To regulate or not: a social dynamics analysis of the race for AI supremacy

The Anh Han\textsuperscript{1,*}, Luís Moniz Pereira \textsuperscript{2}, Francisco C. Santos\textsuperscript{3,4}, Tom Lenaerts\textsuperscript{4,5,*}

\textsuperscript{1} School of Computing and Digital Technologies, Teesside University, Middlesbrough, UK TS1 3BA
\textsuperscript{2} NOVA Laboratory for Computer Science and Informatics (NOVA LINCS), Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
\textsuperscript{3} INESC-ID and Instituto Superior Tecnico, Universidade de Lisboa
\textsuperscript{4} Machine Learning Group, Université Libre de Bruxelles, Boulevard du Triomphe CP212, Brussels, Belgium
\textsuperscript{5} Artificial Intelligence Lab, Vrije Universiteit Brussel, Boulevard de la Plaine 2, 1050 Ixelles, Belgium

* Corresponding authors: The Anh Han (T.Han@tees.ac.uk) and Tom Lenaerts (Tom.Lenaerts@ulb.ac.be)
Abstract

Rapid technological advancements in Artificial Intelligence (AI) as well as the growing deployment of intelligent technologies in new application domains are currently driving the competition between businesses, nations and regions. This race for technological supremacy creates a complex ecology of choices that may lead to negative consequences, in particular, when ethical and safety procedures are underestimated or even ignored. As a consequence, different actors are urging to consider both the normative and social impact of these technological advancements. As there is no easy access to data describing this AI race, theoretical models are necessary to understand its dynamics, allowing for the identification of when, how and which 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 AI supremacy can be achieved plays a crucial role. When this supremacy 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 blindly imposing regulations 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, certain conditions require the promotion of risk-taking as opposed to compliance to 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 AI, a clear understanding about the time-scale of the race is thus required, as this may induce important non-trivial effects.
Keywords: AI race modelling, safety, game theory, regulation, time-scale.
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\textsuperscript{9,10,35,36,43} which have been exhibited. Although successful in highly specialised tasks, these AI success stories appear in the imagination of the general public 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\textsuperscript{11}. 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\textsuperscript{31}. All these (un)announced business, societal and political ambitions indicate that an AI race or bidding war has been triggered\textsuperscript{1,2,12}, where stake-holders in both private and public sectors are competing to be the first to cross the finish line and hence the leader in the development and deployment of powerful, transformative AI\textsuperscript{3,6,7,12}.

Irrespectively of the anticipated benefits, 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\textsuperscript{40} and ii) whilst attempting to be the first/best, some ethical consequences as well as safety procedures may be underestimated or even ignored\textsuperscript{3,12} (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\textsuperscript{15,16}, the blogs of AI experts requesting careful communications\textsuperscript{8} and the proclamations on ethical use of AI in the world\textsuperscript{24,27,32,38}.

While potential AI disaster scenarios are many\textsuperscript{3,30,33,37}, the uncertainties in accurately predicting these risks and outcomes are high\textsuperscript{4}. 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\textsuperscript{13}. Sufficient data is therefore not yet available, requiring a modelling approach to grasp what can be expected in a race for AI supremacy (AIS). 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. This manuscript focuses on defining a baseline model that discusses when to expect unsafe or safe AI development behaviour and when this is disruptive, i.e. when it harms social welfare. Subsequently, it can be employed to evaluate the impact of regulatory mechanisms on the behavioural preferences. We resort to the framework of evolutionary game theory\textsuperscript{22,34} to address this issue.

Concretely, the model assumes that in order to achieve AIS in a domain $X$, a number of development steps or rounds ($W$) are required. Large-scale surveys and analysis of AI experts on their beliefs and predictions about progress in AI suggest that the perceived time-scale for AIS is highly diverse across domains and regions\textsuperscript{4,18}. The model therefore aims to capture these different time-scales of AIS occurrence: When $W$ is small, AIS can be expected to happen in the near future (early AIS regime) while when $W$ is large, AIS will only be achieved far away in time (late AIS 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 AIS 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.

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 require-
ments 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 AI 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 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 their acquired payoffs. The risk is personal for UNSAFE players in the current model. Although the threat is greater for the creator\textsuperscript{3,30}, there may also be repercussions for the other participants or society as a whole, a matter discussed in detail in the Supporting Information (SI). 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 follows safety half of the time will incur only half of the time the risk of disaster over all rounds.

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 AIS Race (AISR) 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 evolutionary game theory 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 $\text{AS}$, the SAFE players) or completely ignore them (denoted by $\text{AU}$, the UNSAFE players). Additionally, we assume that, upon realising that UNSAFE players ignore safety precautions to gain a greater development speed, leading to the winning 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 $p_f$), 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\(^2\). We therefore consider, in line with previous literature on repeated games\(^5\),\(^19\),\(^34\),\(^42\), a conditional strategy (denoted by $\text{CS}$), 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\(^5\),\(^34\). Alternative strategies can be imagined but for the sake of simplicity we focus (for now) on these three.

In the following, we will examine, across different time-scales of the AISR, 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 supremacy (i.e., acquire $B$), we will examine whether (and under what time-scale of the AISR 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 ($N \geq 2$) AISR in SI. Additionally, the situation where the effects of a setback or disaster are no longer just personal are analysed in depth in SI.

## 2 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 AISR 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$^{22,34}$. 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 AISR 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 AIS**: 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 AIS**: In this regime, AIS 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 AIS 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
Figure 1. Different regimes of AIS: when $W$ is small (early AIS) vs when $W$ is larger (late AIS). 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 AIS regime, AU dominates the population, while AS and CS outperform AU in the late AIS 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: $c = 1, b = 4, s = 1.5, B = 10^4, \beta = 0.1, Z = 100$.

panels a and b, where $p_r = 0.6$). In the second AIS 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 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 SI). 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 AIS: 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.

In the current AIS regime, the first condition occurs when (see SI for the proof)

$$p_r > 1 - \frac{1}{s}.\quad (1)$$

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 SI) when

$$p_r > 1 - \frac{1}{3s}.\quad (2)$$
Figure 2. Early AIS 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: $c = 1, b = 4, W = 100, p_{fo} = 0.5, B = 10^4, \beta = 0.1, Z = 100$. 
The two boundary conditions in Equations 1 and 2 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 **AIS 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 **AIS 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 **AIS 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 SI.

As can be observed, in regions (I) and (III), the preferred collective outcomes are also selected by the social dynamics. Whereas in the AIS 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 AIS innovation zone, the benefit of quickly reaching AIS is everything and speed ensures that one arrives first, with limited risk for a setback or even shared disaster (see SI). In terms of social welfare, i.e. the average benefits spread over the population, the AIS innovation zone produces the largest benefits, especially for low risk and high speed combinations (see SI, Figure S13). In the AIS 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 are thus required.
Region (II), the AIS 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 AIS compliance zone and are potentially detrimental (in terms of the missed social welfare) effects in the AIS innovation zone. It is therefore essential to know, when the time-scale to reach AIS is short, what risks can be expected and what speed is acceptable to avoid the AIS dilemma zone and ensure a positive effect for society.

Looking back at the observation in Figure 1 that in the early AIS regulation is necessary, the current analysis reveals that this is only a necessity when risk and development speed put the race in the AIS dilemma zone since the effects would be counterproductive in the two other zones. Yet stimuli to promote risk-taking in the AIS innovation zone and following safety protocols in the AIS 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 $\beta$ in our model (cf. Figure S4 in SI) — in those zones.

Note that the boundaries established by Equations 1 and 2 are applicable for both CS and AS when playing against AU. Thus, similar results are obtained if we consider a population of just two strategies AS and AU (cf. Figure S5 in SI). 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 SI (also see Figure S10). The results are furthermore robust to changes in the number of participants in the race. When considering the AI race among $N$ development teams (see SI), the main difference is that the upper bound of
Figure 3. Late AIS 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 AIS 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 AIS compliance, the AIS dilemma and the AIS 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: $c = 1$, $b = 4$, $s = 1.5$, $B = 10^4$, $\beta = 0.1$, $Z = 100$.

region (II) increases. That is, the AIS dilemma zone increases and the AIS compliance zone disappears. Regulation is thus required for a larger part of the speed-disaster space (cf. Figures S7 and S8 in SI). 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 AIS: risk-taking as opposed to safety compliance may need to be promoted

When AIS 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 AIS regime (see SI). In Figure 3, one can again distinguish three zones, i.e. the AIS compliance, AIS dilemma and AIS 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 AIS compliance and late AIS 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 AIS dilemma zone, regulation should be put in place to enforce behaviour that improves social welfare. However, in contrast to the early AIS where safety should be promoted, in this late AIS regime, unsafe behaviour (speedy innovation) should be promoted to increase social welfare (see Figure S14 in SI). 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 SI 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 SI).

As shown in the SI, the observations remain valid if, instead of pairwise interactions, one considers a race with $N > 2$ teams in the late AIS regime, i.e. all three zones reappear. Moreover, when $N$ increases, while the innovation zone size remains unchanged, the AIS 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 SI). 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 SI). That is, shared risk in the late AIS regime improves the overall social welfare (by allowing more beneficial innovation to happen), reducing the need for regulatory actions to handle the late AIS dilemma zone.

3 Discussion

Our results reveal that knowing the exact timing of reaching AIS 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 AIS 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 AIS zone the social dynamics is selecting for (undesired) behaviour, requiring regulation of risk-taking in the early AIS and safety compliance in the late AIS.

We show furthermore, both in the early and late regimes, that although the three AIS 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 AIS, speedy development
(s) is everything, the race outcome in the late AIS 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 AIS 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 AIS 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.

The AISR model and associated analysis provides thus an instrument for policy makers to think about the supporting mechanisms (such as suitable rewards and sanctions)\(^{20,21,34,37,39,43}\) needed to mediate a given AI race. In the early AIS, controlling the development speed of AI teams appears essential. Yet, policy makers should carefully consider whether it will have the expected outcome, i.e. whether the race is actually occurring in the AIS dilemma zone. In the late AIS, 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.

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 SI the results remain the same in early AIS, while in the late AIS, the outcome is equivalent to the results one obtains when full monitoring (i.e. \(p_{fo} = 1\)) is in effect. Intuitively, longer consequence associated with being detected is equivalent to having a higher probability
of being detected in each round in the current AISR model.

There are of course limitations to the current model, which will require further analysis. 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 AIS, thus effectively increasing $W$. Moreover, the time to reach AIS in a domain $W$ may also be affected by the trust that people have in AI techniques, 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 AI race \cite{6,12}. 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 AIS in some domain, the results of multiple races may not 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. Future models of AISR will thus need to consider that multiple AISR games to study what regulatory actions are most beneficial for this kind of goals.

Notwithstanding all additional features one can imagine that are interesting for framing the AISR, the current work provides a thoroughly analysed base-line model that can be used to answer relevant questions on the regulation of innovation and research activities in the current races for different kinds of AI supremacy.

In conclusion, we have provided here a first plausible AISR model directly useful for policy makers and researchers 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.
4 Methods

AI race model definition. The AI development race is modeled as a repeated two-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 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)

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

(3)

For instance, 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:

$$(1 - p_{fo}) [(1 - p_{fo})(b/2) + p_{fo}b] = (1 - p_{fo}^2) \frac{b}{2}$$. The payoff matrix defining averaged payoffs
for the three strategies reads

\[
\begin{bmatrix}
\pi_{11} & \pi_{12} & \pi_{21} \\
\pi_{22} & \pi_{11} & \pi_{21} \\
\pi_{22} & \pi_{21} & \pi_{11}
\end{bmatrix}
\]

Evolutionary Dynamics in Finite Populations. We adopt here evolutionary game theory (EGT) methods for finite populations to derive analytical results and numerical observations\textsuperscript{23,28,29}. In a repeated games, players’ average payoff over all the game rounds (see the payoff matrix in Equation 4) represents their fitness or social success, and evolutionary dynamics is shaped by social learning\textsuperscript{22,34}, whereby the most successful players will tend to be imitated more often by the other players. In the current work, social learning is modeled using the so-called pairwise comparison rule\textsuperscript{41}, 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, \( (1 + e^{-\beta(f_B-f_A)})^{-1} \), where \( \beta \) 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\textsuperscript{14,23,29}. 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\textsuperscript{14,23,29}. 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,

\[
P_A(k) = \frac{(k-1)\Pi_{A,A} + (Z-k)\Pi_{A,B}}{Z-1}, \quad P_B(k) = \frac{k\Pi_{B,A} + (Z-k-1)\Pi_{B,B}}{Z-1}.
\] (5)

The fixation probability that a single mutant A taking over a whole population with \((Z-1)\) B players is as follows\(^{26,29,41}\)

\[
\rho_{B,A} = \left(1 + \sum_{i=1}^{Z-1} \prod_{j=1}^{i} T^{-}(j) T^{+}(j) \right)^{-1},
\] (6)

where \( T^\pm(k) = \frac{Z-k}{Z} \left[1 + e^{\pm \beta[P_A(k) - P_B(k)]} \right]^{-1} \) 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\(^{14,23}\). Namely, considering a set of \( s \) strategies, \( \{1, \ldots, s\} \), 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\(^{17,25}\), which holds for any intensity of selection and in the limit of large \( N \) when

\[
\pi_{A,A} + \pi_{A,B} > \pi_{B,A} + \pi_{B,B}.
\] (7)
5 Acknowledgements

T.A.H., L.M.P. and T.L. are supported by Future of Life Institute grant RFP2-154. L.M.P. acknowledges support from FCT/MEC NOVA LINCS PEst UID/CEC/04516/2019. F.C.S. acknowledges support from FCT Portugal (grants PTDC/EEI-SII/5081/2014, PTDC/MAT/STA/3358/2014). T.L. acknowledges support by the FuturICT2.0 (www.futurict2.eu) project funded by the FLAG-ERA JCT 2016.

References

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

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

3. Stuart Armstrong, Nick Bostrom, and Carl Shulman. Racing to the precipice: a model of artificial intelligence development. AI & society, 31(2):201–206, 2016.

4. Stuart Armstrong, Kaj Sotala, and Seán S Ó hÉigeartaigh. 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, 2014.

5. Robert Axelrod. The Evolution of Cooperation. Basic Books, ISBN 0-465-02122-2, 1984.

6. Seth D Baum. On the promotion of safe and socially beneficial artificial intelligence. AI & SOCIETY, 32(4):543–551, 2017.

7. Nick Bostrom. Strategic implications of openness in AI development. Global Policy, 8(2):135–148, 2017.
8. Rodney Brooks. The Seven Deadly Sins of Predicting the Future of AI, 2017. [https://rodneybrooks.com/the-seven-deadly-sins-of-predicting-the-future-of-ai/; Online posted 7-September-2017].

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

10. Noam Brown and Tuomas Sandholm. Superhuman AI for multiplayer poker. *Science*, page eaay2400, 2019.

11. Stephen Cave and Kanta Dihal. Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence*, 1(2):74, 2019.

12. Stephen Cave and Seán ÓhÉigeartaigh. An AI Race for Strategic Advantage: Rhetoric and Risks. In AAAI/ACM Conference on Artificial Intelligence, Ethics and Society, pages 36–40, 2018.

13. David Collingridge. *The social control of technology*. New York : St. Martin’s Press, 1980.

14. D. Fudenberg and L. A. Imhof. Imitation processes with small mutations. *Journal of Economic Theory*, 131:251–262, 2005.

15. Future of Life Institute. Autonomous Weapons: An Open Letter from AI & Robotics Researchers. Technical report, Future of Life Institute, Cambridge, MA, 2015.

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

17. Chaitanya S. Gokhale and Arne Traulsen. Evolutionary games in the multiverse. *Proc. Natl. Acad. Sci. U.S.A.*, 107(12):5500–5504, March 2010.
18. Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research*, 62:729–754, 2018.

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

20. The Anh Han, Luís Moniz Pereira, and Tom Lenaerts. Avoiding or Restricting Defectors in Public Goods Games? *J. Royal Soc Interface*, 12(103):20141203, 2015.

21. The Anh Han, Luís Moniz Pereira, and Tom Lenaerts. Modelling and Influencing the AI Bidding War: A Research Agenda. In *Proceedings of the AAAI/ACM conference AI, Ethics and Society*, pages 5–11, 2019.

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

23. L. A. Imhof, D. Fudenberg, and Martin A. Nowak. Evolutionary cycles of cooperation and defection. *Proc. Natl. Acad. Sci. U.S.A.*, 102:10797–10800, 2005.

24. Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, pages 1–11, 2019.

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

26. S. Karlin and H. E. Taylor. *A First Course in Stochastic Processes*. Academic Press, New York, 1975.
27. Montreal Declaration. 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/, 2018.

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

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

30. Dennis Pamlin and Stuart Armstrong. Global challenges: 12 risks that threaten human civilization. *Global Challenges Foundation, Stockholm*, 2015.

31. PwC. Sizing the prize: What’s the real value of ai for your business and how can you capitalise? Technical report, PwC, London, United Kingdom, 2017.

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

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

34. Karl Sigmund. *The Calculus of Selfishness*. Princeton University Press, 2010.

35. David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.
36. David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. Nature, 550(7676):354, 2017.

37. Kaj Sotala and Roman V Yampolskiy. Responses to catastrophic AGI risk: a survey. Physica Scripta, 90(1):018001, 2014.

38. Luc Steels and Ramon Lopez de Mantaras. The barcelona declaration for the proper development and usage of artificial intelligence in europe. AI Communications, (Preprint):1–10, 2018.

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

40. Mariarosaria Taddeo and Luciano Floridi. Regulate artificial intelligence to avert cyber arms race. Nature, 556(7701):296–298, 2018.

41. A. Traulsen, M. A. Nowak, and J. M. Pacheco. Stochastic dynamics of invasion and fixation. Phys. Rev. E, 74:11909, 2006.

42. Sven Van Segbroeck, Jorge M. Pacheco, Tom Lenaerts, and Francisco C. Santos. Emergence of fairness in repeated group interactions. Physical review letters, 108(15):158104, 2012.

43. Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Langhans, Max Tegmark, and Francesco Fuso Nerini. The role of artificial intelligence in achieving the sustainable development goals. Nature Communications, 11(233), 2020.
Supporting Information:
To regulate or not: a social dynamics analysis of the race for AI supremacy

The Anh Han\textsuperscript{1,⋆}, Luís Moniz Pereira \textsuperscript{2}, Francisco C. Santos\textsuperscript{3,4}, Tom Lenaerts\textsuperscript{4,5}

January 16, 2020

\textsuperscript{1} School of Computing and Digital Technologies, Teesside University, Middlesbrough, UK TS1 3BA
\textsuperscript{2} NOVA Laboratory for Computer Science and Informatics (NOVA LINCS), Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
\textsuperscript{3}INESC-ID and Instituto Superior Tecnico, Universidade de Lisboa
\textsuperscript{4} Machine Learning Group, Université Libre de Bruxelles, Boulevard du Triomphe CP212, Brussels, Belgium
\textsuperscript{5} Artificial Intelligence Lab, Vrije Universiteit Brussel, Boulevard de la Plaine 2, 1050 Ixelles, Belgium
Contents

1 Deriving conditions for viability of safety behaviour 3
   1.1 When safety behaviour is the preferred collective outcome 3
   1.2 When safety behaviour is selected by evolution 4

2 Multiplayer AI race 10
   2.1 N-player AI Race definition 10
   2.2 Analytical conditions and AIS zones in $N$-team interactions 12
   2.3 Methods: Payoffs over group samplings 16

3 Disaster scenarios: personal vs collective risks 19

4 Risk of being found out with longer repercussions 24

5 Average population payoffs 25
1 Deriving conditions for viability of safety behaviour

1.1 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 or average payoff than that of a population of players never following safety precautions, that is, \( \Pi_{AS,AS} > \Pi_{AU,AU} \):

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

(1)

Thus,

\[
p_r > 1 - \frac{B + 2W\pi_{11}}{sB + 2W\pi_{22}}.
\]

(2)

Following the definitions of different AIS regimes in the main texts, we simplify this condition for the two regimes. First, in the early AI regime where \( B/W \gg b \), Equation 2 is equivalent to

\[
p_r > 1 - \frac{1}{s}.
\]

(3)

Now, in the late AIS regime where \( W \to \infty \) (i.e. \( B/W \ll b \)), Equation 2 is equivalent to:

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

(4)

We can see that the development speed \( s \) is the crucial factor in the early AIS regime while it does not play any role in the late AIS, where for fixed \( b \) and \( c \), \( p_{fo} \) is the only influencing factor.
1.2 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,

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

(5)

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

(6)

In the early AI regime where \( B/W \gg b \), both equations are simplified to

\[
p_r > 1 - \frac{1}{3s}.
\]  

(7)

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

\[
\pi_{11} + \pi_{12} > (1 - p_r)(\pi_{21} + \pi_{22}).
\]  

(8)

\[
\pi_{11} > (1 - 2p_r)\pi_{22}.
\]  

(9)

which are equivalent to, respectively

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

(10)

\[
p_r > \frac{1}{2} - \frac{b - 2c}{2b(1 - p_{fo}^2)}.
\]  

(11)
Thus, for safety behaviour to be both selected and the preferred outcome, all the \( p_r \) must satisfy all the Eqs (11), (10) and (4).

It is clear to see that the left hand sides of Eqs (11) and (4) are decreasing functions of \( p_{fo} \) whenever \( b \geq 2c \). We now show that it is also the case for the left hand side of Eq 10. 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

\[
(4s + p_{fo}^2 s + p_{fo}(3 + s)) - 2(p_{fo} + s + p_{fo}s) = 2s + p_{fo}^2 s + p - p_{fo}s > 0
\]

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

\[
p_{r} > \max\left\{ 1 - \frac{(b - 2c)(s + 1)}{2sb}, \frac{4c(s + 1) + 2b(s - 1)}{b(1 + 3s)}, \frac{c}{b} \right\}. \quad (12)
\]

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

\[
p_r > \frac{4c(s + 1) + 2b(s - 1)}{b(1 + 3s)} \quad (13)
\]
Figure S1. Across AIS regimes: Frequency of AU as for varying \( p_r \) and different values of \( p_{fo} \) and \( \beta \): when \( W \) is small (early AIS) vs when \( W \) is large (late AIS). \( \beta = 0.01, 0.1, 1 \) 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 2). In general, we observe that AU is dominant for a larger range of \( p_r \) in the early than the late regime. Parameters: \( c = 1, b = 4, s = 1.5, B = 10000, Z = 100 \).
Figure S2. Late AIS ($W = 10^6$). 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 $\beta$. AU has high frequencies in regions below both the blue and green lines, especially for larger $\beta$. Parameters: $c = 1$, $b = 4$, $B = 10000$, $Z = 100$. 
Figure S3. Late AIS: 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 AIS 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: $c = 1, b = 4, s = 1.5, B = 10000, Z = 100$. 
Figure S4. Early AIS: 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 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: $c = 1, b = 4, W = 100, p_{fo} = 0.5, B = 10000, Z = 100$.

which is the condition for AS to be risk-dominant against AU (see Figure S2 for an example when $s = 1.5$).
Figure S5. Early AIS: 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: $c = 1, b = 4, W = 100, p_{fo} = 0.5, B = 10000, Z = 100$.

2 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.

2.1 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
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)

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

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

\(^1\)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 AIS: early AIS (small \( W \)) vs late AIS (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 14). Other parameters: \( c = 1 \), \( b = 6 \), \( s = 1.5 \), \( B = 10000 \), \( N = 5 \), \( Z = 100 \), \( \beta = 0.1 \).

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

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

### 2.2 Analytical conditions and AIS zones in \( N \)-team interactions

**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:

\[
\frac{B}{NW} + \pi(N)_{SAFE} > (1-p_r) \left( \frac{sB}{NW} + \pi(0)_{UNSAFE} \right).
\]
It can be rewritten as
\[ p_r > 1 - \frac{B + W(b - Nc)}{sB + W(1 - p_f)b} \]  
(14)

In **early AIS** (i.e. \( B/W \gg b \)), it is equivalent to:
\[ p_r > 1 - \frac{1}{s} \]  
(15)

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

While in **late AIS** (i.e. \( B/W \ll b \)), it is equivalent to:
\[ p_r > 1 - \frac{b - Nc}{(1 - p_f)b} \]  
(16)

It can be seen that, for this condition to happen in the late AIS, 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 \( N \)).

Figures S6 shows the results for \( N \)-player games across different regimes of AIS (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

\[ \sum_{k=0}^{N-1} \pi(k)_{AU,AS} < \sum_{k=1}^{N} \pi(k)_{AS,AU} \]  
(17)

\[ \sum_{k=0}^{N-1} \pi(k)_{AU,CS} < \sum_{k=1}^{N} \pi(k)_{CS,AU} \]  
(18)
In the early AIS (i.e. $B/W \gg b$), both conditions are reduced to

$$p_r > 1 - \frac{1}{(NH_N)s},$$

(19)

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 15 and 19 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 $1 - \frac{1}{(NH_N)s} > p_r > 1 - \frac{1}{s}$: 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)
Figure S8. Early AIS. 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: $c = 1, b = 4, W = 100, s = 1.5, p_f = 0.5, B = 10000, Z = 100$.

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 AIS, the conditions AS and CS are reduced to

$$p_r > 1 - \frac{\sum_{i=1}^{N} \pi(i)_{SAFE}}{\sum_{i=0}^{N-1} \pi(i)_{UNSAFE}} \quad (20)$$

$$p_r > 1 - \frac{(N-1)\pi(0)_{UNSAFE} + \pi(N)_{SAFE}}{N\pi(0)_{UNSAFE}} = \frac{1}{N} \left( 1 - \frac{\pi(N)_{SAFE}}{\pi(0)_{UNSAFE}} \right) \quad (21)$$
Figure S9. Late AIS 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: \( c = 1, b = 6, s = 1.5, W = 10^6, B = 10000, \beta = 0.1, Z = 100 \).
the average payoffs to \(i\) and \(j\) strategists are\(^{2,4}\):

\[
P_{ij}(x) = \sum_{k=0}^{N-1} H(k, N-1, x-1, Z-1) \Pi_{ij}(k+1) \\
= \sum_{k=0}^{N-1} \binom{x-1}{k} \left( \frac{Z-x}{N-1-k} \right) \Pi_{ij}(k+1),
\]

\[
P_{ji}(x) = \sum_{k=0}^{N-1} H(k, N-1, x, Z-1) \Pi_{ji}(k) \\
= \sum_{k=0}^{N-1} \binom{x}{k} \left( \frac{Z-1-x}{N-1-k} \right) \Pi_{ji}(k).
\] (22)

Now, the probability to change the number \(k\) of agents using strategy \(i\) by \(\pm 1\) in each time step can be written as

\[
T^\pm(k) = \frac{Z-k}{Z} k \left[ 1 + e^{\mp \beta [P_{ij}(k)-P_{ji}(k)]} \right]^{-1},
\] (23)

with \(T^+\) corresponding to an increase from \(k\) to \(k+1\) and \(T^-\) corresponding to the opposite. As before, \(\beta\) 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\(^{1,3}\). 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 $j^1$, which holds for any intensity of selection and in the limit of large population size $Z$ when

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

(24)
3 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 $\gamma$ ($0 \leq \gamma \leq 1$), 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

$$\Pi = \begin{pmatrix}
\text{AS} & \text{AU} & \text{CS} \\
\text{AS} & \frac{B}{2W} + \pi_{11} & (1 - p_r \gamma) \pi_{12} & \frac{B}{2W} + \pi_{11} \\
\text{AU} & (1 - p_r) \left( \frac{sB}{W} + \pi_{21} \right) & (1 - p_r) \left( \frac{sB}{W} + \pi_{22} \right) & \left(1 - p_r\right) \left[ \frac{sB}{W} + \frac{s}{W} \left( \pi_{21} + \left( \frac{W}{s} - 1 \right) \pi_{22} \right) \right] \\
\text{CS} & \frac{B}{2W} + \pi_{11} & (1 - p_r \gamma) \frac{s}{W} \left( \pi_{12} + \left( \frac{W}{s} - 1 \right) \pi_{22} \right) & \frac{B}{2W} + \pi_{11} \\
\end{pmatrix}.$$ (25)

Figure S10 shows the results for different values of $\gamma$ 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 $\gamma$ increases the frequency of AU (The condition under which SAFE is the more beneficial collective outcome, does not depend at all on $\gamma$). They can be
written as follows, respectively

\[ \pi_{11} + (1 - pr\gamma)\pi_{12} > (1 - pr)(\pi_{21} + \pi_{22}). \]  
(26)

\[ (1 - pr\gamma)\pi_{22} + \pi_{11} > 2(1 - pr)\pi_{22}. \]  
(27)

which are equivalent to, respectively

\[ pr > \frac{\pi_{21} + \pi_{22} - \pi_{11} - \pi_{12}}{\pi_{21} + \pi_{22} - \gamma\pi_{12}} \]  
(28)

\[ pr > \frac{\pi_{22} - \pi_{11}}{\pi_{22}(2 - \gamma)} = \frac{1}{2 - \gamma} - \frac{\pi_{11}}{\pi_{22}(2 - \gamma)} \]  
(29)

We can see that the right hand side of the condition of CS is an increasing function of \( \gamma \), 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 \( \gamma \) enlarges the innovation zones (see the red parts) and reduces the dilemma zone.

Next, similar analysis can be done for the \textit{N-team AI race}. The payoffs of AS and CS when playing with AU is scaled by a factor \( (1 - pr\gamma) \) 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 AIS 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 AIS, AU is dominant for a larger range for increasing \( \gamma \), see Figure S12.
Figure S10. Different regimes of AIS 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 $\gamma$ slightly 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 $\gamma$.

Parameters: $p_r = 0.6$ (first two columns); $c = 1, b = 4, B = 10000, \beta = 0.1, Z = 100$. 
Figure S11. Late AIS: 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 $\gamma$ 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 AIS for N-player race: Frequency of AU when $\gamma = 0$ (top row); $\gamma = 0.5$ (middle row) and $\gamma = 1$ (bottom row). Increasing $\gamma$ enlarges the innovation zones (red parts).
Parameters: $c = 1$, $b = 6$, $s = 1.5$, $W = 10^6$, $B = 10000$, $\beta = 0.1$, $Z = 100$. 
4 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

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

(30)

where

\[
\tilde{\pi}_{21} = \frac{1}{u} \sum_{i=1}^{u} (1 - p_{fo})^i \frac{sB}{s+1} = H_u \tilde{\pi}_{21},
\]

\[
\tilde{\pi}_{22} = \frac{1}{u} \sum_{i=1}^{u} (1 - p_{fo})^i \frac{(1+p_{fo})b}{2} = H_u \tilde{\pi}_{22},
\]

\[
\tilde{\pi}_{12} = -c + \frac{1}{u} \sum_{i=1}^{u} (1 - p_{fo})^{i-1} \left( (1 - p_{fo}) \frac{sB}{s+1} + p_{fo}(u + 1 - i)b \right)
= -c + H_u (1 - p_{fo}) \frac{b}{s+1} + \left( p_{fo}(u + 1)H_u + \frac{1-(1-p_{fo})^u}{up_{fo}} - (1 - p_{fo})^u \right) b
\]

where \( H_u = \sum_{i=0}^{u-1} \frac{(1-p_{fo})^i}{u} \leq 1 \)

Thus, exactly the same results are obtained in the early AIS since changing \( u \) does not influence the chance of winning the prizes for all strategies.

In the late AIS (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

\[
\tilde{\pi}_{21} \to 0, \quad \tilde{\pi}_{22} \to 0, \quad \tilde{\pi}_{12} \to -c + b
\]
5 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 AIS 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.

References

1. Chaitanya S. Gokhale and Arne Traulsen. Evolutionary games in the multiverse. *Proc. Natl. Acad. Sci. U.S.A.*, 107(12):5500–5504, March 2010.

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

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

4. Karl Sigmund. *The Calculus of Selfishness*. Princeton University Press, 2010.
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: $c = 1$, $b = 4$, $B = 10000$, $\beta = 0.01$, $Z = 100$. 
Figure S14. Late AIS: Average population payoff (social welfare). Same parameter settings 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: $c = 1$, $b = 4$, $B = 10000$, $Z = 100$. 
Figure S15. Late AIS: Average population payoff (social welfare) for varying $p_{fo}$ and different values of $p_r$. 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: $c = 1, b = 4, B = 10000, s = 1.5, \beta = 0.1, Z = 100$. 