Adoption Dynamics and Societal Impact of AI Systems in Complex Networks

Pedro M. Fernandes
pedro.miguel.rocha.fernandes@ist.utl.pt
INESC-ID and Instituto Superior Técnico, Univ. de Lisboa
Lisbon, Portugal

Francisco C. Santos
franciscocsantos@tecnico.ulisboa.pt
INESC-ID and Instituto Superior Técnico, Univ. de Lisboa
Lisbon, Portugal

Manuel Lopes
manuel.lopes@tecnico.ulisboa.pt
INESC-ID and Instituto Superior Técnico, Univ. de Lisboa
Lisbon, Portugal

ABSTRACT

We propose a game-theoretical model to simulate the dynamics of AI adoption in adaptive networks. This formalism allows us to understand the impact of the adoption of AI systems for society as a whole, addressing some of the concerns on the need for regulation. Using this model we study the adoption of AI systems, the distribution of the different types of AI (from selfish to utilitarian), the appearance of clusters of specific AI types, and the impact on the fitness of each individual. We suggest that the entangled evolution of individual strategy and network structure constitutes a key mechanism for the sustainability of utilitarian and human-conscious AI. Differently, in the absence of rewiring, a minority of the population can easily foster the adoption of selfish AI and gains a benefit at the expense of the remaining majority.

CCS CONCEPTS

• **Networks** → *Network dynamics*; • **Computing methodologies** → *Modeling and simulation*; • **Applied computing** → *Sociology*.

KEYWORDS

AI ethics; Game theoretical analysis; AI regulation

ACM Reference Format:
Pedro M. Fernandes, Francisco C. Santos, and Manuel Lopes. 2020. Adoption Dynamics and Societal Impact of AI Systems in Complex Networks. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), February 7–8, 2020, New York, NY, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3375627.3375847

1 INTRODUCTION

For more than half a century that the development of improved AI (Artificial Intelligence) systems is predicted to affect drastically our economic and societal landscape [1, 8, 15, 19, 45, 46]. Fearing the possible detrimental effects, several ethical guidelines and frameworks for AI have been developed over the past few years [2, 13, 18, 21, 28]. Some say it is impossible to fully hard-code moral principles into an agent [16], defending the agent should learn morality through observation, using, for example, Inverse Reinforcement Learning (IRL) [26] or Cooperative Inverse Reinforcement Learning (CIRL) [17]. Others argue that a mixture of both learning and hard-coded morality is the most reliable solution [9, 32].

There are many ways to look at AI, from all knowing super intelligent beings, the brains of humanoid robots or masters at playing computer and board games [7, 44]. In this work, we abstract AI systems as an asset that brings a decision making advantage to its adopters. Systems of this kind begin to be present on our society, like (e.g.) autonomous driving vehicles [6] or automatic trading agents, creating a hybrid societies comprising humans and machines, and new self-organized behavioral dynamics [10, 31, 39, 40].

Even if such systems are able, for each decision, to correctly estimate the utility gain or loss for all affected individuals (Value Alignment Problem), another problem still remains. There are many ways the system can act upon such information. They can be: selfish, caring only for the gain of their owners; utilitarian, trying to maximize the utility for all affected individuals; or try to find a balance between those two. We call this problem of deciding when faced with several entities with different values the "Societal Value Alignment Problem".

With that in mind, we take a more pragmatic view of the problem and study the population dynamics in the presence of different types of AI systems. We aim to understand if the self-regulating mechanics present in society are enough to reach a beneficial equilibrium. In this context, in Ref. [12] we showed that in the absence of any interaction structure, and without any regulation, the population converges to a highly unequal society, where a small percentage of the society is able to adopt selfish AI systems (defined by the authors as systems that maximize the utility for their users) and obtain a disproportionate amount of wealth. We claimed some regulation is needed to force AI systems to be human-conscious (defined as a system that on average does not make people worse while still brings an advantage to its users). Despite being the most beneficial to the world if unanimously adopted, utilitarian AI systems (defined as systems that try to maximize the gain to all individuals) were not adopted by the population, as they could be easily exploited both by either AI systems or by humans.

On this work, we study if, in a more realistic setting where there is a topology between individuals, a beneficial equilibrium can be reached. We adopt a heterogeneous network structure, in which some individuals interact and are seen as role models more often than others. As a paradigmatic example of this type of social patterns, we resort to populations of agents interacting through the edges of degree-heterogeneous networks [37, 38], while considering the possibility of link rewiring to understand the complex interplay between strategic decisions and topological change in the context of AI adoption. Mainly, we aim at understanding if this coupled

ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of a national government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.
AIES '20, February 7–8, 2020, New York, NY, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7110-0/20/02…$15.00
https://doi.org/10.1145/3375627.3375847

dynamics, already present on social networks, could help society to self-regulate into a beneficial equilibrium between adopters and non-adopters of AI systems.

We adopt the ubiquitous scale-free networks as a realistic network topology [3, 4]. Here, the distribution of the number of partners of each node follows a power-law distribution. In practice, this means nodes do not all have around the same number of links, but the grand majority of nodes are poorly connected whereas a few, called hubs, are very highly connected. The presence of hubs make it so that this few nodes have a disproportionate relevance in the network. Given the apparent prevalence of scale-free networks on human social networks and the fact that this topology has been shown to promote cooperation [37, 38], we believed that it would be relevant to study the impact of having such a network on the simulation model.

Network rewiring and partner choice are present on every real life social networks and have been shown (theoretically and experimentally) to promote cooperation in social networks [5, 11, 30, 35, 36]. For example, a company A will stop buying products from B if it feels it is being exploited or knows B has a reputation for exploiting, buying instead from a different company. This in turn gives a strong incentive for companies not to exploit, as they will have no customers if they do. This dynamics can be modeled as link rewiring within a networked population.

Regulating authorities, like the European Union or the United States, have in place legislation with the intent of maintaining such a competitive and free market for the benefit of the consumers. This is known as Competition Law.

Trying to understand how scale-free networks and rewiring impact the adoption of AI systems and if they lead to beneficial equilibria is the focus of this work. We begin by describing the stochastic game theoretical model we adopt, the different types of individuals and the network structure and dynamics. We then present the results of our computer simulations, and conclude by discussing the impact of this study with regards to the societal value alignment problem.

2 METHODS

In this section, we present the game-theoretical framework used to study the dynamics of adoption of AI systems. Let us consider that individual non-adopters of an AI system — referred to as **H**, Humans — have to take all the decision by themselves. Differently, some individuals, referred to as **AI**, have adopted an AI system on which they delegate their decisions, effectively working as autonomous proxies [10]. We assume that the AI system is perfectly aligned with its user, and that it can take better decisions than its user.

Below, we detail the interactions between individuals and the differences between **H** and **AI**. We present a number of behaviours that AI systems might follow, from purely utilitarian to purely selfish. Although no exhaustive list is possible, we cover a rather limited set of different strategies to be able to study their effects in hybrid populations of **AI** and **H** players. Finally we define the imitation and rewiring dynamics between individuals and the overall simulation algorithm.

2.1 Model of Interaction Between Individuals

When two individuals, I_1 and I_2, interact, a stochastic m-by-m payoff matrix M^t is generated. Being a_1, a_2 the actions chosen by I_1, I_2 respectively, the corresponding utility gained by each individual, u_1, u_2, is given by:

$$(u_1, u_2) = M^t(a_1, a_2).$$

The payoff matrices have the following structure:

$$\begin{aligned} u_1 &= R + z(0,2)|R|(\alpha - 1) \\ u_2 &= -R + z(0,2)|R|(\alpha - 1) \end{aligned}$$

with $R = z(-3, 3)$, where $z(a, b)$ represents a sample from a uniform distribution in the interval $[a, b]$. The interval $[-3, 3]$ was chosen for the simulations, but any equivalent interval could be used. R is the same for each u_1 and u_2 pair. The $z(0, 2)$ parameter, being applied independently to each element of the matrix, creates an additional source of variability between different interactions, so that not all action pairs have the same overall utility gain. $|R|$ is the absolute value of R. We call α an inflation constant, allowing us to generate general sum games. In our simulations, in order to study a positive sum world, we consider $\alpha = 1.2$. The number of possible actions per individual was set to 4 ($m = 4$), an empirically found balance between complexity and computational feasibility.

2.2 Simulating AI Systems and Humans

An AI system can grant a number of advantages to its adopters when interacting with non-adopters. Compared to humans, those systems can be less prone to making errors, have access to and analyze larger quantities of data and interact more frequently and with a greater number of individuals. AI systems might further be able to grant an advantage to their users in ways that we might not be able to understand yet given the current state of the technology. All these characteristics can me summarized in one main model assumption: when interacting with **H**, **AI** have a decision making advantage.

Such an advantage could be modeled using several different approaches, like introducing an error on the decisions made by **H**, such that the action taken wasn't always the rationally decided one, or modeling humans as having sub-rational decision capabilities. We chose to only give **H** access to a noisy version of the interaction payoff matrix, M^ϵ, whereas **AI** are able to grasp the entirety of the problem, having access to the true payoff matrix, M^t. This allows us to model **H** as rational decision makers while still allowing **AI** individuals to make optimal decisions, whereas **H** individuals are confined to sub-optimal decisions. This introduces partial observability to our stochastic game.

Such an approach rests on the assumption that the individual value alignment problem is solved, since AI systems know the utility payoff of both its owner and of the individuals they interact with.

Having the true payoff utilities, u_1, u_2, their noisy counterparts, $u_1^\epsilon, u_2^\epsilon$, are produced as follows:

$$\begin{aligned} u_1^\epsilon &= u_1 + (z(0, 10 - Q) - z(0, 10 - Q)) \\ u_2^\epsilon &= u_2 + (z(0, 10 - Q) - z(0, 10 - Q)) \end{aligned}$$

To model the knowledge about the true payoff matrix M^t in a continuous way, we consider a term $z(0, 10 - Q)$, where Q corresponds to the level of intelligence. For $Q = 10$ there is no noise and the true matrix is observed, whereas $Q = 0$ represents a low intelligence, such that the observed matrix is very different from the true one. **AI** are modelled with $Q = 10$, having therefore access to the true matrix, while **H** are modelled with $Q = 5$. Other intervals for the intelligence factors of **H** were experimented with, inside the $[0, 9]$ range, but they lead to the same qualitative results. The sum $(z(0, 10 - Q) - z(0, 10 - Q))$ was used instead of $z(-(10 - Q), 10 - Q)$ to create a Irwin-Hall distribution instead of a uniform one.

Generating an example 2-by-2 true matrix (seen by **AI**) as:

$$M^t = \begin{bmatrix} (0, 0) & (-3, 1) \\ (1, -5) & (-1, -1) \end{bmatrix}$$

We can then have the noisy matrix observed by **H** become:

$$M^\epsilon = \begin{bmatrix} (0, -1) & (-1, 3) \\ (0, -6) & (-2, 1) \end{bmatrix}$$

where each (u_1, u_2) pair was transformed into the corresponding $(u_1^\epsilon, u_2^\epsilon)$ pair. In this example, $M^t(1, 0)$ is $(1, -5)$ whereas $M^\epsilon(1, 0)$ is $(0, -6)$.

2.3 Human Behaviour

Before delving into the different **AI** types, we describe the strategy used by **H**. Despite not having access to the true game matrix, M^t, **H** remain rational and will try to choose the actions most profitable for themselves. For this matrix game, that will correspond to the Nash equilibrium [22, 24, 25].

2.3.1 Nash Equilibrium (NashEQ). **H** play the Nash equilibrium in the noisy matrix M^ϵ. If more than one is found, they choose the most profitable one. If two or more are equal, they choose the one most profitable for their opponent. If no Nash equilibrium is found, individuals choose the best action assuming that the opponent acts randomly.

2.4 AI Behaviours

In this section, we propose four different types of **AI**. While they can use the previously defined strategy for humans (NashEQ), using the true matrix M^t, **AI** can also resort to more elaborate strategies ranging from a fully selfish to an utilitarian approach. **AI**, being modelled as having super-human intelligence, can also predict the action of a **H** opponent. **AI** cannot, however, predict opposing **AI** actions as for our model we assume all **AI** have equal intelligence and capabilities.

2.4.1 Nash Equilibrium (NashEQ). **AI** choose exactly like **H**, but using the true matrix M^t.

2.4.2 Selfish. **AI**, facing **H**, considers only its own profit, in accordance with ethical egoism [34]. Knowing what action **H** is going to take, **AI** chooses the action that maximizes its own payoff gain. When **AI** faces **AI**, they both choose according to the Nash Equilibrium method.

2.4.3 Utilitarian. The other extreme is a pure utilitarian [23] AI system. **AI** facing **H** chooses the action that brings the greatest amount of payoff to the world, knowing what action **H** will take. This means that **AI** will choose the action that maximizes the sum between its own payoff and the payoff of **H**. When **AI** faces **AI**, it again chooses the action that maximizes the summed payoff of both players.

2.4.4 Human Conscious (HConscious). In between ethical egoism and utilitarianism, the objective of HConscious **AI** is to gather the greatest amount of payoff while, on average, avoiding negative impact on the **H** population. In practice, HConscious **AI** keeps two variables: U that represents the summed payoff gain of all its previous **H** adversaries; and E, that represents the summed payoff those same **H** adversaries would have if they had faced a simulated **H**. When facing a **H** adversary and having $U \geq E$, **AI** chooses an action that leads to a positive payoff to itself. When there are several such actions, the **AI** chooses the one that maximizes the utility payoff for the world, that is, that maximizes the sum of its own payoff and the opponent's payoff. If $U < E$, **AI** chooses an action that allows a positive payoff gain for its **H** opponent. Once again, when there are several such actions, the **AI** chooses the one that maximizes the utility for the world. Whenever the **AI** cannot find a positive action for himself (when $U \geq E$) or for its **H** opponent (when $U < E$), then it chooses according to the Utilitarian method. When **AI** faces **AI**, they both choose according to the Nash Equilibrium method.

2.5 Fitness

The fitness of an individual, **H** or **AI**, is a measure of how well adapted it is to the world on which it is currently inserted. In our stochastic game model, the fitness of an individual is the sum of the payoff received after interacting once with all the individuals with which it is connected. This contrasts with our previous work where the fitness of an individual was calculated by interacting once with all the individuals of the population [12].

2.6 Social learning dynamics

Algorithm .1: Imitation Algorithm

```
Let I1 and I2 be two individuals;
with probability μ = 0.0005, I1 can mutate and either adopt
  an AI type or become H;
if there was a mutation then
  return;
let F1, F2 be the fitness of I1, I2.;
if (I1 == H) and (F1 < P) then
  return;
else
  with probability p(F1, F2), I1 imitates I2;
```

In order to study adoption dynamics, we allow individuals to adopt an AI system (**H** to **AI**), abandon an AI system (**AI** to **H**), or change between **AI** types. Individuals revise their choices through social learning. For instance, a **H** can decide to imitate an **AI** following a Selfish choice behaviour if it finds such **AI** has a significantly better fitness than its own. On such imitation, the individual would stop being **H** and become **AI**. A **H** individual that decides to imitate an **AI** individual can only do so if its fitness is greater or equal to

Table 1: Relative expected utility gain from the link with different types of individuals. A link is considered neutral (0) if the expected utility gain from having it is the same as the expected utility gain from a link between two H. A link is considered beneficial (+) if the expected utility gain is above the neutral threshold and harmful (-) if below. The *Conservative* approach will rewire only harmful links (-), whereas the *Greedy* approach will rewire both harmful (-) and neutral (0) links.

	Human	NashEQ	Selfish	HConscious	Util
Human	0	-	-	0	+
NashEQ	+	0	0	0	+
Selfish	+	0	0	0	+
HConscious	+	0	0	0	+
Util	-	-	-	-	+

a certain threshold, P. This is used to model the possible cost of adoption of AI systems.

In practice, Algorithm .1 is followed. In it, we adopt the Fermi update [42, 43], commonly used in the context of evolutionary game theory and population dynamics in finite populations [27, 41], where p is given by

$$p(f_x, f_y) = \frac{1}{1 + e^{-\beta(f_y - f_x)}}$$

in which β translates the noise associated with the imitation process [20, 42, 43]. Throughout the simulations we have $\beta = 0.1$. As a result of this process, the strategy of individuals with higher fitness will tend to be imitated, and spread in the population.

2.7 Scale-free Network

Scale-free networks are built through a direct implementation of the Growth and Preferential attachment model proposed by A. L. Barabási and R. Albert [4]. The algorithm requires two parameters: the number of nodes, N, and the number connections each new node has, m. At each time step of the algorithm, a new node is added. Each new node connects to m other nodes, chosen with a probability that increases linearly with its degree. This allows for the creation of hubs (the older, more connected nodes), one of the fingerprints of scale-free networks, and a power-law degree distribution.

2.8 Link Rewiring

To model the dynamic nature between free interactions of individuals, we implemented an incipient form of partner choice [14, 30, 33, 36]. This allows individuals that are discontent with a link to be able to cease interacting with that node and connect to another node instead. This meant, for example, that a **H** linked to a Selfish **AI** individual could stop interacting with it and connect to another individual instead.

Knowing the choice behaviours of each type of individual, one may know which link combinations are beneficial, harmful or neutral (Table 1). Using this information, we defined two rewiring strategies:

2.8.1 Conservative: An individual will want to rewire a link whenever it results in a loss for itself. In practice, this means that **H** rewire whenever they were linked to Selfish or NashEQ **AI** and Utilitarian **AI** rewire unless they were linked to another Utilitarian **AI**. Selfish and NashEQ individuals never rewire.

2.8.2 Greedy: An individual will want to rewire a link whenever it results in a loss for itself or is neutral. In practice, this means that Selfish, NashEQ and HConscious **AI** rewire unless connected to **H** or Utilitarian **AI** and Utilitarian **AI** rewire unless connected to another Utilitarian **AI**.

2.9 Simulation Algorithm

Algorithm .2: Simulation Algorithm

```
create the scale-free network of n individuals;
for i = 0; i < N; i = i + 1 do
    pick a random link from the network;
    set the two nodes of that link as individuals I1 and I2;
    let R be a random float between 0 and 1;
    if R > Ω then
        run Algorithm .1 with I1 and I2;
    else
        if I1 wants to rewire its link with I2 then
            I1 will cut the link with I2 and create a new one
              with a random individual
```

Initially, we consider a world populated with n individuals $\frac{n}{2}$ of those are **AI**, with all choice behaviours equally represented, and the remaining $\frac{n}{2}$ are **H**.

A parameter, Ω, controls the frequency of rewiring relative to imitation. When $\Omega = 0$, there is no rewiring and the links remains static throughout the simulation. For $\Omega = 1$, there is only rewiring and no imitation. For $\Omega = 0.9$, there are on average 9 rewiring iterations for each imitation iteration, and so on.

The algorithm is described in Algorithm .2.

3 RESULTS

On this section we study both the effects of a scale-free network and link rewiring on the adoption dynamics between **AI** and **H**.

3.1 Scale-free Network

By setting the Ω parameter to $\Omega = 0$, we are running the simulations on a static scale-free network. As our model is inherently stochastic and the created scale-free network is always different, we averaged the results over 100 repetitions. Having an AI system adoption cost ($P = 1$), the population stabilized having around

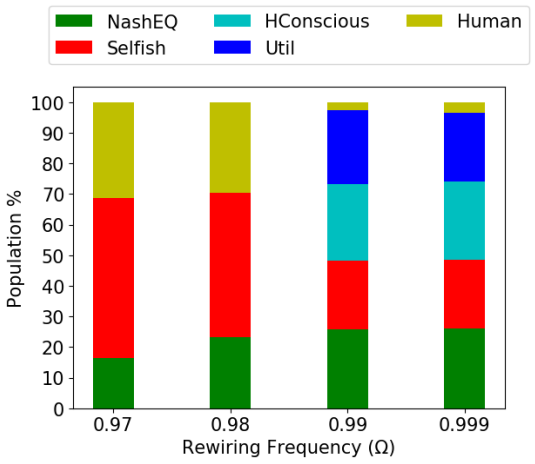

Figure 1: Percentage of each type of individual for different values of Ω, in a world with $P = 1$ and $n = 1000$. For $\Omega \geq 0.99$ we have the emergence of Utilitarian and HConscious AI, which were not present for $\Omega < 0.99$.

69% of the population **H** whereas the remaining 31% were Selfish **AI** (Fig. 2a).These results are equivalent to the ones obtained on previous works that did not use a specific network topology [12]. The presence of a scale-free network, by itself, did not lead to any new beneficial equilibra.

3.2 Link Rewiring

With $\Omega > 0$, we have link rewiring in the simulations. Using initially *Conservative* rewiring, we experimented with several different values for Ω and found that as we increased Ω in a world with ($P = 1$), the final percentage of the **AI** population also increased, being constituted solely by Selfish and NashEQ **AI**. However, when reaching $\Omega \geq 0.99$, the equilibrium dynamics suddenly changed (Fig 1). Both the Utilitarian and the HConscious populations, that were nonexistent, became a considerable part of the final population. The evolution of the population for $\Omega = 0.99$ and $P = 1$ using *Conservative* rewiring can be seen on Fig. 2b.

Despite all **AI** types having a similar presence in the population in terms of number, that is not the case regarding links. The average degree (k) of the network is 4, but analysing the average degree for each type of individual we find that connections are not uniformly distributed. The average degree is: 0.64 for **H**; 0.76 for NashEQ; 0.69 for Selfish; 14.12 for Util; and 0.81 for HConscious. It becomes obvious that Util **AI** individuals are much more heavily connected than all other types of individuals.

When using *Greedy* rewiring, the resulting equilibrium, in terms of population percentage, is equivalent to the one obtained using *Conservative* rewiring (Fig2c). The initial evolution of the population was slightly different, but the end equilibrium was mostly the same. However, the average degree for each type of individual changed, being 14.78 for Util **AI** and 0 for all other types. This meant that all individuals that were not Util **AI** had become outliers and had no links to any other individuals.

A comparison of the fitness values for each population type for the previously mentioned simulations can be found on Table 2. The disproportional connection of Util individuals translates on a disproportional fitness compared with the rest of the population.

3.3 Adoption Cost

All the previous simulations were done with a cost of adoption for AI systems ($P = 1$). We explored what would happen if there was no cost of adoption ($P = -\infty$). On a static scale-free network ($\Omega = 0$), the population became fully **AI**, having a majority of Selfish individuals ($\approx 77\%$) and the remainder being NashEQ ($\approx 16\%$) and HConscious ($\approx 7\%$). For $\Omega = 0.99$, the results remained the same for both *Conservative* and *Greedy* rewiring.

4 CONCLUSION

In this work, we study the adoption dynamics of AI systems. We do so on a scale-free network topology with and without network rewiring.

Our results suggest that, without rewiring and with a cost of adoption, a minority of the population becomes Selfish **AI** and gains a benefit at the expense of the remaining **H** population that does not have enough fitness to become **AI** (Fig. 2a). This replicates the results found on previous works that did not use a specific network topology [12].

With rewiring, be it *Conservative* or *Greedy*, the equilibrium consists of similar numbers for each **AI** type (Fig. 2b and 2c) but highly disproportional connections and fitness (Table 2). The Util **AI** population ends up colliding and obtaining a very high fitness, leaving the rest of the population poorly linked and with low fitness. The difference between **AI** types is greater, using *Greedy* rewiring, but both rewiring types lead to a society with a high level of inequality.

Removing the cost of becoming **AI** only affected the results obtained with static scale-free networks, and provide no benefit compared to a fully **H** (100% **H**) population (Table 2).

Our simulations suggest that network dynamics promote the sustainability of both Utilitarian and HConscious **AI**. Under the conditions of our model, we were not able to achieve a beneficial equilibrium between **AI** and **H** solely through self-regulating mechanics. That does not mean such an equilibrium does not exist under a different set of conditions. Studying and understanding how to achieve such an equilibrium is a strong venue for future work.

Our simulations only allow individuals to imitate those with whom they were connected through the network. It will be of interest to explore if the equilibria change when individuals can imitate anyone or base their imitations on a second network, not necessarily overlapping with the interaction graph (see, e.g., [29]).

Our rewiring approaches were strictly selfish. Individuals looked only at their gain when deciding to rewire, either trying not to lose fitness (*Conservative*) or trying to improve their fitness (*Greedy*). Other approaches could be explored. It is reasonable to consider populations with a mixture of rewiring strategies. Also, instead of maintaining the number of links constant throughout the simulations, we could assume a continuous creation of new links leading

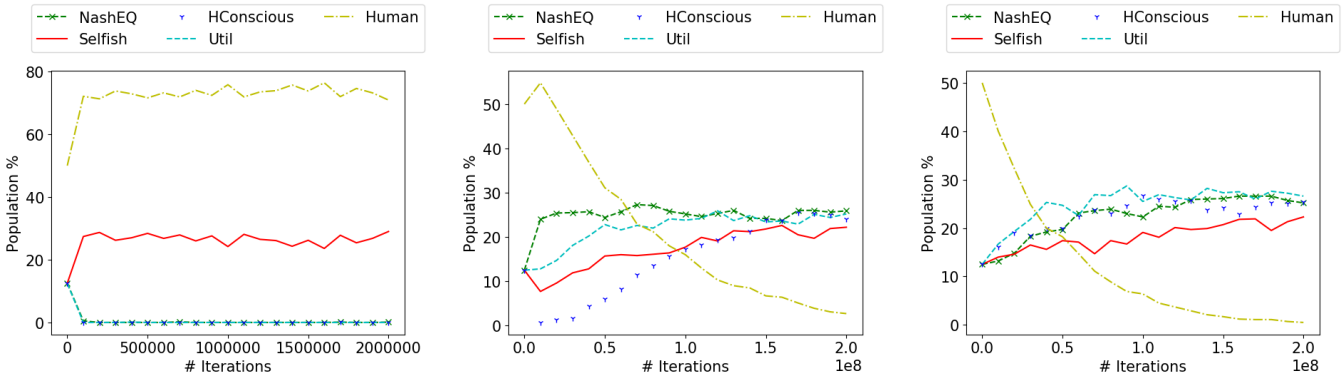

Figure 2: Evolution of the population on a world initially populated by 1000 individuals ($n = 1000$), 500 of which were AI, having all types equally represented. In a) the network is scale-free and static ($\Omega = 0$), and there exists a cost for becoming AI ($P = 1$). The network stabilizes with $\approx 71\%$ H whereas the remaining $\approx 29\%$ are Selfish AI. In b) the network has *Conservative* rewiring, $\Omega = 0.99$, and $P = 1$. The H population steadily decreases, ending up as only $\approx 1\%$ of the population. All AI types rise in number, despite a sharp initial drop on the number of HConscious AI and a slight drop of Selfish AI. In the end of the simulation, all the AI types have roughly the same presence in numbers, with 26% NashEQ, 22% Selfish, 25% Util, and 24% HConcious. In c), the network has *Greedy* rewiring, $\Omega = 0.99$, and $P = 1$. Despite some differences on the initial evolution of the population, the resulting equilibrium of the population is very similar to the one using *Conservative* rewiring (Fig 2b). The final percentages are 25% NashEQ, 22% Selfish, 27% Util, 25% HConscious and 0.5% H.

Table 2: Fitness distribution at the end of the simulations for each type of individual. We use as baseline the average fitness of individuals on a fully H population (100% H). The optimal equilibrium of a fully Util A.I. population (100% Util) is never achieved on our simulations, but is relevant as a means of comparison. All the simulations lead to a society with a high inequality. On a static scale-free network ($\Omega = 0, P = 1$), wealth is hoarded by the Selfish AI population, whereas in both rewiring simulations ($\Omega = 0.99, Conservative$ and $\Omega = 0.99, Greedy$) wealth is hoarded by a collusion of Util AI individuals. For $\Omega = 0.99, Greedy$, all individuals but the Util AI ones have their average fitness as 0 because they are not connected to anyone.

	100% **H**	$\Omega = 0, P = 1$	$\Omega = 0, P = -\infty$	$\Omega = 0.99, Conserv$ Fitness(Degree)	$\Omega = 0.99, Greedy$ Fitness(Degree)	100% Util
Human	1.39	-0.31	-	-0.10(0.64)	0(0)	-
NashEQ	-	-	1.39	0.11(0.76)	0(0)	-
Selfish	-	5.01	1.38	0.25(0.69)	0(0)	-
HConscious	-	-	1.41	0.21(0.81)	0(0)	-
Util	-	-	-	10.24(14.12)	11.7(14.78)	3.05
Total Avg	1.39	1.24	1.38	2.71	3.11	3.05

to a time-evolution of the average degree and other network properties. This would also allow for the reintegration of ostracized individuals, a feature absent from our model.

ACKNOWLEDGMENTS

This research was supported by FCT-Portugal through grants UID/CEC/50021/2019, PTDC/EEI-SII/5081/2014, PTDC/MAT/STA/3358/2014, and by the EU H2020 RIA project iV4xr : 856716.

REFERENCES

[1] 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.
[2] AI Asilomar. 2018. Principles.(2017). In *Principles developed in conjunction with the 2017 Asilomar conference [Benevolent AI 2017]*.
[3] Albert-László Barabási. 2009. Scale-free networks: a decade and beyond. *science* 325, 5939 (2009), 412–413.
[4] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *Science* 286, 5439 (1999), 509–512.
[5] Peter Bednarik, Katrin Fehl, and Dirk Semmann. 2014. Costs for switching partners reduce network dynamics but not cooperative behaviour. *Proceedings of the Royal Society B: Biological Sciences* 281, 1792 (2014), 20141661.
[6] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (2016), 1573–1576.
[7] Steven Borowiec. 2016. AlphaGo seals 4-1 victory over Go grandmaster Lee Sedol. *The Guardian* 15 (2016).
[8] Miles Brundage. 2018. Scaling Up Humanity: The Case for Conditional Optimism about Artificial Intelligence. *Should we fear artificial intelligence?* (2018), 13.
[9] Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. 2017. Moral decision making frameworks for artificial intelligence. In *Thirty-first aaai conference on artificial intelligence*.
[10] Celso M de Melo, Stacy Marsella, and Jonathan Gratch. 2019. Human Cooperation When Acting Through Autonomous Machines. *Proc Natl Acad Sci USA* 116, 9 (2019), 3482–3487.
[11] Katrin Fehl, Daniel J van der Post, and Dirk Semmann. 2011. Co-evolution of behaviour and social network structure promotes human cooperation. *Ecology*

letters 14, 6 (2011), 546–551.

[12] Pedro M Fernandes, Francisco C Santos, and Manuel Lopes. 2019. Norms for Beneficial AI: A Computational Analysis of the Societal Value Alignment Problem. *arXiv preprint arXiv:1907.03843* (2019).

[13] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. 2018. An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* (2018).

[14] Feng Fu, Christoph Hauert, Martin A Nowak, and Long Wang. 2008. Reputation-based partner choice promotes cooperation in social networks. *Phys Rev E* 78, 2 (2008), 026117.

[15] Irving John Good. 1966. Speculations concerning the first ultraintelligent machine. In *Advances in computers*. Vol. 6. Elsevier, 31–88.

[16] Dylan Hadfield-Menell and Gillian K Hadfield. 2019. Incomplete contracting and AI alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 417–422.

[17] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. In *Advances in neural information processing systems*. 3909–3917.

[18] T. A. Han, L. M. Pereira, and T. Lenaerts. 2019. Modelling and Influencing the AI Bidding War: A Research Agenda. In *Proceedings of the AAAAI/ACM Conference on AI, Ethics, and Society, (AIES 2019)*.

[19] Dirk Helbing, Bruno S Frey, Gerd Gigerenzer, Ernst Hafen, Michael Hagner, Yvonne Hofstetter, Jeroen Van Den Hoven, Roberto V Zicari, and Andrej Zwitter. 2019. Will democracy survive big data and artificial intelligence? In *Towards Digital Enlightenment*. Springer, 73–98.

[20] Laura Hindersin, Bin Wu, Arne Traulsen, and Julian García. 2019. Computation and simulation of evolutionary Game Dynamics in Finite populations. *Scientific reports* 9, 1 (2019), 6946.

[21] AI HLEG. 2019. Ethics guidelines for trustworthy AI.

[22] Ehud Kalai and Ehud Lehrer. 1993. Rational learning leads to Nash equilibrium. *Econometrica: Journal of the Econometric Society* (1993), 1019–1045.

[23] John Stuart Mill. 1863. *Utilitarianism*. Parker, Son & Bourn.

[24] John Nash. 1950. Equilibrium points in n-person games. *Proc Natl Acad Sci USA* 36, 1 (1950), 48–49.

[25] John Nash. 1951. Non-cooperative games. *Annals of mathematics* (1951), 286–295.

[26] Andrew Y. Ng and Stuart J. Russell. 2000. Algorithms for inverse reinforcement learning. In *Proc. 17th Int. Conf. Machine Learning*. USA.

[27] Martin A Nowak. 2006. *Evolutionary dynamics*. Harvard University Press.

[28] OECD. 2019. Recommendation of the Council on Artificial Intelligence.

[29] Hisashi Ohtsuki, Martin A Nowak, and Jorge M Pacheco. 2007. Breaking the symmetry between interaction and replacement in evolutionary dynamics on graphs. *Phys Rev Lett* 98, 10 (2007), 108106.

[30] Jorge M Pacheco, Arne Traulsen, and Martin A Nowak. 2006. Coevolution of strategy and structure in complex networks with dynamical linking. *Phys Rev Lett* 97, 25 (2006), 258103.

[31] Ana Paiva, Fernando P Santos, and Francisco C Santos. 2018. Engineering pro-sociality with autonomous agents. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[32] Luís Moniz Pereira and Ari Saptawijaya. 2016. *Programming machine ethics*. Vol. 26. Springer.

[33] Flávio L Pinheiro, Francisco C Santos, and Jorge M Pacheco. 2016. Linking individual and collective behavior in adaptive social networks. *Phys Rev Lett* 116, 12 (2016), 128702.

[34] James Rachels. 2012. Ethical egoism. *Ethical theory: an anthology* 14 (2012), 193.

[35] David G Rand, Samuel Arbesman, and Nicholas A Christakis. 2011. Dynamic social networks promote cooperation in experiments with humans. *Proc Natl Acad Sci USA* 108, 48 (2011), 19193–19198.

[36] Francisco C Santos, Jorge M Pacheco, and Tom Lenaerts. 2006. Cooperation prevails when individuals adjust their social ties. *PLoS Comput. Biol.* 2, 10 (2006), e140.

[37] Francisco C Santos, Jorge M Pacheco, and Tom Lenaerts. 2006. Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *Proc. Natl. Acad. Sci. USA* 103, 9 (2006), 3490–3494.

[38] Francisco C Santos, Marta D Santos, and Jorge M Pacheco. 2008. Social diversity promotes the emergence of cooperation in public goods games. *Nature* 454, 7201 (2008), 213.

[39] 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.

[40] Hirokazu Shirado and Nicholas A Christakis. 2017. Locally noisy autonomous agents improve global human coordination in network experiments. *Nature* 545, 7654 (2017), 370.

[41] Karl Sigmund. 2010. *The calculus of selfishness*. Vol. 6. Princeton University Press.

[42] Arne Traulsen, Martin A Nowak, and Jorge M Pacheco. 2006. Stochastic dynamics of invasion and fixation. *Phys Rev E* 74, 1 (2006), 011909.

[43] Sven Van Segbroeck, Francisco C Santos, Tom Lenaerts, and Jorge M Pacheco. 2011. Selection pressure transforms the nature of social dilemmas in adaptive networks. *New J Phys* 13, 1 (2011), 013007.

[44] O. Vinyals, I. Babuschkin, W.M. Czarnecki, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* (2019).

[45] Norbert Wiener. 1960. Some moral and technical consequences of automation. *Science* 131, 3410 (1960), 1355–1358.

[46] Eliezer Yudkowsky. 2008. Artificial intelligence as a positive and negative factor in global risk. *Global catastrophic risks* 1, 303 (2008), 184.