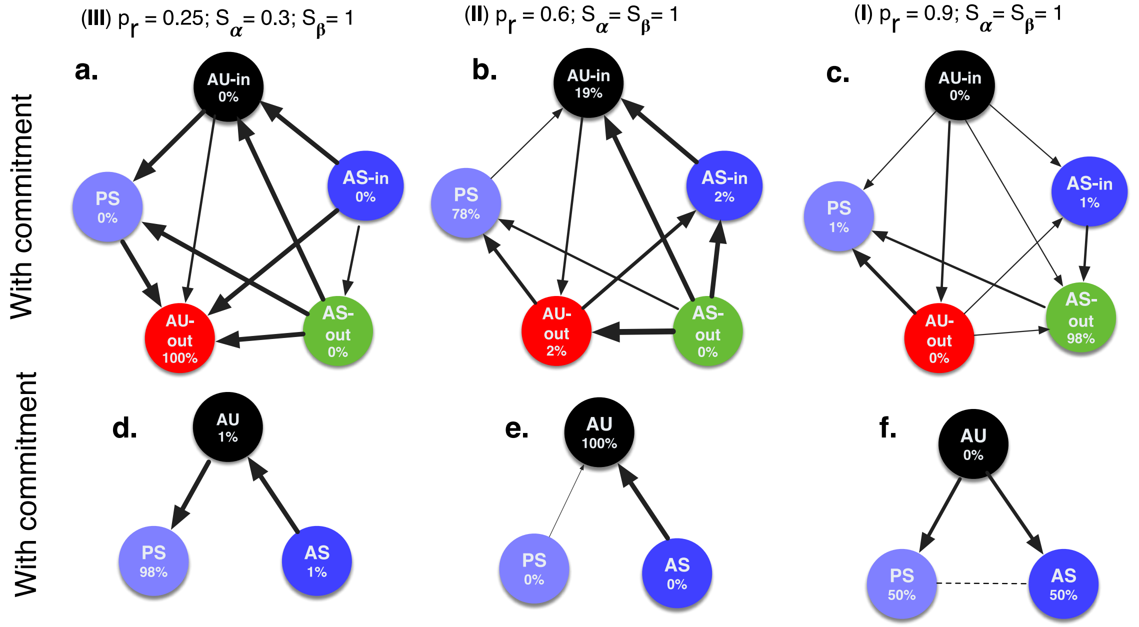

Fig. 3. Transitions and stationary distributions when agreement is present (top row) against when it is absent (bottom row), for three regions. For clarity, only stronger transitions (than the ones in the opposite directions) are shown; no transition either way means neutral. The probability of each transition can be also assessed by the numerical values next to each arrow, indicated using the fixation of a neutral mutant ($\rho_N = 1/Z$) as a reference. The choice of s_α and s_β values were chosen to illustrate the main difference between with vs without commitment scenarios, in the three zones. Parameters: $b = 4$, $c = 1$, $s = 1.5$, $W = 100$, $B = 10^4$, $\beta = 1$, $Z = 100$.

3.1. Self-enforcing bilateral commitments

In Fig. 2 (top row), we show the stationary distributions of the five strategies for varying p_r across the three zones **I**, **II** and **III**. We show results for two important scenarios: when efficient punishment can be made for a small cost (*left column:* $s_\alpha = 0.3$, $s_\beta = 1$) and when punishment is not highly efficient (*right column:*, $s_\alpha = 1$, $s_\beta = 1$). Punishment is considered efficient when its effect (s_α) on the development speed of the player performing the punishment is significantly smaller than the effect (s_β) it has on the player undergoing the punishment.

In both cases, AU-out dominates when p_r is small (zone **III**), PS dominates when it is intermediate (first part of zone **II**), while AS-out dominates when it is large (part of zone **II** and zone **I**). It is important to notice that these results reflect the most desirable outcomes for all three zones: risky innovation in **III** and safety compliance in **I** and **II** (see the black line in the bottom row of Fig. 2). Moreover, a more efficient punishment leads to better safety outcome in zone **II** (see panels c and d). It is worth noting that inefficient punishment may result in the presence of AU-in strategists, in other words cheaters. It is thus important in this self-organised commitment solution that participants have

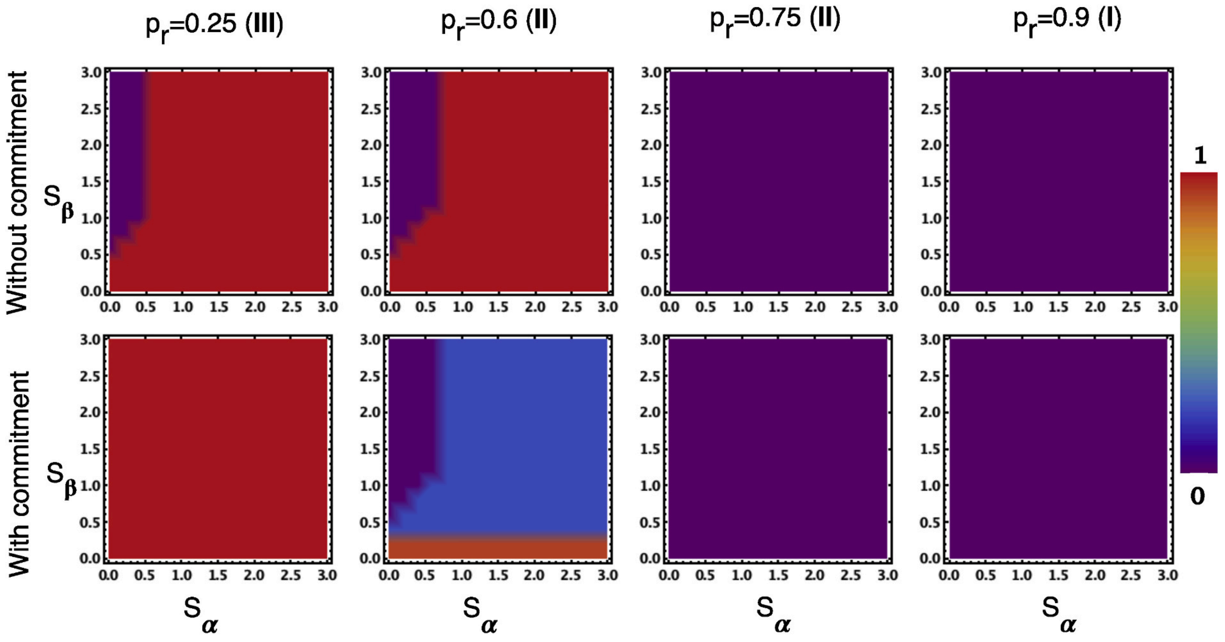

Fig. 4. Frequency of unsafe in all regions for varying s_α and s_β, when commitment is absent (top row) vs when it is in use (bottom row). Parameters: $b = 4$, $c = 1$, $s = 1.5$, $W = 100$, $B = 10^4$, $\beta = 1$, $Z = 100$.

an effective mechanism to punish commitment violators. One solution could be directly affecting the player's publicly available reputation.

To clarify the benefit of voluntary bilateral commitment, Fig. 2 (bottom row) compares the overall frequency of unsafe/risk-taking when commitment is possible (black line) vs when it is absent (red line). We note that the system where commitment is absent can be interpreted as one where commitment is not voluntary. That is, it is implicitly assumed that every participant in the race agreed to play SAFE, without a choice, and they would be sanctioned if they play UNSAFE in the interaction. In the absence of bilateral commitments, over-regulation occurs, i.e. safety compliance is abundant but not desired for a large part of zone **III** when efficient punishment can be carried out for a small cost (panel c), while a safety dilemma occurs (i.e. safety compliance is desired but infrequent) for a large part of zone **II** when sanctioning is not highly efficient.

These observations can be better understood by examining the transitions and stationary distributions in Fig. 3. In region **III** (first column, low risk, where risky innovation is desired), low-cost highly-efficient punishment can lead to significant reduction of innovation since PS dominates AU in absence of bilateral commitments (panel d). Addition of AU-out provides an escape as this strategy is not punished since it never commits to safe course of actions, thereby dominating PS (panel a). In region **II** (dilemma zone, safety is wanted but needs to be enforced), when punishment is not highly efficient, there is a large amount of unwanted AU in the population (panel e). As AU-out is dominated by PS, since PS plays also UNSAFE when not having a commitment partner, it leads to a lower frequency of unsafe behaviour (comparing panels b and e). In region **III**, addition AU-out does not change the desired outcome of the dominance of safety compliance.

In Fig. 4, we show that these remarkable observations regarding voluntary bilateral commitment are robust for different regimes of punishment effectiveness. In particular, when comparing the unsafe frequency with commitments against when it is absent for varying s_α and s_β. In region **III**, over-regulation occurs when there is only punishment whenever s_α is sufficiently small and s_β is large (purple area, top row, first column), but when an agreement is in place that is not the case (bottom row, first column). In region **II** with the lower range of p_r (second column, $p_r = 0.6$), unsafe frequency is lower in the latter in for most s_α and s_β. In the higher range of region **II** and region **III** (third and fourth columns, with $p_r = 0.75$ and $p_r = 0.9$), desired safe behaviour is dominant in both cases.

It is noteworthy that all results are robust also for other regimes of weaker or stronger selection intensities, i.e. $\beta = 0.1$ or $\beta = 10$ (see e.g. Figs. A1 and A2 in Appendix).

3.2. *Institutionally governed bilateral commitments*

When assuming that it is an institution that governs the dynamics of the race to supremacy through AI in some domain, the punishment of unsafe behaviour violating a bilateral commitment is carried out by the institution instead of a peer punisher. Similarly to PP, IP reduces the development speed of a participant that selected the UNSAFE action by s_β.

On the one hand, when no bilateral commitments can be made, the population consists of two strategies, AS and AU, where AU's speed will be $s - s_\beta$ when being punished (by the institution). On the other hand, when voluntary commitments are possible, there will only be the four strategies, AS-in, AU-in, AS-out, and AU-out. There is no PS strategy as there is no peer sanctioning, which implies also no parameter s_α given that the institution is not part of the population[3] Concretely, the institution can only sanction AU-in when both players committed. In that case, its speed will be reduced, from s to $s - s_\beta$. AU-out as before will be free of any sanction.

In Fig. 5, we compare the effect of institutional sanctioning with and without the option of having prior voluntary bilateral commitments between the two players. The outcomes are similar to the case of PP discussed in the previous section. In particular, one can observe that the presence of commitments can significantly improve safety outcome in region **II** without being detrimental to the desired behaviour and thus social outcome in the other two regions. Without commitments, institutional punishment leads again to over-regulation of unsafe/innovation in region **III**, which, as was mentioned before, is not a desired outcome. Moreover, with a smaller s_β (compare top and bottom rows of the figure), the outcomes remain similar, with a lower level of over-regulation but weaker results for region **II**.

In short, we observe that freedom of choice between bilateral commitments to safety regulations (which can be self-enforcing or institutionally enforced) and unilateral risky/unsafe endeavours provides an efficient solution to ensure desirable outcomes in all scenarios of the DSAIR model. First, it alleviates the problem of over-regulation when the risk is low (which happens when punishment can be very efficient), while at the same time it reduces the frequency of unsafe behaviour in the dilemma zone (region **II**). Furthermore, the desirable safety compliance outcomes in region **I** remain unaffected.

4. Discussion and conclusions

This paper proposes and analyses, using a multi-agent and population dynamics modelling approach, a novel solution that promotes desirable safety compliance in a technology development race, while at the same time avoiding stifling beneficial innovation due to over-regulation. We base our study on a previously proposed EGT model [29] that describes the dynamics of a competition between safety and risk-taking (unsafe) behaviours, within an (alleged) race for supremacy through AI in a marketable domain. We show that, by allowing race participants the freedom of choosing to 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 as a whole, are achieved in all regions of the parameter space.

This system of voluntary bilateral commitments, either depending on sanctioning actions of peers or by an institution, provides a mechanism to overcome the problems associated with over-regulation, which might occur whenever risk taking behaviour, which may be perceived as unsafe, is unconditionally penalised without taking into account the true risk level (in relation to the cost and benefit of the safety compliance and of the opting for risky behaviour), as shown in Ref. [30]. On the one hand, allowing for participants to explicitly commit to safety precautions in their development process ensures higher levels of safety compliance when that is desired, because then unsafe/risk-taking strategies become either clearly identified (AU-out) and suitably coped with, or punished through the binding commitment (AU-in). On the other hand, the commitment option enables the (honest) risk-taking behaviour (namely, AU-out) to prevail whenever it is collectively preferred, since then it is not punished, as a result of not joining the commitment in the first place.

There have been several theoretical modelling studies based on EGT, showing the benefit of establishing prior commitments or agreements for promoting the evolution of certain positive behaviours, such as cooperation and coordination in a population of self-regarding agents, see e.g. Refs. [41–45]. Behavioral experiments with human subjects have also been performed showing the promoting role of commitments for cooperative behaviours, see e.g. Refs. [33,46]. In all such contexts, it is well-defined what is the correct (positive) behaviour most beneficial collectively and thus the one that should be promoted. In our case that no longer holds, since whether or not a behaviour is the most beneficial collectively depends on the behavioural region in which the race occurs at the time. This region is not well defined, especially when data is not abundantly available. For example, we might need to wait for a

[3] Future work might look at how to minimise the cost and efforts from the institution while ensuring a desired level of safety compliance [58–63].

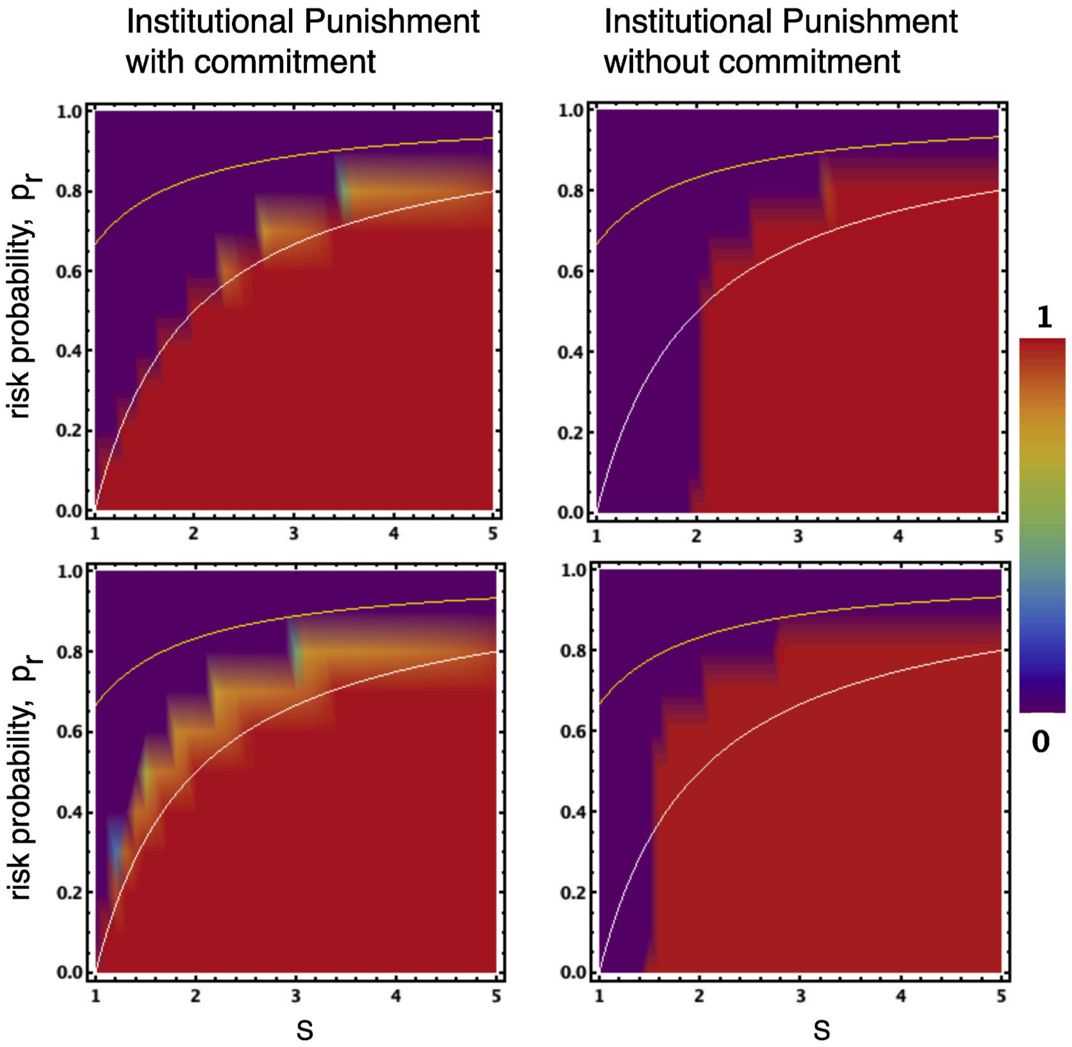

Fig. 5. Institutional punishment, with (left column) or without a safety agreement (right column). We report the frequency of unsafe behaviour (UNSAFE) in a population of AU and AS in the latter case and a population of AS-in, AU-in, AS-out, AU-out, in the former case. An institutional punishment reduces the speed of AU in the latter and AU-in (when an agreement is formed) in the former, by $s_\beta = 1$ (**top row**) and by $s_\beta = 0.5$ (**bottom row**). The lines within the figures are as in Fig. 1. Parameters: $b = 4$, $c = 1$, $W = 100$, $B = 10^4$, $\beta = 1$, $Z = 100$.

technology to be audited and even largely adopted by users to know the level of risk associated with a particular technology (but then it might be too late already to provide pertinent regulations). Such level of uncertainty can also lead to other externalities, such as the emergence of strongly polarized positions [47], an effect not yet characterized by our model.

Interestingly, we show here that establishing the possibility of a voluntary bilateral commitments is as well highly effective in promoting collectively desired/preferred behaviour in all cases. Commitment to cooperative actions in social dilemmas work because they allow cooperators to avoid free-riding strategies, for the case of an AI race, because they permit players the freedom to follow their preferred course of development without being punished (as long as they are being honest). One can also see it as a particular form of binding signal or pledge. The results of this work are in line with Hadfield's arguments [21] in the general context of technology globalisation; that is, legal contracts and institutions should allow freedom (via coordination) instead of exacting it from individuals in a population. For the sake of clarity, here we illustrate this idea with a pairwise race model; however, as shown in Ref. [29], this model can be easily extended to N-player interactions, with a concomitant increase in complexity, yet leading to the same qualitative messages. Future work may explicitly consider N-player commitments and more complex communication and bottom-up dynamics typical of this type of agreements, see e.g. Refs. [42,48,49].

In [16] the authors propose the so-called Windfall Clause for mediating the tension in AI competition. It is proposed there to arrange prior commitments or agreements from the race participants (e.g., companies, governments) for sharing a large part of the windfall benefit (i.e. the large prize B in our model) with society. As discussed in that paper, it requires significant legal infrastructures to be put in place to enable such a mechanism. Our approach does not so require because legal contracts can be used to ensure compliance with prior safety agreements.

Last but not least, our proposed mechanism resolves an important limitation of soft law for AI governance [26,28]. As in soft law, the model assumes there is a set of safety standards to which players voluntarily commit. Yet, our voluntary commitment approach assumes also that there is an entity/institution that facilitates either peer or institutional sanctioning, against those who dishonour a commitment that they voluntarily agreed to. Participants in our artificial race are fully aware of this. As we have shown, this sanctioning is essential to ensure beneficial outcomes while maintaining the potential of innovation through rapid developments. Given that participants in an AI race can be global, the institution is likely to be an international one. Consequently, our proposed governance approach aligns more closely to theories on global law [27], requiring further exploration of models to AI governance in that direction. One avenue could be to examine how the EU's draft proposal aligns with our observations, and vice versa, how the proposed measures can be operationalised in the current model.

The voluntary commitment with enforcement we proposed here shows also that by having a pre-commitment stage, the safe players can identify those who do not want to commit to safety standards, clarifying their intentions [41,50,51], providing thus additional information to act upon. In general, our model reveals that a voluntary "hard law" mechanism may have the potential to provide the necessary protection against mis-aligned AI developments, steering away from potential societal or existential risk problems [52].

It is noteworthy that, although we focus in this paper on an AI development race, the model proposed can be more generally applicable to other kinds of long-term situations of competition, such as the development of technological innovation and its racing for patents in which there is a significant advantage (i.e. a large B) to be gotten in being one of the first ever to reach an important target [53–55]. Other important domains involve pharmaceutical and vaccines development race, where companies might attempt to cut a few corners by not following strictly the safe clinical trial protocols, with the view to be the first ones to develop and put on the market some pharmaceutical product, in order to reap the highest possible share of benefit in the market [56,57]. However, certain aspects or factors of the current model might need to be revised. For example, a significant profit B for a vaccine development race would be harder to achieve than in the case of AI, since a developed vaccine needs to be approved by suitable authorities before users have trust in it and use it. Vaccine developers can only generate the profit B if they could show evidence of the vaccine's safety and closely followed all the safety procedures. In the case of AI, the winning developers can simply deploy the technologies and get the profit B (in most cases). One can expand the current model to capture also the vaccine race (and also generalised the current AI race model) by for example having a new parameter to capture when B can be generated for the race winners, which depends on the nature of the race (e.g. vaccine vs AI) and also the frequency of safety compliance in the past. Our future works will address these issues.

Acknowledgements

The authors acknowledge the feedback from Mireille Hildebrandt and Gregory Lewkowicz on issues related with soft law and the necessity of "harder" law in AI governance. T.A.H., L.M.P., T.L. are supported by Future of Life Institute grant RFP2-154. T.A.H. is also supported by a Leverhulme Research Fellowship (RF-2020-603/9). LM.P. support by NOVA-LINCS (UIDB/04516/2020) with the financial support of FCT-Funda\c c\~ao para a Ci\^encia e a Tecnologia, Portugal, through national funds. F.C.S. acknowledges support from FCT Portugal (grants UIDB/50021/2020, PTDC/MAT-APL/6804/2020 and PTDC/CCI-INF/7366/2020). T.L benefits from the support by the Flemish Government through the AI Research Program and F.C.S and T.L. by TAILOR, a project funded by EU Horizon 2020 research and innovation program under GA No 952215. T.L. is furthermore supported by the F.N.R.S. project with grant number 31257234, the F.W.O. project with grant nr. G.0391.13N, the FuturICT 2.0(\url{www.futurict2.eu}) project funded by the FLAG-ERA JCT 2016 and the Service Public de Wallonie Recherche under grant n\textdegree\ 2010235–ARIAC by DigitalWallonia4.ai.

5. Appendix

5.1. Effect of larger intensity of selection

See Figures A1 and A2.

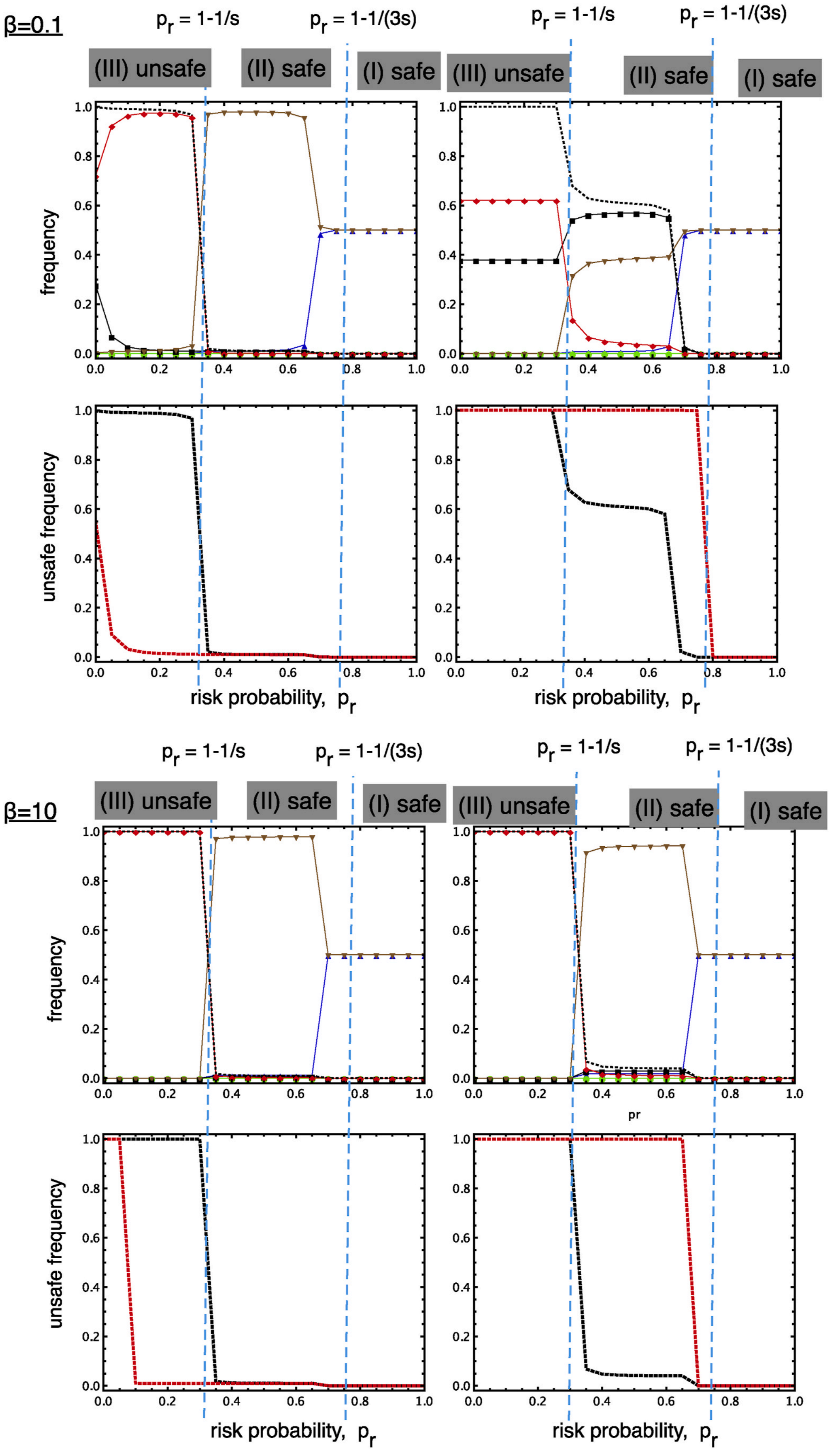

Fig. A1. Same as Fig. 2, with other values of β.

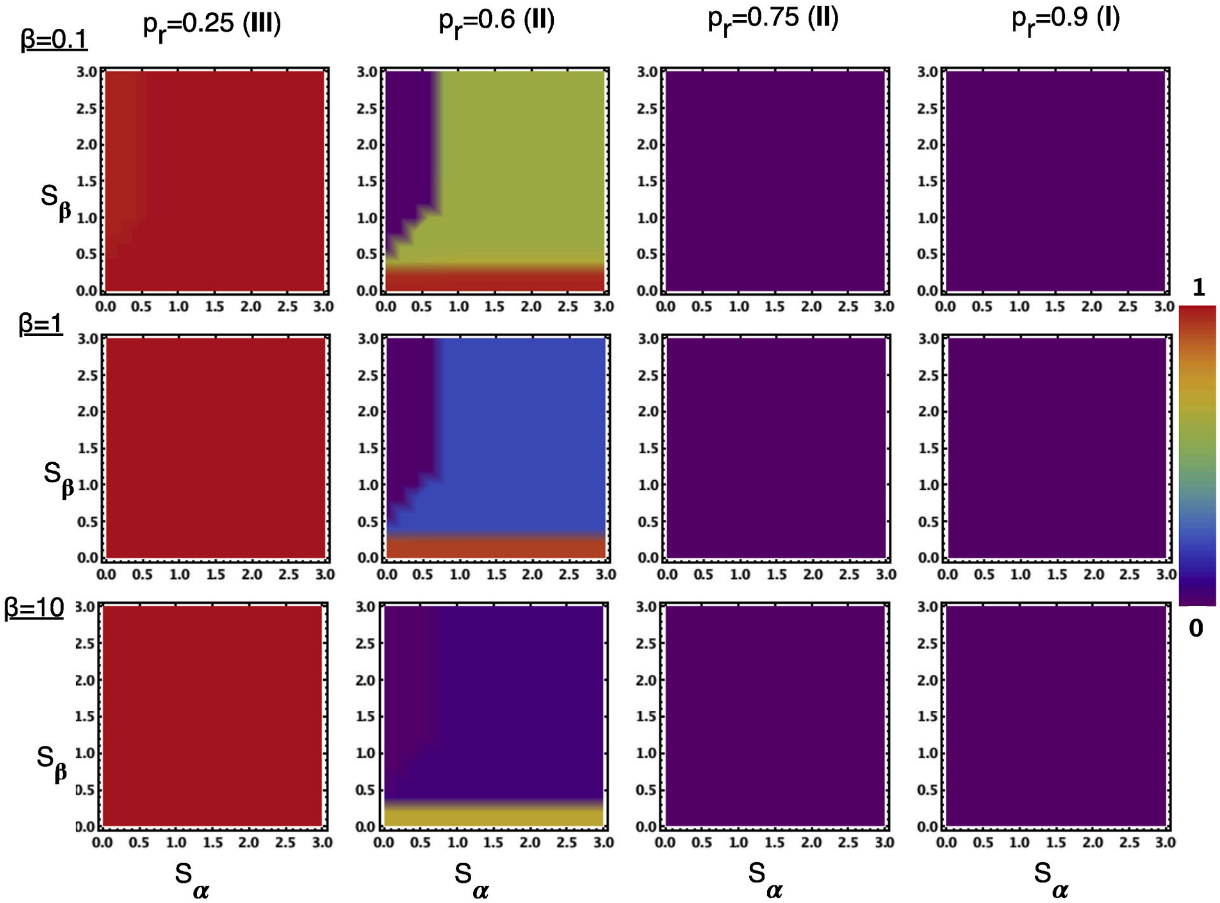

Fig. A2. Frequency of unsafe in all regions for varying s_α and s_β, when safety agreement is in use, for different values of intensities of selection β. Similar observations as in the main text. Parameters: $b = 4$, $c = 1$, $s = 1.5$, $W = 100$, $B = 10^4$, $Z = 100$.

5.2. Other scenarios of modelling

See Figures A3 and A4.

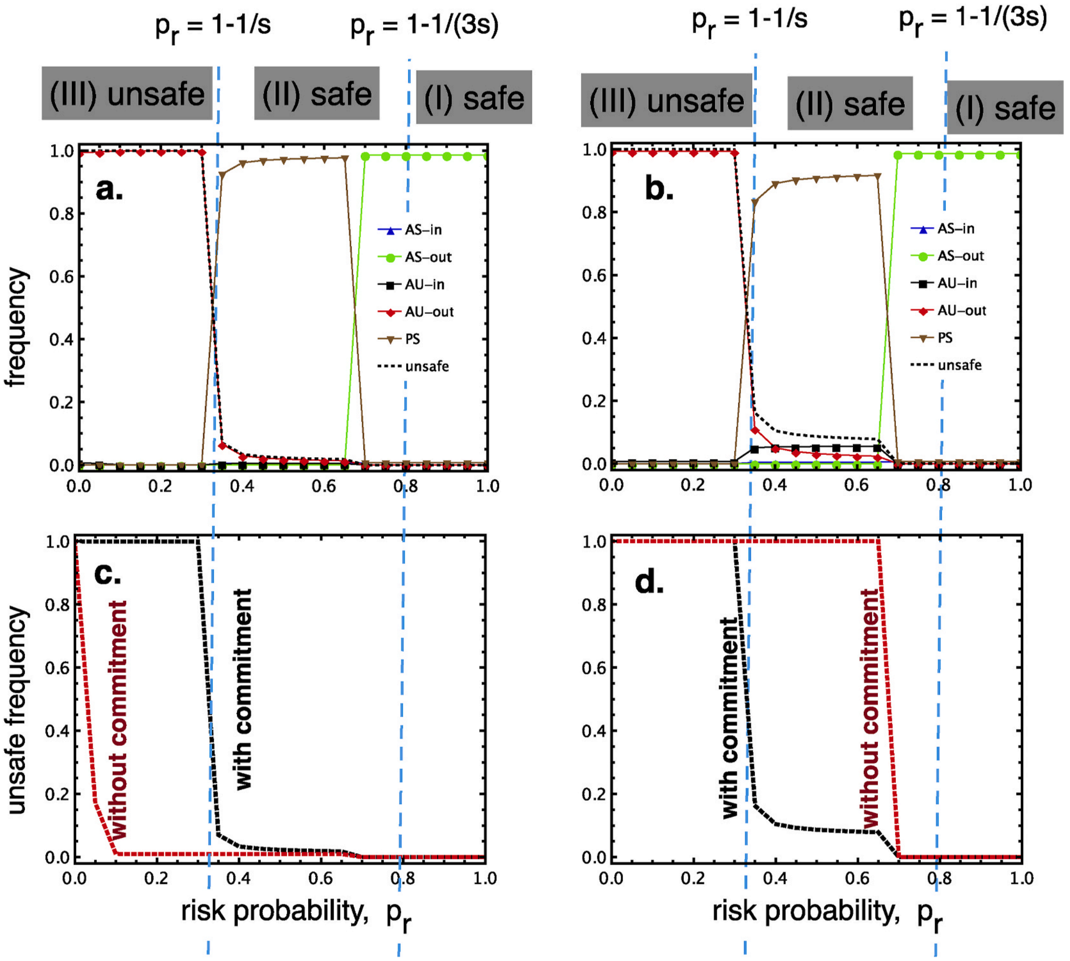

Fig. A3. When AU-in always plays SAFE in an interaction. Behavioural dominance of in different zones for varying p_r, in presence of prior commitment (top row: panels a, b) and comparison of its overall unsafe behaviour against when commitment is absent (bottom row: panels c, d). We show results for two important scenarios: when efficient punishment can be made for a small cost (*left column:* $s_\alpha = 0.3$, $s_\beta = 1$) and when punishment is not highly efficient (*right column:*, $s_\alpha = 1$, $s_\beta = 1$). In both cases, AU-out dominates when p_r is small (zone **III**), PS dominates when it is intermediate (first part of zone **II**), while PS and AU-in together dominate when it is large (part of zone **II** and zone **I**). That results in the fact that when a commitment is present desirable outcomes are achieved in all three zones: unsafe/innovation in **III** and safe in **I** and **II**, see bottom row. In absence of a commitment, over-regulation occurs (i.e. safe is abundant but not desired) for a large part of zone **I** when efficient punishment can be done for a small cost (panel c), while safety dilemma occurs (i.e. safe is desired but infrequent) for a large part of zone **II** when punishment is not highly efficient. Parameters: $b = 4$, $c = 1$, $s = 1.5$, $W = 100$, $B = 10^4$, $\beta = 1$, $Z = 100$.

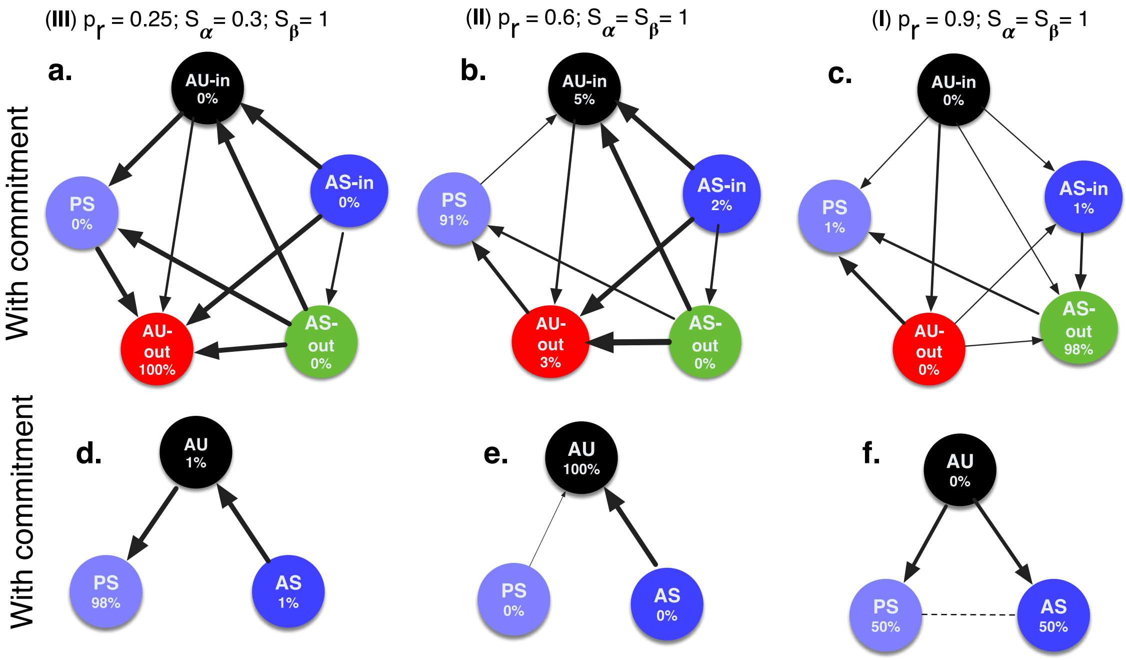

Fig. A4. When AU-in always play SAFE in an interaction. Transitions and stationary distributions when a commitment is present (top row) against when it is absent (bottom row), for three regions (only stronger transitions are shown; no transition either way means neutral). In region **III** (first column, low risk, where UNSAFE or innovation is wanted), low-cost highly efficient punishment can lead to significant reduction of innovation since PS dominates AU in absence of a commitment (panel d). Addition of AU-out provides an escape as this strategy is not punished, thereby dominating PS (panel a). In region **II** (dilemma zone, safety is wanted), when punishment is not highly efficient, there is still some amount of unwanted AU in the system. As AU-out is dominated by PS since PS plays UNSAFE in absence of a commitment, it leads to lower unsafe behaviour. In region **III**, addition AU-out does not change the desired outcome of safety. Parameters: $b = 4$, $c = 1$, $s = 1.5$, $W = 100$, $B = 10^4$, $\beta = 1$, $Z = 100$.

References

[1] N. Bostrom, Superintelligence: Paths, Dangers, Strategies, 2014.

[2] A. Holzinger, E. Weippl, A.M. Tjoa, P. Kieseberg, Digital transformation for sustainable development goals (sdgs)-a security, safety and privacy perspective on ai, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, 2021, pp. 1–20.

[3] C. Stix, M. Maas, Bridging the Gap: the Case for an 'incompletely Theorized Agreement'on Ai Policy, 2020.

[4] A. Tzachor, J. Whittlestone, L. Sundaram, et al., Artificial intelligence in a crisis needs ethics with urgency, Nat. Machine Intell. 2 (7) (2020) 365–366.

[5] M. Coeckelbergh, AI Ethics, MIT Press, 2020.

[6] S.D. Baum, On the promotion of safe and socially beneficial artificial intelligence, AI Soc. 32 (4) (2017) 543–551.

[7] S. Cave, S. ÓhÉigeartaigh, An AI race for strategic advantage: rhetoric and risks, in: AAAI/ACM Conference on Artificial Intelligence, Ethics and Society, 2018, pp. 36–40.

[8] R. de Neufville, S.D. Baum, Collective action on artificial intelligence: a primer and review, Technol. Soc. 66 (2021) 101649.

[9] M. Taddeo, L. Floridi, Regulate artificial intelligence to avert cyber arms race, Nature 556 (7701) (2018) 296–298.

[10] European Commission, White Paper on Artificial Intelligence – an European Approach to Excellence and Trust, Technical report, European Commission, 2020.

[11] K. Stöger, D. Schneeberger, A. Holzinger, Medical artificial intelligence: the european legal perspective, Commun. ACM 64 (11) (2021) 34–36.

[12] A. Askell, M. Brundage, G. Hadfield, The Role of Cooperation in Responsible AI Development, 2019 arXiv preprint arXiv:1907.04534.

[13] E.M. Geist, It's already too late to stop the ai arms race: we must manage it instead, Bull. At. Sci. 72 (5) (2016) 318–321.

[14] T.A. Han, L.M. Pereira, T. Lenaerts, Modelling and influencing the AI bidding war: a research agenda, in: Proceedings of the AAAI/ACM Conference AI, Ethics and Society, 2019, pp. 5–11.

[15] P. Nemitz, Constitutional democracy and technology in the age of artificial intelligence, Phil. Trans. Math. Phys. Eng. Sci. 376 (2133) (2018) 20180089.

[16] C. O'Keefe, P. Cihon, B. Garfinkel, C. Flynn, J. Leung, A. Dafoe, The windfall clause: distributing the benefits of ai for the common good, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 327–331.

[17] C. Shulman, S. Armstrong, Arms control and intelligence explosions, in: 7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, July, 2009, pp. 2–4.

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

[19] J.O. McGinnis, UL Rev., Accelerating AI. *Nw*, vol. 104, 2010, p. 1253.

[20] J.K. Gurney, Sue My Car Not Me: Products Liability and Accidents Involving Autonomous Vehicles, 2013, p. 247. U. Ill. JL Tech. & Pol'y.

[21] G.K. Hadfield, Rules for a Flat World: Why Humans Invented Law and How to Reinvent it for a Complex Global Economy, Oxford University Press, 2017.

[22] K.-F. Lee, AI Superpowers: China, Silicon Valley, and the New World Order, Houghton Mifflin Harcourt, 2018.

[23] EDRI, Civil Society Calls for AI Red Lines in the European Union's Artificial Intelligence Proposal, 2021. Technical report, European Commission. Accessed January-29-2021.

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

[25] C.I. Gutierrez, G.E. Marchant, A. Carden, K. Hoffner, A. Kearl, Preliminary Results of a Global Database on the Soft Law Governance of Artificial Intelligence, 2020 (Available at: SSRN).

[26] R. Hagemann, J. Huddleston Skees, A. Thierer, Soft law for hard problems: the governance of emerging technologies in an uncertain future, Colo. Tech. LJ 17 (2018) 37.

[27] B. Frydman, G. Lewkowicz, Les codes de conduite: source du droit global? Rev. Trimest. Droits Homme 73 (2009) 73.

[28] G. Marchant, "soft Law" Governance of Artificial Intelligence. UCLA: the Program on Understanding Law, Science, and Evidence (PULSE), 2019.

[29] T.A. Han, L.M. Pereira, F.C. Santos, T. Lenaerts, To regulate or not: a social dynamics analysis of an idealised AI race, J. Artif. Intell. Res. 69 (2020) 881–921.

[30] T.A. Han, L.M. Pereira, T. Lenaerts, F.C. Santos, Mediating artificial intelligence developments through negative and positive incentives, PLoS One 16 (1) (2021), e0244592.

[31] D. Collingridge, The Social Control of Technology, St. Martin's Press, New York, 1980.

[32] S. Barrett, Environment and Statecraft: the Strategy of Environmental Treaty-Making: the Strategy of Environmental Treaty-Making, Oxford University Press, 2003.

[33] T.L. Cherry, D.M. McEvoy, Enforcing compliance with environmental agreements in the absence of strong institutions: an experimental analysis, Environ. Resour. Econ. 54 (1) (2013) 63–77.

[34] A. Tavoni, A. Dannenberg, G. Kallis, A. Löschel, Inequality, communication and the avoidance of disastrous climate change in a public goods game, Proc. Natl. Acad. Sci. U.S.A. 108 (2011) 11825–11829.

[35] L. Hindersin, B. Wu, A. Traulsen, J. García, Computation and simulation of evolutionary game dynamics in finite populations, Sci. Rep. 9 (1) (2019) 1–21.

[36] K. Sigmund, The Calculus of Selfishness, Princeton University Press, 2010.

[37] A. Traulsen, M.A. Nowak, J.M. Pacheco, Stochastic dynamics of invasion and fixation, Phys. Rev. E 74 (2006) 11909.

[38] J. Grujić, T. Lenaerts, Do people imitate when making decisions? evidence from a spatial prisoner?s dilemma experiment, R. Soc. Open Sci. 7 (7) (2020) 200618.

[39] L.A. Imhof, D. Fudenberg, M.A. Nowak, Evolutionary cycles of cooperation and defection, Proc. Natl. Acad. Sci. U.S.A. 102 (2005) 10797–10800.

[40] M.A. Nowak, A. Sasaki, C. Taylor, D. Fudenberg, Emergence of cooperation and evolutionary stability in finite populations, Nature 428 (2004) 646–650.

[41] T.A. Han, T. Lenaerts, A synergy of costly punishment and commitment in cooperation dilemmas, Adapt. Behav. 24 (4) (2016) 237–248.

[42] T.A. Han, L.M. Pereira, T. Lenaerts, Evolution of Commitment and Level of Participation in Public Goods Games, Autonomous Agents and Multi-Agent Systems, 2017, pp. 1–23.

[43] T.A. Han, L.M. Pereira, F.C. Santos, T. Lenaerts, Good agreements make good friends, Sci. Rep. 3 (2695) (2013).

[44] N.B. Ogbo, A. Elgarig, T.A. Han, Evolution of Coordination in Pairwise and Multi-Player Interactions via Prior Commitments, 2021. Adaptive Behavior (In Press). Preprint arXiv:2009.11727.

[45] T. Sasaki, I. Okada, S. Uchida, X. Chen, Commitment to cooperation and peer punishment: its evolution, Games 6 (4) (2015) 574–587.

[46] X.-P. Chen, S.S. Komorita, The effects of communication and commitment in a public goods social dilemma, Organ. Behav. Hum. Decis. Process. 60 (3) (1994) 367–386.

[47] E.F. Domingos, J. Grujić, J.C. Burguillo, G. Kirchsteiger, F.C. Santos, T. Lenaerts, Timing uncertainty in collective risk dilemmas encourages group reciprocation and polarization, iScience 23 (12) (2020) 101752.

[48] T.A. Han, L.M. Pereira, L.A. Martinez-Vaquero, T. Lenaerts, Centralized vs. personalized commitments and their influence on cooperation in group interactions, in: AAAI, 2017, pp. 2999–3005.

[49] V.V. Vasconcelos, F.C. Santos, J.M. Pacheco, A bottom-up institutional approach to cooperative governance of risky commons, Nat. Clim. Change 3 (9) (2013) 797–801.

[50] T.A. Han, Intention Recognition, Commitments and Their Roles in the Evolution of Cooperation: from Artificial Intelligence Techniques to Evolutionary Game Theory Models, ume 9, Springer SAPERE series, 2013.

[51] T.A. Han, F.C. Santos, T. Lenaerts, L.M. Pereira, Synergy between intention recognition and commitments in cooperation dilemmas, Sci. Rep. 5 (9312) (2015).

[52] S. Russell, J. Bohannon, Artificial intelligence. fears of an ai pioneer, Science (New York, NY) 349 (6245) (2015), 252–252.

[53] S. Campart, E. Pfister, Technological races and stock market value: evidence from the pharmaceutical industry, Econ. Innovat. N. Technol. 23 (3) (2014) 215–238.

[54] V. Denicolò, L.A. Franzoni, On the winner-take-all principle in innovation races, J. Eur. Econ. Assoc. 8 (5) (2010) 1133–1158.

[55] M.A. Lemley, The Myth of the Sole Inventor, Michigan Law Review, 2012, pp. 709–760.

[56] F.M. Abbott, M.N.G. Dukes, G. Dukes, Global Pharmaceutical Policy: Ensuring Medicines for Tomorrow's World, Edward Elgar Publishing, 2009.

[57] R. Burrell, C. Kelly, The Covid-19 Pandemic and the Challenge for Innovation Policy, 2020. Available at: SSRN 3576481.

[58] X. Chen, T. Sasaki, Å. Brännström, U. Dieckmann, First carrot, then stick: how the adaptive hybridization of incentives promotes cooperation, J. R. Soc. Interface 12 (102) (2015) 20140935.

[59] M.C. Couto, J.M. Pacheco, F.C. Santos, Governance of risky public goods under graduated punishment, J. Theor. Biol. 505 (2020) 110423.

[60] M.H. Duong, T.A. Han, Cost efficiency of institutional incentives for promoting cooperation in finite populations, Proc. Math. Phys. Eng. Sci. 477 (2254) (2021) 20210568.

[61] T.A. Han, S. Lynch, L. Tran-Thanh, F.C. Santos, Fostering cooperation in structured populations through local and global interference strategies, in: IJCAI-ECAI'2018, 2018, pp. 289–295.

[62] T.A. Han, L. Tran-Thanh, Cost-effective external interference for promoting the evolution of cooperation, Sci. Rep. 8 (1) (2018) 1–9.

[63] S. Wang, X. Chen, A. Szolnoki, Exploring optimal institutional incentives for public cooperation, Commun. Nonlinear Sci. Numer. Simulat. 79 (2019) 104914.