scientific reports

Check for updates

OPEN

Artificial intelligence development races in heterogeneous settings

Theodor Cimpeanu[1], Francisco C. Santos[3], Luís Moniz Pereira[2], Tom Lenaerts[4,5,6,7] & The Anh Han[1]✉

Regulation of advanced technologies such as Artificial Intelligence (AI) has become increasingly important, given the associated risks and apparent ethical issues. With the great benefits promised from being able to first supply such technologies, safety precautions and societal consequences might be ignored or shortchanged in exchange for speeding up the development, therefore engendering a racing narrative among the developers. Starting from a game-theoretical model describing an idealised technology race in a fully connected world of players, here we investigate how different interaction structures among race participants can alter collective choices and requirements for regulatory actions. Our findings indicate that, when participants portray a strong diversity in terms of connections and peer-influence (e.g., when scale-free networks shape interactions among parties), the conflicts that exist in homogeneous settings are significantly reduced, thereby lessening the need for regulatory actions. Furthermore, our results suggest that technology governance and regulation may profit from the world's patent heterogeneity and inequality among firms and nations, so as to enable the design and implementation of meticulous interventions on a minority of participants, which is capable of influencing an entire population towards an ethical and sustainable use of advanced technologies.

Researchers and stakeholders alike have urged for due diligence in regard to AI development on the basis of several concerns. Not least among them is that AI systems could easily be applied to nefarious purposes, such as espionage or cyberterrorism[1]. Moreover, the desire to be at the foreground of the state-of-the-art or the pressure imposed by upper management might tempt developers to ignore safety procedures or ethical consequences[2,3]. Indeed, such concerns have been expressed in many forms, from letters of scientists against the use of AI in military applications[4,5], to blogs of AI experts requesting careful communications[6], and proclamations on the ethical use of AI[7–10].

Regulation and governance of advanced technologies such as Artificial Intelligence (AI) has become increasingly more important given their potential implications, such as associated risks and ethical issues[4,5,7–12]. With the great benefits promised from being first able to supply such technologies, stake-holders might cut corners on safety precautions in order to ensure a rapid deployment, in a race towards AI market supremacy (AIS)[2,3]. One does not need to look very far to find potentially disastrous scenarios associated with AI[2,13–15], but accurately predicting outcomes and accounting for these risks is exceedingly difficult in the face of uncertainty[16]. As part of the double-bind problem put forward by the Collingridge Dilemma, the impact of a new technology is difficult to predict before it has been already extensively developed and widely adopted, and also difficult to control or change after it has become entrenched[17]. Given the lack of available data and the inherent unpredictability involved in this new field of technology, a modelling approach is therefore desirable to provide a better grasp of any expectations with regard to a race for AIS. Such modelling allows for dynamic descriptions of several key features of the AI race (or its parts), providing an understanding of possible outcomes, considering external factors and conditions, and the ramifications of any policies that aim to regulate such race.

With this aim in mind, a baseline model of an innovation race has been recently proposed[18], in which innovation dynamics are pictured through the lens of Evolutionary Game Theory (EGT) and all race participants are equally well-connected in the system (well-mixed populations). The baseline results showed the importance of accounting for different time-scales of development, and also exposed the dilemmas that arise when what is

[1]School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough TS1 3BA, UK. [2]NOVA Laboratory for Computer Science and Informatics (NOVA-LINCS), Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal. [3]INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal. [4]Machine Learning Group, Université Libre de Bruxelles, 1050 Brussels, Belgium. [5]Artificial Intelligence Lab, Vrije Universiteit Brussel, 1050 Brussels, Belgium. [6]Center for Human-Compatible AI, University of California, Berkeley 94702, USA. [7]FARI Institute, Université Libre de Bruxelles-Vrije Universiteit Brussel, 1050 Brussels, Belgium. ✉email: T.Han@tees.ac.uk

individually preferred by developers differs from what is globally beneficial. When domain supremacy could be achieved in the short-term, unsafe development required culling for to promote the welfare of society, and the opposite was true for the very long term, to prevent excessive regulation at the start of exploration. However, real-world stakeholders and their interactions are far from homogeneous. Some individuals are more influential than others, or play different roles in the unfolding of new technologies. Technology races are shaped by complex networks of exchange, influence, and competition where diversity abounds. It has been shown that particular networks of contacts can promote the evolution of positive behaviours in various settings, including cooperation[19–24], fairness[25–29] and trust[30]. In this paper, we take inspiration from the disconnect between the previous line of research and the heterogeneity observed in real-world interactions, and ask whether network topology can influence the adoption of safety measures in innovation dynamics, and shape the tensions of the AI development race.

The impact of network topology is particularly important in the context of technology regulation and governance. Technology innovation and collaboration networks (e.g. among firms, stakeholders and AI researchers) are highly heterogeneous[31,32]. Developers or development teams interact more frequently within their groups than without, forming alliances and networks of followers and collaborators[33,34]. Many companies compete in several markets while others compete in only a few, and their positions in inter-organisational networks strongly influence their behaviour (such as resource sharing) and innovation outcome[34,35]. It is important to understand how diversity in the network of contacts influences race dynamics and the conditions under which regulatory actions are needed. Therefore, we depart from a minimal AI race model[18], examining instead how network structures influence safety decision making within an AI development race.

In a structured population, players are competing with co-players in their network neighbourhoods. Firms interact or directly compete through complex ties of competition, such that some players may play a pivotal role in a global outcome. Here we abstract these relationships as a graph or a network. We compare different forms of network structures, from homogeneous ones—such as complete graphs (equivalent to well-mixed populations), and square lattices—to different types of scale-free networks[36] (see "Methods"), representing different levels of diversity in the number of co-player races a player can compete in. Our results show that when race participants are distributed in a heterogeneous network, the conflicting tensions arising in the well-mixed case are significantly reduced, thereby softening the need for regulatory actions. This is, however, not the case when the network is not accompanied by some degree of relational heterogeneity, even in different types of spatial lattice networks.

In the following sections, we describe the models in detail, then present our results.

Models and summary of previous results

We first define the AI race game[18] and recall relevant results from previous works in the well-mixed populations setting.

AI race model definition. Assuming that winning the race towards supremacy is the goal of the development teams (or players) and that a number of development steps (or advancements/rounds) are required, the players have two strategic options in each step: to follow safety precautions (denoted by strategy SAFE) or to ignore them (denoted by strategy UNSAFE)[18]. As it takes more time and effort to comply with the precautionary requirements, playing SAFE is not only costlier, but also implies a slower development speed, compared to playing UNSAFE. Let us also assume that to play SAFE, players need to pay a cost c, whereas the opposite strategy is free. The increase in speed when playing UNSAFE is given by a free parameter $s > 1$, while the speed when playing SAFE is normalised to 1. The interactions are iterated until one or more teams achieve a designated objective, after having completed W development steps. As a result, the players obtain a large benefit B, shared among those who reach the target objective at the same time. However, a setback or disaster can happen with some probability, which is assumed to increase with the number of times the safety requirements have been omitted by the winning team(s). Although many potential AI disaster scenarios have been sketched[2,14], the uncertainties in accurately predicting these outcomes are high. When such a disaster occurs, risk-taking participants lose all their benefits. We denote by p_r the risk probability of such a disaster occurring when no safety precaution is followed at all.

We model an AI development race 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 share this benefit proportionally to their development speed. Moreover, we assume that with some probability p_{fo} those playing UNSAFE might be found out, wherein their disregard for safety precautions is exposed, leading to their products not being adopted due to safety concerns, thus receiving 0 benefit. Thus, in each round of the race, we can write the payoff matrix as follows (with respect to the row player)

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

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 (with probability $p_f o$) the full benefit b in case the UNSAFE co-player is found out, and obtains (with probability $1 - p_f o$) a small part of the benefit $b/(s+1)$ otherwise, dependent on the co-player's speed of development s. When playing with a SAFE

Parameter	Symbol	Range analysed
Population size	Z	{100, 1000, 1024}
Intensity of selection	β	{1}
Average connectivity of a scale-free network	z	{4}
Number of new edges for each new node in SF networks	m	{2}
Probability of being found out when playing unsafe	p_{fo}	{0, 0.05, 0.1, ..., 1}
Probability of disaster occurring due to unsafe development	p_r	{0, 0.05, 0.1, ..., 1}
Benefit of winning the race (reaching AI supremacy)	B	$\{10^4\}$
Benefit of intermediate AI advancements	b	{4}
Cost of adhering to safety standards	c	{1}
Speed of development (due to disregarding safety)	s	{1, 1.25, 1.5, ..., 5}
Number of development rounds until AI supremacy is reached	W	$\{100, 10^6\}$

Table 1. Model parameters and parameter space analysed.

player, the UNSAFE one 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 one, it obtains the shared benefit $b/2$ when both are not found out, but the full benefit b when it is not found out while the co-player is found out, and 0 otherwise. The corresponding average payoff is thus: $(1-p_f o)\left[(1-p_f o)(b/2)+p_f o b\right] = (1-p_f^2 o)\frac{b}{2}$.

In the AI development process, players repeatedly interact (or compete) with each other using the *innovation* game described above. In order to clearly examine the effect of population structures on the overall outcomes of the AI race, in line with previous network reciprocity analyses (e.g. in social dilemma games[21,37,38]), we focus in this paper on two unconditional strategies[18]:

- AS (always complies with safety precautions)
- AU (never complies with safety precautions)

Denoting by Π_{ij} $(i,j \in \{1,2\})$ the entries of the matrix Π above, the payoff matrix defining the averaged payoffs for AU vs AS reads

$$\begin{array}{cc} & \begin{array}{cc} AS & \quad AU \end{array} \\ \begin{array}{c} AS \\ AU \end{array} & \left(\begin{array}{cc} \frac{B}{2W} + \Pi_{11} & \Pi_{12} \\ (1-p_r)\left(\frac{sB}{W} + \Pi_{21}\right) & (1-p_r)\left(\frac{sB}{2W} + \Pi_{22}\right) \end{array} \right). \end{array} \tag{2}$$

As described in Eq. (2), we encounter the following scenarios. When only two safe players interact, they complete the race simultaneously after an average of W development rounds, thereby obtaining the averaged split of the full prize $\frac{B}{2W}$ per round; furthermore, the safe players also obtain the intermediate benefit per round (π_{11}, see Eq. (1)). When a safe player only encounters an unsafe player, the only benefit obtained by the safe player is the intermediary benefit in each round, whereas the unsafe player receives the full prize B; moreover, the unsafe player completes the race in $\frac{W}{s}$ development rounds, so it receives an extra average of $\frac{sB}{W}$ of the full prize per round. Furthermore, the unsafe behaviour attracts the possibility of a disaster occurring, causing them to lose all gains, with probability p_r, which is reflected in the payoff matrix (consider π_{22} in Eq. (1)). Similarly, we can extract the average payoffs for solely two unsafe players interacting, by considering that they finish the race at the same time and get the appropriate intermediate benefit π_{22} (see Eq. 1).

Summary of previous results in well-mixed settings. In order to clearly present the contribution of the present work, we next recall the analytical conditions derived in[18] and how these will be used to inform the analysis that follows. Our analysis will differentiate between two development regimes: an early/short-term regime and a late/long-term one. The difference in time-scale between the two regimes plays a key role in identifying which regulatory actions are needed and when. This distinction is in line with previous works adopting analytical approaches using stochastic population dynamics[18]. The early regime is underpinned by the race participants' ability to readily reach the ultimate prize B in the shortest time frame available. In other words, winning the ultimate prize in W rounds is much more important than any benefits achieved in single rounds until then, i.e. $B/W >> b$. Contrarily, a late regime is defined by a desire to do well in each development round, as technological supremacy cannot be achieved in the foreseeable future. That is, singular gains b, even when accounting for the safety cost c, become more tempting than aiming towards winning the ultimate prize, i.e. $B/W << b$. For a reminder of the meanings of the parameters described above, see Table 1.

We have also made use of the previous analytical results[18] which identify the risk-dominant boundaries of the AI race game for both early and late development regimes in well-mixed populations. These are useful as a baseline or reference model, determining the regions in which regulatory actions are needed or otherwise, and moreover, if needed, which behaviour should be promoted. In the early regime, the two dotted lines mark region

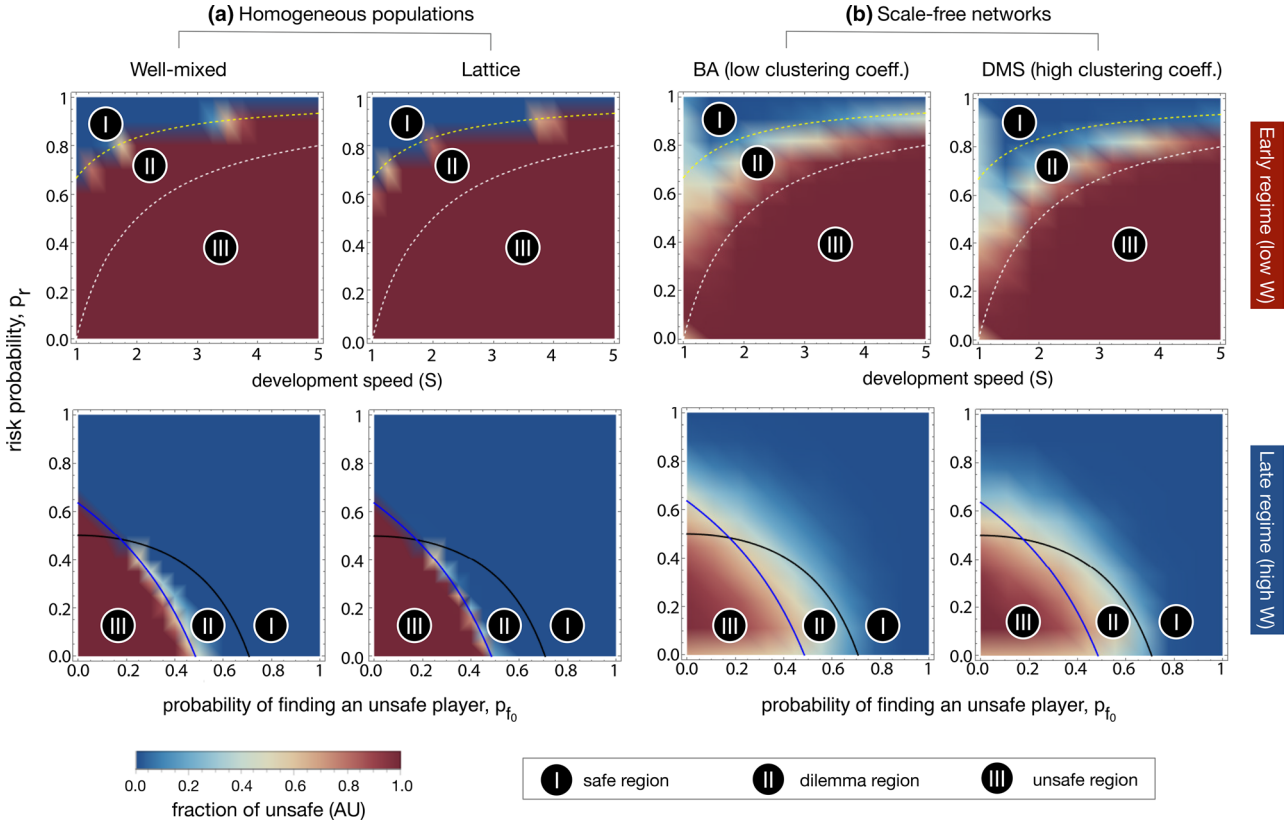

Figure 1. Color gradients indicating the average fraction of AU (unsafe strategy) for (**a**) homogeneous (well-mixed and lattices) populations and (**b**) scale-free networks (BA and DMS models). The top row addresses the early regime (low W) for varying development speed (s) and risk probability (p_r). The bottom row addresses the late regime (high W) for varying p_{fo} (the chances that an UNSAFE player is found out) and risk probability (p_r). Dotted and full lines indicate the phase diagram obtained analytically[18]. In the early regime (upper panels), region II indicates the parameters in which safe AI development is the preferred collective outcome, but unsafe development is expected to emerge and regulation may be needed—thus the dilemma. In regions I and III, safe and unsafe AI development, respectively, are both the preferred collective outcomes and the ones expected to emerge from self-organization, hence not requiring regulation. In the late regime (lower panels), the solid black line marks the boundary above which safety is the preferred outcome, whereas the blue line indicates the boundary above which safety becomes risk dominant against unsafe development. The results obtained for well-mixed populations and lattices (**a**) suggest that, for both early and late regimes, the nature of the dilemma, as represented by the analytical phase diagram, remains unaltered. Moreover, homogeneous interaction structures cannot reduce the need for regulation in the early regime. Differently, we show that heterogeneous interaction structures (scale-free networks, (**b**)) are able to significantly reduce the prevalence of unsafe behaviors for almost all parameter regions, including both late and early regimes. This effect is enhanced whenever scale-free networks are combined with high clustering coefficient (i.e., in the DMS model). Other parameters: $p_{fo} = 0.5$, and $W = 100$ (top panels); $s = 1.5$ and $W = 10^6$ (bottom panels); $c = 1, b = 4, B = 10^4$, and $\beta = 1$, in all panels.

(II) within the boundaries $p_r \in [1 - 1/s,\ 1 - 1/(3s)]$ for which safety development is the preferred collective outcome, but where unsafe development is selected for by social dynamics (see e.g. Fig. 1, first row). Thus, in this region (II), regulation is required to improve safety compliance. Outside of these boundaries, safe (in region I) and unsafe (in region III), respectively, are both the preferred collective outcomes and the ones selected for by social dynamics, hence requiring no regulatory actions. For the late AI race (e.g. Fig. 1, bottom row), the solid black line marks the boundary above which safety is the preferred collective outcome, where $p_r < 1 - \frac{b-2c}{b(1-p_{fo}^2)}$, whereas the blue line indicates where AS becomes risk-dominant against AU, where $p_r < \frac{4c(s+1)+2b(s-1)}{b(1+3s)}$. Again, in this regime three regions can be distinguished, with (I) and (III) having similar meanings to those in the early regime. However, differently from the early regime, in region (II) regulatory actions are needed to improve (unsafe) innovation instead of safety compliance, due to the low risk. These regions are derived from the analytical conditions described in[18], where these are explained in further detail.

Results

Based on extensive computer simulations (see "Methods"), our analysis identifies the prevalence of individuals adopting unsafe procedures after reaching a stationary state and infer the most likely behavioural trends and patterns associated with the agents taking part in the AI race game for distinct network topologies. The main findings from this work are described in this section, whereby each subsection will provide a key insight followed by the results and intuitions which motivate each claim.

Heterogeneous interactions reduce the viability of unsafe development in both short and long-term races. We examine the impact of different network structures, homogeneous and heterogeneous, on the safety outcome of the evolutionary dynamics for the two different development regimes descried above. To commence our analysis, we first study the role of degree-homogeneous graphs (here illustrated by structural spatiality) in the evolution of strategies in the AI race game. First, we simulated the AI race game in well-mixed populations (see Fig. 1, first column). We then explored the same game on a square lattice, where each agent can interact with its four edge neighbours, in Fig. 1 (second column). We show that the trends remain the same when compared with well-mixed populations, with very slight differences in numerical values between the two. Specifically, towards the top of area (Region II), at the risk-dominant boundary between AS and AU players in the case of an early AI race, we see some safe developmental activity where previously there was none. In practice, this shifts the boundary very slightly towards an optimal conclusion.

Thus, except for minute atypical situations, we may argue that homogeneous spatial variation is not enough to influence safe technological development, with minimal improvement when compared with a well-mixed population (complete network). To further increase our confidence that such structures have very small effects on the AI race game, we confirm that 8-neighbour lattices (where agents can also interact with corner neighbours) yield very similar trends, with negligible differences when compared to either the regular square lattice or well-mixed populations (see Supplementary Information, Fig. S2).

As a means of investigating beyond simple homogeneous structures and their roles in the evolution of appropriate developmental practices in the AI race, we make use of the previously defined BA and DMS network models (see "Methods"). Contrary to the findings on homogeneous networks, scale-free interaction structures produce marked improvements in almost all parameter regions of the AI race game (see Fig. 1).

Previously, it has been suggested that different approaches to regulation were required, subject to the time-line and risk region in which the AI development race is placed, after inferring the preferences developers would have towards safety compliance[18]. Given that innovation in the field of AI (or more broadly, technological advancement as a whole), should be profitable (and robust) to developers, shareholders and society altogether, we must therefore discuss the analytical loci where these objectives can be fulfilled. Assuredly, we see that diversity in players introduces two marked improvements in both early and late safety regimes. Firstly and most importantly, we note that very little regulation is required in the case of a late AI race (large W), principally concerning the existing observations in homogeneous settings (e.g., well-mixed populations and lattices). Intuitively, this suggests that there is little encouragement needed to promote risk-taking in late AIS regimes: Diversity enables beneficial innovation. Secondly, the region for early AIS regimes in which regulation must be enforced is diminished, but not completely eliminated. Consequently, governance should still be prescribed when developers are racing towards an early or otherwise unidentified AI regime (based on the number of development steps or risk of disaster). It stands to reason that insight into what regime type the AI race operates in, is therefore paramount to the success of any potential regulatory actions. The following sections will attempt to look further into assessing these observations.

Figure 1 (top panels) presents a fine-grained glimpse into the early regime. In region (II), the safety dilemma zone, social welfare is once more conspicuously improved by heterogeneity. Concerted safe behaviour is favoured, even in the face of being disregarded by social dynamics in the analytical sense. We discern the clear improvements discussed earlier, but also echo the messages put forward in[18]. We contend that it is vital for regulators to intervene in these conditions, for encouraging pro-social, safe conduct, and in doing so avert conceivably dangerous outcomes. Heterogeneity lessens the burden on policy makers, allowing for greater freedom in the errors and oversights that could occur in governing towards the goal of safe AI development.

While the difference between heterogeneous and homogeneous networks is evident, there also exists a distinction between the different types of heterogeneous networks. In this paper we discuss the BA and DMS models, and also their normalised counterparts, in which individuals' payoffs are divided by the number of neighbours. In such scenarios one could assume that there is an inherent cost to maintaining a link to another agent. In this sense, there exists some levelling of the payoffs, seemingly increasing fairness and reducing wealth inequality. But we confirm that normalising the network leads to similar dynamics as observed in homogeneous populations (see Fig. S3), with only very slight differences.

In order to accurately depict the measured differences between the different types of networks, we varied the risk probability (p_r) for both the early and late regime. We report the results of this analysis in Fig. 2, where we also show the preferred collective outcome, using the different regions described earlier in this section. These figures help expose the effect of heterogeneity on the frequency of unsafe behaviour in the different dilemma zones. In particular, we notice a mediating effect in the requirements for regulation, for both regimes and types of scale-free networks.

Specifically, in the case of the early regime (see Fig. 2, left column), we observe the presence of safety for a much broader range of risk probability values, than in the case of either well-mixed or structured populations. In the late regime (see Fig. 2, right column), however, we also highlight an increase in unsafe behaviour even beyond the boundary for which safety would have been the preferred collective outcome. In this case, heterogeneity has its drawbacks. On the one hand, innovative behaviour sees some improvement when it is in the interest

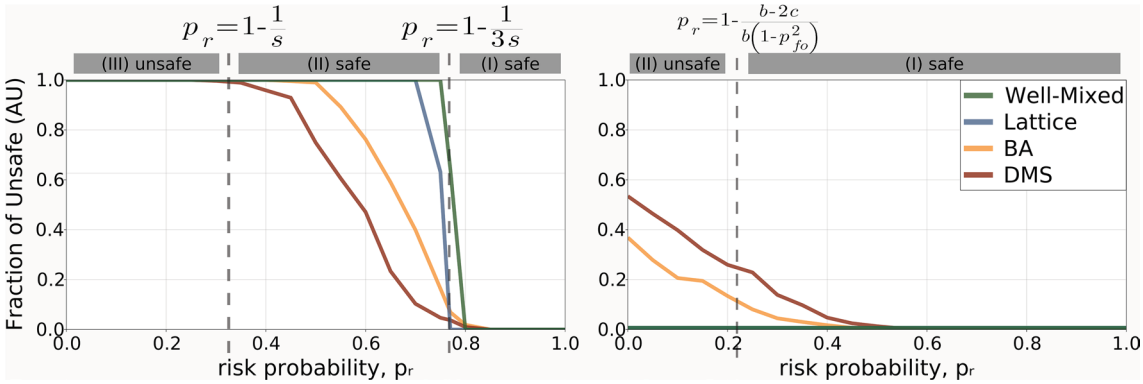

Figure 2. Heterogeneous networks moderate the need for regulation, shown by measuring frequency of unsafe developments across a range of different risk probabilities. The boundaries between zones are indicated with blue dashed lines, whereas the grey-highlighted texts on top of the figures indicate the collectively desired behaviour in each zone. The left panel reports the results for the early regime ($p_{fo} = 0.5,\ W = 100$), while the right panel does so for the late regime ($p_{fo} = 0.6,\ W = 10^6$) (parameter values are chosen for a clear illustration). Parameters: $c = 1,\ b = 4,\ B = 10^4,\ s = 1.5,\ \beta = 1$.

of the common good for it to be so, but the same is true, albeit rarely, when it is not. We also note that the effects described above are amplified in the case of DMS networks, in comparison to their BA counterparts. Observing a high degree of inter-group interactions (clustering) may play a key role in determining if intervention is required in the AI race. Moreover, we confirm these findings by producing typical runs showing the time evolution of unsafe behaviour for each network type (please see Fig. S1).

Hubs and their role in decelerating the race. Highly connected individuals (hubs) typically play a key role in many real-world networks of contacts and change the dynamics observed in heterogeneous populations[21,24,37,39]. In order to study the role that hubs play in the AI race, in the context of scale-free networks, we classify nodes into three separate connectivity classes[21]. We obtain three classes of individuals, based on their number of contacts (links) k_i and the average network connectivity z:

1. Low degree, whenever $k_i < z$,
2. Medium degree, whenever $z \leq k_i < \frac{k_{max}}{3}$ and
3. High degree (hubs), whenever $\frac{k_{max}}{3} \leq k_i \leq k_{max}$.

Dedicated minorities are often identified as major drivers in the emergence of collective behaviours in social, physical and biological systems, see[40–43]. Given the relative importance of hubs in other systems, we explore whether highly connected, committed individuals are prime targets for safety regulation in the AI race. By introducing individuals with pathologically safe tendencies (fixed behaviours)[41]—these are sometimes referred to as zealots, see[30,40,41,43]—in the network, we can better understand the power of influential devotees in the safe development of a general AI.

We progressively introduce pathological safe players based on their degree centrality (i.e. number of connections). In other words, the most connected nodes will be the first to be targeted. The benefits of this approach are twofold, as they allow us to study the relative differences between the three classes of individuals, but also the effect of regulating the key developers in the AI race. For a full analysis of the differences between high, medium and low degree individuals in the baseline case, please see Fig. S5.

Hubs prefer slower, safer developments in the early AI race, and this can be further exploited by introducing safety zealots in key locations in the network (see Fig. 3). When safety is the preferred collective outcome, hubs can drive the population away from unsafe development, and this effect is even more apparent in the case of highly clustered scale-free networks (Fig. 3, right column). Following the sharp increase in global safety after the conversion of high degree players to zealotry, we also observe a similar, but not as pronounced influence as the most connected medium degree individuals follow suit, an effect which plateaus shortly thereafter. We further confirm these results by selecting the same targets (the top 10% of individuals based on degree centrality), but introducing them in reverse order (i.e. starting with the highest connected medium degree individuals and ending with the most connected high degree ones; see Fig. S9).

Whereas the capacity of hubs to drive the population towards safety is evident in region II of the early regime (when safety is the collective preferred outcome), the opposite is true for region III. High degree individuals are more capable at influencing the overall population than medium degree individuals, even the most highly connected ones, but we see a much more gradual decrease in innovation as the most connected nodes are steadily converted to zealotry. Even in the presence of great uncertainty, a small percentage of very well connected developers can ensure safety with very little negative impact on innovation.

Conversely, we show that highly connected individuals prefer innovation in the late regime, irrespective of the preferred collective outcome. But even the introduction of one pathological safe player (0.1% of the population) in the largest hub is enough to ensure that the entire population converges to safe development in most instances.

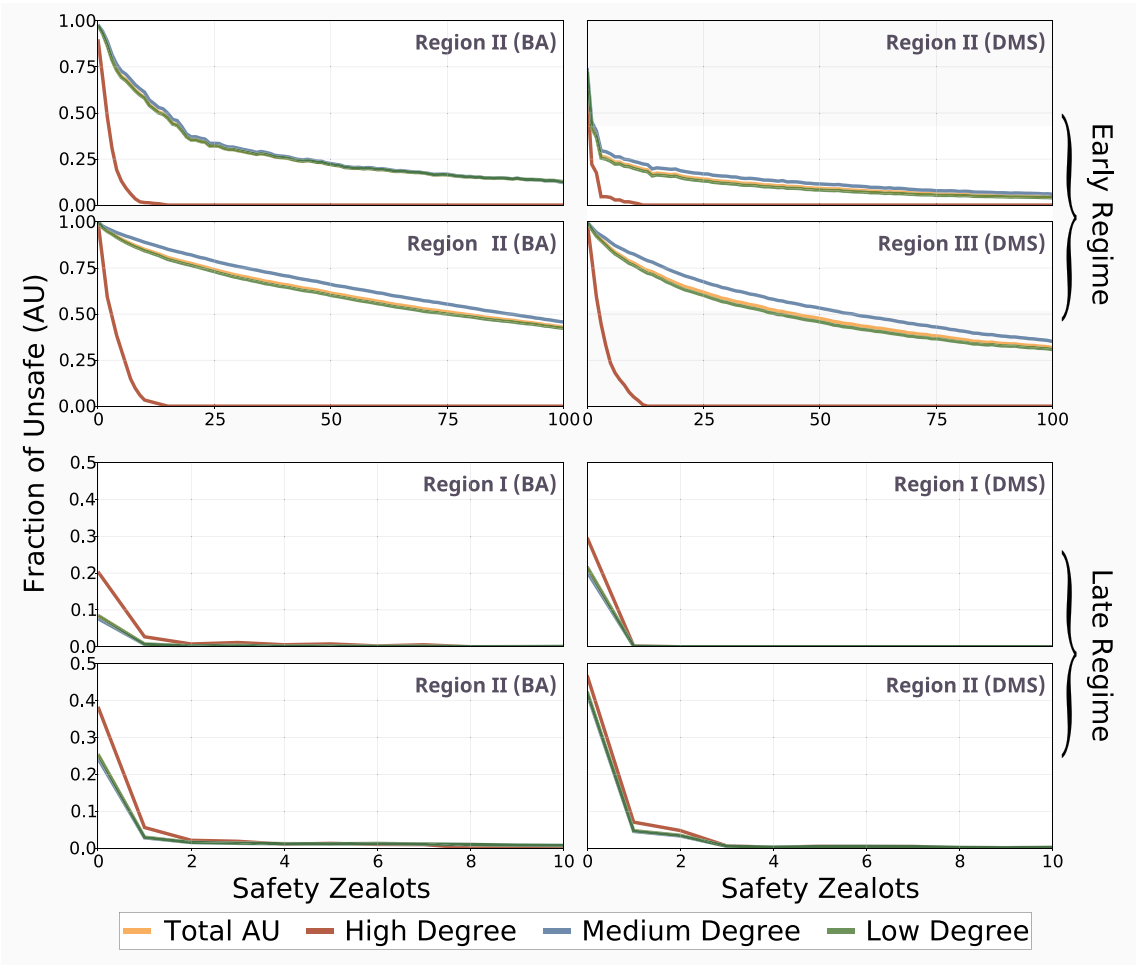

Figure 3. Hubs prefer slower, thus safer developments in the early race, and this can be further exploited by progressively introducing safety zealots in highly connected nodes. We show the results for both regimes, as well as the appropriate regions where safety (early region II and late region I), and conversely where innovation (early region III and late region II) are the preferred collective outcomes. The top four panels report the results for the early regime ($p_{fo} = 0.5,\ W = 100$ with $p_r = 0.5$ for region II and $p_r = 0.1$ for region III), and the bottom four do so for the late regime ($p_{fo} = 0.6,\ W = 10^6$ with $p_r = 0.3$ for region I and $p_r = 0.1$ and region II). We show a subset of the results in the late regime for clear representation; see Fig. S8 for a comprehensive view. Other parameters: $c = 1,\ b = 4,\ B = 10^4,\ s = 1.5,\ \beta = 1$.

In the case when safety is socially preferred (third row of Fig. 3), the successful regulation of the AI race requires a very small minority of individuals to dedicate themselves to safety, but in cases of uncertainty, innovation is very easy to stifle in the late regime, even when it would be beneficial not to do so (region II).

A small minority of highly connected individuals can help mitigate race tensions under uncertainty. Uncertainty can limit the options of regulatory agencies in the quest towards the development of safe AI, and narrow solutions to regulation could have potentially disastrous consequences, given the existential risk that general AI poses to humanity[44]. Moreover, the promised benefit of such technology is great enough that stifling innovation could be nearly as harmful as the catastrophic consequences themselves, given the potential solutions that such technology could provide to problems in the context of existential risk, healthcare, politics, and many other fields (e.g.[45,46]).

To provide general solutions to the problem of regulating the AI race, we explore the impact of safety zealots (as discussed in the previous section) across the whole range of possible scenarios. We cannot be sure of the nature of the network of contacts that governs real-world AI developers, nor the actual timeline of the race. We show that by enforcing safety for a very select minority of highly connected individuals, race tensions can be mitigated in nearly all cases (see Fig. 4). We provide a full analysis of the effect of zealots in well-mixed networks in Fig. S4, and note that the lack of heterogeneity produces nearly identical results to lattice networks.

Slowing key individuals in the early regime can dramatically reduce existential risk in the case of heterogeneous interactions. For both regions, hubs in DMS networks can drive the other nodes towards safety (see Fig. 4, top panels), but the reduction in unsafe developments in region II is significantly higher than in region III for low numbers of safety zealots. Outside of the few individuals that are converted to zealots, other nodes maintain

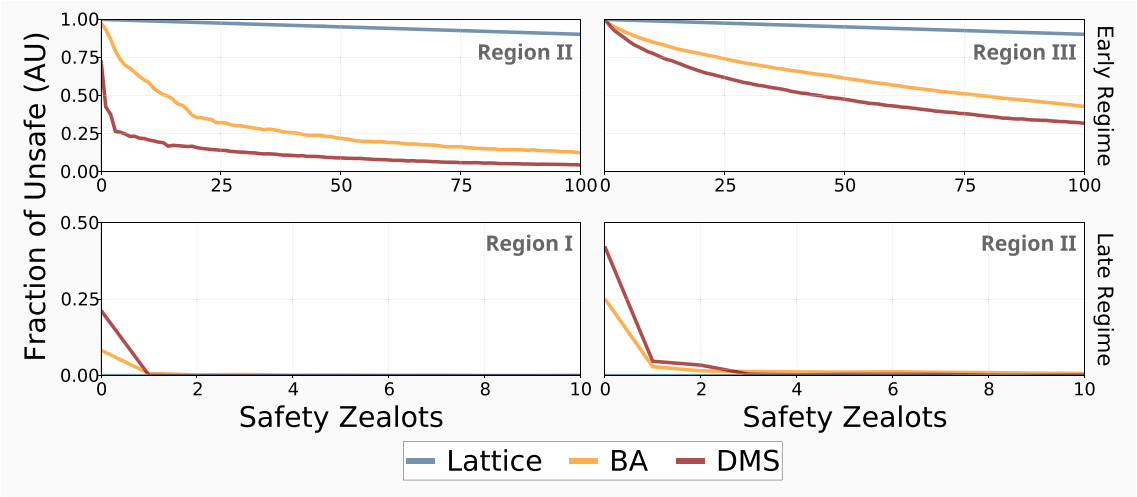

Figure 4. Introducing a small number of safety zealots can mitigate race tensions under uncertainty. We show the results for both regimes, as well as the appropriate regions where safety (early region II and late region I), and conversely where innovation (early region III and late region II) are the preferred collective outcomes. The top panels report the results for the early regime ($p_{fo} = 0.5,\ W = 100$ with $p_r = 0.5$ for region II and $p_r = 0.1$ for region III), and the bottom do so for the late regime ($p_{fo} = 0.6,\ W = 10^6$ with $p_r = 0.3$ for region I and $p_r = 0.1$ and region II). We note that these values were chosen for clear representation. Other parameters: $c = 1,\ b = 4,\ B = 10^4,\ s = 1.5,\ \beta = 1$.

their speed and continually innovate in region III, which suggests that this approach could be fundamental to the governance of developmental races. We note that if the proportion of safety zealots is not high enough, this effect cannot be reproduced, even in the presence of additional interference (such as artificially funding zealots or accelerating their development); For a more thorough analysis, please see Figs. S6 and S7.

Given a drawn-out race, this small minority of zealots can negatively impact innovation in the late regime (region II), where the relative increase that heterogeneous interactions provided rapidly disappears as pathological players are introduced. On the other hand, conditional strategies have been shown to further diminish the need to promote innovation in these conditions[18], and the introduction of these advanced strategies in this model could eliminate the negative effects of safety zealots in this region.

Discussion

In this paper, we have considered the implications of network dynamics on a technological race for supremacy in the field of AI, with its implied risks and hazardous consequences[2,13,14]. We make use of a previously proposed evolutionary game theoretic model[18] and study how the tension and temptation resulting from the race can can be mediated, for both early and late development regimes.

Network reciprocity has been shown to promote the evolution of various positive outcomes in many settings[21,26,27,47] and, given the high levels of heterogeneity identified in the networks of firms, stakeholders and AI researchers[31,32], it is important to understand the effects of reciprocity and how it shapes the dynamics and global outcome of the development race. It is just as important to ensure that appropriate context-dependent regulatory actions are provided. This modelling approach and the associated results are applicable to other technologies and competitions, such as patent races or the development of biotechnology, pharmaceuticals, and climate change mitigation technology, where there is a significant advantage to be achieved by reaching some target first[48–52]. Given a sufficiently tempting potential gain, individuals are more likely to invest in high-risk technology[53], which suggests that these insights could be applicable to many similar fields in which risk and innovation must be constantly balanced.

It is noteworthy that, despite a number of proposals and debates on how to prevent, regulate, or resolve an AI race[1,3,54–59], only a few formal modelling studies have been proposed[2,18,60,61]. These works focus on homogeneous populations, where there are no inherent structures indicating the network of contacts among developing teams. Innovation dynamics (including AI) emerge from complex systems marked by a strong diversity in influence and companies' power. Firms create intricate networks of concurrent development, in which some develop a higher number of products, influencing and competing with a significant number of others. Our work advances this line of research, revealing the impact of these network structures among race participants, on the dynamics and global outcome of the development race.

We began by validating the analytical results obtained as a baseline in a completely homogeneous population[18], using extensive agent-based simulations. We then adopted a similar methodology to analyse the effects of gradually increasing network heterogeneity, equivalent to diversifying the connectivity and influence of the race participants. This was accomplished by studying square lattices, and later two types of scale-free networks with varying degrees of clustering, with and without normalised payoffs (i.e. wealth inequality). Our findings suggest that the race tensions previously found in homogeneous networks are lowered, but that this effect only occurs in the presence of a certain degree of relational heterogeneity. In other words, spatial complexity by itself is

not sufficient for the expectation of tempering the need for regulatory actions. Amongst all the network types studied, we found that scale-free networks with high clustering are the least demanding in terms of regulatory need, closely followed by regular scale-free networks.

The questions of how network structures and diversity influence the outcomes of behavioural dynamics, or the roles of network reciprocity, have been studied extensively in many fields, including Computer Science, Physics, Evolutionary Biology and Economics[21,22,34,37,38,47,62–65]. Network reciprocity can promote the evolution of positive behaviours in various settings including cooperation dilemmas[21,22,37,47], fairness[25–27] and trust[30]. Their applications are diverse, ranging from healthcare[32], to network interference and influence maximization[66–68], and to climate change[69]. The present work contributes new insights to this literature by studying the role of network reciprocity in the context of a technology development race. This strategy scenario is more intricate than the above-mentioned game theoretical scenarios (i.e., cooperation, trust and fairness) because, on the one hand, whether a social dilemma arises (where a collectively desired behaviour is not selected by evolutionary dynamics) depends on external factors (e.g., risk probability p_r in the early regime and monitoring probability p_{fo} in the late regime)[18]. On the other hand, the collectively desired behaviour in the arisen social dilemma is different depending on the time-scale in which the race occurs. Interestingly, regardless of this more complex nature of the scenario, the different desirable behaviours can always be promoted in heterogeneous networks.

As an avenue of exploring the role of prominent players in the development race, we make use of a previously proposed model of studying the influence of nodes based on their degrees of connectivity[21]. These highly connected individuals have a tendency towards safety compliance in comparison to their counterparts. In an attempt to exploit this effect, as well as to better understand the impact of such seemingly significant nodes, we introduced several pathological players[30,40,43] in key locations of the network (highly connected nodes). We showed the role of hubs in slowing development and promoting safety, and argue that a small minority of influential developers can drastically reduce race tensions in almost all cases. The addition of pathological participants in these important locations can play a key role in the emergence of safety, without sacrificing innovation, and this effect is robust under uncertain race conditions. Our contribution explains the effects of heterogeneity in the networks that underlie the interactions between developers and teams of developers. We contend that there exist several ways in which this type of network heterogeneity could be promoted by relevant decision-makers, but argue that such mechanisms merit a dedicated body of research. Some examples of this could include dynamical linking[70], whereby the relationship between two nodes could be altered by an outside decision-maker or the parties involved, or modifying the stakeholders' access to information, thereby amplifying selection dynamics[71].

We note that our analyses focus on the binary extremes of developer behaviour, safe or unsafe development, in an effort to focus an already expansive problem into a manageable discussion. The addition of conditional, mixed, or random strategies could provide the basis for a novel piece of work. As observed with conditionally safe players in the well-mixed scenario[18], we envisage that these additions would show little to no effect in the early regime, with the opposite being true for the late regime, at least in homogeneous settings.

In short, our results have shown that heterogeneous networks can significantly mediate the tensions observed in a well-mixed world, in both early and late development regimes[18], thereby reducing the need for regulatory actions. Since a real-world network of contacts among technological firms and developers/researchers appears to be highly non-homogeneous, our findings provide important insights for the design of technological regulation and governance frameworks (such as the one proposed in the EU White Paper[11]). Namely, the underlying structure of the relevant network (among developers and teams) needs to be carefully investigated to avoid for example unnecessary actions (i.e. regulating when that is not needed, as would have been otherwise suggested in homogeneous world models). Moreover, our findings suggest to increase heterogeneity or diversity in the network as a way to escape tensions arisen from a race for technological supremacy.

Methods

Below we describe different network structures and the details of how simulations on those networks are carried out.

Population dynamics. We consider a population of agents distributed on a network (see below for different network types), who are randomly assigned a strategy AS or AU. At each time step or generation, each agent plays the game with its immediate neighbours. The sucesss of each agent (i.e., its fitness) is the sum of the payoffs in all these encounters. In the SI, we also discuss the limit where scores are normalised by the number of interactions (i.e., the connection *degree* of a node)[72]. Each individual fitness, as detailed below, defines the time-evolution of strategies, as successful choices will tend to be imitated by their peers.

At the end of each generation, a randomly selected agent A with a fitness f_A chooses to copy the strategy of a randomly selected neighbour, agent B, with fitness f_B with a probability p that increases with their fitness difference. Here we adopt the well-studied Fermi update or pairwise comparison rule, where[73,74]:

$$p = (1 + e^{\beta(f_A - f_B)})^{-1}. \tag{3}$$

In this case, β conveniently describes the selection intensity—i.e., the importance of individual success in the imitations process: $\beta = 0$ represents neutral drift while $\beta \to \infty$ represents increasingly deterministic imitation[73]. Varying β allows capturing a wide range of update rules and levels of stochasticity, including those used by humans, as measured in lab experiments[75–77]. In line with previous works and lab experiments, we set $\beta = 1$ in our simulations, ensuring a high intensity of selection[78]. This update rule implicitly assumes an asynchronous update rule, where at most one imitation occurs at each time-step. We have nonetheless confirmed that similar results are obtained with a synchronous update rule.

Network topologies. Links in the network describe a relationship of proximity both in the interactional sense (whom the agents can interact with), but also observationally (whom the agents can imitate). Ergo, the network of interactions coincides with the imitation network[79]. As each network type converges at different rates and naturally presents with various degrees of heterogeneity, we choose different population sizes and maximum numbers of runs in the various experiments to account for this while optimising run-time.

Specifically, to study the effect of network structures on the safety outcome, we will analyse the following types of networks, from simple to more complex:

1. Well-mixed population (WM) (complete graph): each agent interacts with all other agents in a population,
2. Square lattice (SL) of size $Z = L \times L$ with periodic boundary conditions—a widely adopted population structure in population dynamics and evolutionary games (for a survey, see[38]). Each agent can only interact with its four immediate edge neighbours. We also study the 8-neighbour lattice for confirmation (see SI),
3. Scale-free (SF) networks[36,80,81], generated through two growing network models—the widely-adopted Barabási-Albert (BA) model[36,82] and the Dorogovtsev-Mendes-Samukhin (DMS) model[80,83], the latter of which allows us to assess the role of a large number of triangular motifs (i.e. high clustering coefficient). Both BA and DMS models portray a power-law degree distribution $P(k) \propto k^{-\gamma}$ with the same exponent $\gamma = 3$. In the BA model, graphs are generated via the combined mechanisms of growth and preferential attachment where new nodes preferentially attach to m existing nodes with a probability that is proportional to their already existing number of connections[36]. In the case of the DMS model, new connections are chosen based on an edge lottery: each new vertex attaches to both ends of randomly chosen edges, also connecting to m existing nodes. As such, we favour the the creation of triangular motifs, thereby enhancing the clustering coefficient of the graph. In both cases, the average connectivity is $z = 2m$.

Overall, WM populations offer a convenient baseline scenario, where interaction structure is absent. With the SL we introduce a network structure, yet one where all nodes can be seen as equivalent. Finally, the two SF models allow us to address the role of heterogeneous structures with low (BA) and high (DMS) clustering coefficients. The SF networks portray a heterogeneity which mimics the power-law distribution of wealth (and opportunities) of real-world settings.

Computer simulations. For well-mixed populations and lattice networks, we chose populations of $Z = 100$ agents and $Z = 32 \times 32$ agents, respectively. In contrast, for scale-free networks, we chose $Z = 1000$, while also pre-seeding with agents 10 different networks (of each type) on which to run all the experiments in an effort to minimise the effect of network topology and the initial, stochastic distributions of players. We chose an average connectivity of $z = 4$ for our SF networks, to coincide with the regular average connectivity in square lattices for the sake of comparison.

We simulated the evolutionary process for 10^4 generations (a generation corresponds to Z time-steps) in the case of scale-free networks and 10^3 generations otherwise. The equilibrium frequencies of each strategy were obtained by averaging over the final 10^3 steps. Each data point shown below was obtained from averaging over 25 independent realisations, for each of the 10 different instances used in each network topology.

Data availability

Code for simulations is available upon request.

Received: 25 November 2021; Accepted: 24 December 2021
Published online: 02 February 2022

References

1. Taddeo, M. & Floridi, L. Regulate artificial intelligence to avert cyber arms race. *Nature* **556**(7701), 296–298 (2018).
2. Armstrong, S., Bostrom, N. & Shulman, C. Racing to the precipice: A model of artificial intelligence development. *AI Soc.* **31**(2), 201–206 (2016).
3. Cave, S. & ÓhÉigeartaigh, S. An AI race for strategic advantage: rhetoric and risks. In *AAAI/ACM Conference on Artificial Intelligence, Ethics and Society* 36–40, (2018).
4. Future of Life Institute. *Autonomous Weapons: An Open Letter from AI & Robotics Researchers* (Technical report, Future of Life Institute, 2015).
5. Future of Life Institute. Lethal autonomous weapons pledge. https://futureoflife.org/lethal-autonomous-weapons-pledge/, (2019).
6. Brooks, R. 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.
7. Declaration, Montreal. 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).
8. Steels, L. & Lopez de Mantaras, R. The barcelona declaration for the proper development and usage of artificial intelligence in Europe. *AI Commun.* (Preprint):1–10, (2018).
9. Russell, S., Hauert, S., Altman, R. & Veloso, M. Ethics of artificial intelligence. *Nature* **521**(7553), 415–416 (2015).
10. Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **1**, 389–399. https://doi.org/10.1038/s42256-019-0088-2 (2019).
11. European Commission. *White paper on Artificial Intelligence - An European approach to excellence and trust* (Technical report, European Commission, 2020).
12. Perc, M., Ozer, M. & Hojnik, J. Social and juristic challenges of artificial intelligence. *Palgrave Commun.* **5**(1), 1–7 (2019).
13. Sotala, K. & Yampolskiy, R. V. Responses to catastrophic AGI risk: A survey. *Physica Scripta* **90**(1), 018001 (2014).
14. Pamlin, D. & Armstrong, S. *Global Challenges: 12 Risks that Threaten Human Civilization* (Global Challenges Foundation, 2015).
15. O'neil, C. Weapons of math destruction: How big data increases inequality and threatens democracy. *Crown* (2016).

16. Armstrong, S., Sotala, K. & Ó hÉigeartaigh, S. The errors, insights and lessons of famous AI predictions-and what they mean for the future. *J. Exp. Theor. Artif. Intell.* **26**(3), 317–342 (2014).
17. Collingridge, D. *The Social Control of Technology* (St. Martin's Press, 1980).
18. Han, T. A., Pereira, L. M., Santos, F. C. & Lenaerts, T. To regulate or not: A social dynamics analysis of an idealised AI race. *J. Artif. Intell. Res.* **69**, 881–921 (2020).
19. Santos, F. C., Pacheco, J. M. & Lenaerts, T. Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *Proc. Natl. Acad. Sci. USA* **103**(9), 3490–3494 (2006).
20. Ohtsuki, H., Hauert, C., Lieberman, E. & Nowak, M. A. A simple rule for the evolution of cooperation on graphs and social networks. *Nature* **441**(7092), 502–505 (2006).
21. Santos, F. C., Santos, M. D. & Pacheco, J. M. Social diversity promotes the emergence of cooperation in public goods games. *Nature* **454**, 214–216 (2008).
22. Perc, M. *et al.* Statistical physics of human cooperation. *Phys. Rep.* **687**, 1–51 (2017).
23. Chen, X., Sasaki, T., Brännström, Å. & Dieckmann, U. First carrot, then stick: How the adaptive hybridization of incentives promotes cooperation. *J. R. Soc. Interface* **12**(102), 20140935 (2015).
24. Perc, M. & Szolnoki, A. Coevolutionary games—A mini review. *BioSystems* **99**(2), 109–125 (2010).
25. Page, K. M., Nowak, M. A. & Sigmund, K. The spatial ultimatum game. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* **267**(1458), 2177–2182 (2000).
26. Szolnoki, A., Perc, M. & Szabó, G. Defense mechanisms of empathetic players in the spatial ultimatum game. *Phys. Rev. Lett.* **109**(7), 078701 (2012).
27. Te, W., Feng, F., Zhang, Y. & Wang, L. Adaptive role switching promotes fairness in networked ultimatum game. *Sci. Rep.* **3**, 1550 (2013).
28. Santos, F. P., Pacheco, J. M., Paiva, A. & Santos, F. C. Structural power and the evolution of collective fairness in social networks. *PLoS ONE* **12**(4), e0175687 (2017).
29. Cimpeanu, T., Perret, C. & Han, T. A. Cost-efficient interventions for promoting fairness in the ultimatum game. *Knowl. Based Syst.* **233**, 107545 (2021).
30. Kumar, A., Capraro, V. & Perc, M. The evolution of trust and trustworthiness. *J. R. Soc. Interface* **17**(169), 20200491 (2020).
31. Schilling, M. A. & Phelps, C. C. Interfirm collaboration networks: The impact of large-scale network structure on firm innovation. *Manag. Sci.* **53**(7), 1113–1126 (2007).
32. Newman, M. E. J. Coauthorship networks and patterns of scientific collaboration. *Proc. Natl. Acad. Sci.* **101**(suppl 1), 5200–5205 (2004).
33. Barabasi, A.-L. *Linked-how Everything is Connected to Everything Else and what it Means F* (Perseus Books Group, 2014).
34. Ahuja, G. Collaboration networks, structural holes, and innovation: A longitudinal study. *Adm. Sci. Q.* **45**(3), 425–455 (2000).
35. Shipilov, A. & Gawer, A. Integrating research on interorganizational networks and ecosystems. *Acad. Manag. Ann.* **14**(1), 92–121 (2020).
36. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**(5439), 509–512 (1999).
37. Santos, F. C., Pacheco, J. M. & Lenaerts, T. Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *Proc. Natl. Acad. Sci. USA* **103**, 3490–3494 (2006).
38. Szabó, G. & Fáth, G. Evolutionary games on graphs. *Phys. Rep.* **446**(4–6), 97–216 (2007).
39. Pastor-Satorras, R. & Vespignani, A. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86**(14), 3200 (2001).
40. Pacheco, J. M. & Santos, F. C. The messianic effect of pathological altruism. *Pathological Altruism*. pp. 300–310. (New York, NY, US: Oxford University Press, 2012).
41. Santos, F. P., Pacheco, J. M., Paiva, A. & Santos, F. C. Evolution of collective fairness in hybrid populations of humans and agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 6146–6153, (2019).
42. Paiva, A., Santos, F., & Santos, F. Engineering pro-sociality with autonomous agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, (2018).
43. Cardillo, A. & Masuda, N. Critical mass effect in evolutionary games triggered by zealots. *Phys. Rev. Res.* **2**(2), 023305 (2020).
44. Scherer, M. U. Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *SSRN Electron. J.* **29**, 353 (2015).
45. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**(7788), 89–94 (2020).
46. Rolnick, D. et al. Tackling climate change with machine learning, (2019). Preprint available at arXiv:1906.05433.
47. Ranjbar-Sahraei, B., Bou Ammar, H., Bloembergen, D., Tuyls, K., & Weiss, G. Evolution of cooperation in arbitrary complex networks. In *AAMAS'2014*, 677–684, (2014).
48. Denicolò, V. & Franzoni, L. A. On the winner-take-all principle in innovation races. *J. Eur. Econ. Assoc.* **8**(5), 1133–1158 (2010).
49. Campart, S. & Pfister, E. Technological races and stock market value: Evidence from the pharmaceutical industry. *Econ. Innov. New Technol.* **23**(3), 215–238 (2014).
50. Lemley, M. The myth of the sole inventor. *Mich. Law Rev.* **110**, 709–760 (2012).
51. Abbott, F. M., Dukes, M. N. G. & Dukes, G. *Global Pharmaceutical Policy: Ensuring Medicines for Tomorrow's World* (Edward Elgar Publishing, 2009).
52. Burrell, R. & Kelly, C. The covid-19 pandemic and the challenge for innovation policy. Available at SSRN 3576481, (2020).
53. Andrews, T. M., Delton, A. W. & Kline, R. High-risk high-reward investments to mitigate climate change. *Nat. Clim. Change* **8**(10), 890–894 (2018).
54. Baum, S. D. On the promotion of safe and socially beneficial artificial intelligence. *AI Soc.* **32**(4), 543–551 (2017).
55. Geist, E. M. It's already too late to stop the AI arms race: We must manage it instead. *Bull. Atom. Sci.* **72**(5), 318–321 (2016).
56. Shulman, C. & Armstrong, S. Arms control and intelligence explosions. In *7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, July*, 2–4, (2009).
57. Vinuesa, R. *et al.* The role of artificial intelligence in achieving the sustainable development goals. *Nat. Commun.* **11**(233), 1–10 (2020).
58. Askell, A., Brundage, M., & Hadfield, G. The Role of Cooperation in Responsible AI Development. arXiv preprint arXiv:1907.04534, (2019).
59. Han, T. A., Pereira, L. M., & Lenaerts, T. Modelling and influencing the AI binding war: a research agenda. In *Proceedings of the AAAI/ACM conference AI, Ethics and Society*, 5–11, (2019).
60. Han, T. A., Pereira, L. M., Lenaerts, T. & Santos, F. C. Mediating artificial intelligence developments through negative and positive incentives. *PLoS ONE* **16**(1), e0244592 (2021).
61. Han, T. A., Lenaerts, T., Santos, F. C., & Pereira, L. M. Voluntary safety commitments provide an escape from over-regulation in AI development. Technology in Society (In Press), (2022).
62. Perc, M., Gómez-Gardenes, J., Szolnoki, A., Floría, L. M. & Moreno, Y. Evolutionary dynamics of group interactions on structured populations: A review. *J. R. Soc. Interface* **10**(80), 20120997 (2013).
63. Han, T. A. Lynch, S., Tran-Thanh, L. & Santos, F. C. Fostering cooperation in structured populations through local and global interference strategies. In *IJCAI-ECAI'2018*, 289–295, (2018).
64. Raghunandan, M. A. & Subramanian, C. A. Sustaining cooperation on networks: an analytical study based on evolutionary game theory. In *AAMAS*, Vol. 12, 913–920 (Citeseer, 2012).

65. Perc, M. The social physics collective. *Sci. Rep.* **9**, 1–3 (2019).
66. Wilder, B., Immorlica, N., Rice, E., & Tambe, M.. Maximizing influence in an unknown social network. In *AAAI Conference on Artificial Intelligence (AAAI-18)*, (2018).
67. Bloembergen, D., Sahraei, B. R., Bou-Ammar, H., Tuyls, K. & Weiss, G. Influencing social networks: an optimal control study. In *ECAI*, Vol. 14, 105–110, (2014).
68. Cimpeanu, T., Han, T. A., & Santos, F. C. Exogenous rewards for promoting cooperation in scale-free networks. In *Artificial Life Conference Proceedings*, 316–323 (MIT Press, 2019).
69. Santos, F. C. & Pacheco, J. M. Risk of collective failure provides an escape from the tragedy of the commons. *PNAS* **108**(26), 10421–10425 (2011).
70. Pacheco, J. M., Traulsen, A. & Nowak, M. A. Coevolution of strategy and structure in complex networks with dynamical linking. *Phys. Rev. Lett.* **97**, 258103 (2006).
71. Tkadlec, J., Pavlogiannis, A., Chatterjee, K. & Nowak, M. A. Fast and strong amplifiers of natural selection. *Nat. Commun.* **12**(1), 1–6 (2021).
72. Santos, F. C. & Pacheco, J. M. A new route to the evolution of cooperation. *J. Evol. Biol.* **19**(3), 726–733 (2006).
73. Traulsen, A., Nowak, M. A. & Pacheco, J. M. Stochastic dynamics of invasion and fixation. *Phys. Rev. E* **74**, 11909 (2006).
74. Santos, F. C., Pinheiro, F. L., Lenaerts, T. & Pacheco, J. M. The role of diversity in the evolution of cooperation. *J. Theor. Biol.* **299**, 88–96 (2012).
75. Zisis, I., Di Guida, S., Han, T. A., Kirchsteiger, G. & Lenaerts, T. Generosity motivated by acceptance—Evolutionary analysis of an anticipation games. *Sci. Rep.* **5**(18076), 1–11 (2015).
76. Rand, D. G., Tarnita, C. E., Ohtsuki, H. & Nowak, M. A. Evolution of fairness in the one-shot anonymous ultimatum game. *Proc. Natl. Acad. Sci. USA* **110**, 2581–2586 (2013).
77. Grujić, J. & Lenaerts, T. Do people imitate when making decisions? Evidence from a spatial prisoner's dilemma experiment. *R. Soc. Open Sci.* **7**(7), 200618 (2020).
78. Pinheiro, F. L., Santos, F. C. & Pacheco, J. M. How selection pressure changes the nature of social dilemmas in structured populations. *New J. Phys.* **14**(7), 073035 (2012).
79. Ohtsuki, H., Nowak, M. A. & Pacheco, J. M. Breaking the symmetry between interaction and replacement in evolutionary dynamics on graphs. *Phys. Rev. Lett.* **98**(10), 108106 (2007).
80. Dorogovtsev, S. *Complex Networks* (Oxford University Press, 2010).
81. Newman, M. E. J. The structure and function of complex networks. *SIAM Rev.* **45**(2), 167–256 (2003).
82. Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).
83. Dorogovtsev, S. N., Mendes, J. F. F. & Samukhin, A. N. Size-dependent degree distribution of a scale-free growing network. *Phys. Rev. E* **63**(6), 062101 (2001).

Acknowledgements

T.A.H., L.M.P., T.L. and T.C., are supported by Future of Life Institute grant RFP2-154. T.C. and T.A.H. also acknowledge support from Berkeley Existential Risk Initiative (BERI) collaboration fund. LM.P. support 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 UIDB/50021/2020, PTDC/MAT-APL/6804/2020 and PTDC/CCI-INF/7366/2020). T.L benefits from the support by the Flemish Government through the AI Research Program and F.C.S and T.L.by TAILOR, a project funded by EU Horizon 2020 research and innovation program under GA No 952215. T.L. is furthermore supported by the F.N.R.S. project with grant number 31257234, the F.W.O. project with grant nr. G.0391.13N, the FuturICT 2.0 (http://www.futurict2.eu) project funded by the FLAG-ERA JCT 2016 and the Service Public de Wallonie Recherche under grant n° 2010235-ARIAC by DigitalWallonia4.ai. T.A.H. is also supported by a Leverhulme Research Fellowship (RF-2020-603/9).

Author contributions

T.C., F.C.S., L.M.P., T.L. and T.A.H. designed the research; T.C. implemented the software and prepared the figures; T.C., F.C.S., L.M.P., T.L. and T.A.H. wrote the manuscript; T.C., F.C.S., L.M.P., T.L. and T.A.H. reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-05729-3.

Correspondence and requests for materials should be addressed to T.A.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.