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

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

Abstract

Rapid technological advancements in Artificial Intelligence (AI), as well as the growing deployment of intelligent technologies in new application domains, have generated serious anxiety and a fear of missing out among different stake-holders, fostering a racing narrative. Whether real or not, the belief in such a race for domain supremacy through AI, can make it real simply from its consequences, as put forward by the Thomas theorem. These consequences may be negative, as racing for technological supremacy creates a complex ecology of choices that could push stake-holders to underestimate or even ignore ethical and safety procedures. As a consequence, different actors are urging to consider both the normative and social impact of these technological advancements, contemplating the use of the precautionary principle in AI innovation and research. Yet, given the breadth and depth of AI and its advances, it is difficult to assess which technology needs regulation and when. As there is no easy access to data describing this alleged AI race, theoretical models are necessary to understand its potential dynamics, allowing for the identification of when procedures need to be put in place to favour outcomes beneficial for all. We show that, next to the risks of setbacks and being reprimanded for unsafe behaviour, the time-scale in which domain supremacy can be achieved plays a crucial role. When this can be achieved in a short term, those who completely ignore the safety precautions are bound to win the race but at a cost to society, apparently requiring regulatory actions. Our analysis reveals that imposing regulations for all risk and timing conditions may not have the anticipated effect as only for specific conditions a dilemma arises between what is individually preferred and globally beneficial. Similar observations can be made for the long-term development case. Yet different from the short-term situation, conditions can be identified that require

the promotion of risk-taking as opposed to compliance with safety regulations in order to improve social welfare. These results remain robust both when two or several actors are involved in the race and when collective rather than individual setbacks are produced by risk-taking behaviour. When defining codes of conduct and regulatory policies for applications of AI, a clear understanding of the time-scale of the race is thus required, as this may induce important non-trivial effects.

1. Introduction

Interest in AI has exploded in academia and businesses in the last few years. This excitement is, on the one hand, due to a series of superhuman performances generated by particular breakthrough technologies (Silver et al., 2017; Brown & Sandholm, 2018; Silver et al., 2018; Brown & Sandholm, 2019). Although successful in highly specialised tasks, these AI success stories appear in the imagination of the general public as well as many media, as Hollywood-like Artificial General Intelligence (AGI), able to perform a broad set of intellectual tasks while continuously improving itself, generating thus unrealistic expectations and unnecessary fears (Cave & Dihal, 2019). On the other hand, this excitement is further promoted by political and business leaders alike, for both anticipate important gains from turning previously idle data into active assets within business plans (PwC, 2017). All these (un)announced business, societal and political ambitions reveal a certain level of anxiety, driving these stake-holders to quickly jump on an accelerating wagon just to make sure they will not stay behind. This anxiety is further stimulated by an AI race narrative (AI-Roadmap-Institute, 2017; Cave & ÓhÉigeartaigh, 2018; Apps, 2019; Cave, Dihal, & Dillon, 2020), where stake-holders in both private and public sectors are allegedly competing in an arms-race to lead the development and deployment of powerful, transformative AI (Armstrong, Bostrom, & Shulman, 2016; Baum, 2017; Bostrom, 2017; Cave & ÓhÉigeartaigh, 2018; Lee, 2018).

Irrespectively of the reality of such a race, whose existence and terms are contested (Dignum, Muller, & Theodorou, 2020; Sotala & Yampolskiy, 2014; Lee, 2018; Cave & ÓhÉigeartaigh, 2018), there exist indeed, even if not within a global race towards an AGI, several salient competition races to the market, regarding elaborate AI tools of wide-ranging use, for example, sophisticated flexible image recognition, natural speech and language understanding and interaction, or a combination of vision and language (Taddeo & Floridi, 2018; Lee, 2018). Consequently, many actors have urged for due diligence as i) these AI systems can also be employed for more nefarious activities, e.g. espionage and cyberterrorism (Taddeo & Floridi, 2018) and ii) whilst attempting to be the first/best, some ethical consequences as well as safety procedures may be underestimated or even ignored (Armstrong et al., 2016; Cave & ÓhÉigeartaigh, 2018) (notwithstanding the issue that certain claims about achieving AGI may be overly optimistic or just oversold). These concerns are highlighted by the many letters of scientists against the use of AI in military applications (Future of Life Institute, 2015, 2019), the blogs of AI experts requesting careful communications (Brooks, 2017) and the proclamations on ethical use of AI in the world (Montreal Declaration, 2018; Steels & Lopez de Mantaras, 2018; Russell et al., 2015; Jobin, Ienca, & Vayena, 2019; European Commission, 2020).

While potential AI disaster scenarios are many (Sotala & Yampolskiy, 2014; Armstrong et al., 2016; Pamlin & Armstrong, 2015; Schubert, Caviola, & Faber, 2019), the uncer-

tainties in accurately predicting these risks and outcomes are high (Armstrong, Sotala, & ÓhÉigeartaigh, 2014). As put forward by the Collingridge Dilemma, the impact of a new technology is difficult to predict unless large steps have been taken in its development and it becomes generally adopted (Collingridge, 1980). Sufficient data is therefore not yet available, requiring a modelling approach to grasp what can be expected in a race for domain supremacy through AI (DSAI). Models provide dynamic descriptions of the key features of this race (or parts thereof) allowing one to understand what outcomes are possible under certain conditions and what may be the effect of policies that aim to regulate the race. Especially, this latter issue is important as regulations are put into place for a race that may not even exist, producing outcomes that were actually not intended in the first place. To realise such ambitions, one first needs a baseline model that describes the racing dynamics and how its parameters control the observations, which is the main ambition of this manuscript.

The idealised model proposed here models the decision-process of each race participant where she can choose between unsafe or safe development (or deployment) of AI technology steps to reach DSAI and discusses when this may become disruptive, i.e. when social welfare is harmed. The model reveals when (and when not) regulations may be required and what actions should be stimulated to promote social welfare. It thus provides a tool useful for researchers and practitioners in law and technology, on the one hand, and researchers involved in topics related to AI policy-making, on the other hand, to understand the implications and the necessity of the regulations they intend to propose (see also *Lessons-learned box*). For instance, the model can be employed to evaluate the impact of regulatory mechanisms like rewards for safety compliance or fines for unsafe actions on the behavioural preferences. We resort to the framework of evolutionary game theory (EGT) (Smith, 1982; Hofbauer & Sigmund, 1998; Sigmund, 2010) to define the model. Note that even though the focus here is on AI technology, the proposed model is generally applicable to many competitive situations wherein a *winner-take-all* scenario is possible, which includes all technological innovation developments and patent races where there is a significant advantage to be achieved by reaching a target first (Denicolò & Franzoni, 2010; Campart & Pfister, 2014; Lemley, 2012). Other domains include pharmaceutical development where firms could try to cut corners by not following safe clinical trial protocols in an effort to be the first to develop a pharmaceutical produce (e.g. consider the current race for a COVID-19 vaccine), in order to take the highest possible share of the market benefit (Abbott, Dukes, & Dukes, 2009); Besides tremendous economic advantage, a winner of such a vaccine race can also gain significant political and reputation influence (Burrell & Kelly, 2020).

Concretely, the model assumes that in order to achieve DSAI in a domain X, a number of development steps or rounds are required. We assume, upon completion of each round, that there is a probability ω that yet another development round is required to reach DSAI—which results in an average number $W = (1-\omega)^{-1}$ of rounds per competition/race (Sigmund, 2010). It is thus important to understand that, while the average time-scale of development to reach DSAI can be defined, there is no explicit finish line in the model proposed here. Large-scale surveys and analysis of AI experts on their beliefs and predictions about progress in AI suggest that the perceived time-scale for DSAI is highly diverse across domains and regions (Armstrong et al., 2014; Grace et al., 2018). The model therefore aims to capture these different time-scales of DSAI occurrence: When W is small, DSAI can be

expected to happen in the near future (early DSAI regime) while when W is large, DSAI will only be achieved far away in time (late DSAI regime).

Because this is a race, each participant acts by herself during each step in order to reach the target and differs in the speed (s) with which she can complete each of the subtasks. The race thus consists of multiple rounds and the fastest participant will reap the benefit (b) at each round when she finishes before the others, winning the ultimate prize ($B \gg b$) once she carries out the final step achieving DSAI in the domain X. When multiple participants reach the end of an intermediate round or the final target at the same time they share the benefits, i.e. b and B, respectively. Other factors could play a role into why a participant wins, e.g. access to more qualified staff or larger budgets to start with. As we propose here a baseline/idealised model, we assume that all participants arrive at the start with equal wealth and resources. We focus here specifically on the choice of acting safely or not in trying to achieve DSAI in a domain first. Future variations of this model can then explore the impact of these additional characteristics and how they influence outcomes.

In this race, higher s may only be achievable by cutting corners, implying that some ethical or safety procedures are ignored. It takes time and effort to comply to precautionary requirements or acquire ethical approvals. Following a safe development process is thus not only more costly, it also results in a slower development speed. One can therefore consider that i) participants in the race that act safely (SAFE) pay a cost $c > 0$, which is not paid by participants that ignore safety procedures (UNSAFE) and ii) the speed of development of UNSAFE participants is faster ($s > 1$), compared to the speed of SAFE participants being normalised to $s = 1$. So essentially a SAFE player needs W rounds (on average) to complete the task, whereas an UNSAFE player will only need W/s.

Yet, UNSAFE strategists may suffer a personal setback or disaster during the race, losing the acquired payoffs. Concretely, a disaster or setback, removes the intermediate (b) and final (B) gains (see earlier) with a certain probability. The risk is personal for UNSAFE players in the current model. Although the threat is greater for the creator (Armstrong et al., 2016; Pamlin & Armstrong, 2015), there may also be repercussions for the other participants or society as a whole, a matter discussed in detail in the Appendix. As will be shown, this extension of spreading repercussions does not influence the results discussed in the next sections. The probability that the personal setback occurs is denoted by p_r and assumed to increase linearly with the frequency the participant violates the safety precautions. For example, if a participant always plays SAFE then disaster will not occur, given that

$$\left(\frac{|UNSAFE|}{|SAFE| + |UNSAFE|} \right) p_r = 0,$$

with $|UNSAFE|$ and $|SAFE|$ indicating the number of SAFE and UNSAFE actions respectively. A participant that only performs SAFE actions half of the time will incur only half of the risk of disaster over all rounds. The way we define the probability for a setback of disaster assumes that the risk is part of the development process and is thus not external: It is a direct function of the UNSAFE actions taken by the participants. We discuss implications of this assumption later on.

Finally, the model incorporates the possibility that an UNSAFE player is found out at each step of the race, which is an additional risk for UNSAFE players that corresponds to a simple form of regulation. We therefore assume that with some probability p_{fo} those

playing UNSAFE might be detected and their unsafe behavior disclosed, leading to 0 payoff in that round.

Given these different characteristics of the DSAI Race (DSAIR) model, we can now explore which strategies, involving SAFE an UNSAFE actions, are dominant under which conditions, i.e. the parameters defined by this model. Since we resort to EGT to answer this question, we consider a population of size Z in which players engage in a pairwise (or N-player) race. Each player can choose to consistently follow safety precautions (denoted by **AS**, the SAFE players) or completely ignore them (denoted by **AU**, the UNSAFE players). Additionally, we assume that, upon realising that UNSAFE players ignore safety precautions to gain a greater development speed, leading to the wining of the prize B (and a larger share of the intermediate benefit in each round, b, especially in the regime of weak monitoring or low p_{fo}), SAFE players might adopt unsafeness as well to avoid further disadvantage. It is indeed observed that competing countries or companies might engage in such a safety corner-cutting behaviour in deploying unsafe AI to avoid falling behind (Apps, 2019). We therefore consider, in line with previous literature on repeated games (Axelrod, 1984; Nowak & Sigmund, 1993; Sigmund, 2010; Han, Pereira, & Santos, 2011; Van Segbroeck et al., 2012), a conditional strategy (denoted by **CS**), which plays SAFE in the first round and then adopts the move its co-player used in the previous round. This so-called direct reciprocity strategy has been shown to promote cooperation in the context of repeated social dilemmas, outperforming consistently defective individuals (Axelrod, 1984; Sigmund, 2010). Alternative strategies can be imagined but for the sake of simplicity we focus (for now) on these three.

Importantly, our modelling approach seeks for a balance between the complexities of AI governance and the abstraction power that an idealised model of a technological race may offer. Such abstraction, however, should not be seen as an argument for oversimplified visions where all technological races are perceived as equivalent. Instead, the aim is to identify some of the key elements of the social dilemmas pertaining an idealised AI race, despite the specificities of each particular AI product and application, and different visions on the problem. Moreover, our insights into when regulatory requirements are necessarily related to the impact of the few factors (translated into different parameters) included in the model, and not their whole interplay with many others which may be included in future iterations of this framework. With this disclaimer in mind, in the following we will examine, across different time-scales of the DSAIR, under which conditions (for instance, regarding the disaster probability), safety behaviour should be promoted or externally enforced. Similarly, we address when one should omit the safety precautions for a larger social welfare to arise faster, when the benefits gained in doing so exceed the risk of a setback or personal disaster. Moreover, given the first-mover advantage of UNSAFE players in the race to AI -driven domain supremacy (i.e., acquire B), we will examine whether (and under what time-scale of the DSAIR model) conditional behaviours can still act as a promoting mechanism to achieve safety when required, or otherwise other mechanisms are needed. For the sake of clarity, we investigate here the pairwise race model and perform the analysis for the N-player ($N \geq 2$) DSAIR in the Appendix. Additionally, the situation where the effects of a setback or disaster are no longer just personal is also reported in depth in the Appendix.

2. Materials and Methods

We first describe our AI race model, then provide details of the EGT method being used for analysing the model.

2.1 Race Model Definition

The race for domain supremacy through AI is modeled as a repeated two-player game, consisting of W development rounds (on average). In each round, the players can collect benefits from their intermediate AI products, depending on whether they choose to play SAFE or UNSAFE. Assuming a fixed benefit, b, from the AI market, teams will share this benefit proportionally to their development speed. Moreover, we assume that with some probability p_{fo} those playing UNSAFE might be found out about their unsafe development and their products won't be used, leading to 0 benefit. Thus, in each round of the race, we can write the payoff matrix as follows (with respect to the row player)

$$\pi = \begin{array}{c} \\ SAFE \\ UNSAFE \end{array} \begin{array}{c} \begin{array}{cc} SAFE & \quad UNSAFE \end{array} \\ \begin{pmatrix} -c+\frac{b}{2} & -c+(1-p_{fo})\frac{b}{s+1}+p_{fo}b \\ (1-p_{fo})\frac{sb}{s+1} & (1-p_{fo}^2)\frac{b}{2} \end{pmatrix} \end{array}. \tag{1}$$

The payoff matrix in Equation 1 should be interpreted as follows. When two SAFE players interact, each needs to pay the cost c and they share the benefit b. When a SAFE player interacts with an UNSAFE one the SAFE player pays a cost c and obtains the full benefit b in case the UNSAFE co-player is found out (with probability p_{fo}), and obtains a small part of the benefit $b/(s+1)$ otherwise (i.e. with probability $1-p_{fo}$). When playing with a SAFE player, the UNSAFE does not have to pay any cost and obtains a larger share $bs/(s+1)$ when not found out. Finally, when an UNSAFE player interacts with another UNSAFE, it obtains the shared benefit $b/2$ when both are not found out and the full benefit b when it is not found out while the co-player is found out, and 0 otherwise. The payoff is thus: $(1-p_{fo})\left[(1-p_{fo})(b/2)+p_{fo}b\right] = (1-p_{fo}^2)\frac{b}{2}$.

The payoff matrix defining averaged payoffs for the three strategies AS, AU and CS, reads

$$\Pi = \begin{array}{c} \\ AS \\ AU \\ CS \end{array} \begin{array}{c} \begin{array}{ccc} AS & \quad AU & \quad CS \end{array} \\ \begin{pmatrix} \frac{B}{2W}+\pi_{11} & \pi_{12} & \frac{B}{2W}+\pi_{11} \\ (1-p_r)\left(\frac{sB}{W}+\pi_{21}\right) & (1-p_r)\left(\frac{sB}{2W}+\pi_{22}\right) & (1-p_r)\left[\frac{sB}{W}+\frac{s}{W}\left(\pi_{21}+(\frac{W}{s}-1)\pi_{22}\right)\right] \\ \frac{B}{2W}+\pi_{11} & \frac{s}{W}\left(\pi_{12}+(\frac{W}{s}-1)\pi_{22}\right) & \frac{B}{2W}+\pi_{11} \end{pmatrix} \end{array}. \tag{2}$$

The payoff matrix in Equation 2 should be understood as follows. When two AS players interact (and similarly for when an AS interacts with a CS player or when two CS players interact), they complete the race at the same time after, on average, W development rounds, thus obtaining on average $\frac{B}{2W}$ per round; moreover, these players always play SAFE in each round thus obtaining π_{11} (see first row and first column in the payoff matrix given by Equation 1) as the intermediate benefit per round. When an AS player interacts with a AU player, the AU wins the race and obtains the full prize B while AS gains nothing. Thus,

AS's average payoff only comes from the intermediate benefit in each round, which is equal to π_{12}. Yet, since AU completes the race in W/s development rounds, it obtains on average $\frac{B}{W/s} = \frac{sB}{W}$ per round from the prize. Additionally, it obtains π_{21} as the intermediate benefit per round. However, since AU plays UNSAFE in every round, a disaster may occur with probability p_r, thus the average payoff of AU when interacting with AS is $(1-p_r)\left(\frac{sB}{W} + \pi_{21}\right)$. Similarly, AU's payoff when interacting with CS follows from the fact that AU wins the race, completing it, on average, in W/s development rounds, where it earns π_{21} in the first round (since CS starts with SAFE) and π_{22} in the subsequent rounds (since CS plays UNSAFE after the first round to reciprocate what AU played). Yet again, a disaster may occur with probability p_r. All elements in the payoff matrix in Equation 2 can be explained in a similar fashion.

2.2 Evolutionary Dynamics in Finite Populations

We adopt here EGT methods for finite populations to derive analytical results and numerical observations (Nowak et al., 2004; Imhof, Fudenberg, & Nowak, 2005; Nowak, 2006). In a repeated games, players' average payoff over all the game rounds (see the payoff matrix in Equation 2) represents their *fitness* or social *success*, and evolutionary dynamics is shaped by social learning (Hofbauer & Sigmund, 1998; Sigmund, 2010), whereby the most successful players will tend to be imitated more often by the other players (Grujić & Lenaerts, 2020). In the current work, social learning is modeled using the so-called pairwise comparison rule (Traulsen, Nowak, & Pacheco, 2006), assuming that a player A with fitness f_A adopts the strategy of another player B with fitness f_B with probability given by the Fermi function, $\left(1 + e^{-\beta(f_B - f_A)}\right)^{-1}$, where β conveniently describes the selection intensity ($\beta = 0$ represents neutral drift while $\beta \to \infty$ represents increasingly deterministic selection). For convenience of numerical computations, but without affecting analytical results, we assume here small mutation limit (Fudenberg & Imhof, 2005; Imhof et al., 2005; Nowak et al., 2004). As such, at most two strategies are present in the population simultaneously, and the behavioural dynamics can thus be described by a Markov Chain, where each state represents a homogeneous population and the transition probabilities between any two states are given by the fixation probability of a single mutant (Fudenberg & Imhof, 2005; Imhof et al., 2005; Nowak et al., 2004). The resulting Markov Chain has a stationary distribution, which describes the average time the population spends in an end state. In two-player game, the average payoffs in a population of k A players and $(Z-k)$ B players can be given as below (recall that Z is the population size), respectively,

$$P_A(k) = \frac{(k-1)\Pi_{A,A} + (Z-k)\Pi_{A,B}}{Z-1}, \quad P_B(k) = \frac{k\Pi_{B,A} + (Z-k-1)\Pi_{B,B}}{Z-1}. \tag{3}$$

The fixation probability that a single mutant A taking over a whole population with $(Z-1)$ B players is as follows (Traulsen et al., 2006; Karlin & Taylor, 1975)

$$\rho_{B,A} = \left(1 + \sum_{i=1}^{Z-1} \prod_{j=1}^{i} \frac{T^-(j)}{T^+(j)}\right)^{-1}, \tag{4}$$

where $T^{\pm}(k) = \frac{Z-k}{Z}\frac{k}{Z}\left[1+e^{\mp\beta[P_A(k)-P_B(k)]}\right]^{-1}$ describes the probability to change the number of A players by $\pm$ one in a time step. Specifically, when $\beta = 0$, $\rho_{B,A} = 1/Z$, representing the transition probability at neutral limit.

Having obtained the fixation probabilities between any two states of a Markov chain, we can now describe its stationary distribution (Fudenberg & Imhof, 2005; Imhof et al., 2005). Namely, considering a set of s strategies, $\{1, ..., s\}$, their stationary distribution is given by the normalised eigenvector associated with the eigenvalue 1 of the transposed of a matrix $M = \{T_{ij}\}_{i,j=1}^{s}$, where $T_{ij,j\neq i} = \rho_{ji}/(s-1)$ and $T_{ii} = 1 - \sum_{j=1,j\neq i}^{s} T_{ij}$.

Risk-dominant conditions. We can determine which selection direction is more probable: an A mutant fixating in a homogeneous population of individuals playing B or a B mutant fixating in a homogeneous population of individuals playing A. When the first is more likely than the latter, A is said to be *risk-dominant* against B (Kandori, Mailath, & Rob, 1993; Gokhale & Traulsen, 2010), which holds for any intensity of selection and in the limit of large N when

$$\pi_{A,A} + \pi_{A,B} > \pi_{B,A} + \pi_{B,B}. \tag{5}$$

3. Results

We calculate the long-term frequency of each possible behavioural composition of the population, the so-called stationary distribution (cf. Methods), as this will reveal the action preferences (i.e. behaving safely or not) of a finite set of virtual players within the context of the DSAIR game defined above. This stochastic social dynamics of the population occurs in the presence of errors, both in terms of errors of imitation and of behavioural changes, the latter representing an open exploration of the possible strategies by the virtual participants (Hofbauer & Sigmund, 1998; Sigmund, 2010). As can be observed in Figure 1, the preference for the strategies AS, AU and CS changes for different lengths of the race. We distinguish two regimes in the DSAIR that depend on the relationship between the number of rounds W needed to achieve the ultimate benefit B and the revenue that can be achieved at every round, i.e. b:

i) **Early DSAI**: This regime is characterised by the observation that the ultimate prize of winning the race in W rounds strongly outweighs the benefits that can be achieved in a single round, i.e. $B/W \gg b$. Being fast is thus a key driver here.

ii) **Late DSAI**: In this regime, DSAI will not be achieved in a foreseeable future, making the gains at each round b, even when having to pay the safety cost c, more attractive than the ultimate prize of winning the race B, i.e. $B/W \ll b$.

We observe that in the first DSAI regime, AU dominates the population whenever the probability that an AI disaster occurs due to unsafe development (p_r) is not too high (see Figure 1c; also panels a and b, where $p_r = 0.6$). In the second DSAI regime, AS and CS take over (Figure 1a-b). When an AI disaster is more likely to occur due to unsafe developments (i.e. large p_r, see Figure 1c), AU disappears in both regimes.

Given the difference in behavioural preferences toward safety developments in the early and late regimes, different kinds of regulation may be required. Since AI developments

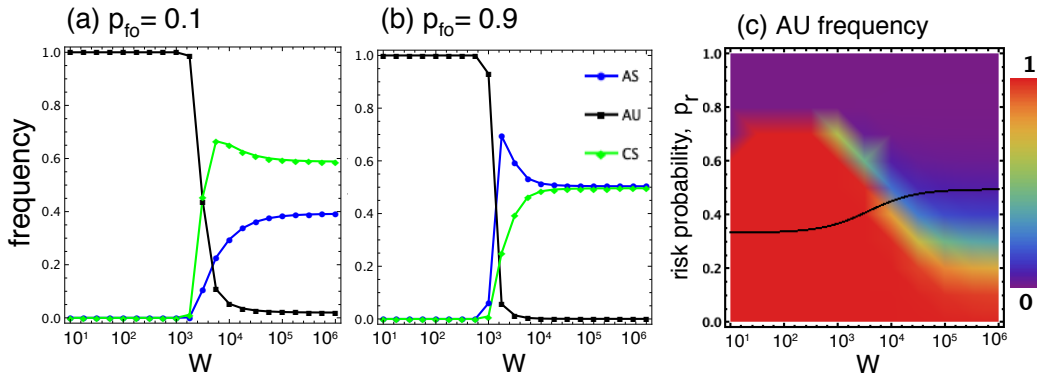

Figure 1: **Different regimes of DSAI: when W is small (early DSAI) vs when W is larger (late DSAI).** Panels (a) and (b) show the frequency of each strategy, i.e. AS, AU and CS, in a population ($p_r = 0.6$). In the early DSAI regime, AU dominates the population, while AS and CS outperform AU in the late DSAI regime. The former observation is valid for p_r values lower than 0.8, see panel (c) ($p_{fo} = 0.1$). For a high risk probability of disaster occurring due to ignoring safety precautions (high p_r), AU disappears in both regimes. The black line in (c) indicates the threshold of p_r above which SAFE is the preferred collective action and below which UNSAFE is the preferred one. Parameters: $c = 1$, $b = 4$, $s = 1.5$, $B = 10^4$, $\beta = 0.1$, $Z = 100$.

should at least provide a beneficial outcome for the individual developers and interested users in society, we first investigate under which conditions they can achieve their ambitions by acting safely, thus avoiding the risk of personal setbacks or shared disaster (see Appendix). When the benefits of all making safe developments ($\Pi_{AS,AS}$) outweigh the benefits of all doing things unsafely ($\Pi_{AU,AU}$), i.e. when $\Pi_{AS,AS} > \Pi_{AU,AU}$, this goal can be achieved (see Methods). The black line in Figure 1c depicts this threshold in function of p_r, revealing that there is a large part in the early regime (red area above the black line) where regulation should be put in place to restrain unsafe development behaviour. On the other hand, in the late regime (beyond 10^4 development steps), risk-taking should be promoted as this will improve social welfare (area below the black line).

Figure 1 thus underlines the importance of knowing in which regime the race is operating, since this would affect the type of regulation that one should introduce. In order to assess these observations in detail, we carry out a more in-depth analysis in the following sections.

Early DSAI: only under specific conditions will regulation improve welfare

We first focus again on the analytical conditions under which $\Pi_{AS,AS} > \Pi_{AU,AU}$ and then determine when the safe and reciprocal strategies are more likely to be imitated as this shows what behaviour to expect when participants can alter their actions in function of the benefits they can gain.

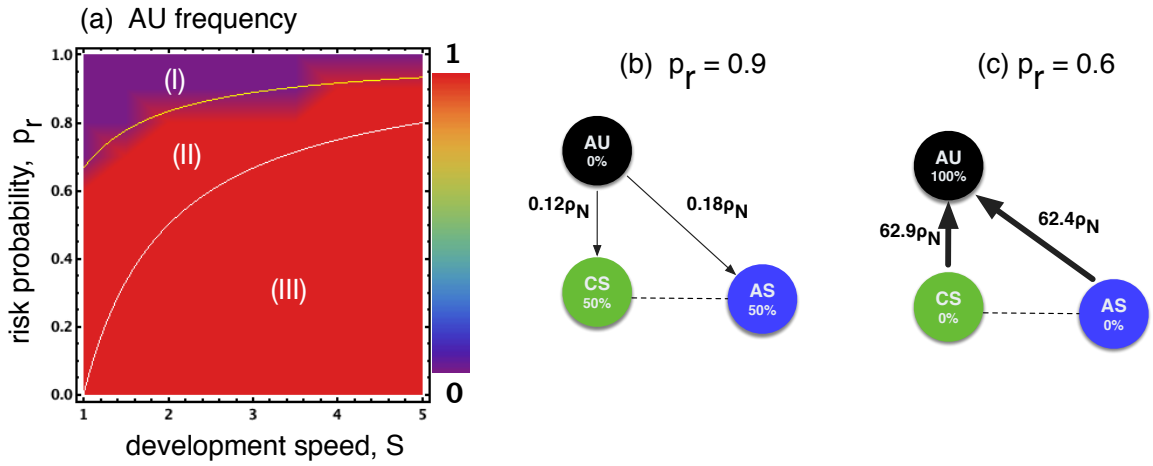

Figure 2: **Early DSAI regime**. **(a)** Frequency of AU as a function of the speed gained, s, and the probability of AI disaster occurring, p_r, when ignoring safety. In general, we observe that when the risk probability is small, AU is dominant. The larger s is, AU dominates for a larger range of p_r. Region (**II**): The two solid lines inside the plots indicate the boundaries $p_r \in [1-1/s, 1-1/(3s)]$ where safety development is the preferred collective outcome but unsafe development is selected by social dynamics. Regions (**I**) (resp., (**III**)) indicate where safe (resp., unsafe) development is both the preferred collective outcome and the one selected by social dynamics. Panels **(b)** ($p_r = 0.9$) and **(c)** ($p_r = 0.6$): transition probabilities and stationary distribution in a population of AS, AU, and CS, with $s = 1.5$. AU dominates in panel (c), corresponding to region (**II**), while AS and CS dominate in panel (b), corresponding to region (**I**). We only show the stronger directions. Parameters: $c = 1$, $b = 4$, $W = 100$, $p_{fo} = 0.5$, $B = 10^4$, $\beta = 0.1$, $Z = 100$.

In the current DSAI regime, the first condition occurs when (see Appendix for the proof)

$$p_r > 1 - \frac{1}{s}. \tag{6}$$

That is, when the risk of a personal setback (p_r) is larger than the gain one can get from a greater development speed, then safe development is the preferred collective action in the population, and vice versa.

Analysis of the second question, i.e. when safe (AS) and conditionally safe (CS) strategies are more likely to be imitated, reveals that both are preferred over AU by the social learning dynamics we use here (see risk-dominance analysis in the Appendix) when

$$p_r > 1 - \frac{1}{3s}. \tag{7}$$

The two boundary conditions in Equations 6 and 7 divide the space defined by the speed of development (s) and the risk of disaster (p_r) into three regions, as shown in Figure 2a:

(**I**) when $p_r > 1 - \frac{1}{3s}$: This is the *DSAI compliance zone*, where safe AI development is both the preferred collective outcome and fully safe or conditionally safe behaviour is the social norm (see Figure 2b for an example: for $s = 1.5$ the condition becomes $p_r > 0.78$);

(**II**) when $1 - \frac{1}{3s} > p_r > 1 - \frac{1}{s}$: This intermediate zone captures a dilemma since, collectively, safe AI developments are preferred, yet the social dynamics pushes the population to the state where everyone develops AI in an unsafe manner. We will refer to this zone as the *DSAI dilemma zone* (see Figure 2c for an example: for $s = 1.5$ the condition becomes $0.78 > p_r > 0.33$);

(**III**) when $p_r < 1 - \frac{1}{s}$: This is the *DSAI innovation zone*, where unsafe development is both the preferred collective outcome and the one selected by the social dynamics.

The results visualised in Figure 2 remain present for different parameter settings as is shown in Figure S4 in the Appendix.

As can be observed, in regions (**I**) and (**III**), the preferred collective outcomes are also selected by the social dynamics. Whereas in the DSAI compliance zone, the high risk of disaster motivates participants to adopt a safe strategy even when the final benefit B outweighs marginal benefits per round. In the latter, the DSAI innovation zone, the benefit of quickly reaching DSAI is everything and speed ensures that one arrives first, with limited risk for a setback or even shared disaster (see Appendix). In terms of social welfare, i.e. the average benefits spread over the population, the DSAI innovation zone produces the largest benefits, especially for low risk and high speed combinations (see Appendix, Figure S13). In the DSAI compliance zone, the social welfare is stable no matter the speed, yet lower than in (**III**). Yet switching to unsafe actions here would only lead to a worse outcome, so compliance to safety and ethical regulations is thus required.

Region (**II**), the DSAI dilemma zone, is somewhat peculiar as collective safe behaviour is preferred, yet social dynamics selects for unsafe behaviour. As a consequence, social welfare is lower than what can be seen in the two other zones. Regulation of unsafe behaviour is thus required here as it will nudge the social dynamics towards safe behaviour and, consequently, greater overall social welfare. Such regulation activities will have no effect in the DSAI compliance zone and are potentially detrimental (in terms of the missed social welfare) effects in the DSAI innovation zone. It is therefore essential to know, when the time-scale to reach DSAI is short, what risks can be expected and what speed is acceptable to avoid the DSAI dilemma zone and ensure a positive effect for society.

Looking back at the observation in Figure 1 that in the early DSAI regulation is necessary, the current analysis reveals that this is only a necessity when risk and development speed put the race in the DSAI dilemma zone since the effects would be counterproductive in the two other zones. Yet stimuli to promote risk-taking in the DSAI innovation zone and following safety protocols in the DSAI compliance zone are potentially useful when participants in the race are unsure about the importance of following those actions, i.e. when participants are still exploring and not imitating enough the most beneficial behaviours — expressed by imitation strength β in our model (cf. Figure S4 in Appendix) — in those zones.

Note that the boundaries established by Equations 6 and 7 are applicable for both CS and AS when playing against AU. Thus, similar results are obtained if we consider a

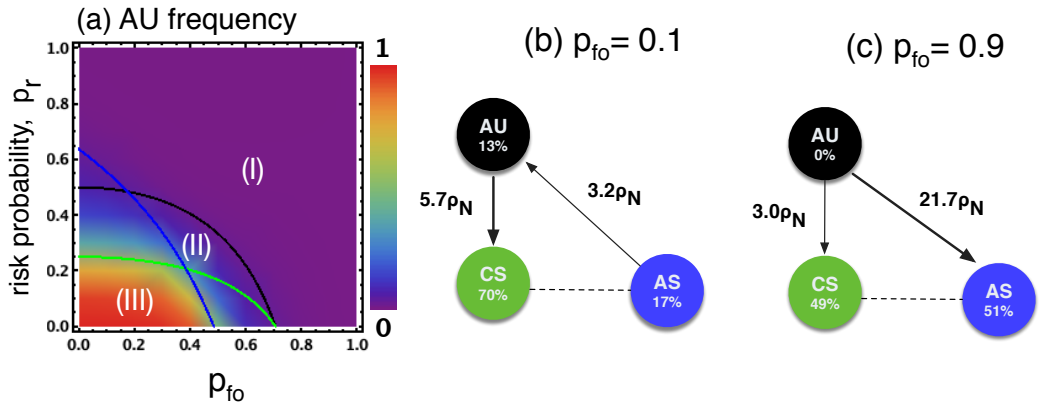

Figure 3: **Late DSAI regime**. **(a)** Frequency of AU as a function of the probability of unsafe development being found out, p_{fo}, and the probability of AI disaster occurring p_r, when the number of development steps to reach DSAI is large ($W = 10^6$). AU has a low frequency whenever p_{fo} or p_r are sufficiently high. The lines indicate the conditions above which safety behavior is the preferred collective outcome (black line) and when AS and CS are risk-dominant against AU (blue and green lines, respectively). CS is risk-dominant for a larger range of p_r than AS for small p_{fo}, which is reversed for large p_{fo}. The numbers refer again to the three zones, i.e. the DSAI compliance, the DSAI dilemma and the DSAI innovation zones. **(b-c)**: transition probabilities and stationary distribution ($p_r = 0.4$). Against AU, AS performs better than CS when p_{fo} is large, which is reversed when p_{fo} is small. Parameters: $c = 1$, $b = 4$, $s = 1.5$, $B = 10^4$, $\beta = 0.1$, $Z = 100$.

population of just two strategies AS and AU (cf. Figure S5 in the Appendix). Adding CS does not change the overall outcome and conditions for safe AI development to be selected. These results also remain unchanged when the risk of setbacks is not just personal, i.e. being shared among the race participants (whether equally or not), as shown analytically in the Appendix (also see Figure S10). The results are furthermore robust to changes in the number of participants in the race. When considering the race among N development teams (see Appendix), the main difference is that the upper bound of region (**II**) increases. That is, the DSAI dilemma zone increases and the DSAI compliance zone disappears. Regulation is thus required for a larger part of the speed-disaster space (cf. Figures S7 and S8 in the Appendix). The reason is, the larger the group size the greater the chance that there is at least one AU player in the group with other AS and CS players, who would then win the development race.

Late DSAI: risk-taking as opposed to safety compliance may need to be promoted

When DSAI is unachievable in the short term, AS and CS are the dominant social norms, as was shown in Figure 1. However, when the probability of disaster is rather small, unsafe behaviour would lead to a relatively greater welfare, yet overall much less than in the early

DSAI regime (see Appendix). In Figure 3, one can again distinguish three zones, i.e. the DSAI compliance, DSAI dilemma and DSAI innovation zones, based on conditions for which safety behaviour is the preferred collective outcome and when AS and CS are risk-dominant against AU (see the black, blue and green lines, respectively, in Figure 3).

In both the late DSAI compliance and late DSAI innovation zones, regulation is not required as before. Although, as also pointed out in the previous section, stimulating a faster acquisition of the required behaviour in those zones can potentially be useful. In the late DSAI dilemma zone, regulation should be put in place to enforce behaviour that improves social welfare. However, in contrast to the early DSAI where safety should be promoted, in this late DSAI regime, unsafe behaviour (speedy innovation) should be promoted to increase social welfare (see Figure S14 in the Appendix). This zone covers the area in-between intermediate p_r with low p_{fo}, and low p_r with intermediate p_{fo}. In both areas decreasing the level of monitoring leads to better social welfare. In the latter where p_r is low, decreasing p_{fo} would move it into the innovation zone. In the former, despite not completely removing the dilemma, decreasing p_{fo} increases the frequency of AU and the overall social welfare. Interestingly, high levels of detection risk removes the dilemma zone, moving both areas into the compliance zone, as also can be observed in Figures S1 and S2 in the Appendix for other parameter settings, yet lower social welfare is obtained. Note however that in the compliance zone where p_r is high, social welfare is highest for intermediate levels of monitoring (see Figure S15 in the Appendix).

As shown in the Appendix, the observations remain valid if, instead of pairwise interactions, one considers a race with $N > 2$ teams in the late DSAI regime, i.e. all three zones reappear. Moreover, when N increases, while the innovation zone size remains unchanged, the DSAI dilemma zone again increases. Also in this case AU becomes the preferred collective outcome for a wider range of p_r and p_{fo} (see Figure S9 in the Appendix). Additionally, when the risk of disaster is not just personal but is rather shared among the race participants, we observe that the preference boundary between collectively safe and unsafe behaviour remains the same yet the individual preference towards risky development increases, i.e. the innovation zone becomes larger while the dilemma zone becomes smaller and disappears (see Figure S11 in the Appendix). That is, shared risk in the late DSAI regime improves the overall social welfare (by allowing more beneficial innovation to happen), reducing the need for regulatory actions to handle the late DSAI dilemma zone.

4. Discussion

This paper studies the dynamics associated with an alleged race for leadership in a domain using AI and its associated technologies. The model applies to other innovation dynamics in which the objective is to become the first to bring the technology to market. We focus on the conflict between safety compliant and rapid risky development (Denicolò & Franzoni, 2010; Abbott et al., 2009), assuming that only the actions of the participant influence the outcome and that the time-window associated with this race is unknown. This problem is examined through a multiagent/complex systems approach, adopting well-established methods from population dynamics and EGT to achieve the goals. The adoption of practices originated from evolutionary biology, when combined with social learning dynamics, allows us to grasp the full complexity of the ecology of choices in innovation dynamics without a single pre-

defined path. Akin to other evolutionary races, see e.g., the Red Queen Hypothesis (Leigh, 1973), individuals adapt their choices while facing ever-evolving populations of opponents. As a result of this (open-ended) socio-technical dynamics, our analytical model describes the long-term prevalence of each strategy in time.

Our results reveal that knowing the exact timing of reaching DSAI in a domain is not crucial, only whether it can be achieved early or late, as this will influence what regulations are potentially suitable. We identified three different DSAI zones in both the early and late regimes, i.e. the safety compliance, the dilemma and the innovation zones. They are respectively characterised by high risk, intermediate risk and low risk for personal as well as shared setbacks. In the compliance and innovation zones, regulatory actions that reverse the behaviour selected by social dynamics should be avoided, as they would be detrimental to the overall social welfare. Stimulating, on the other hand, a faster acquisition of the required behaviour in those zones can potentially be useful. In the dilemma zone, however, regulatory actions promoting the collectively beneficial outcome are essential since the behaviour selected by social dynamics goes against society's interest, lowering social welfare. In this DSAI zone the social dynamics is selecting for (undesired) behaviour, requiring regulation of risk-taking in the early DSAI and safety compliance in the late DSAI.

We show furthermore, both in the early and late regimes, that although the three DSAI zones are determined by similar ranges of the risk for setbacks (p_r), they differ in the secondary factors that control the extent of these zones. While in the early DSAI, speedy development (s) is everything, the race outcome in the late DSAI is mainly determined by the efficiency and level of monitoring of unsafe behaviour (p_{fo}). Although speed in the early regime appears to handle some levels of disaster risk, it may lead participants to enter the dilemma zone where individual interests counter societal welfare, and this area increases in function of the number of participants in the race. As is shown in Figure S2, speed does not influence the regions in the late DSAI regime. The risk of being detected actually limits unsafe behaviour to the area of low risk situations. Yet more participants will increase again the area (see Figure S9) as well as sharing the effects of a disaster (see Figure S11). It appears thus that holding unsafe players responsible for bad outcomes of the DSAI race will ensure, at least in the late regime, that unsafe actions remain limited. Moreover, the presence of conditionally safe players, i.e. the threat that others may also start behaving unsafely, limits the unsafe actions to lower risk areas.

Moreover, one should consider the possibility that the risk of being identified as an unsafe player may not just affect a single development round, but may also have repercussions on subsequent rounds, i.e. the unsafe player may also loose b for instance in all subsequent rounds. As shown in the Appendix the results remain the same in early DSAI, while in the late DSAI, the outcome is equivalent to the results one obtains when full monitoring (i.e. $p_{fo} = 1$) is in effect. Intuitively, longer consequences associated with being detected is equivalent to having a higher probability of being detected in each round in the current DSAIR model.

Lessons-learned box: This box summarises some essential observations that may be useful to researchers working on regulating the use of AI in real-world application domains.

Introducing novel technologies like AI-supported products in society is considered to be lucrative for many stake-holders. This worldwide potential is used to urge people to transform their business into one supported by AI and related technologies. In a competitive market, and relative to the anticipated gain, persons or groups will be more or less eager to enter. As there is a fear of missing out or not being in the lead, a race may be perceived by the stakeholders wherein unsafe actions (no proper testing, hackable software, data leaks, etc.) are considered acceptable risks in order to be first or get the biggest share of the pie. Regulatory bodies and policy makers are aware and are proposing regulatory actions. Yet regulation may have no effect or inverse effects when there is no complete understanding, nor data, of the technology race. As put forward by the Collingridge Dilemma, the impact of new technology is difficult to predict unless large steps have been taken in its development and it becomes generally adopted.

To introduce effective regulations, a model is therefore required that can show the effect of some of the main parameters driving such a race. Here, we identify the conditions in which regulation may be required given the perceived risks and gains, and which type of actions, risk taking or compliance to safety requirements, should be stimulated. Our idealised model and associated analysis reveal that the **timeline** in which one can reach the supremacy in a development race/competition in a given domain determines **(a)** what behaviour needs to be promoted to ensure the maintenance of societal welfare; and **(b)** when regulatory actions are required that promote benefits for society.

- For **(a)**, we show that for an early/short-term timeline, safety compliance may need to be promoted when risks of setbacks are intermediate, while in a late/long-term timeline, innovative, risk-taking behaviour will improve the benefits to society (Figure 1c).
- To support **(b)**, for both timelines, our analysis shows that only under the conditions that define a dilemma zone (Figures 2a and 3a) regulation to promote the corresponding behaviour as determined in **(a)** is necessary. Outside these regions, there is either no effect or regulations can even be detrimental for advancements and society.

Uninformed promotion of compliance or risk-taking behaviour without knowing the timeline and associated dilemma zones, may lead to unwanted consequences, such as disastrous outcomes for businesses and society by promoting risk-taking in the early timeline and over-regulation of innovation when enforcing compliance to extensive protocols in the late timeline. It may be useful to first test regulatory ideas within the context of abstract models, or an extension of it, to fully grasp what the effect of certain regulatory decisions.

The DSAIR model and associated analysis provides thus an instrument for researchers interested in AI regulation and policy making to think about the supporting mechanisms (such as suitable rewards and sanctions) (Sigmund, 2010; Sotala & Yampolskiy, 2014; Szolnoki & Perc, 2013; Han, Pereira, & Lenaerts, 2015, 2019; Vinuesa et al., 2020) needed to mediate a given race; for preliminary results, see our recent work in (Han et al., 2020). In the early DSAI, controlling the development speed of AI teams appears essential. Yet, policy researchers should carefully consider whether it will have the expected outcome, i.e. whether the race is actually occurring in the DSAI dilemma zone. In the late DSAI, monitoring was perceived to be crucial. Decreasing the level of monitoring can reduce the

dilemma zone and increases social welfare, increasing speedy innovation. Intermediate levels of monitoring lead to highest social welfare in the compliance zone. We summarised the lessons learnt from our modelling and analysis in a *Lessons-learned box*.

In particular, consider the recently published (in February 2020) White Paper on AI by the European Commission (European Commission, 2020), our study can be applied and provide insights to shape the future EU regulatory framework on products and services relying on AI (see Section C of the White Paper)—particularly, in determining the scope of its application. The Commission considers a risk-based approach to determine when an AI application is high-risk, in light of what is at stake, and therefore needs to be targeted with regulatory actions. It focuses on whether the sector/domain and the intended use involve significant risks. An AI application can be considered to be high-risk depending on the sector it is being used in. Our results show that, in order to determine the scope of regulation, the timeline of the technology to reach supremacy also needs to be taken into account.

There are of course limitations to the current model and different simplifying assumptions were made, which will require further analysis. A first simplifying assumption is that a higher speed can only be achieved by cutting ethical/safety corners. While it is a natural one to make when agents are homogeneous in terms of their resources and wealth, it may not remain true once agents become heterogeneous. For example, some teams may revise their choices or act as role models more often than others (Santos, Pacheco, & Lenaerts, 2006; Santos, Santos, & Pacheco, 2008), or react to uncertainty in different ways, leading to polarized behaviors (Domingos et al., 2020; Ross & Portugali, 2018). One may also consider that stronger teams with more resources at hand or having more support from others might ensure a greater speed than those without such advantage, even without cutting ethical/safety corners. Under wealth inequality (see, e.g., (Tavoni et al., 2011; Vasconcelos et al., 2014)), these stronger teams might be able to comply regardless of the competition from weaker teams and still win the race. As the objective of the current paper is to propose and examine an initial model, these issues were not examined and were left for future work. One might also consider whether regulatory mechanisms like a speed penalty proportionate to a participant's resources would benefit society, since greater resources entail accrued responsibility.

A second simplification, mentioned also in the introduction, is that safety compliant actions do not lead to a disaster. Realistically, on the one hand, even when someone aims to comply, they may make mistakes when implementing a certain safety or ethical guideline, whether or not being aware of it. A consequence is that, disaster will always be possible even for unconditionally compliant players, impacting thus all types in the same manner. On the other hand, one could argue that, regardless of individual mistakes, SAFE actions of highly regulated and safety conscious industries may also result in some unforeseen failure. Including this possibility would require the introduction of an additional parameter (p_g), different from p_r, that expresses some general risk of failure. This general risk will nonetheless be lower than the risk after taking UNSAFE actions, as clearly the risk for a disaster increases with how well safety regulations are followed. Moreover, as this general risk affects every player type, meaning that the effect on their payoffs is equivalent, the payoff matrix in Equation 2 would become $(1-p_g)\Pi$. Thus, the introduction of this general risk would be equivalent to having a scaled (namely, weaker) imitation strength $(1-p_g)\times\beta$.

We have shown that our results are robust for different values of β, both analytically (see Equation 31) and numerically (in the Appendix, Figures S1-5, S11 and S14). A weaker imitation strength will decrease the frequency difference between all behaviour types, but the results we presented regarding time-scales and behaviour zones, will remain unchanged.

In addition, safely developed products may also be misused by a third party. Such an externality is not considered in the current idealised model, yet could be of interest for regulatory agencies. Such errors (Nowak & Sigmund, 1993; Sigmund, 2010; Han, 2013; Van Segbroeck et al., 2012) as well as participant heterogeneity (Hauser et al., 2019; Vasconcelos et al., 2014) have been shown to play an important role in long-term interactions in the context of social dilemmas, which can also be found in the current DSAIR model. These additions can be explored in future work, either by ourselves or teams interested by this model for their own purposes.

Next to these simplifying assumptions other limitations were introduced. On the one hand, the effect of unsafe behavior on W has not been considered. It may well be that accumulated detected unsafe behaviour, whether by a single player or jointly accumulated by a number of them, may expand the time necessary to reach the DSAI, thus effectively increasing W. Moreover, the time to reach DSAI in a domain may also be affected by the trust that people have in AI techniques (Andras et al., 2018), even when deliberate unsafe behaviour is not the issue. Rhetoric and framing of the AI development race and how close it is to achieve the AGI might strongly influence the dynamics and outcome of the race (Cave & ÓhÉigeartaigh, 2018; Baum, 2017; Cave et al., 2020). In future work, such phenomena should be examined and introduced on top of the base model presented.

One the other hand, the model also did not consider that to achieve DSAI in some domain, the results of multiple races may need to be combined. Here long-term targets like AGI are considered to be achievable in one race. Clearly AGI will require solutions to multiple subproblems, which by themselves may be achieved in development efforts occurring at different time scales. It is even more relevant to consider separate domains or sectors given the fact that a technology might have different levels of risk depending on where it is being used. This issue has been highlighted in the White Paper on AI by the European Commission (European Commission, 2020), where a risk-based approach was considered for determining the scope of regulatory actions to be applied. Future models of DSAIR will thus need to consider that multiple DSAIR games to study what regulatory actions are most beneficial for this kind of goals.

In conclusion, we have provided here a first plausible DSAIR model directly useful for researchers interested in AI regulation and policy making to evaluate the risks associated with the ongoing AI development and applications race, and have shown and analysed its reasonably acceptable behavioural consequences. Our results indicate the crucial need of clarifying the time-scale of digital innovation supremacy and the risks in relation to ignoring safety and ethical precautions in speeding up innovation, in order to determine suitable regulations of AI safety behaviour beneficial for all.

5. Acknowledgements

T.A.H., L.M.P. and T.L. are supported by Future of Life Institute grant RFP2-154. L.M.P. is supported by NOVA LINCS (UIDB/04516/2020) with the financial support of FCT-

Fundação para a Ciência e a Tecnologia, Portugal, through national funds. F.C.S. acknowledges support from FCT Portugal (grants PTDC/EEI-SII/5081/2014, PTDC/MAT/STA/3358/2014, and UIDB/50021/2020). T.L. acknowledges support of the F.N.R.S. project with grant number 31257234 and the FuturICT2.0 (www.futurict2.eu) project funded by the FLAG-ERA JCT 2016.

Appendix A. Deriving Conditions for Viability of Safety Behaviour

When Safety Behaviour is The Preferred Collective Outcome

We derive analytical condition for which a population of players always following safety precautions has a greater social welfare (i.e. average payoff) than that of a population of players never following safety precautions, that is, $\Pi_{AS,AS} > \Pi_{AU,AU}$:

$$\frac{B}{2W} + \pi_{11} > (1 - p_r)\left(\frac{sB}{2W} + \pi_{22}\right). \tag{8}$$

Thus,

$$p_r > 1 - \frac{B + 2W\pi_{11}}{sB + 2W\pi_{22}}. \tag{9}$$

Following the definitions of different DSAI regimes in the main texts, we simplify this condition for the two regimes. First, in the **early DSAI regime** where $B/W \gg b$, Equation 9 is equivalent to

$$p_r > 1 - \frac{1}{s}. \tag{10}$$

Now, in the **late DSAI regime** where $W \to \infty$ (i.e. $B/W \ll b$), Equation 9 is equivalent to:

$$p_r > 1 - \frac{\pi_{11}}{\pi_{22}} = 1 - \frac{b - 2c}{b(1 - p_{fo}^2)}. \tag{11}$$

We can see that the development speed (s) is the crucial factor in the early DSAI regime while it does not play any role in the late DSAI, where for fixed b and c, p_{fo} is the only influencing factor.

When Safety Behaviour is Selected by Evolution

We now derive conditions for which AS and CS are risk-dominant against AU, which are the case if and only if, respectively,

$$\frac{B}{2W} + \pi_{11} + \pi_{12} > (1 - p_r)\left(\frac{3sB}{2W} + \pi_{21} + \pi_{22}\right). \tag{12}$$

$$\frac{s}{W}\left(\pi_{12} + (\frac{W}{s} - 1)\pi_{22}\right) + \frac{B}{2W} + \pi_{11} > (1 - p_r)\left[\frac{sB}{2W} + \frac{sB}{W} + \frac{s}{W}\left(\pi_{21} + (\frac{W}{s} - 1)\pi_{22}\right) + \pi_{22}\right]. \tag{13}$$

In the **early DSAI regime** where $B/W \gg b$, both equations are simplified to

$$p_r > 1 - \frac{1}{3s}. \tag{14}$$

On the other hand, in the **late DSAI regime** where $W \to \infty$ (i.e. $B/W \ll c$), they are simplified to, respectively

$$\pi_{11} + \pi_{12} > (1 - p_r)(\pi_{21} + \pi_{22}), \tag{15}$$

$$\pi_{11} > (1 - 2p_r)\pi_{22}, \tag{16}$$

which are equivalent to, respectively

$$p_r > \frac{4c(1+s) - b\left(2 + p_{fo}^2 + (-2 + p_{fo}(4 + p_{fo}))s\right)}{b(1 - p_{fo})(1 + p_{fo} + (3 + p_{fo})s)}, \tag{17}$$

$$p_r > \frac{1}{2} - \frac{b - 2c}{2b(1 - p_{fo}^2)}. \tag{18}$$

Thus, for safety behaviour to be both selected and the preferred outcome, p_r must satisfy all the conditions in Equations (18), (17) and (11).

It is clear to see that the right hand sides of Equations (18) and (11) are decreasing functions of p_{fo} whenever $b \geq 2c$. We now show that it is also the case for the right hand side of Equation 17. Indeed, its first order derivative by p_{fo} gives

$$-\frac{2(1+s)\left[b\left(4s + p_{fo}^2 s + p_{fo}(3+s)\right) - 4c(p_{fo} + s + p_{fo}s)\right]}{b(1 - p_{fo})^2(1 + p_{fo} + 3s + p_{fo}s)^2},$$

which is negative whenever $b \geq 2c$ because

$$\left(4s + p_{fo}^2 s + p_{fo}(3+s)\right) - 2(p_{fo} + s + p_{fo}s) = 2s + p_{fo}^2 s + p - p_{fo}s > 0.$$

In short, we have shown that for $b \geq c$, the larger p_{fo} the easier the conditions for the safety behaviour to be both selected and the preferred outcome. Figure S2 validates these observations numerically. Similarly, we also can show that these conditions are harder to achieve the larger s is.

Thus, the hardest conditions are obtained when $p_{fo} = 0$, which is equivalent to

$$p_r > \max\{1 - \frac{(b-2c)(s+1)}{2sb},\ \frac{4c(s+1) + 2b(s-1)}{b(1+3s)},\ \frac{c}{b}\}. \tag{19}$$

It is easily seen that the right hand side is greater than 1 iff $b < 2c$, i.e. this condition would not be achieved (since $p_r \leq 1$) in that case. Assuming $b \geq 2c$, since $\frac{4c(s+1)+2b(s-1)}{b(1+3s)} > 1 - \frac{(b-2c)(s+1)}{2sb} > \frac{c}{b}$, it can be further simplified to

$$p_r > \frac{4c(s+1) + 2b(s-1)}{b(1+3s)}, \tag{20}$$

which is the condition for AS to be risk-dominant against AU (see Figure S2 for an example when $s = 1.5$).

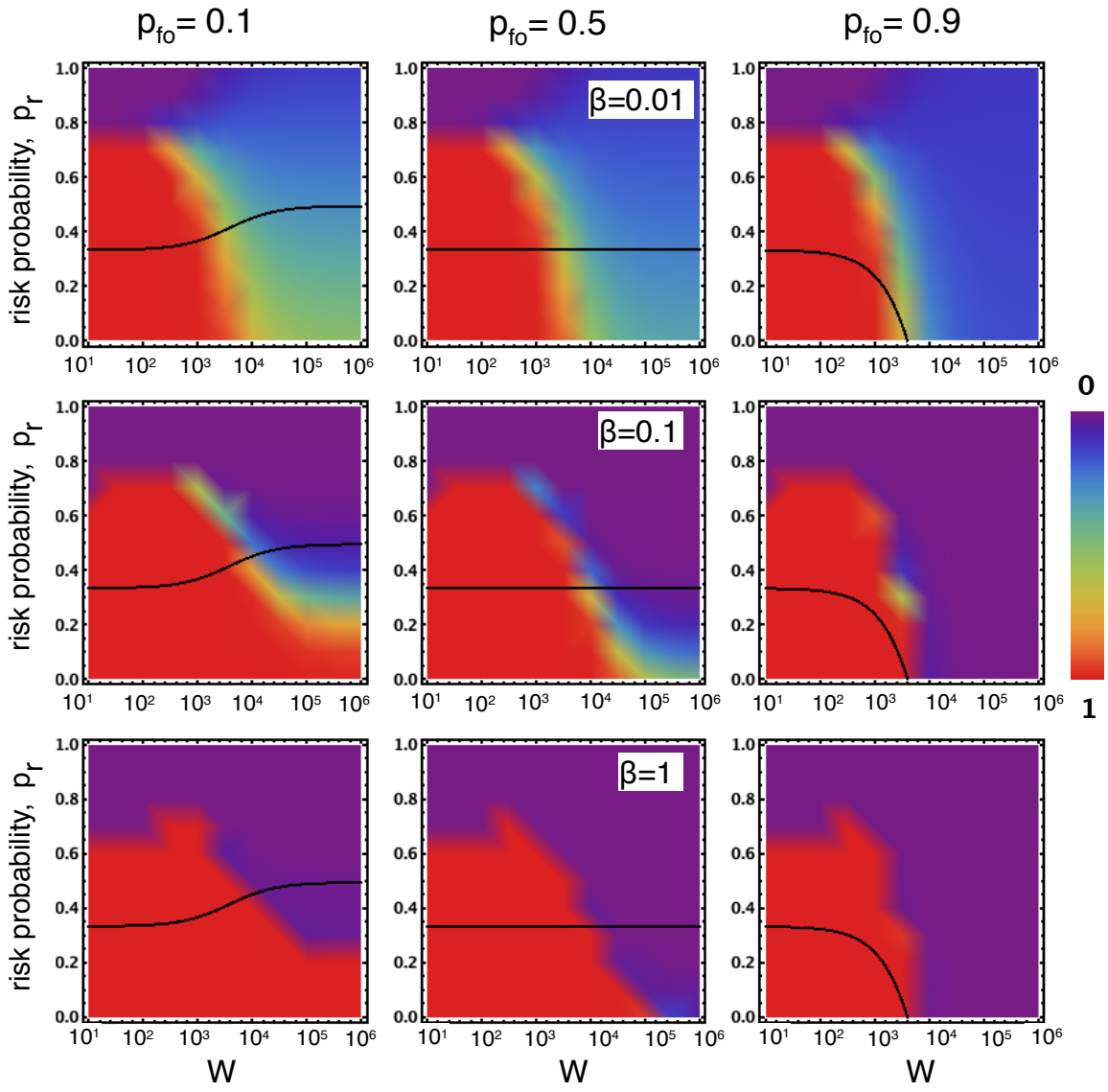

Figure S1: **Across DSAI regimes. Frequency of AU as for varying p_r and different values of p_{fo} and β**: when W is small (early DSAI) vs when W is large (late DSAI). $\beta = 0.01,\ 0.1,\ 1$ for top, middle and bottom rows, respectively. The *black lines* indicate the threshold of p_r above which SAFE is the preferred collective action and below which UNSAFE is the preferred one (see Equation 9). In general, we observe that AU is dominant for a larger range of p_r in the early than the late regime. Parameters: $c = 1$, $b = 4$, $s = 1.5$, $B = 10^4$, $Z = 100$.

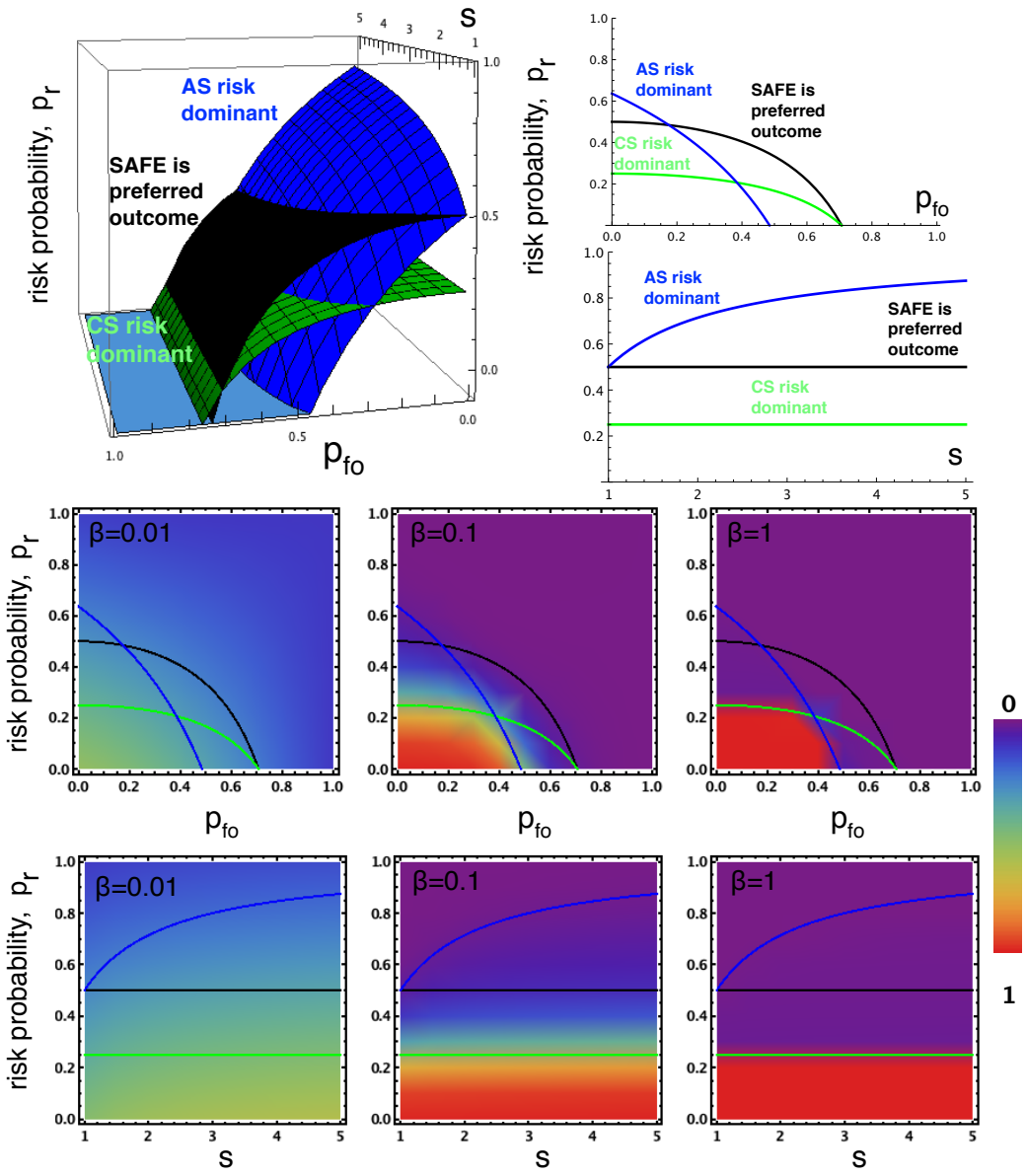

Figure S2: **Late DSAI** ($W = 10^6$). The curves/lines indicate the conditions above which safety behavior is the preferred collective outcome (black ones) and when AS and CS are risk-dominant against AU (green and blue ones, respectively). The threshold for AS is greater than than CS when p_{fo} is small, which is reversed when p_{fo} is large **(Top row)**. **(Middle and bottom rows)** Frequency of AU as a function of p_r and p_{fo} (bottom: $s = 1.5$) or s (middle: $p_{fo} = 0$), respectively, for different values of β. AU has high frequencies in regions below both the blue and green lines, especially for larger β. Parameters: $c = 1$, $b = 4$, $B = 10^4$, $Z = 100$.

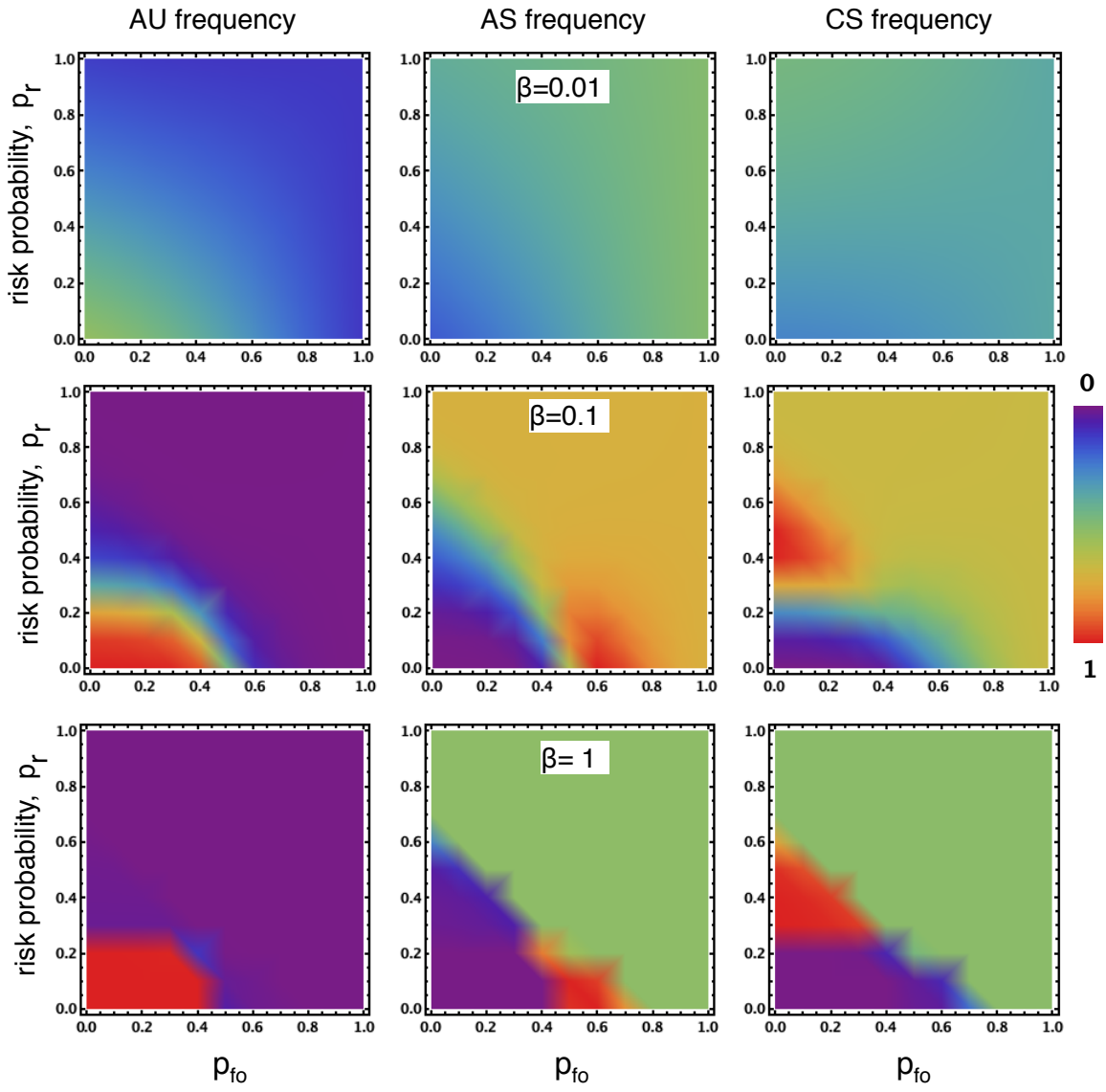

Figure S3: **Late DSAI: Frequency of AU, AS and CS** as a function of the probability of unsafe development being found out, p_{fo}, and the probability of AI disaster occurring p_r, when the number of development steps to reach DSAI is very large ($W = 10^6$). $\beta = 0.01,\ 0.1,\ 1$ for top, middle and bottom rows, respectively. AU has a low frequency whenever p_{fo} or p_r are sufficiently high. AS performs best when p_{fo} is large. Parameters: $c = 1$, $b = 4$, $s = 1.5$, $B = 10^4$, $Z = 100$.

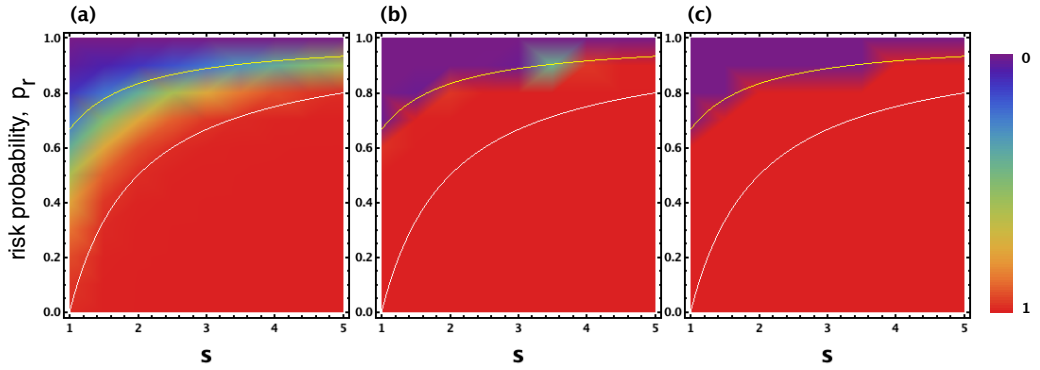

Figure S4: **Early DSAI: Frequency of AU in a population of three strategies, AS, AU, and CS**, as a function of the speed gained when ignoring safety, s, and the the risk probability p_r. In general, we observe that when the risk probability is small, AU is dominant. Also, the larger B and s, AU dominates for a larger range. The two solid lines inside the plots indicate the boundaries $p_r \in [1-1/(3s), 1-1/s]$ where safety development is preferred but non-safety development is preferable (risk-dominant against CS and AS). The observations are valid for varying the selection intensities: $\beta = 0.001,\ 0.01,\ 0.1$ for panels (a), (b) and (c), respectively. Other parameters: $c = 1$, $b = 4$, $W = 100$, $p_{fo} = 0.5$, $B = 10^4$, $Z = 100$.

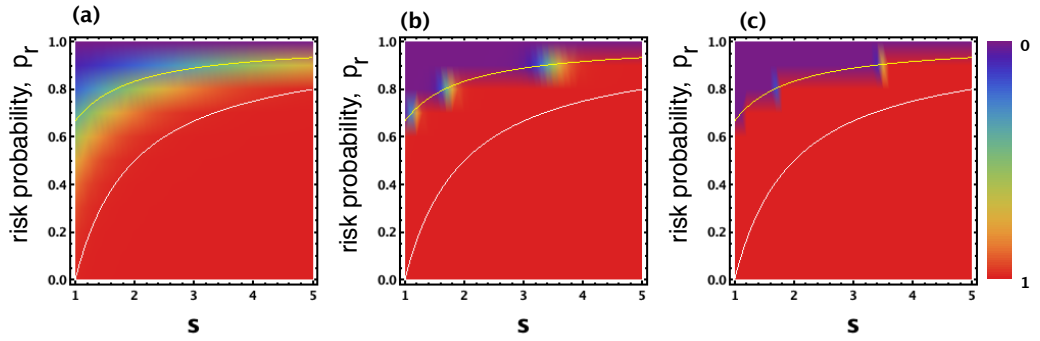

Figure S5: **Early DSAI: Frequency of AU in a population of two strategies, AS and AU**, as a function of the speed gained when ignoring safety, s, and the the risk probability p_r. In general, we observe that when the risk probability is small, AU is dominant. Also, the larger B and s, AU dominates for a larger range. The two solid lines inside the plots indicate the boundaries $p_r \in [1-1/(3s), 1-1/s]$ where safety development is preferred but non-safety development is preferable (risk-dominant against CS and AS). The observations are valid for varying the selection intensities: $\beta = 0.001,\ 0.01,\ 0.1$ for panels (a), (b) and (c), respectively. Other parameters: $c = 1$, $b = 4$, $W = 100$, $p_{fo} = 0.5$, $B = 10^4$, $Z = 100$.

Appendix B. Multiplayer AI Race

In this section we describe the N-team model of the AI race, extending the two-team model in the main text. We then describe the Methods used for analysing multi-player games.

N-player AI Race Definition

The AI development race is modeled as a repeated N-player game, consisting of W development rounds. In each round, the players can collect benefits from their intermediate AI products, depending on whether they choose to play SAFE or UNSAFE. Assuming a fixed benefit, b, from the AI market, teams will share this benefit proportionally to their development speed. Moreover, we assume that with some probability p_{fo} those playing UNSAFE might be found out [1]about their unsafe development and their products won't be used, leading to 0 benefit.

In a group of where k players choosing SAFE and $(N-k)$ choosing UNSAFE, the payoffs for players adopting SAFE and UNSAFE in each round of the race are, respectively

$$\pi(k)_{SAFE} = \begin{cases} -c + (1-p_{fo})\frac{b}{k+s(N-k)} + p_{fo}\frac{b}{k} & \text{if } \ 1 \leq k < N \\ -c + \frac{b}{N} & \text{if } k = N \end{cases}$$

$$\pi(k)_{UNSAFE} = (1-p_{fo})\frac{sb}{k+s(N-k)} \text{ for } 0 \leq k < N$$

We consider a well-mixed, finite population of size Z, where players repeatedly interact with each other in the AI development process, using one of the following three strategies :

- AS (always complies with safety precaution)
- AU (never complies with safety precaution)
- CS (conditionally safe, plays SAFE in the first round; then plays SAFE if everyone in the group plays SAFE in the previous round and plays UNSAFE otherwise)

The average payoffs for the repeated games (k denotes the number of AS or CS when playing with AU)

$$\Pi_{AS,AU}(k) = \begin{cases} \pi(k)_{SAFE} & \text{if } \ 1 \leq k < N \\ \frac{B}{NW} + \pi(N)_{SAFE} & \text{if } k = N \end{cases}$$

$$\Pi_{AU,AS}(k) = p\left(\frac{sB}{W(N-k)} + \pi(k)_{UNSAFE}\right) \text{for } 0 \leq k < N$$

$$\Pi_{CS,AU}(k) = \begin{cases} \frac{s}{W}\left(\pi(k)_{SAFE} + (\frac{W}{s}-1)\pi(0)_{UNSAFE}\right) & \text{if } \ 1 \leq k < N \\ \frac{B}{NW} + \pi(N)_{SAFE} & \text{if } k = N \end{cases}$$

$$\Pi_{AU,CS}(k) = p\left[\frac{sB}{W(N-k)} + \frac{s}{W}\left(\pi(k)_{UNSAFE} + (\frac{W}{s}-1)\pi(0)_{UNSAFE}\right)\right] \text{for } 0 \leq k < N$$

1. For simplicity of calculation, we assume that all the UNSAFE players will be found out or not together, e.g. whenever investigation is done then they are found out; otherwise they are not.

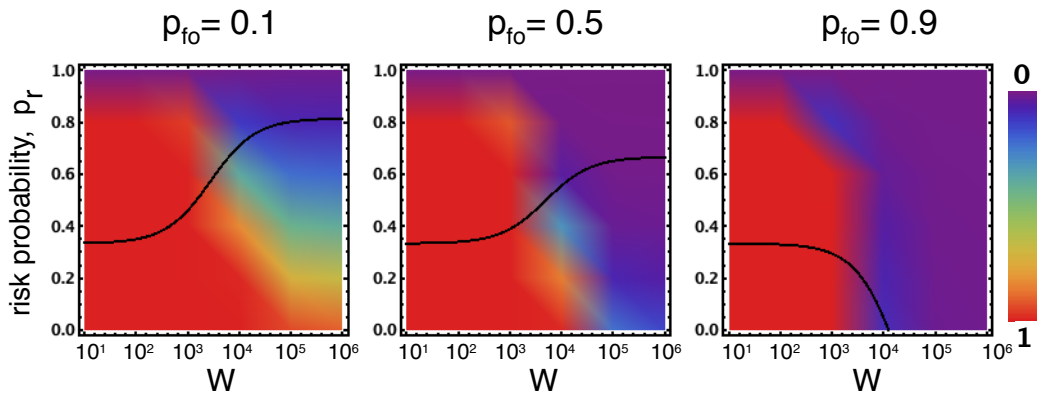

Figure S6: **Different regimes of DSAI: early DSAI (small W) vs late DSAI (large W), in multi-team game**. Frequency AU in a population of the three strategies AS, AU and CS in co-presence, as a function of p_r and W. The black lines indicate the conditions above which SAFE is the preferred collective outcome and below which UNSAFE is (see Equation 21). Other parameters: $c = 1$, $b = 6$, $s = 1.5$, $B = 10000$, $N = 5$, $Z = 100$, $\beta = 0.1$.

Analytical Conditions and DSAI Zones in N-team Interactions

First, the condition for $\Pi_{AS,AU}(N) > \Pi_{AU,AS}(0)$, ensuring that a population of players following safety precautions has a greater social welfare or average payoff than that of a population of players never following safety precautions, reads

$$\frac{B}{NW} + \pi(N)_{SAFE} > (1 - p_r)\left(\frac{sB}{NW} + \pi(0)_{UNSAFE}\right).$$

It can be rewritten as

$$p_r > 1 - \frac{B + W(b - Nc)}{sB + W(1 - p_{fo})b}. \tag{21}$$

In **early DSAI** (i.e. $B/W \gg b$), it is equivalent to:

$$p_r > 1 - \frac{1}{s}. \tag{22}$$

which is exactly the same as the condition for pairwise game, and does not depend on the group size N.

While in **late DSAI** (i.e. $B/W \ll b$), it is equivalent to:

$$p_r > 1 - \frac{b - Nc}{(1 - p_{fo})b}. \tag{23}$$

It can be seen that, for this condition to happen in the late DSAI, it is necessary that $b > Nc$. Moreover, the left hand size is an increasing function of N (compare the black lines in Figure S9 for different values of N).

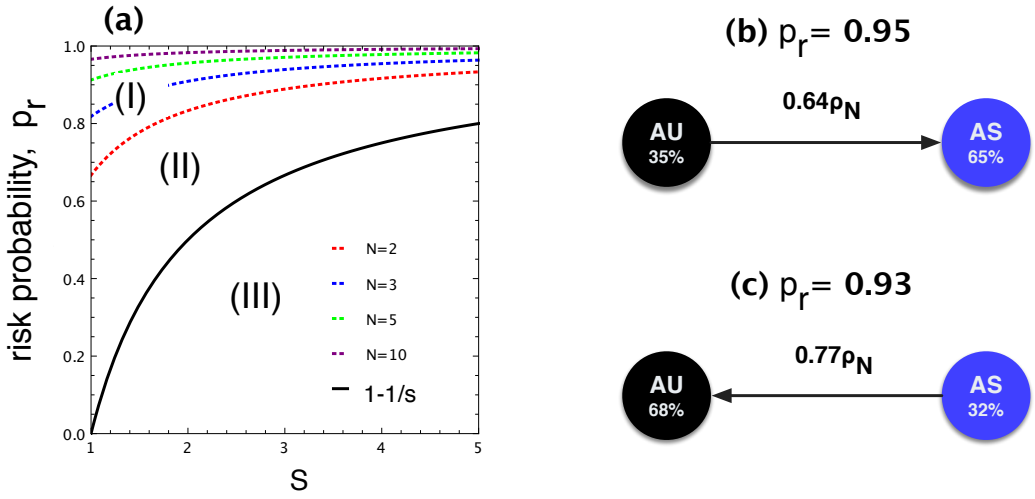

Figure S7: **Early DSAI zones in N-team interactions.** Dotted lines indicate the condition in Equation 26 for different values of group size N. The solid black line indicates the condition in 21. The larger N the larger the region (II) and smaller the region (I). In panels (b), (c): $N = 5$. Other parameters: $c = 1$, $b = 4$, $W = 100$, $s = 1.5$, $p_{fo} = 0.5$, $B = 10^4$, $Z = 100$.

Figures S6 shows the results for N-player games across different regimes of DSAI (i.e. varying W). Similar observation is obtained as in the pairwise game in the main text.

Risk-dominance of AS and CS against AU: On the other hand, AS and CS are risk-dominant against AU, respectively, iff

$$\sum_{k=0}^{N-1} \pi(k)_{AU,AS} < \sum_{k=1}^{N} \pi(k)_{AS,AU} \tag{24}$$

$$\sum_{k=0}^{N-1} \pi(k)_{AU,CS} < \sum_{k=1}^{N} \pi(k)_{CS,AU} \tag{25}$$

In the **early DSAI** (i.e. $B/W \gg b$), both conditions are reduced to

$$p_r > 1 - \frac{1}{(NH_N)s}. \tag{26}$$

where $H_N = \sum_{i=1}^{N} 1/i$. Since $H_N > \log N$ we can see that the left hand side of the inequality approaches 1 for increasingly large group size, $N \to \infty$.

Thus, the two boundary conditions in Equations 22 and 26 divide the parameter space s-p_r into three regions, see Figure S7a: (**I**) when $p_r > 1 - \frac{1}{(NH_N)s}$: safety development is both the preferred collective outcome and selected by evolution (see Figure S7b for an example: for $s = 1.5$ the condition becomes $p_r > 0.94$); (**II**) when $1 - \frac{1}{(NH_N)s} > p_r > 1 - \frac{1}{s}$:

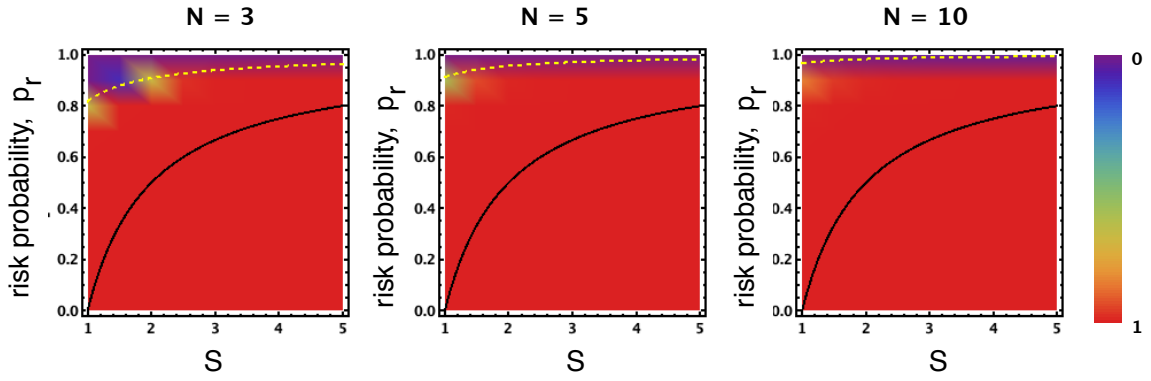

Figure S8: **Early DSAI.** Frequency of AU as a function of the speed gained, s, and the probability of AI disaster occurring p_r, when ignoring safety. Other parameters: $c = 1$, $b = 4$, $W = 100$, $s = 1.5$, $p_{fo} = 0.5$, $B = 10000$, $Z = 100$.

although it is more desirable to ensure safety development as the collective outcome, natural selection/social learning would drive the population to the state where safety precaution is mostly ignored (see Figure S7c for an example: for $s = 1.5$ the condition becomes $0.94 > p_r > 0.33$); (**III**) when $p_r < 1 - \frac{1}{s}$, unsafe development is both the preferred collective outcome and selected by evolution. Numerical results (cf. Methods below) in Figure S7 confirm this division of the regions.

We observed that, the larger s is, the greater the threshold for p_r. Moreover, a larger group size leads to a larger region (II) – AU is selected for a larger range of the parameter space s-p_r. The reason is that, the larger the group size, the greater the chance that there is at least one AU player in the group (with other AS/CS players), who would win the development race.

Now, for the **late DSAI**, the conditions AS and CS are reduced to

$$p_r > 1 - \frac{\sum_{i=1}^{N} \pi(i)_{SAFE}}{\sum_{i=0}^{N-1} \pi(i)_{UNSAFE}}, \tag{27}$$

$$p_r > 1 - \frac{(N-1)\pi(0)_{UNSAFE} + \pi(N)_{SAFE}}{N\pi(0)_{UNSAFE}} = \frac{1}{N}\left(1 - \frac{\pi(N)_{SAFE}}{\pi(0)_{UNSAFE}}\right). \tag{28}$$

Methods: Payoffs Over Group Samplings

In finite populations, the groups engaging in a N-player game are given by multivariate hypergeometric sampling. For transition between two pure states (small mutation), this reduces to sampling (without replacement) from a hypergeometric distribution (Hauert et al., 2007; Sigmund, 2010). Namely, in a population of size Z with x individuals of type i and $Z - x$ individuals of type j, the probability to select k individuals of type i and $N - k$

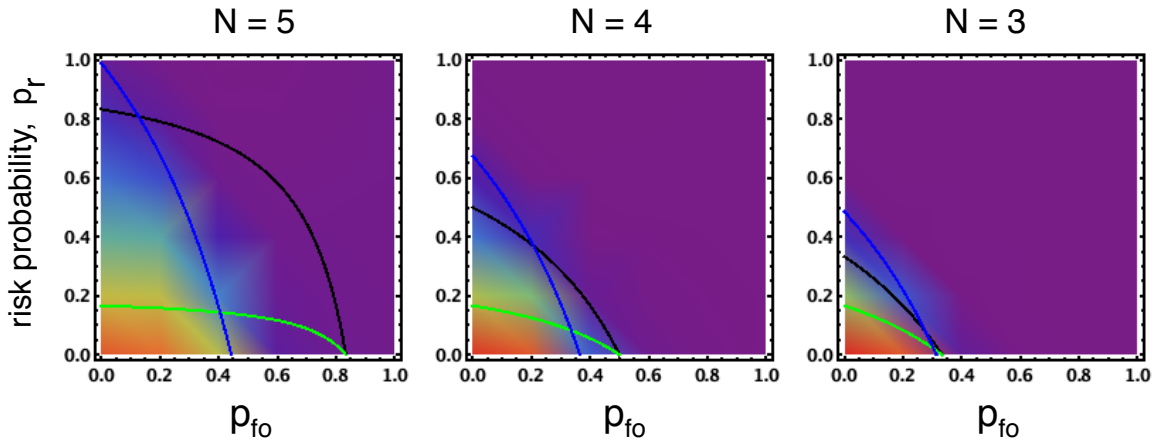

Figure S9: **Late DSAI in N-player interactions.** Frequency of AU as a function of p_{fo} and p_r for different competition size N. The three lines indicate the conditions as in the main texts (Figure 3). The size of the innovation zone is quite similar for different N, but since the larger N the larger the region below the black line (see also analysis), the size of the dilemma zone is increased. Other parameters: $c = 1$, $b = 6$, $s = 1.5$, $W = 10^6$, $B = 10000$, $\beta = 0.1$, $Z = 100$.

individuals of type j in N trials is (Hauert et al., 2007)

$$H(k, N, x, Z) = \frac{\binom{x}{k}\binom{Z-x}{N-k}}{\binom{Z}{N}}.$$

Recall that $\Pi_{ij}(k)$ and $\Pi_{ji}(k)$ (see the section above) denote the payoff of a strategist of type i and j, respectively, when the random sampling consists of k players of type i and $N - k$ players of type j (as derived above). Hence, in a population of x i-strategists and $(Z - x)$ j-strategists, the average payoffs to i and j strategists are (Hauert et al., 2007;

Sigmund, 2010; Pacheco, Santos, Souza, & Skyrms, 2009):

$$\begin{aligned} P_{ij}(x) &= \sum_{k=0}^{N-1} H(k, N-1, x-1, Z-1)\ \Pi_{ij}(k+1) \\ &= \sum_{k=0}^{N-1} \frac{\binom{x-1}{k}\binom{Z-x}{N-1-k}}{\binom{Z-1}{N-1}}\ \Pi_{ij}(k+1), \\ P_{ji}(x) &= \sum_{k=0}^{N-1} H(k, N-1, x, Z-1)\ \Pi_{ji}(k) \\ &= \sum_{k=0}^{N-1} \frac{\binom{x}{k}\binom{Z-1-x}{N-1-k}}{\binom{Z-1}{N-1}}\ \Pi_{ji}(k). \end{aligned} \tag{29}$$

Now, the probability to change the number k of agents using strategy i by ± 1 in each time step can be written as (Traulsen et al., 2006)

$$T^{\pm}(k) = \frac{Z-k}{Z}\frac{k}{Z}\left[1 + e^{\mp\beta[P_{ij}(k) - P_{ji}(k)]}\right]^{-1}, \tag{30}$$

with T^+ corresponding to an increase from k tot $k+1$ and T^- corresponding to the opposite. As before, β expresses the unavoidable noise associated with errors in the imitation process. Fixation probability and stationary distribution are calculated in the same way as in two-player games.

Risk-Dominance Condition

An important analytical criteria to determine the evolutionary viability of a given strategy is whether it is risk-dominant with respect to other strategies (Nowak, 2006; Gokhale & Traulsen, 2010). Namely, one considers which selection direction is more probable: an i mutant fixating in a homogeneous population of agents playing j or a j mutant fixating in a homogeneous population of agents playing i. When the first is more likely than the latter, i is said to be *risk-dominant* against j (Gokhale & Traulsen, 2010), which holds for any intensity of selection and in the limit of large population size Z when

$$\sum_{k=1}^{N} \Pi_{ij}(k) \geq \sum_{k=0}^{N-1} \Pi_{ji}(k). \tag{31}$$

Appendix C. Disaster Scenarios: Personal vs Collective Risks

In the main text we consider that AI risk is personal, i.e. when a disaster occurs due to omitting safety requirements, only UNSAFE players suffer. Here we consider that AI disaster also affects co-players of the interactions. Namely, when a disaster occurs, the UNSAFE players lose their payoffs as before but now their SAFE co-players would lose a

fraction of their payoffs, denoted by γ ($0 \leq \gamma \leq 1$), with $\gamma = 0$ corresponding to personal risk (as in the main text) and $\gamma = 1$ representing collective risk. So the payoff of AS when playing with AU becomes, in *two-team AI race*: $\pi_{12}(1 - p_r + p_r(1-\gamma)) = \pi_{12}(1 - p_r\gamma)$. Similarly for CS when playing with AU. Thus, the payoff matrix defining averaged payoffs for the three strategies becomes

$$\Pi = \begin{array}{c} \\ AS \\ AU \\ CS \end{array} \begin{array}{c} \begin{array}{ccc} AS & AU & CS \end{array} \\ \begin{pmatrix} \frac{B}{2W} + \pi_{11} & (1 - p_r\gamma)\pi_{12} & \frac{B}{2W} + \pi_{11} \\ (1 - p_r)\left(\frac{sB}{W} + \pi_{21}\right) & (1 - p_r)\left(\frac{sB}{2W} + \pi_{22}\right) & (1 - p_r)\left[\frac{sB}{W} + \frac{s}{W}\left(\pi_{21} + (\frac{W}{s} - 1)\pi_{22}\right)\right] \\ \frac{B}{2W} + \pi_{11} & (1 - p_r\gamma)\frac{s}{W}\left(\pi_{12} + (\frac{W}{s} - 1)\pi_{22}\right) & \frac{B}{2W} + \pi_{11} \end{pmatrix} \end{array}. \tag{32}$$

Figure S10 shows the results for different values of γ across regimes. In the early regime, little difference is observed when moving from completely personal risk ($\gamma = 0$, as in the main text) to mixed risk ($\gamma = 0.5$) and collective risk ($\gamma = 1$). It is also easily seen (similar to the analysis in Section 1 of this SI), the same conditions are obtained in this regime for when AS and CS are risk-dominant against AU as well as when SAFE is the more beneficial collective outcome.

In the late regime, a larger γ increases the frequency of AU (The condition under which SAFE is the more beneficial collective outcome, does not depend at all on γ). They can be written as follows, respectively

$$\pi_{11} + (1 - p_r\gamma)\pi_{12} > (1 - p_r)(\pi_{21} + \pi_{22}), \tag{33}$$

$$(1 - p_r\gamma)\pi_{22} + \pi_{11} > 2(1 - p_r)\pi_{22}. \tag{34}$$

which are equivalent to, respectively

$$p_r > \frac{\pi_{21} + \pi_{22} - \pi_{11} - \pi_{12}}{\pi_{21} + \pi_{22} - \gamma\pi_{12}}, \tag{35}$$

$$p_r > \frac{\pi_{22} - \pi_{11}}{\pi_{22}(2 - \gamma)} = \frac{1}{2 - \gamma} - \frac{\pi_{11}}{\pi_{22}(2 - \gamma)}. \tag{36}$$

We can see that the right hand side of the condition of CS is an increasing function of γ, and when $\gamma = 1$ (shared or collective risk), the condition for CS is the same as for when SAFE is the preferred collective outcome.

Figure S11 shows the frequency of AU in the late regime and the corresponding conditions obtained (see black, blue and green lines). We observe that increasing γ enlarges the innovation zones (see the red parts) and reduces the dilemma zone.

Next, similar analysis can be done for the *N-team AI race*. The payoffs of AS and CS when playing with AU is scaled by a factor $(1 - p_r\gamma)$ and all other payoffs remain the same. Similar observations are obtained as in the two-player case. Namely, the same conditions are obtained in the early DSAI regime for when AS and CS are risk-dominant against AU as well as when SAFE is the more beneficial collective outcome. For the late DSAI, AU is dominant for a larger range for increasing γ, see Figure S12.

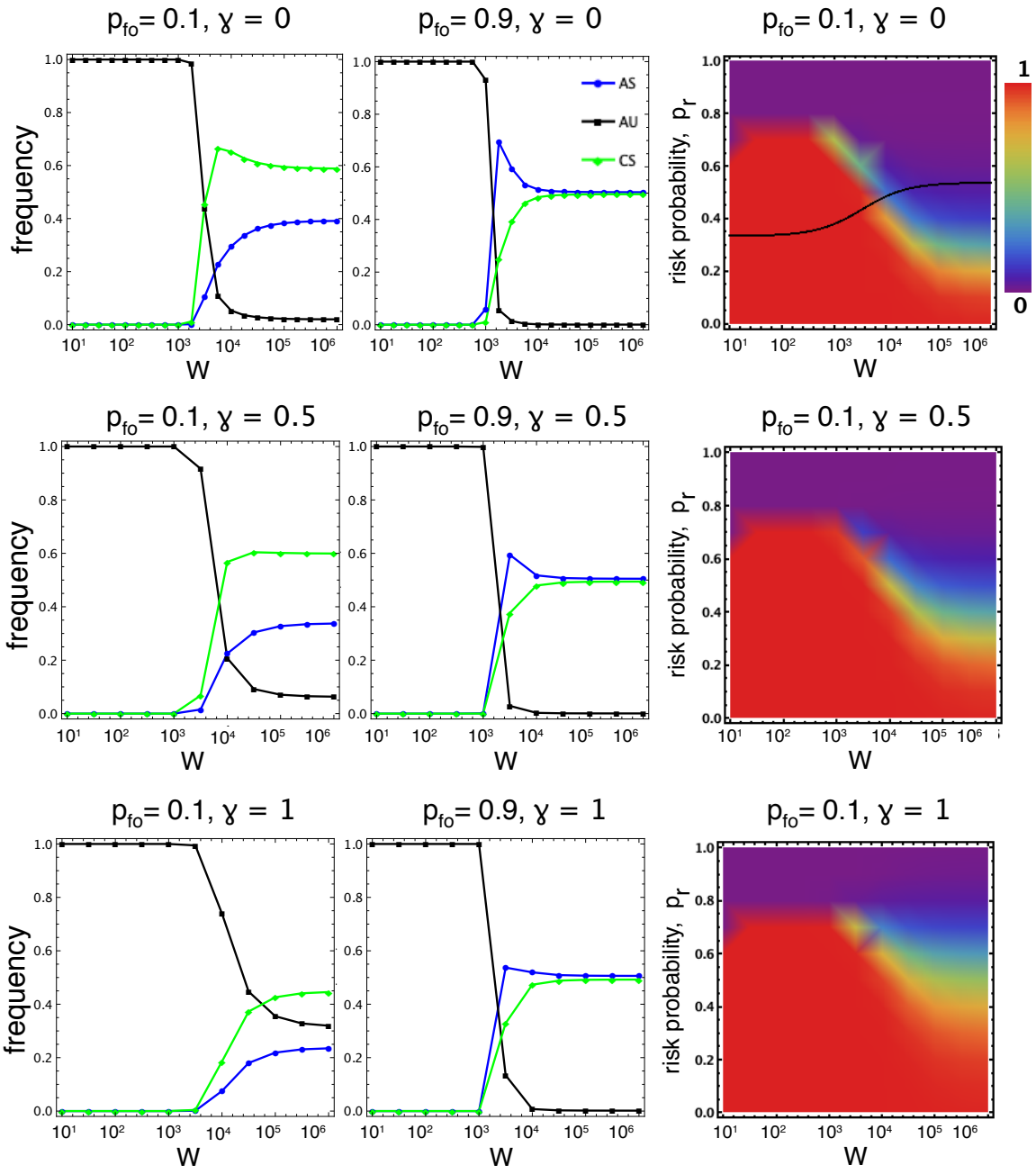

Figure S10: **Different regimes of DSAI for different types of risk: when $\gamma = 0$ (top row); $\gamma = 0.5$ (middle row) and $\gamma = 1$ (bottom row).** Little difference is observed when moving from completely personal risk ($\gamma = 0$) to mixed types of risk ($\gamma = 0.5$) and collective risk ($\gamma = 1$), especially in the early regime. In the late regime, larger γ slights increases the frequency of AU. Note that the conditions for which SAFE generates a larger social welfare than UNSAFE behaviour (the black line in the top left panel), does not change with γ. Parameters: $p_r = 0.6$ (first two columns); $c = 1$, $b = 4$, $B = 10000$, $\beta = 0.1$, $Z = 100$.

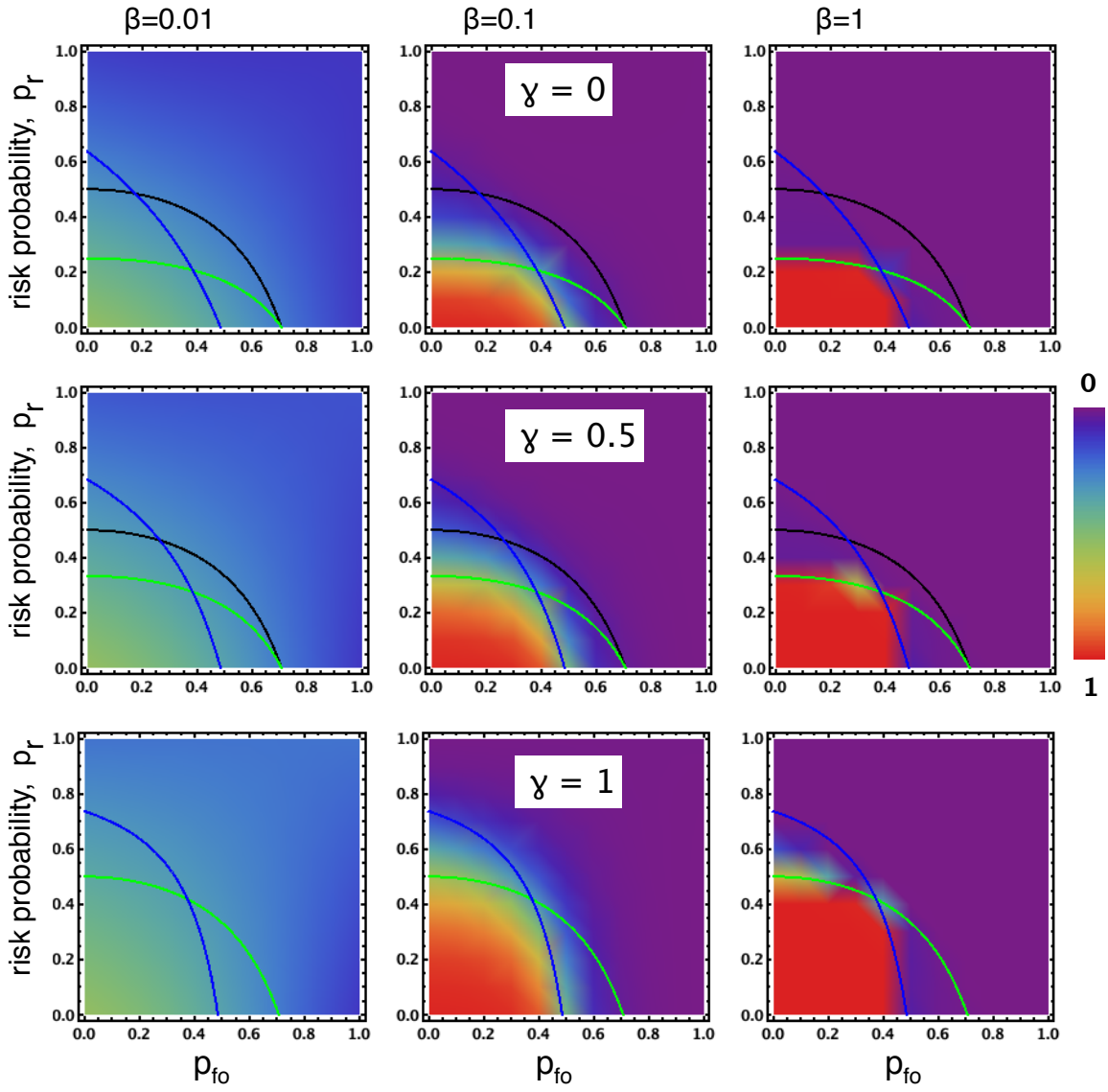

Figure S11: **Late DSAI: Frequency of AU when $\gamma = 0$ (top row); $\gamma = 0.5$ (middle row) and $\gamma = 1$ (bottom row).** The three lines (black, blue and green) are the same as in the main text (Figure 3) (in the bottom line the black and green lines are the same). Increasing γ enlarges the innovation zones (red parts). Parameters: $c = 1$, $b = 4$, $s = 1.5$, $W = 10^6$, $B = 10000$, $\beta = 0.1$, $Z = 100$.

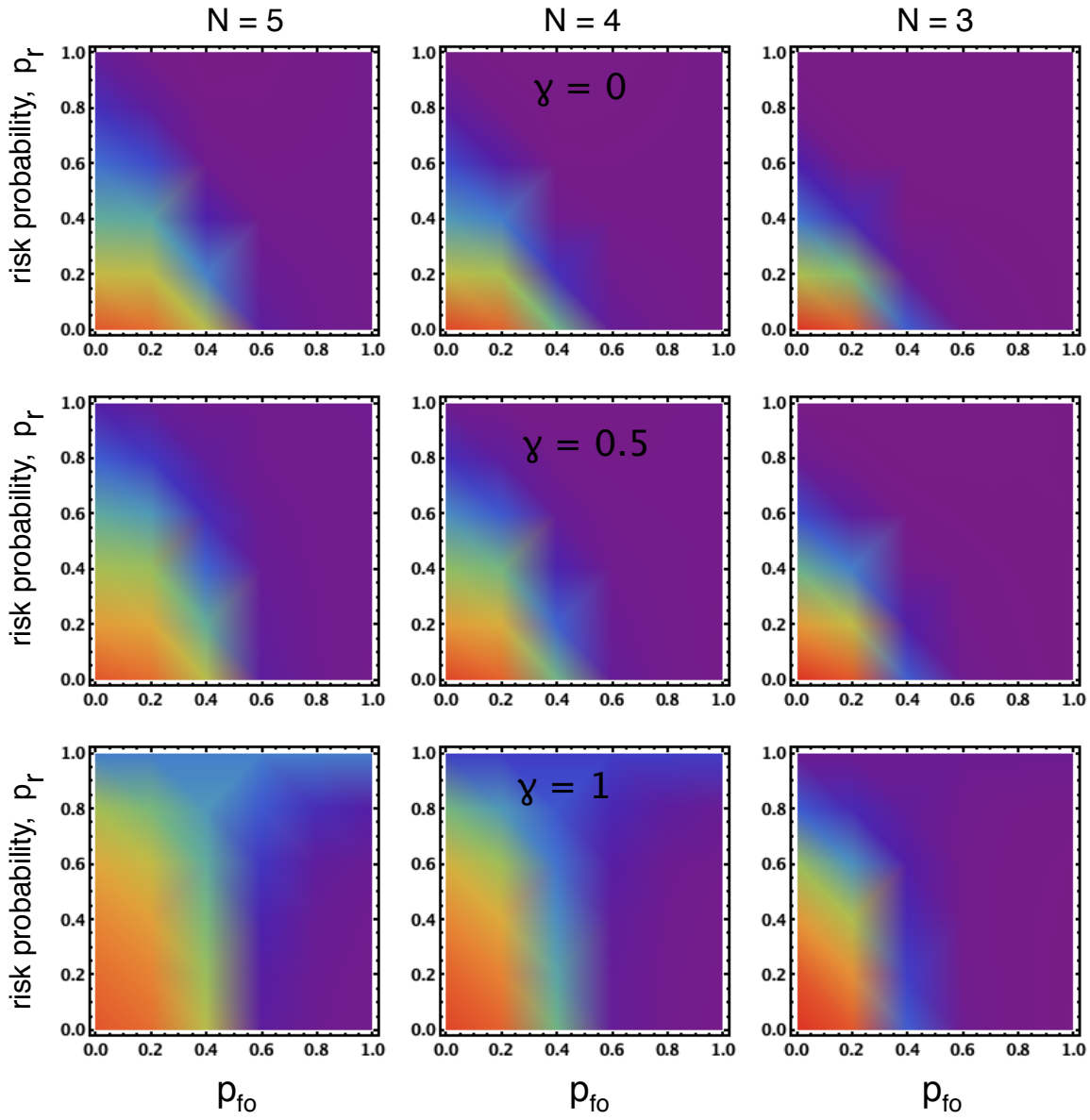

Figure S12: **Late DSAI for N-player race: Frequency of AU when $\gamma = 0$ (top row); $\gamma = 0.5$ (middle row) and $\gamma = 1$ (bottom row).** Increasing γ enlarges the innovation zones (red parts). Parameters: $c = 1$, $b = 6$, $s = 1.5$, $W = 10^6$, $B = 10000$, $\beta = 0.1$, $Z = 100$.

Appendix D. Risk of Being Found Out With Longer Repercussions

We analyse here the case that the risk of unsafe development being disclosed induces that the found-out unsafe player does not gain her share of b in the subsequent $(u-1)$ (where $1 \leq u \leq W$) rounds. That would clearly reduce the payoffs of AU when interacting with others and increase their payoffs when interacting with AU.

The new payoff matrix defining averaged payoffs for the three strategies reads

$$\Pi = \begin{array}{c} \\ AS \\ AU \\ CS \end{array} \begin{array}{c} \begin{array}{ccc} AS & AU & CS \end{array} \\ \begin{pmatrix} \frac{B}{2W} + \pi_{11} & \tilde{\pi}_{12} & \frac{B}{2W} + \pi_{11} \\ (1-p_r)\left(\frac{sB}{W} + \tilde{\pi}_{21}\right) & (1-p_r)\left(\frac{sB}{2W} + \tilde{\pi}_{22}\right) & (1-p_r)\left[\frac{sB}{W} + \frac{s}{W}\left(\pi_{21} + (\frac{W}{s}-1)\tilde{\pi}_{22}\right)\right] \\ \frac{B}{2W} + \pi_{11} & \frac{s}{W}\left(\pi_{12} + (\frac{W}{s}-1)\tilde{\pi}_{22}\right) & \frac{B}{2W} + \pi_{11} \end{pmatrix} \end{array}. \tag{37}$$

where $\tilde{\pi}_{21} = \frac{1}{u}\sum_{i=1}^{u}(1-p_{fo})^i \frac{sb}{s+1} = H_u \pi_{21}$, $\tilde{\pi}_{22} = \frac{1}{u}\sum_{i=1}^{u}(1-p_{fo})^i \frac{(1+p_{fo})b}{2} = H_u \pi_{22}$, $\tilde{\pi}_{12} = -c + \frac{1}{u}\sum_{i=1}^{u}(1-p_{fo})^{i-1}\left((1-p_{fo})\frac{b}{s+1} + p_{fo}(u+1-i)b\right) = -c + H_u(1-p_{fo})\frac{b}{s+1} + \left(p_{fo}(u+1)H_u + \frac{1-(1-p_{fo})^u}{up_{fo}} - (1-p_{fo})^u\right)b$, and $H_u = \frac{\sum_{i=0}^{u-1}(1-p_{fo})^i}{u} \leq 1$. Thus, exactly the same results are obtained in the early DSAI since changing u does not influence the chance of winning the prizes for all strategies.

In the late DSAI (i.e. $W \to +\infty$), considering the limit of $u/W \gg 0$ (when found out, a significant portion of the the subsequent rounds are influenced), we have that $H_u \to 0$ and $p_{fo}(u+1)H_u \to 1$ (assuming $p_{fo} > 0$). That has the same effect as having $p_{fo} = 1$ since

$$\tilde{\pi}_{21} \to 0, \quad \tilde{\pi}_{22} \to 0, \quad \tilde{\pi}_{12} \to -c + b$$

Appendix E. Average Population Payoffs

In Figure S13 we show the average population payoffs, representing its social welfare. For the early regime (see again Figure 1a in main text), in regions (I) and (III) of the s-p_r space the best possible average payoffs are achieved since SAFE (resp., UNSAFE) population is the one generating a larger payoff than the other and they are also dominating (close to 100% frequency). So no additional mechanism/regulation is required that would change this preferred outcome. In region (II), while SAFE is the outcome with the larger average payoff, since UNSAFE dominates, a significantly lower payoff is obtained. Thus, regulation is crucial to be put in place herein. Note that the highest social welfare is achieved for low p_r and high s (successful innovation), with the dominance of UNSAFE. A misplaced regulation (to achieve SAFE) would destroy this significant social welfare gained through innovation.

In the late DSAI regime, see Figure S15, a significant lower social welfare is obtained in this dilemma zone, compared to the one in the unsafe zone, to which regulation can be used to achieve.

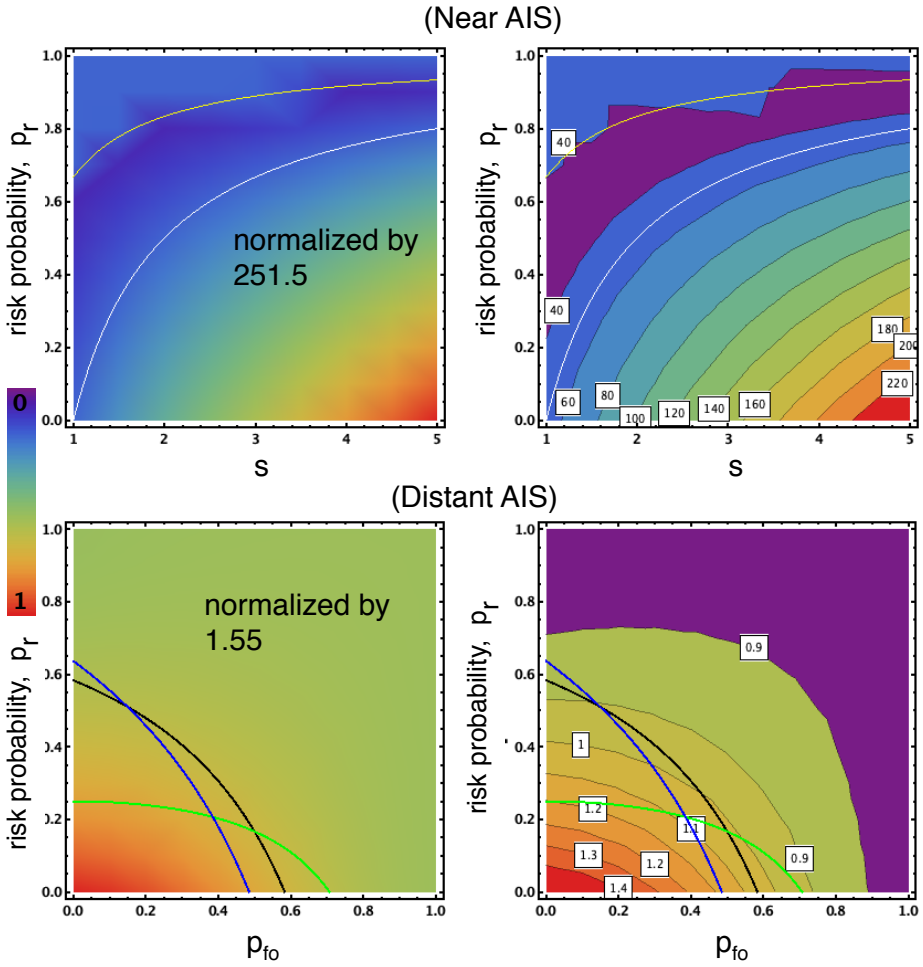

Figure S13: **Average population payoff (social welfare)**. Top row: early ($p_{fo} = 0.5$); Bottom row: late regimes ($s = 1.5$). The lines indicate the conditions above which safety behavior is the preferred collective outcome and when AS and CS are risk-dominant against AU. Parameters: $c = 1$, $b = 4$, $B = 10^4$, $\beta = 0.1$, $Z = 100$.

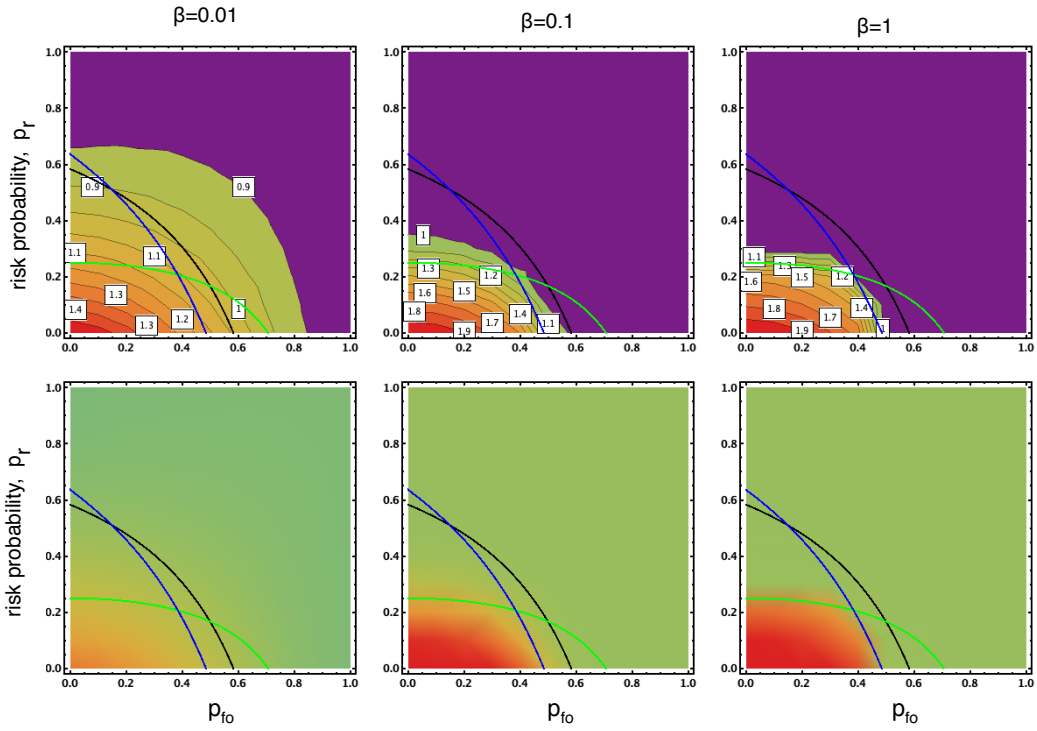

Figure S14: **Late DSAI: Average population payoff** . Same parameter settings in in Figure S2. The lines indicate the conditions above which safety behavior is the preferred collective outcome and when AS and CS are risk-dominant against AU. This welfare is significantly lower in the dilemma zone (below black line and above blue and green lines), see also main text discussion. Parameters: $c = 1$, $b = 4$, $B = 10^4$, $Z = 100$.

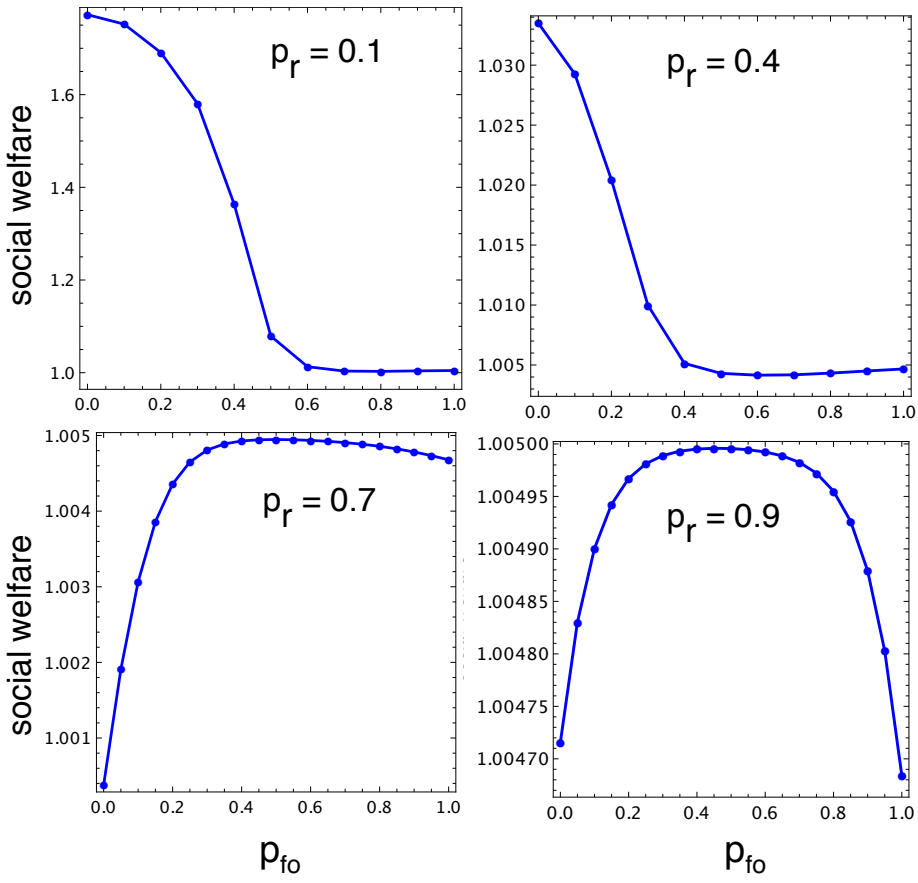

Figure S15: **Late DSAI: Average population payoff (social welfare) for varying p_{fo} and different values of p_r.** Note that the y-axes of the four panels are not made homogenous (e.g. of the same value range), for a clear observation of how social welfare changes as a function of p_{fo}, for different p_r. When p_r is small to intermediate, social welfare decreases with p_{fo}; while when it is larger, an intermediate p_{fo} leads to the highest social welfare. Parameters: $c = 1$, $b = 4$, $B = 10000$, $s = 1.5$, $\beta = 0.1$, $Z = 100$.

References

Abbott, F. M., Dukes, M. N. G., & Dukes, G. (2009). *Global pharmaceutical policy: ensuring medicines for tomorrow's world.* Edward Elgar Publishing.

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

Andras, P., Esterle, L., Guckert, M., Han, T. A., Lewis, P. R., Milanovic, K., Payne, T., Perret, C., Pitt, J., Powers, S. T., et al. (2018). Trusting intelligent machines: Deepening trust within socio-technical systems. *IEEE Technology and Society Magazine*, *37*(4), 76–83.

Apps, P. (2019). Are China, Russia winning the AI arms race?.. [Reuters; Online posted 15-January-2019].

Armstrong, S., Bostrom, N., & Shulman, C. (2016). Racing to the precipice: a model of artificial intelligence development. *AI & society*, *31*(2), 201–206.

Armstrong, S., Sotala, K., & Ó hÉigeartaigh, S. S. (2014). The errors, insights and lessons of famous ai predictions–and what they mean for the future. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*(3), 317–342.

Axelrod, R. (1984). *The Evolution of Cooperation.* Basic Books, ISBN 0-465-02122-2.

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.

Brooks, R. (2017). The Seven Deadly Sins of Predicting the Future of AI.. [https://rodneybrooks.com/the-seven-deadly-sins-of-predicting-the-future-of-ai/; Online posted 7-September-2017].

Brown, N., & Sandholm, T. (2018). Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, *359*(6374), 418–424.

Brown, N., & Sandholm, T. (2019). Superhuman AI for multiplayer poker. *Science*, eaay2400.

Burrell, R., & Kelly, C. (2020). The covid-19 pandemic and the challenge for innovation policy. *Available at SSRN 3576481.*

Campart, S., & Pfister, E. (2014). Technological races and stock market value: evidence from the pharmaceutical industry. *Economics of Innovation and New Technology*, *23*(3), 215–238.

Cave, S., & Dihal, K. (2019). Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence*, *1*(2), 74.

Cave, S., Dihal, K., & Dillon, S. (2020). *AI narratives: A history of imaginative thinking about intelligent machines.* Oxford University Press.

Cave, S., & ÓhÉigeartaigh, S. (2018). An AI Race for Strategic Advantage: Rhetoric and Risks. In *AAAI/ACM Conference on Artificial Intelligence, Ethics and Society*, pp. 36–40.

Collingridge, D. (1980). *The social control of technology*. New York : St. Martin's Press.

Denicolò, V., & Franzoni, L. A. (2010). On the winner-take-all principle in innovation races. *Journal of the European Economic Association*, *8*(5), 1133–1158.

Dignum, V., Muller, C., & Theodorou, A. (2020). Final analysis of the EU Whitepaper on AI.. [https://allai.nl/wp-content/uploads/2020/06/ALLAI-Final-Analysis-of-the-EU-Whitepaper-on-AI-consultation.pdf; Accessed on 14-June-2020].

Domingos, E. F., Grujić, J., Burguillo, J. C., Kirchsteiger, G., Santos, F. C., & Lenaerts, T. (2020). Timing uncertainty in collective risk dilemmas encourages group reciprocation and polarization. *iScience, in press, arXiv preprint arXiv:2003.07317*.

European Commission (2020). White paper on Artificial Intelligence – An European approach to excellence and trust. Tech. rep., European Commission.

Fudenberg, D., & Imhof, L. A. (2005). Imitation processes with small mutations. *Journal of Economic Theory*, *131*, 251–262.

Future of Life Institute (2015). Autonomous Weapons: An Open Letter from AI & Robotics Researchers. Tech. rep., Future of Life Institute, Cambridge, MA.

Future of Life Institute (2019). Lethal autonomous weapons pledge. https://futureoflife.org/lethal-autonomous-weapons-pledge/.

Gokhale, C. S., & Traulsen, A. (2010). Evolutionary games in the multiverse. *Proc. Natl. Acad. Sci. U.S.A.*, *107*(12), 5500–5504.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research*, *62*, 729–754.

Grujić, J., & Lenaerts, T. (2020). Do people imitate when making decisions? evidence from a spatial prisoner?s dilemma experiment. *Royal Society Open Science*, *7*(7), 200618.

Han, T. A., Pereira, L. M., & Santos, F. C. (2011). Intention recognition promotes the emergence of cooperation. *Adaptive Behavior*, *19*(3), 264–279.

Han, T. A. (2013). *Intention Recognition, Commitments and Their Roles in the Evolution of Cooperation: From Artificial Intelligence Techniques to Evolutionary Game Theory Models*, Vol. 9. Springer SAPERE series.

Han, T. A., Pereira, L. M., & Lenaerts, T. (2015). Avoiding or Restricting Defectors in Public Goods Games?. *J. Royal Soc Interface*, *12*(103), 20141203.

Han, T. A., Pereira, L. M., & Lenaerts, T. (2019). Modelling and Influencing the AI Bidding War: A Research Agenda. In *Proceedings of the AAAI/ACM conference AI, Ethics and Society*, pp. 5–11.

Han, T. A., Pereira, L. M., Lenaerts, T., & Santos, F. C. (2020). Mediating Artificial Intelligence Developments through Negative and Positive Incentives.. arXiv: 2010.00403.

Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A., & Sigmund, K. (2007). Via freedom to coercion: The emergence of costly punishment. *Science*, *316*, 1905–1907.

Hauser, O. P., Hilbe, C., Chatterjee, K., & Nowak, M. A. (2019). Social dilemmas among unequals. *Nature*, *572*(7770), 524–527.

Hofbauer, J., & Sigmund, K. (1998). *Evolutionary Games and Population Dynamics*. Cambridge University Press.

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

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1–11.

Kandori, M., Mailath, G. J., & Rob, R. (1993). Learning, mutation, and long run equilibria in games. *Econometrica*, *61*, 29–56.

Karlin, S., & Taylor, H. E. (1975). *A First Course in Stochastic Processes*. Academic Press, New York.

Lee, K.-F. (2018). *AI superpowers: China, Silicon Valley, and the new world order*. Houghton Mifflin Harcourt.

Leigh, V. V. (1973). A new evolutionary law. *Evol. Theory*, 1–30.

Lemley, M. A. (2012). The myth of the sole inventor. *Michigan Law Review*, 709–760.

Montreal Declaration (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/.

Nowak, M. A. (2006). *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard University Press, Cambridge, MA.

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

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*(5805), 1560.

Nowak, M. A., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in prisoner's dilemma. *Nature*, *364*, 56–58.

Pacheco, J. M., Santos, F. C., Souza, M. O., & Skyrms, B. (2009). Evolutionary dynamics of collective action in n-person stag hunt dilemmas.. *Proc. R. Soc. B*, *276*, 315–321.

Pamlin, D., & Armstrong, S. (2015). Global challenges: 12 risks that threaten human civilization. *Global Challenges Foundation, Stockholm*.

PwC (2017). Sizing the prize: What's the real value of ai for your business and how can you capitalise?. Tech. rep., PwC, London, United Kingdom.

Ross, G. M., & Portugali, J. (2018). Urban regulatory focus: a new concept linking city size to human behaviour. *Royal Society Open Science*, *5*(5), 171478.

Russell, S., Hauert, S., Altman, R., & Veloso, M. (2015). Ethics of artificial intelligence. *Nature*, *521*(7553), 415–416.

Santos, F. C., Pacheco, J. M., & Lenaerts, T. (2006). Evolutionary dynamics of social dilemmas in structured heterogeneous populations.. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 3490–3494.

Santos, F. C., Santos, M. D., & Pacheco, J. M. (2008). Social diversity promotes the emergence of cooperation in public goods games. *Nature*, *454*(7201), 213–216.

Schubert, S., Caviola, L., & Faber, N. (2019). The psychology of existential risk: Moral judgments about human extinction. *Scientific Reports*, *9*(15100).

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

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, *362*(6419), 1140–1144.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354.

Smith, J. M. (1982). *Evolution and the Theory of Games*. Cambridge university press.

Sotala, K., & Yampolskiy, R. V. (2014). Responses to catastrophic AGI risk: a survey. *Physica Scripta*, *90*(1), 018001.

Steels, L., & Lopez de Mantaras, R. (2018). The barcelona declaration for the proper development and usage of artificial intelligence in europe. *AI Communications*, pp. 1–10.

Szolnoki, A., & Perc, M. (2013). Correlation of positive and negative reciprocity fails to confer an evolutionary advantage: Phase transitions to elementary strategies. *Physical Review X*, *3*(4), 041021.

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

Tavoni, A., Dannenberg, A., Kallis, G., & Löschel, A. (2011). Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *Proceedings of the National Academy of Sciences*, *108*(29), 11825–11829.

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

Van Segbroeck, S., Pacheco, J. M., Lenaerts, T., & Santos, F. C. (2012). Emergence of fairness in repeated group interactions. *Physical review letters*, *108*(15), 158104.

Vasconcelos, V. V., Santos, F. C., Pacheco, J. M., & Levin, S. A. (2014). Climate policies under wealth inequality. *Proceedings of the National Academy of Sciences*, *111*(6), 2212–2216.

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