AI Development Race Can Be Mediated on Heterogeneous Networks

PaperID: 434

ABSTRACT

The field of Artificial Intelligence (AI) has been introducing a certain level of anxiety in research, business and also policy. Tensions are further heightened by an AI race narrative which makes many stakeholders fear that they might be missing out. Whether real or not, a belief in this narrative may be detrimental as some stakeholders will feel obliged to cut corners on safety precautions or ignore societal consequences. Starting from a game-theoretical model describing an idealised technology race in a well-mixed world, 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 to design and implement meticulous interventions on a minority of participants capable of influencing an entire population towards an ethical and sustainable use of AI.

KEYWORDS

AI Safety, Complex Networks, Evolutionary Game Theory, Agent-based Simulation

ACM Reference Format:

PaperID: 434. 2021. AI Development Race Can Be Mediated on Heterogeneous Networks . In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), London, UK, May 3–7, 2021*, IFAAMAS, 11 pages.

1 INTRODUCTION

With the current business and governmental anxiety about AI and the promises made about the impact of AI technology, there is a risk for stake-holders to cut corners, preferring rapid deployment of their AI technology over an adherence to safety and ethical procedures, or a willingness to examine their societal impact [3, 12]. Many stakeholders/researchers have urged for due diligence as i) these AI systems can also be employed for more nefarious activities, e.g. espionage and cyberterrorism [53] and ii) whilst attempting to be the first/best, some ethical consequences as well as safety procedures may be underestimated or even ignored [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 [29, 30], the blogs of AI experts requesting careful communications [10] and the proclamations on ethical use of AI in the world [16, 25, 40, 50].

While potential AI disaster scenarios are many [3, 33, 49], the uncertainties in accurately predicting these risks and outcomes are high [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 [15]. 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.

With this aim in mind, a baseline model [24] using methods from Evolutionary Game Theory (EGT), has been proposed. One of the core, simplified assumptions made in that work is that interactions among the race participants are uniform. However, real-world interactions are not random. Some interactions are more frequent than others, and some individuals have many more contacts than others. It has been shown that network reciprocity can promote the evolution of positive behaviours in various settings including cooperation (in the one-shot prisoner's dilemma and public goods game) [35, 39, 44], fairness [32, 52, 57] and trust [26]. The idea is that, in spatial populations, disadvantageous good/positive behaviours can form clusters to protect themselves from the invasion of wrongdoers. In this paper, we take inspiration from this line of research and we ask whether network reciprocity can mediate the tension of the AI development race.

This issue is particularly important in the context of technology regulation and governance. Indeed, technology innovation and collaboration networks (e.g. among firms, stakeholders and AI researchers) are highly heterogenous [28, 46] . Developers or development teams interact more frequently within their groups, forming alliances and networks of followers and collaborators [1, 6]. Many companies compete in several markets while others compete in only a few, and their positions in interorganisational networks strongly influence their behaviour (such as resource sharing) and innovation outcome [1, 47]. Hence, it is important to understand how such diversity in the network of contacts underlying technology development competition and collaboration, influences race dynamics and conditions under which regulatory actions are needed.

Therefore, in this paper, we examine how network structures influence safety decision making within an AI development race, by generalising the AI race model proposed in [24] for a non-spatial setting (i.e. well-mixed world). Namely, in order to achieve that goal, a number of development steps or technological advancements (or rounds) are required, where in each round the development teams (or players) have two strategic options: to follow the safety precaution (SAFE) or to ignore this safety precaution (UNSAFE), as was defined in [24]. Since it takes more time and effort to comply with the precautionary requirements, playing SAFE is not only costlier,

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, London, UK.

but also implies a slower development speed, compared to playing UNSAFE. Let us assume that to play SAFE players need to pay a cost $c > 0$, while playing UNSAFE does not cost them anything. Also, the development speed when playing UNSAFE is $s > 1$ while the speed when playing SAFE is normalised to 1. The interaction is iterated until one or more teams achieve the designated objective, after having completed W development steps. As the result, they obtain a large benefit or prize B, which is shared among those who reach the target at the same time. However, a development setback or disaster might 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 [3, 33], the uncertainties in accurately predicting these outcomes are high. When such a disaster occurs, the winning team risk-taking participant loses all its benefits. We denote by p_r the risk probability of such a disaster occurring when no safety precaution is followed at all (see Section 3 for further details).

In a spatial setting, players are competing with the co-players in their neighbourhood. We compare different forms of network structures, from homogenous ones such as complete graph and square lattice to different types of scale-free networks (for details, see Methods section), representing different levels of diversity on the number of races a player can compete in. Our results show that when race participants are distributed in a heterogeneous network, the race tension that has been demonstrated in the well-mixed case, is significantly reduced, softening the need for regulatory actions. This is however not the case when the network is not accompanied by a certain degree of relational heterogeneity, even in different types of spatial, lattice networks.

The next section will review relevant literature, including works on AI race modelling and network reciprocity. We then describe in detail our models and methods, in Section 3. The results are described in Section 4, and final concluding remarks in Section 5. Additionally, we attach to this submission a Supporting Information (SI) document with some additional agent-based simulation results to support the robustness of the findings in this work.

2 RELATED WORK

Although there have been a number of proposals and debates on how to prevent, regulate, or resolve an AI race [5, 8, 12, 20, 48, 53, 55], only a few formal modelling studies have been proposed [3, 23, 24]. These works focus on homogenous populations, where there are no inherent structures indicating the network of contacts among developing teams. The current paper advances this line of research, by studying the effect of network structures that underline the network of contacts among race participants, on the dynamics and global outcome of the development race.

The question 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 [1, 22, 34, 35, 37, 39, 42, 44, 51]. Network reciprocity can promote the evolution of positive behaviours in various settings including cooperation [35, 39, 42, 44], fairness [32, 52, 57] and trust [26]. Their applications are diverse, from healthcare [28], network interference and influence maximization [9, 14, 56], climate change [41], etc. Inspired by this literature, this paper studies the role of network reciprocity in the context of technology development race, extending previous modelling works (described above) where spatial structure was not taken into account.

3 MODELS AND METHODS

We first recall the AI race game model as developed in [24], in the context of well-mixed populations. We then describe different spatial structures to be studied and the details of how simulations on networks are carried out.

3.1 AI race model definition

The AI development race is modelled 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, wherein their disregard for safety precautions is exposed, leading to their products not being adopted, 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)

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

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

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 analysis (e.g. in social dilemma games [42, 44, 51]), we focus in this paper on two unconditional strategies [24]:

- 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

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

3.2 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 score for each agent is the sum of the payoffs in these encounters. In the SI, we also discuss the limit where scores are normalised by the number of interactions (i.e., the *degree* of a node). At the end of each generation, a randomly selected agent A with score f_A chooses to copy the strategy of a randomly selected neighbour, agent B, with score f_B with a probability given by the Fermi rule [43, 54]: $(1+e^{\beta(f_A-f_B)})^{-1}$, where 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) [54]. Varying β allows capturing a wide range of update rules and levels of stochasticity, including those used by humans, as measured in lab experiments [21, 38, 58]. In line with previous works and lab experiments, we set $\beta = 1$ in our simulations, ensuring a high intensity of selection [36]. As each network type converges at different rates and naturally presents with various degrees of heterogeneity, we choose different population sizes and maximum number of runs in the various experiments to account for this while optimising run-time. These will be mentioned as appropriate in the following sections.

3.3 Network Topologies

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 network): 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 [51]). 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 [7, 17, 27], generated through two growing network models — the widely-adopted Barabási-Albert (BA) model [2, 7] and a specialised version of the former that produces a large number of triangular motifs (i.e. high clustering coefficient), the Dorogovtsev-Mendes-Samukhin (DMS) model [17, 18]. 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 number of connections [7]. In the case of the DMS model, new connections are chosen based on an edge lottery (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 $z = 2m$. Moreover, for both types we study un-normalised versions (large wealth inequality), but also normalised versions (by the number of connections for each node).

Overall, WM populations offer a convenient baseline scenario, where interaction structure is absent. With the SL we introduce a network structure, yet 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.

3.4 Computer Simulations

At each time step or generation of a simulation, we calculate the averaged payoffs in the AI race as described previously. Links in the network describe a relationship of proximity both in the interactional sense (who the agents can interact with), but also observationally (who the agents can imitate). Ergo, the network of interactions coincides with the network of imitations. We chose an asynchronous update rule, where at most one imitation occurs in each generation (similar results are obtained with synchronous update rules [42, 43]). For well-mixed populations and lattice networks, we chose populations of $n = 100$ agents and $Z = 32 \times 32$ agents, respectively. Contrastingly, for scale-free networks, we chose $n = 1000$, while also pre-seeding 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 simulate the evolutionary process for 10^7 generations in the case of scale-free networks and 10^6 generations otherwise, only measuring the results for the final 10^3 steps for a clear and fair comparison (e.g. due to the fluctuations characteristic of these stationary states). Furthermore, the results for each combination of network and parameter values are obtained by averaging over 25 independent realisations.

4 RESULTS

The simulations described in the previous section (See Section 3) allow us to identify the prevalence of each strategy after reaching a stationary state. From this, we can infer the most likely trends and self-organized behavioural patterns associated with the agents taking part in the AI race game for different network topologies.

As described in [24], it is important to make the distinction between two development regimes: an early/short-term regime and a late/long-term one. The difference in time-scale between the two regimes is key in identifying which regulatory actions are needed and when. The early regime is underpinned by how able the race participants are to 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 technology supremacy will not 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. $B/W << b$.

As a starting point and to enable a clear comparison between our results in spatial settings and those of the previous work in a well-mixed world [24], we will first provide simulation results for the well-mixed case. We will also make use of the results and analytical conditions described therein regarding the risk-dominant boundaries of the AI race game for both early and late development regimes. They are useful to determine the regions in which regulatory actions are needed or not, and moreover, if needed, which behaviour should be promoted. Note that in [24], the results were obtained analytically, while in this work, we adopt extensive agent-based simulations. Hence, our work also contributes in providing a simulation validation of the results obtained in that work.

4.1 Well-mixed populations

In Figure 1 (Left Column) we show three types of density plots, each with specific risk-dominant regions marked explicitly, for a full discussion on the analytical conditions, we refer back to [24]. They are all in close accordance with the analytical and numerical results therein. Indeed, first of all, we consider a comparison between early and late regimes, by varying the number of development steps W and the probability of disaster occurring p_r. In this case, the solid black line indicates the threshold for p_r above which SAFE is the preferred collective action and below which UNSAFE is the desired one. Thus, we depict the boundary for which the benefits of all players acting safely outweigh the profits of all ignoring safety (i.e. $\Pi_{AS,AS} > \Pi_{AU,AU}$), as a function of p_r (see Figure 1, first row). Secondly, we present in more detail the results concerning the early regime, in which the AI race ends sometime in the foreseeable future. The two dotted lines mark region (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 Figure 1, second 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. Finally, we discuss the results concerning the late AI race (large W), see Figure 1, bottom row. In this case the solid black line marks the boundary above which safety is the preferred collective outcome, whereas the blue line indicates where AS becomes risk-dominant against AU (since the formulas are rather complex/long, see [24] for details). Again, in this regime three regions can be distinguished, with (I) and (III) having similar meanings to those in the early regime. However, different from the early regime, in region (II) of the late regime, regulatory actions are need to improve (unsafe) innovation instead of safety compliance, due to low risk.

4.2 Lattices

Here, we analyse the role of spatial structure in the evolution of strategies in the AI race game. We simulate the same simplified game on a square lattice, where each agent can interact with its four edge neighbours, in Figure 1 (Right Column). We show that the trends remain the same when compared with well-mixed populations, with very slight differences in numerical values between the

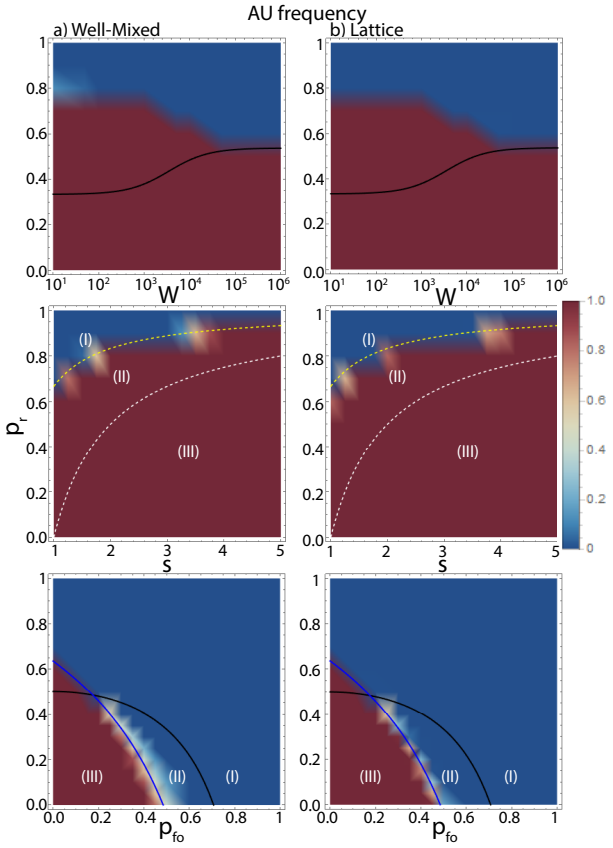

Figure 1: Average fraction of AU (unsafe strategy) for well-mixed populations (Left Column) and lattice networks (Right Column). The top row reports the spectrum between an early and late AI race (for varying W, with $p_{fo} = 0.1$, $s = 1.5$), the middle row addresses the early regime for varying s and p_r ($p_{fo} = 0.5$, $W = 100$), and the bottom row addresses the late regime for varying p_{fo} and p_r ($s = 1.5$, $W = 10^6$). Other parameters: $c = 1$, $b = 4$, $B = 10^4$, $\beta = 1$.

two. Specifically, towards the top of area (**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. We report one such realisation in Figure 2, where the spatial structure leads to a shift in evolutionary outcomes. In practice, this shifts the boundary very slightly towards an optimal conclusion.

Thus, except for minute atypical situations, we may argue that homogenous spatial variation is not enough to mediate and influence a 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, Figure S1).

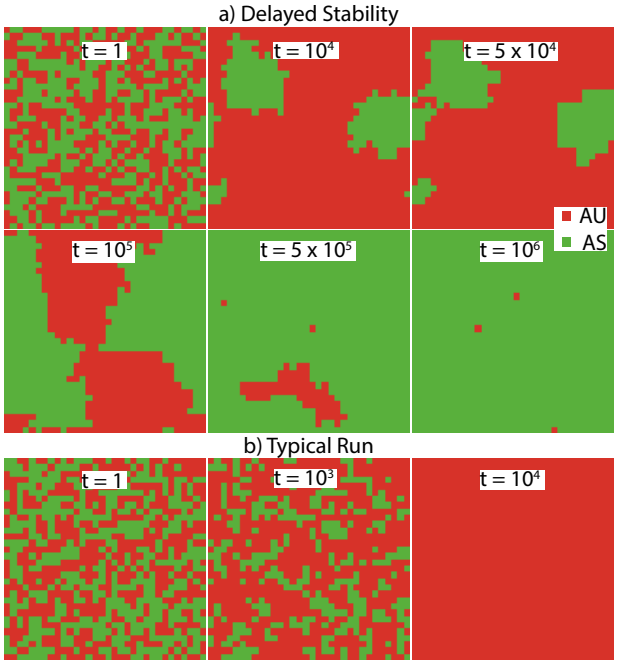

Figure 2: Snapshots of the evolution of strategies in a 32x32 square lattice. For most realisations, in all regions, the system quickly converges to a homogeneous state (on such typical run can be seen in panel b). However, solely in region (II) of the early regime, we observed some atypical realisations in which the system very slowly fluctuates from an almost absorbing state to the opposite strategy. We show one such realisation for illustration purposes, see panel a. Parameters: $p_r = 0.74$, $p_{fo} = 0.5$, $c = 1$, $s = 1.5$, $b = 4$, $W = 100$, $B = 10^4, \beta = 1$.

4.3 Scale-free networks

4.3.1 Network heterogeneity mediates the AI race dilemma. As a means of investigating beyond simple spatial structures and their roles in the evolution of appropriate developmental practice in the AI race, we make use of the previously defined BA and DMS network models (See again Section 3). Contrary to the findings on homogeneous networks, scale-free interaction structures produce marked improvements in almost all relevant dilemmas of the AI race game.

The first consequential divergence is exposed from the very first figure, where we discuss a comparison of the two AI safety regimes (see Figure 3, first row). Indeed, whereas previously, there was a clear delimitation (based on development steps W) wherein unsafe players strongly outperformed safe ones, scale-free networks instead produce a much more balanced outcome, in which AU players only moderately surpass safe players in the early regime (smaller W). These results indicate the benefits of individual influence-diversity (in the form of network heterogeneity), in the AI development race. In some ways, diversity acts as an equalizer, lessening the advantage that unsafe players would usually benefit from, either when developers require speed (early AIS) or when there is a high probability of a disaster occurring due to neglected safety provisions (p_r).

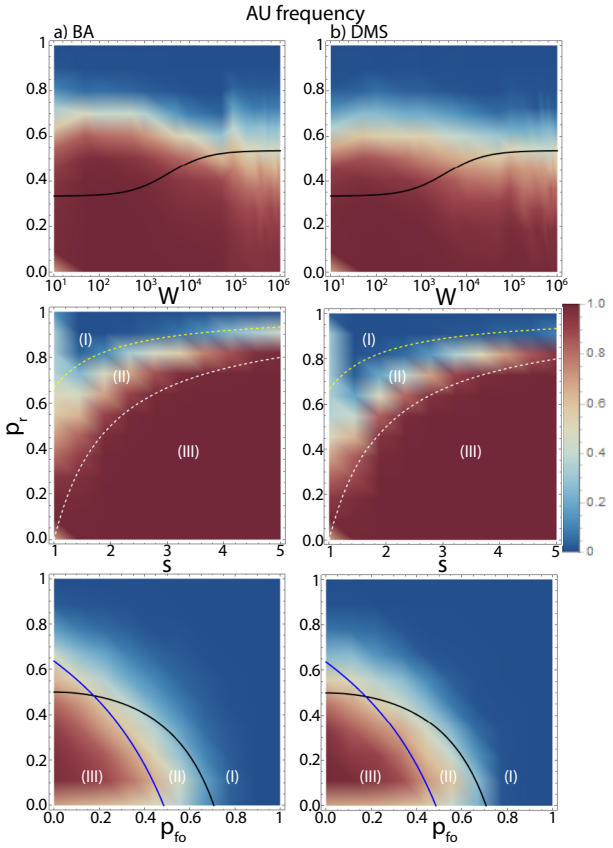

Figure 3: Comparison between the two different scale-free networks, BA and DMS. The top row reports the spectrum between an early and late AI race ($p_{fo} = 0.1$, $s = 1.5$**), the middle row addresses the early regime in more detail (**$p_{fo} = 0.5$, $W = 100$**) and the bottom row considers a late AI race (**$W = 10^6$, $s = 1.5$**). Parameters:** $c = 1$, $b = 4$, $B = 10^4$, $\beta = 1$.

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 belongs to, after inferring the preferences developers would have towards safety compliance [24]. Given that innovation in the field of AI, or more broadly, technological advancements as a whole, should be profitable (and robust) to developers, shareholders and society altogether, we must therefore discuss the analytical locus where these initiatives 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 existing observations on 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 benign audacity. Secondly, the region for early AIS regimes in which regulation must be enforced is lessened, but not completely eliminated. On that account, 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

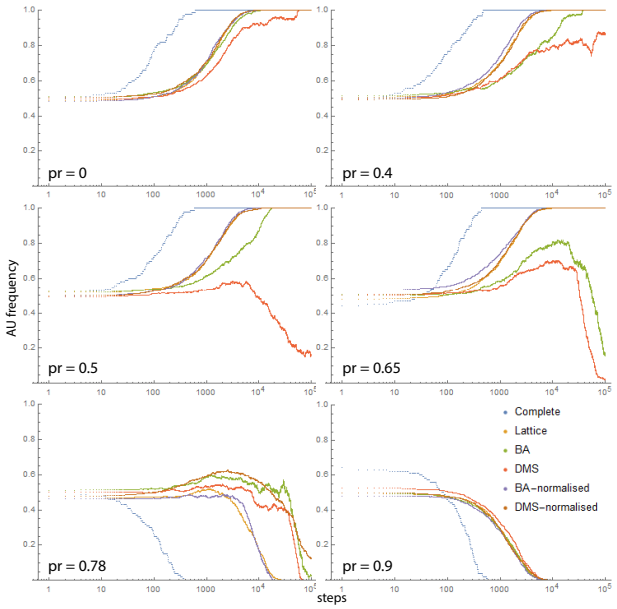

Figure 4: Typical runs for different risk probability values for the early AI race, on each type of network. Parameters: $c = 1,\ b = 4,\ B = 10^4,\ \beta = 1$.

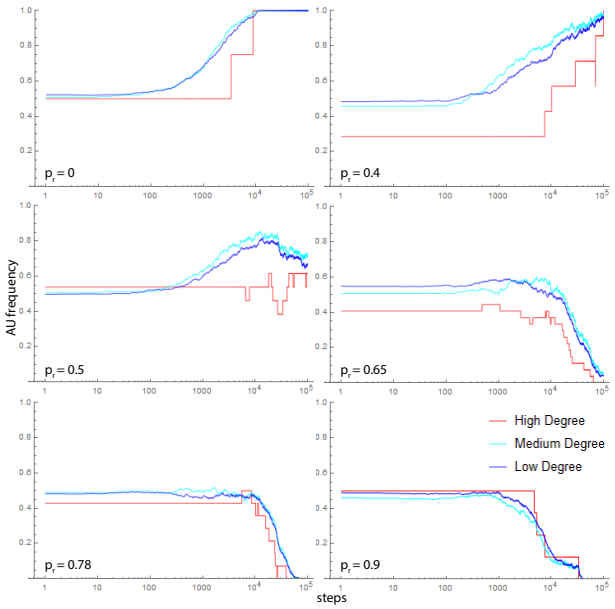

Figure 5: Distribution of unsafe behaviour (AU) in an early AI race, grouped by degree class (connectivity) of the nodes on DMS networks. Parameters: $c = 1,\ b = 4,\ B = 10^4,\ W = 100,\ \beta = 1$.

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 3 (row 2) presents a fine-grained glimpse into the early regime. In region (**II**), the safety dilemma zone, social welfare is once more improved conspicuously by heterogeneity. In this area, concerted safe behaviour is favoured, 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 [24], we contend that it is vital for regulators to intervene when necessary, for encouraging prosocial, safe conduct, and in doing so avert conceivably dangerous outcomes. In many ways, 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, thus seemingly increasing fairness and reducing wealth inequality. We confirm that normalising the network leads to similar dynamics observed on homogeneous populations (see Supplementary Information, Figure S2), with only very slight numerical differences.

To further illustrate the key differences between each type of network, we plot typical simulation runs for different p_r risk probability values in the area (**II**) of the early AI race (see Figure 4). It is immediately apparent that the two un-normalised scale-free networks provide significant improvements in safety compliance in the dilemma zone. This is further compounded by the effect of clustering on the threshold at which safe development becomes evolutionarily stable. Specifically, we note that when the risk of a disaster occurring due to inadequate safety compliance is intermediate (see, e.g. $p_r = 0.5$ and 0.65), we see a definitive improvement in highly clustered networks (i.e. DMS) as opposed to the basic BA model.

In all cases, an increase of complexity in the network structure (i.e. more heterogeneity) leads to a slower convergence to an absorbed state. This is true even despite differing population sizes (for example lattice versus well-mixed populations). We note that this does not affect the stationary states, merely the time it takes to reach them. We may, disregard these differences in convergence time for the purposes of this analysis. However, this time to convergence may deserve a future work studying the cost associated with regulating a progressing population [13, 14, 22].

4.3.2 The role of high-degree nodes: degree analysis. Highly connected individuals (hubs) play a key role in many real-world networks of contacts [42, 44]. In order to study the role hubs play in the AI race, in the context of scale-free networks, we classify nodes into three separate connectivity classes [44]. We obtain three classes of individuals, based on their number of contacts (links) k_i and 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 $\frac{k_{max}}{3} \leq k_i \leq k_{max}$.

Figure 5 shows the evolution over time of unsafe behaviour (AU) in the dilemma zone of an early AI race for different environments (corresponding to varying probability values of a disaster occurring

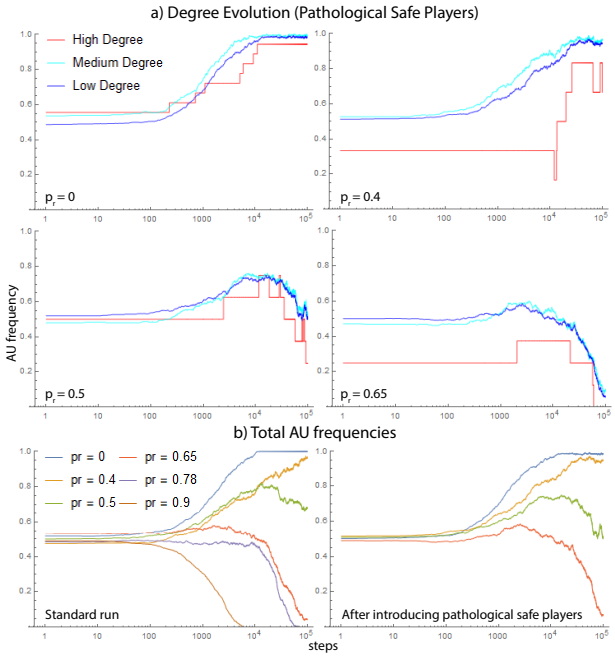

Figure 6: Typical runs exploring the evolutionary degree distribution of unsafe behaviour in an early AI race, following the introduction of safety zealots (pathological safe players) in the population on DMS networks. We randomly allocate 10% of high degree individuals as safety zealots. Note that we measure the frequency for the whole population, including the pathological players. Parameters: $c = 1$, $b = 4$, $B = 10^4$, $W = 100$, $\beta = 1$.

caused by insufficient safety regulation, p_r). High-degree individuals appear to have a higher tendency towards safety compliance (at equilibrium) when compared to their lowly or moderately connected counterparts, except for region (**III**), where highly connected individuals are driving to innovate (optimally so). In spite of this, we see the same trends for regions (**I**) and (**III**). However, in the region (**II**), highly connected individuals become important leaders in the shift from unsafe to safe behaviour in the AI race. Specifically, for large p_r values (see $p_r = 0.65$; $p_r = 0.78$), there is an evident disparity between the high degree individuals and the bulk of the population, and indeed, this is the region in which heterogeneity improves safety compliance the most. For low p_r values, heterogeneity fails to improve the outcome, but it does serve as an equaliser for intermediate risk values ($p_r = 0.5$). Regulatory actions would therefore still be required to constrain developers when heterogeneity cannot improve safety enough in region **II**, in the case of low risk for disaster to occur.

4.3.3 The effect of safe pathological behaviour. Dedicated minorities are often identified as major drivers in the emergence of collective behaviours in social, physical and biological systems, see e.g. [11]. Given the previously emphasised importance that hubs play in the emergence of safety, we then explore whether highly connected, committed individuals are prime targets for safety regulation in the AI race. By introducing individuals with pathological safe tendencies [45] (these are sometimes referred to as zealots, see [11, 26, 31, 45] in hubs, i.e. belonging to the high degree class as described above), we can better understand the power of influential devotees in the safe development of general AI.

We have already established the tendencies of highly influential agents (hubs) towards safe behaviour and it is immediately apparent in Figure 6 that this hinders any potential benefits to be gained by forcefully planting zealots in important nodes. As influential individuals favour safety as a baseline and they comprise only a very small minority in the population, the effect of planting safe zealots is very small indeed. We see a very small following of such hubs for low p_r values (without leaving area (**II**)). This small following might play a key role in the re-emergence of safe behaviour in the presence of noise or high mutation limits, this would make for an interesting topic for future work on this subject. We also explore the potential of unsafe zealots to destabilise a population when safety is the evolutionary outcome, as well as the impact of safe zealots on reducing innovation when it would be harmful to do so (see Supplementary Information, Figure S3). We found no apparent consequence of introducing pathological players in either of these scenarios.

Finally, we briefly investigate potential avenues for a regulatory agent to improve safety compliance in heterogeneous networks (see Figure 7). As an initial step and prospective approach to interference (by external agents such as an international organisation), we artificially increase the speed (increasing their payoffs by $\frac{sB}{W}$) and, alternatively, the wealth of the previously introduced safety zealots in highly connected nodes, by a very large amount that ensures they will always be imitated (10^7). Each approach has its merits in different regions of the early regime, and we see the effectiveness of funding highly connected nodes when the risk for disaster is low. On the other hand, a high risk improves the efficacy of speeding up the development for these dedicated minorities. We note that targeting highly influential players is not sufficient to mitigate the race tensions entirely. Further exploration on this topic would provide more insight into how external interference can be deployed efficiently [14, 22], for example by international organisations and/or local governments, to further mediate the tensions of a race to technological supremacy.

5 CONCLUSIONS

Here, we consider the implications of network dynamics on a technological race for supremacy in the field of AI, with all its implied risks and hazardous consequences [3, 33, 49]. We make use of a previously proposed evolutionary game theoretic model [24] and study how the tension and temptation resulted 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 [39, 44, 52, 57] and, given the high levels of heterogeneity identified in the networks of firms, stakeholders and AI researchers [28, 46], it is very important to understand the effects of reciprocity and how it shapes the dynamics and global outcome of the development race. It is to ensure that appropriate context-dependent regulatory actions are provided.

We begin by validating the analytical results obtained as a baseline in a completely homogeneous population [24], using extensive agent-based simulations. We then adopt similar methodology to

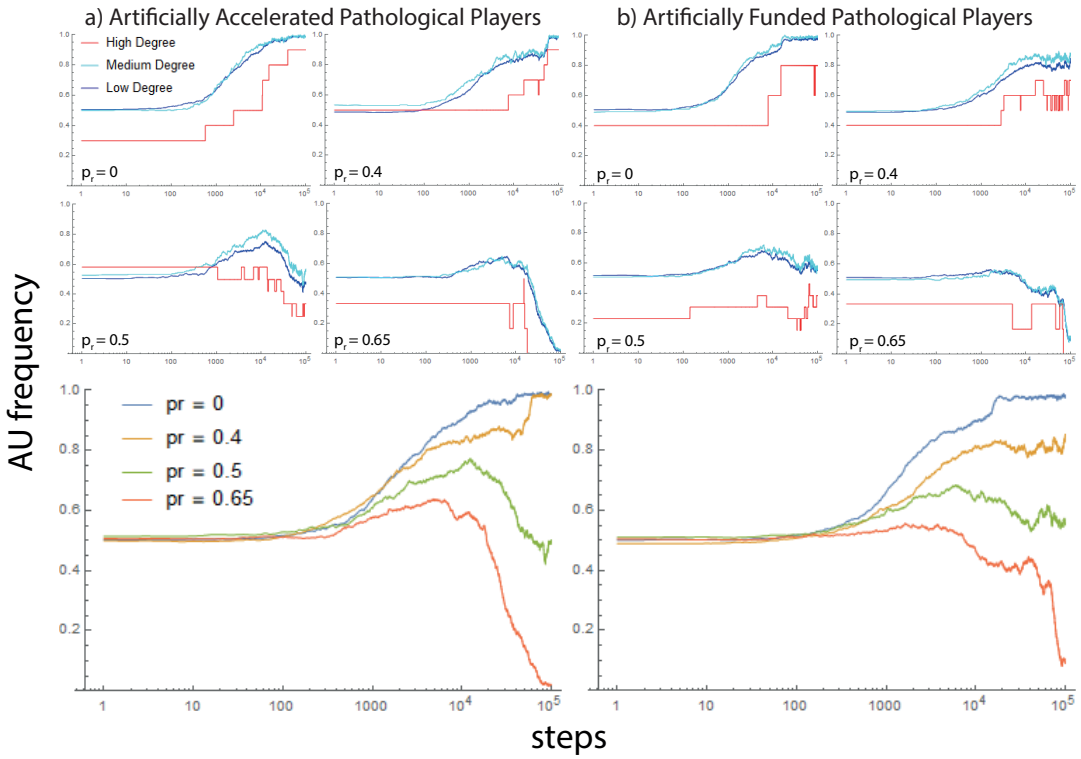

Figure 7: Typical runs exploring the evolutionary degree distribution of unsafe behaviour in an early AI race, following the artificial acceleration (or funding) of safety zealots (pathological safe players) in the population on DMS networks. We randomly allocate 10% of high degree individuals as safety zealots. Note that we measure the frequency for the whole population, including the pathological players. Parameters: $c = 1,\ b = 4,\ B = 10^4,\ W = 100,\ \beta = 1$.

analyse the effects of gradually increasing network heterogeneity, equivalently to diversifying the connectivity and influence of the race participants. By studying square lattice (with four and eight neighbours) 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 is not sufficient in the expectation of tempering the necessity 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.

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 [44]. These highly connected individuals have a tendency towards safety compliance, in comparison to their counterparts. In an attempt to exploit this observed effect, as well as to better understand the impact of such seemingly significant nodes, we introduce several pathological players [11, 26] in key locations of the network (highly connected nodes). We see very little improvement in safety compliance following the addition of such pathological participants and suggest that the presence of these dedicated minorities might play a key role in the re-emergence of safety compliance in the presence of abundant noise and randomness (e.g. when intensity of selection is small [11]). We plan to have a full analysis of these factors in future work.

In short, our results have shown that heterogenous networks can significantly mediate the tensions observed in the well-mixed world, in both early and late development regimes [24], 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-homogenous, our findings provide important insights for the design of technological regulation and governance frameworks (such as the one proposed in the EU White Paper [19]). 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 it is not needed, as would have been otherwise suggested in a homogeneous world models). Moreover, our findings suggest to increase heterogeneity or diversity in this network as a way to escape tensions arisen from a race for technological supremacy.

REFERENCES

[1] Gautam Ahuja. 2000. Collaboration networks, structural holes, and innovation: A longitudinal study. *Administrative science quarterly* 45, 3 (2000), 425–455.

[2] Réka Albert and Albert-László Barabási. 2002. Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74 (Jan 2002), 47–97. Issue 1. https://doi.org/10.1103/RevModPhys.74.47

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

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

[5] Amanda Askell, Miles Brundage, and Gillian Hadfield. 2019. The Role of Cooperation in Responsible AI Development. *arXiv preprint arXiv:1907.04534* (2019).

[6] Albert-Laszlo Barabasi. 2014. *Linked-how Everything is Connected to Everything Else and what it Means F.* Perseus Books Group.

[7] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science* 286, 5439 (1999), 509–512.

[8] Seth D Baum. 2017. On the promotion of safe and socially beneficial artificial intelligence. *AI & Society* 32, 4 (2017), 543–551.

[9] Daan Bloembergen, Bijan Ranjbar Sahraei, Haitham Bou-Ammar, Karl Tuyls, and Gerhard Weiss. 2014. Influencing Social Networks: An Optimal Control Study.. In *ECAI*, Vol. 14. 105–110.

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

[11] Alessio Cardillo and Naoki Masuda. 2020. Critical mass effect in evolutionary games triggered by zealots. *Physical Review Research* 2, 2 (Jun 2020). https://doi.org/10.1103/physrevresearch.2.023305

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

[13] Xiaojie Chen, Tatsuya Sasaki, Åke Brännström, and Ulf Dieckmann. 2015. First carrot, then stick: how the adaptive hybridization of incentives promotes cooperation. *Journal of The Royal Society Interface* 12, 102 (2015), 20140935.

[14] Theodor Cimpeanu, The Anh Han, and Francisco C Santos. 2019. Exogenous Rewards for Promoting Cooperation in Scale-Free Networks. In *Artificial Life Conference Proceedings*. MIT Press, 316–323.

[15] David Collingridge. 1980. *The social control of technology*. New York : St. Martin's Press.

[16] Montreal Declaration. 2018. The Montreal Declaration for the Responsible Development of Artificial Intelligence Launched. https://www.canasean.com/the-montreal-declaration-for-the-responsible-development-of-artificial-intelligence-launched/.

[17] S Dorogovtsev. 2010. *Complex networks*. Oxford: Oxford University Press.

[18] Sergey N Dorogovtsev, Jos FF Mendes, and Alexander N Samukhin. 2001. Size-dependent degree distribution of a scale-free growing network. *Physical Review E* 63, 6 (2001), 062101.

[19] European Commission. 2020. *White paper on Artificial Intelligence – An European approach to excellence and trust*. Technical Report. European Commission. AccessedMay26,2020:https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

[20] Edward Moore Geist. 2016. It's already too late to stop the AI arms race: We must manage it instead. *Bulletin of the Atomic Scientists* 72, 5 (2016), 318–321.

[21] Jelena Grujić and Tom Lenaerts. 2020. Do people imitate when making decisions? Evidence from a spatial Prisoner?s Dilemma experiment. *Royal Society open science* 7, 7 (2020), 200618.

[22] The Anh Han, Simon Lynch, Long Tran-Thanh, and Francisco C. Santos. 2018. Fostering Cooperation in Structured Populations Through Local and Global Interference Strategies. In *IJCAI-ECAI'2018*. 289–295.

[23] The Anh Han, Luis Moniz Pereira, Tom Lenaerts, and Francisco C. Santos. 2020. Mediating Artificial Intelligence Developments through Negative and Positive Incentives. (2020). arXiv: 2010.00403.

[24] The Anh Han, Luis Moniz Pereira, Francisco C. Santos, and Tom Lenaerts. 2020 (In Press). To Regulate or Not: A Social Dynamics Analysis of an Idealised AI Race. *Journal of Artificial Intelligence Research, pre-print available at arXiv:1907.12393 (2020)* (2020 (In Press)).

[25] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* (2019), 1–11.

[26] Aanjaneya Kumar, Valerio Capraro, and Matjaž Perc. 2020. The evolution of trust and trustworthiness. *Journal of The Royal Society Interface* 17, 169 (Aug 2020), 20200491. https://doi.org/10.1098/rsif.2020.0491

[27] Mark EJ Newman. 2003. The structure and function of complex networks. *SIAM review* 45, 2 (2003), 167–256.

[28] Mark EJ Newman. 2004. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences* 101, suppl 1 (2004), 5200–5205.

[29] Future of Life Institute. 2015. *Autonomous Weapons: An Open Letter from AI & Robotics Researchers*. Technical Report. Future of Life Institute, Cambridge, MA.

[30] Future of Life Institute. 2019. Lethal Autonomous Weapons Pledge. https://futureoflife.org/lethal-autonomous-weapons-pledge/.

[31] Jorge M Pacheco and Francisco C Santos. 2011. The messianic effect of pathological altruism. *Pathological Altruism* (2011), 300.

[32] Karen M Page, Martin A Nowak, and Karl Sigmund. 2000. The spatial ultimatum game. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 267, 1458 (2000), 2177–2182.

[33] Dennis Pamlin and Stuart Armstrong. 2015. Global challenges: 12 risks that threaten human civilization. *Global Challenges Foundation, Stockholm* (2015).

[34] Matjaž Perc, Jesús Gómez-Gardenes, Attila Szolnoki, Luis M Floría, and Yamir Moreno. 2013. Evolutionary dynamics of group interactions on structured populations: a review. *Journal of the royal society interface* 10, 80 (2013), 20120997.

[35] Matjaž Perc, Jillian J Jordan, David G Rand, Zhen Wang, Stefano Boccaletti, and Attila Szolnoki. 2017. Statistical physics of human cooperation. *Phys Rep* 687 (2017), 1–51.

[36] Flavio L Pinheiro, Francisco C Santos, and Jorge M Pacheco. 2012. How selection pressure changes the nature of social dilemmas in structured populations. *New Journal of Physics* 14, 7 (2012), 073035.

[37] MA Raghunandan and CA Subramanian. 2012. Sustaining cooperation on networks: an analytical study based on evolutionary game theory.. In *AAMAS*, Vol. 12. Citeseer, 913–920.

[38] David G. Rand, Corina E. Tarnita, Hisashi Ohtsuki, and Martin A. Nowak. 2013. Evolution of fairness in the one-shot anonymous Ultimatum Game. *Proc. Natl. Acad. Sci. USA* 110 (2013), 2581–2586.

[39] Bijan Ranjbar-Sahraei, Haitham Bou Ammar, Daan Bloembergen, Karl Tuyls, and Gerhard Weiss. 2014. Evolution of cooperation in arbitrary complex networks. In *AAMAS'2014*. 677–684.

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

[41] Francisco C. Santos and Jorge M. Pacheco. 2011. Risk of collective failure provides an escape from the tragedy of the commons. *PNAS* 108, 26 (2011), 10421–10425.

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

[43] Francisco C Santos, Flavio L Pinheiro, Tom Lenaerts, and Jorge M Pacheco. 2012. The role of diversity in the evolution of cooperation. *Journal of theoretical biology* 299 (2012), 88–96.

[44] F. C. Santos, M. D. Santos, and J. M. Pacheco. 2008. Social diversity promotes the emergence of cooperation in public goods games. *Nature* 454 (2008), 214–216.

[45] Fernando P Santos, Jorge M Pacheco, Ana Paiva, and Francisco C Santos. 2019. Evolution of collective fairness in hybrid populations of humans and agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6146–6153.

[46] Melissa A Schilling and Corey C Phelps. 2007. Interfirm collaboration networks: The impact of large-scale network structure on firm innovation. *Management science* 53, 7 (2007), 1113–1126.

[47] Andrew Shipilov and Annabelle Gawer. 2020. Integrating research on interorganizational networks and ecosystems. *Academy of Management Annals* 14, 1 (2020), 92–121.

[48] Carl Shulman and Stuart Armstrong. 2009. Arms control and intelligence explosions. In *7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, July*. 2–4.

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

[50] Luc Steels and Ramon Lopez de Mantaras. 2018. The Barcelona declaration for the proper development and usage of artificial intelligence in Europe. *AI Communications* Preprint (2018), 1–10.

[51] G. Szabó and G. Fáth. 2007. Evolutionary games on graphs. *Phys Rep* 97-216, 4-6 (2007).

[52] Attila Szolnoki, Matjaž Perc, and György Szabó. 2012. Defense mechanisms of empathetic players in the spatial ultimatum game. *Physical review letters* 109, 7 (2012), 078701.

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

[54] A. Traulsen, M. A. Nowak, and J. M. Pacheco. 2006. Stochastic Dynamics of Invasion and Fixation. *Phys. Rev. E* 74 (2006), 11909.

[55] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Langhans, Max Tegmark, and Francesco Fuso Nerini. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications* 11, 233 (2020).

[56] Bryan Wilder, Nicole Immorlica, Eric Rice, and Milind Tambe. 2018. Maximizing Influence in an Unknown Social Network. In *AAAI conference on Artificial Intelligence (AAAI-18)*.

[57] Te Wu, Feng Fu, Yanling Zhang, and Long Wang. 2013. Adaptive role switching promotes fairness in networked ultimatum game. *Scientific reports* 3 (2013), 1550.

[58] Ioannis Zisis, Sibilla Di Guida, The Anh Han, Georg Kirchsteiger, and Tom Lenaerts. 2015. Generosity motivated by acceptance - evolutionary analysis of an anticipation games. *Scientific reports* 5, 18076 (2015).

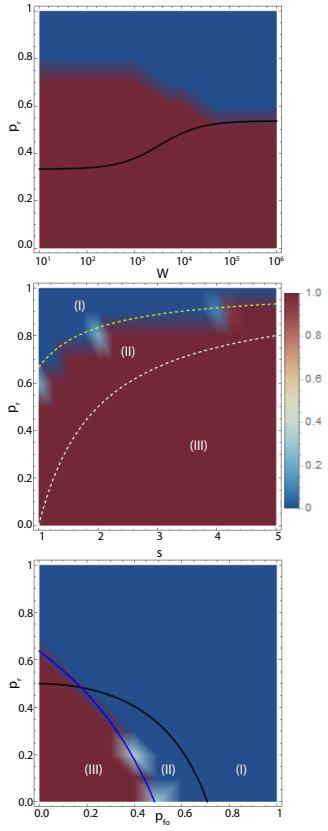

Figure S1: Total AU frequencies for the 8-neighbour lattice. The top row reports the spectrum between an early and late AI race (for varying W, with $p_{fo} = 0.1,\ s = 1.5$), the middle row addresses the early regime for varying s and p_r ($p_{fo} = 0.5,\ W = 100$), and the bottom row addressees the late regime for varying p_{fo} and p_r ($s = 1.5$, $W = 10^6$). Other parameters: $c = 1,\ b = 4,\ B = 10^4,\ \beta = 1$.

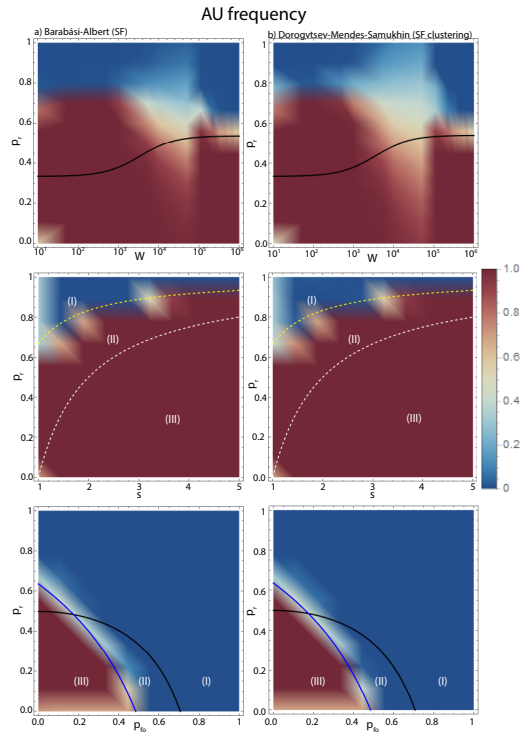

Figure S2: Comparison between the two different scale-free networks, BA and DMS. In this case, the payoffs have been normalised. The top row reports the spectrum between an early and late AI race ($p_{fo} = 0.1,\ s = 1.5$), the middle row addresses the early regime in more detail ($p_{fo} = 0.5,\ W = 100$) and the bottom row considers a late AI race ($W = 10^6,\ s = 1.5$). Parameters: $c = 1,\ b = 4,\ B = 10^4,\ \beta = 1$.

6 SUPPLEMENTARY INFORMATION

Figure S1 confirms the similar trends encountered in the regular square lattice. There are some very minor differences, but there is very little difference between well-mixed, the normal lattice and the eight-neighbour lattice. We confirm the similar late convergence found previously in some cases of the regular lattice.

We see very few improvements over the previously mentioned results on homogeneous populations. Interestingly, there is an area in the late regime where this type of normalised scale-free network produces more unsafe results (undesirably so) than either the well-mixed or lattice variants. We see some slight improvements in area **(II)** of the early regime.

In order to better understand the role and influence of highly connected zealots in the population, as well as to explore any potential for a government or regulatory agency to interfere in the AI race, we artificially accelerate or fund the safe zealots introduced as described in Section 4.3.3. In addition to the introduction of the players following pathological safe behaviour, we either accelerate their development (similarly to how unsafe players gain increased speed, in this case we add $\frac{sB}{W}$ to the influential pathological players payoffs, where $s = 2$), or heavily invest in these players (to the extent that other players will always imitate, by increasing their payoffs by a very large amount 10^7). Figure 7 displays our findings - very little improvement throughout, specifically with speed working more effectively for low p_r, while funding produces better results for high values of p_r.

We study a comprehensive view of pathological players (zealots) planted in a well-mixed lattice (see Figure S3) using similar methodology as described in Section 4.3.3, but in this case modifying 10% of the total population (not just highly connected nodes). We remove the pathological players from the frequency average to show how these affect the remainder of the population. We see very little effect of pathological players and we suggest that much lower β values would be required to see an effect. With the addition of mutation and more stochasticity, it would be possible for these pathological players to have a significant impact on the outcome.

Figure S4 validates the typical runs chosen to display the different trends earlier in the paper. We note the great variability between runs, due to network topology and the inherent stochastic nature of the system.

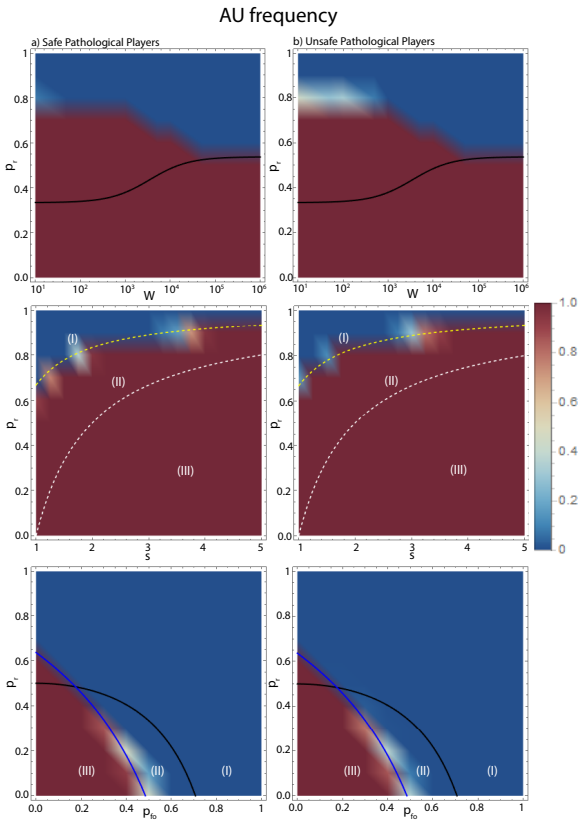

Figure S3: Introducing safe and unsafe zealots in the well-mixed scenario. Please note that the pathological players are excluded from these frequencies. The top row reports the spectrum between an early and late AI race ($p_{fo} = 0.1,\ s = 1.5$), the middle row addresses the early regime in more detail ($p_{fo} = 0.5,\ W = 100$) and the bottom row considers a late AI race ($W = 10^6,\ s = 1.5$). Parameters: $c = 1,\ b = 4,\ B = 10^4,\ \beta = 1$.

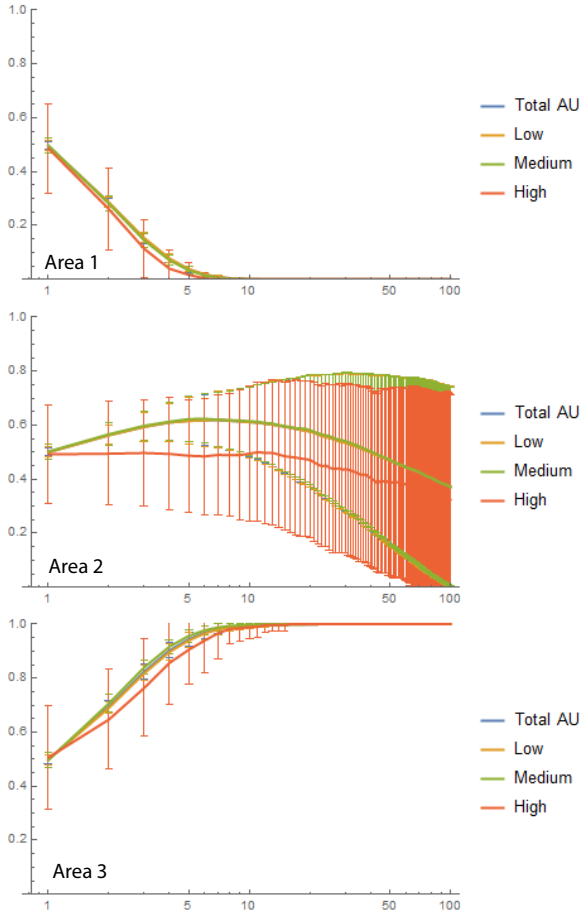

Figure S4: Distribution of unsafe behaviour (AU) in an early AI race, grouped by degree class (connectivity) of the nodes on scale-free networks, with error bars. For area 1 we test for $p_r = 1$, $p_r = 0.65$ for area 2 and $p_r = 0$ for area 3. Other parameters: $c = 1,\ b = 4,\ s = 1.5,\ B = 10^4,\ W = 100,\ \beta = 1$.