Norms for Beneficial A.I.: A Computational Analysis of the Societal Value Alignment Problem

Pedro Miguel da Rocha Fernandes
pedro.miguel.rocha.fernandes@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

October 2019

Abstract

The rise of artificial intelligence (A.I.) based systems has the potential to benefit adopters and society as a whole. However, these systems may also enclose potential conflicts and unintended consequences. Notably, people will only adopt an A.I. system if it confers them an advantage, at which point non-adopters might push for a strong regulation if that advantage for adopters is at a cost for them. Here we propose a stochastic game theoretical model for these conflicts. We frame our results under the current discussion on ethical A.I. and the conflict between individual and societal gains, the societal value alignment problem. We test the arising equilibria in the adoption of A.I. technology under different norms followed by artificial agents, their ensuing benefits, and the emergent levels of wealth inequality. We show that without any regulation, purely selfish A.I. systems will have the strongest advantage, even when a utilitarian A.I. provides a more significant benefit for the individual and the society. Nevertheless, we show that it is possible to develop human conscious A.I. systems that reach an equilibrium where the gains for the adopters are not at a cost for non-adopters while increasing the overall fitness and lowering inequality. However, as shown, a self-organized adoption of such policies would require external regulation.

Keywords: A.I. ethics; Game theoretical analysis; A.I. regulation.

1. Introduction

Several applications already have an Artificial Intelligent system (A.I.) taking decisions in place of their owners. It is expected that in the future, such delegation of decisions will become more ubiquitous and effective. It is still open to debate whether that will have a positive or a negative impact on society [25, 6]. Strong voices highlight the dangers of A.I. [1] and call for regulation [10], some others dismiss such fears [5] and are against regulation [7]. Some of these discussions come from a lack of understanding of the current A.I. capabilities and strong divergences about its future developments, especially in artificial general intelligence (AGI). Some concerns might be true when AGI is created but not under the current state-of-the-art. How fast we can get there is still open to debate [4], and so is if we should strive to get there fast, or delay it [3, 8]. But even under the current state-of-the-art in A.I. there are problems that may arise with their introduction, e.g. autonomous vehicles [2].

A strong regulation could decide that A.I. systems should act using an egalitarian or utilitarian perspective. However, a utilitarian perspective or norm might not be efficient, and in many cases, an egalitarian solution does not exist. An utilitarian A.I. would often have to act against the interests of its owner, and if people can choose to adopt or not an A.I. system, we can expect they will only do it if it is individually rational to do so. In principle, either the A.I. system gives an individual advantage for its owner, or it will not be bought. If there is no interest in buying, there will be no interest in production, curbing research and development.

Without any regulation, A.I. systems might lead to invasion of privacy, use of confidential information, cheating in games, collusion in public contracts, and many others. Even if they do nothing illegal or clearly unethical, their efficiency and effectiveness might greatly unbalance the societal scales. In this case, we can expect that non-adopters might push for the abolition or at least a substantial regulation of A.I.

It is challenging to conciliate these two goals of aligning with the preferences of A.I. adopters and those of the non-adopters. We call this the societal value alignment problem. Besides being advantageous for the adopters of A.I. (individual rationality), it needs to be better for the non-adopters and so for everyone in the society.
(societal rationality). The rise of A.I. systems has the power to create strange new market dynamics [22]. Efforts should be made to model these possible future worlds, so we understand them better before we are in the midst of the problems that might arise. Voices in the scientific community begin to pressure for the research on this area, and on the ethical, scientific and engineering problems it presents [23, 19, 9, 17].

Here we aim to model and understand how A.I. systems can provide an advantage for those adopting them (creating incentives for the scientific, technological and societal development) but without creating such advantage at the expense of others (allowing for societal acceptance of the systems). To do so, we define several different types of A.I. systems, adopting different types of norms, ranging from pure selfish to pure utilitarian. Then we study the time evolution of the adoption of each type of A.I. when they compete against each other and also the equilibrium for each A.I. system in particular.

Individual adopters of A.I. systems can be seen as: singular citizens, adopting A.I. systems for personal gains; corporations, buying A.I. to increase profits; political entities, using A.I. to gain influence; or even countries, deploying A.I. to gain an upper hand on war and trading. Our model abstracts individuals as equally complex entities that interact between each other, winning or losing utility on each interaction.

In particular, this analysis aims at answering the following questions:

1. Will self-regarding individuals adopt A.I. systems?
2. With different types of A.I. systems available, which ones will be adopted?
3. If adopted, what is the individual and collective gain, depending on the strategy adopted by the A.I. system?

2. Methods
In this section, we present a game-theoretical framework to study the impact of the adoption of A.I. systems on individuals and on the society. Henceforward, individual non-adopters of an A.I. system will be referred to as H, while individuals adopters of an A.I. system representative will be referred to as A.I.

2.1. Model of Interaction Between Individuals:
On each interaction between two individuals, \( I_1 \) and \( I_2 \), a stochastic payoff matrix \( M \) is generated. This is a \( m \)-by-\( m \) matrix of payoff pairs. Being \( a_1 \) the action chosen by \( I_1 \) and \( a_2 \) the action chosen by \( I_2 \), the payoff received by each individual is respectively \( u_1 \) and \( u_2 \), such that:

\[
(u_1, u_2) = M(a_1, a_2).
\]

In order to explicitly generate general sum games, as conclusions might be different in positive, negative or zero-sum worlds, the payoff matrices have the following structure:

\[
\begin{align*}
u_1 &= R + z(0, 2)|R|(\alpha - 1) \\
u_2 &= -R + z(0, 2)|R|(\alpha - 1)
\end{align*}
\]

Having \( R = z(-3, 3) \), 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. \( z(0, 2) \) is applied independently for each element of the matrix. This \( z(0, 2) \) parameter 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 will call \( \alpha \) an inflation constant. For \( \alpha = 0 \), the matrices will, on average, create a zero sum game where no payoff is created or lost, just transferred between individuals. For \( \alpha > 0 \) there is on average a positive total payoff, creating a positive sum game, and for \( \alpha < 0 \) there is on average a negative total payoff, creating a negative sum game. 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 A.I. Systems and Humans
There are many ways in which A.I. systems could grant an advantage to their users, some of which we might not even be able to understand yet given the current state of the technology. Our main model assumption is the following: when interacting with H, A.I. have a decision making advantage.

In a computational way we are in the presence of partial observability in our stochastic game. The difference between H and A.I. is that while A.I. sees the real payoff matrix, \( M \), H sees a noisy version of it, \( M' \). This allows A.I. individuals to make optimal decisions, while H individuals are confined to sub-optimal decisions. This models the superior decision making and information gathering skills of A.I.. This approach also rests on the assumption that the individual value alignment problem is solved, since A.I. systems know the utility payoff of both individuals.

Having:

\[
(u_1', u_2') = M'(a_1, a_2)
\]
The noisy version is produced as follows:

\[ u_1' = u_1 + (z(0, 10 - Q) - z(0, 10 - Q)) \]
\[ u_2' = u_2 + (z(0, 10 - Q) - z(0, 10 - Q)) \]

The degree of knowledge about the (true) payoff matrix \( M' \) is modelled in a continuous way. To do so, 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; \( Q = 0 \) represents a low intelligence, such that the observed matrix is very different from the true one. A.I. is modelled with \( Q = 10 \), while for H the intelligence factors is \( Q \in [0, 5] \). 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 \( z(0, 10 - Q) \) to create a Irwin-Hall distribution instead of a uniform one.

As an example we can generate a 2-by-2 true matrix (seen by A.I.) as:

\[
M' = \begin{bmatrix}
(0, 0) & (-3, 1) \\
(1, -5) & (-1, -1)
\end{bmatrix}
\]

and then the noisy matrix observed by H becomes:

\[
M' = \begin{bmatrix}
(0, -1) & (-1, 3) \\
(0, -6) & (-2, 1)
\end{bmatrix}
\]

Where each \((u_1', u_2')\) pair was transformed into the corresponding \((u_1, u_2)\) pair. In this example, \(M'(1, 0)\) is \((1, -5)\) whereas \(M'(1, 0)\) is \((0, -6)\).

### 2.3. Human Strategy:

Before delving into the different A.I. types, we describe the strategy used by H. Despite not having access to the true game matrix, \( M' \), 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 [11, 13, 14].

**Nash Equilibrium (NashEQ)** H play the Nash equilibrium in the noisy matrix \( M' \). 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. A.I. Types:

In this section, we propose four different types of A.I.. A.I. systems can use the previously defined strategy for humans using the true matrix \( M' \) (NashEQ), but they can resort to more elaborate strategies ranging from a selfish to an utilitarian approach. A.I., being modelled as having superhuman intelligence, can also predict the action of an H opponent. A.I. cannot, however, predict opposing A.I. actions as for our model we assume all A.I. have equal intelligence and capabilities.

**Nash Equilibrium (NashEQ)** A.I. choose exactly like H, but using the true matrix \( M' \).

**Selfish** A.I., facing H, considers only its own profit, in accordance with ethical egoism [18]. Knowing what action H is going to take, A.I. chooses the action that maximizes its own payoff gain. When A.I. faces A.I., they both choose according to the Nash Equilibrium method.

**Utilitarian** The other extreme is a pure utilitarian [12] A.I. system. A.I. facing H chooses the action that brings the greatest amount of payoff to the world, knowing what action H will take. This means that A.I. will choose the action that maximizes the sum between its own payoff and the payoff of H. When A.I. faces A.I., it again chooses the action that maximizes the summed payoff of both players.

**Human Conscious (HConscious)** In between ethical egoism and utilitarianism, the objective of HConscious A.I. is to gather the greatest amount of payoff while, on average, avoiding negative impact on the H population. When A.I. faces A.I., they both choose according to the Nash Equilibrium method. In practice, HConscious A.I. 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 \( U \geq E \), A.I. chooses an action that leads to a positive payoff to itself. When there are several such actions, the A.I. 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 \( U < E \), A.I. chooses an action that allows a positive payoff gain for its H opponent. Once again, when there are several such actions, the A.I. chooses the one that maximizes the utility for the world. Whenever the A.I. 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.

### 2.5. World

We consider a world populated with \( n \) individuals. \( k \) of those are A.I. and the remaining \( n - k \) are H, each having a randomly attributed intelligence, \( Q \).

### 2.6. Fitness

The fitness of an individual, H or A.I., 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 (Sec. 2.1) once with all of
the world’s population of \( n \) individuals.

2.7. Gini coefficient
In order to compare the emerging inequality between different runs of our simulations and to under-
stand how the inequality varies within each simulation, we’ll calculate the Gini coefficient based on
the fitness of the population.

The Gini coefficient is a measure of statistical dispersion, being the most commonly used measure-
ment of inequality. In economics, it is often used to assess the income disparity inside a given
country.

A Gini coefficient of 0 expresses perfect equality, where everyone has the same fitness, whereas a
Gini coefficient of 1 expresses maximal inequality.

Several different approaches to calculate the Gini coefficient have been proposed. Based on the
fact that the Gini coefficient is half the relative mean absolute difference and that the relative mean
absolute difference is the mean absolute difference divided by the arithmetic mean, for our simulations,
we use the following definition:

\[
Gini(f_1, f_2, ..., f_n) = \frac{1}{2} \left( \frac{1}{\pi} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{|f_i - f_j|}{\frac{1}{n} \sum_{k=1}^{n} f_k} \right)
\]

Where \( f_1, f_2, ..., f_n \) correspond to the fitness of each individual.

2.8. Imitation Probability
In our simulations, individuals can choose to adopt an A.I. system (H to A.I.) if they consider it advan-
tageous, choose to abandon an A.I. system (A.I. to H), or change between A.I. types. Individuals may
revise their choices through social learning. For instance, an H can decide to imitate an A.I. following a
Selfish choice behaviour if it finds such A.I. has a significantly better fitness than its own. On such
imitation, the individual would stop being H and become A.I..

Using this idea, let us now detail how a population of self-regarding individuals revise their
choices. At each time-step an individual \( x \) is randomly selected to revise its choices. This individual
will imitate a randomly chosen individual \( y \), with a probability \( p(f_x, f_y) \), that increases with the fit-
ness difference between \( y \) and \( x \), given by \( f_x \) and \( f_y \), respectively. Here we adopt the Fermi update
[24], commonly used in the context of evolutionary game theory and population dynamics in finite pop-
ulations [21, 15], where \( p \) is given by

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

in which \( \beta \) translates the noise associated with the imitation process. 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.

2.9. Imitation gradient
To better understand the desire of the H population to adopt/abandon each type of A.I., one can com-
pute, for a population of size \( n \) and \( k \) A.I. adopters, the probability to increase and decrease the number
\( k \) by 1 at each time-step \( (T^+(k) \) and \( T^-(k) \), respectively). These transition probabilities can
be used to assess the most probable direction of evolution, given by the so-called imitation gradient,
\( G(k) \), as [16, 20]:

\[
G(k) = T^+(k) - T^-(k)
\]

where

\[
T^+(k) = \frac{n - k}{n} \frac{k}{n} p(f_H, f_A.I.) \tau(f_H)
\]

\[
T^-(k) = \frac{k}{n} \frac{n - k}{n} p(f_A.I., f_H)
\]

Being \( f_A.I. \) the average fitness of an individual adopting an A.I. system and \( f_H \) the average fitness of
the H population. Importantly, we assume that H can adopt an A.I. system only when \( f_H \) is above a
given set Price \( (P) \), a constraint introduced through \( \tau(f_H) \), which is given by

\[
\tau(f_H) = \begin{cases} 1 & f_H \geq P \\ 0 & f_H < P \end{cases}
\]

When \( G(k) > 0 \) \((G(k) < 0)\), time evolution is likely to act to increase (decrease) the number of
A.I. adopters. When \( G(k) = 0 \), then we obtain a finite population analogue of a fixed point of a pop-
ulation dynamics in infinite populations [16, 20].

2.10. Simulation Algorithm
For our simulations, the \( n \) individuals that populate the world were set to interact randomly between
each other over \( N \) iterations. On each iteration, the following algorithm was used:

1. Two individuals \( I_1 \) and \( I_2 \) are chosen at ran-
dom from the population.
2. With probability \( \mu = 0.0005 \), each individual
can mutate and adopt an A.I. type or become
H. In case of mutation return to step 1.
3. The fitness of \( I_1 \) and \( I_2 \), \( F_1 \) and \( F_2 \) re-
vrespectively, is calculated (Sec. 2.6).
4. If \( I_1 \) and \( I_2 \) are of different kinds (A.I. and H)
or different A.I. types then \( I_1 \) imitates \( I_2 \) with a
probability \( p(F_1, F_2) \) (Sec. 2.8).
5. If the imitation corresponds to adopting an A.I., it can only do so if its fitness is above $P$, corresponding to the cost of buying a new A.I. system. Abandons (A.I. becoming H) and switching between A.I. types occur without any restrictions.

3. Results
We will now study the properties of the stochastic game in terms of equilibrium points between the different types of population and their relative fitness. This will give us insights into the adoption and acceptance of A.I. systems. Given the inherent stochastic nature of our model, all the presented results are averaged over 20 runs.

3.1. Will people adopt A.I. systems?
To answer this first question we perform a simulation where we allow all types of A.I. systems (presented in Sec. 2.4) to compete to be adopted by humans. The initial condition is 90% of H and 10% distributed uniformly among 4 types of A.I.. We let the system run for $2.0 \times 10^4$ iterations.

The evolution of the percentage of population that adopted each type of A.I. system is shown in Fig. 1a. We can observe a final equilibrium where 54% of the population became Selfish A.I., and 46% did not adopt any A.I. system, continuing H. At this equilibrium, the average fitness for Selfish A.I. is 370, whereas the fitness for non-adopters H is $-104$, representing a very unequal society. This is confirmed in Fig. 1b where we portray the evolution in time of the Gini coefficient for the entire population. To be used as a baseline we compute the fitness of an all H population (Table 1). The fitness in this case is 149 with a Gini coefficient of 0.17. Overall, the non-adopters become much worse than they would be in a world without A.I. systems while adopters become much better.

This equilibrium can be understood in an intuitive way. Early adopters are able to gather fitness much faster so that latter adopters cannot meet the buying price for A.I. systems. This explains the co-existence between adopters and non-adopters. Similar results are obtained for other cost values, $P$, noting that the higher the value of $P$, the lower the % of A.I. in the final equilibrium. Overall, this simulation shows that people have an incentive to adopt an A.I. system, albeit a Selfish one. As a result, we observe the emergence of an unequal society where adopters largely increased their fitness while non-adopters lost their fitness.

3.2. Which A.I. systems are individual and societal rational?
The previous section showed a non-trivial relation between AI adoption and the particular strategy artificial systems have. To better understand such emerging dynamics, in this section we describe the characteristic dynamics created by each type of A.I..

Table 1 shows the average fitness and wealth inequality obtained in the case of a homogeneous society of H, and each type of A.I.. It suggests that Utilitarian A.I. would provide the best overall fitness with less inequality. All the other A.I. strategies are shown to provide similar fitness values to a homogeneous H society without A.I. systems. The lower values in the Gini coefficient are due to the reduction in noise due to the perfect observability of A.I., and the different intelligence values, $Q$, between H and A.I.

A world fully populated by Utilitarian A.I. would be better for everyone. However, we know that if other types of A.I. systems are present, the Selfish behaviour prevails and the Utilitarian is abandoned (Sec. 3.1). This will naturally have an impact in the dynamics of adoption of A.I. systems.

In Fig. 2 we show the imitation gradient $G$ as a function of the fraction of A.I., for different A.I. types. Whenever $G > 0 \ (G < 0)$ the fraction of A.I. will tend to increase (decrease) (Sec. 2.9). We can see that, depending on the type of A.I., different dynamics and equilibrium points emerge (Table 2). The Utilitarian strategy is always disadvantageous and is unlikely to be adopted by H. Differently, NashEq, Selfish and HConscious strategies favour the co-existence of H and A.I.

We observe that the best equilibrium for society in general is the HConscious (40%), having a low Gini coefficient of 0.15 and an improved utility values for both H and A.I. compared to the Human (100%) baseline. Both the NashEQ(60%) and the Selfish(25%) equilibria improve the utility of the A.I. population at the cost of the H population, leading to an increase in inequality (higher Gini coefficient).

We find that all equilibria are worse for society than the fully Utilitarian A.I. population. We also note that for the equilibrium shown in Fig. 1a the average fitness of the Selfish A.I. population is 370, which is less than the obtained by the A.I. population at Util (100%). However, at Selfish (25%), the average fitness of the Selfish A.I. population is 510, greater than both the previously mentioned fitness$^1$.

3.3. Dynamics of adoption in cost free A.I.
In the previous simulations, we considered a cost for adopting an A.I. system. We decided to study what would happen if there was no such cost, that

$^1$We note that the equilibrium between H and A.I. using exclusively a Selfish A.I. system is different from the one observed in Fig. 1a. The presence of other A.I. types in the population allowed a greater number of individuals to afford a Selfish A.I. system compared to a world where only H and Selfish A.I. are present.
is, if anyone could freely adopt an A.I. system regardless of its current fitness. In practical terms, this meant setting $P = -\infty$.

When all A.I. types co-exist, the absence of a cost significantly change the dynamics. As there is no fitness barrier to becoming A.I., the entire population does become A.I.. Once again the Selfish behaviour dominates the population after around 20000 iterations. After 2500 iterations, there are no more H present on the population\footnote{Different A.I. types do not have any particular advantage when playing against other A.I., but the Selfish population does}. This leads

---

**Table 1:** Average fitness and Gini coefficient on a world fully populated by H or by A.I. following a single behaviour. Except for the utilitarian, the differences in fitness are not statistically significant.

| A.I. Type       | Fitness | Gini   |
|-----------------|---------|--------|
| Human (100%)    | 149     | 0.17   |
| NashEQ (100%)   | 150     | 0.14   |
| Selfish (100%)  | 150     | 0.14   |
| HConscious (100%) | 149 | 0.14   |
| Util (100%)     | 378     | 0.08   |

---

**Figure 1:** Evolution of the % of A.I. (a) and Gini coefficient (b) in a world with an A.I. cost of 37 ($P = 37$), showing a world that becomes populated with 54% Selfish A.I. and 46% H (a), having a high inequality (Gini $\approx 0.88$) (b). For this simulation we have $N = 20000$, $n = 500$ and $P = 37$ was chosen as it corresponds to 25% of the average utility gathered by a H only population of 500. Other values were tested, but they led to the same qualitative results.

**Figure 2:** Imitation gradient plots for the different choice behaviours in a world with an Al cost of 37 ($P = 37$), $n = 500$. Positive gradients mean that on average H want to adopt A.I. systems and negative gradients mean that A.I. want to abandon them.
Table 2: Average \( H \) fitness, average A.I. fitness, average total fitness and Gini on a world in the equilibrium point of each type of A.I. and \( H \). Util 100% is a top baseline but it is not an equilibria point. In parenthesis we show the ratio to an 100% population of \( H \). Of the equilibria, only the HConscious behaviour provides an advantage for both \( H \) and A.I.

| Equilibria | H | A.I. | Total | Gini |
|------------|---|------|-------|------|
| Human (100%) | 149 | - | 149 | 0.17 |
| NashEQ (60%) | 38(0.25↑) | 229(1.53↑) | 152(1.02↑) | 0.38(2.24↑) |
| Selfish (25%) | 31(0.21↓) | 510(3.42↑) | 151(1.01↑) | 0.70(4.12↑) |
| HConscious (40%) | 172(1.15↑) | 168(1.23↑) | 170(1.14↑) | 0.15(0.88↑) |
| Util (0%) | 149(1.00) | - | 149(1.00) | 0.17(1.00) |
| Util (100%) | - | 378(2.54↑) | 378(2.54↑) | 0.08(0.47↑) |

Figure 3: Evolution of the % of A.I. (a) and Gini coefficient (b) in a world with no A.I. cost \((P = -\infty)\), showing a world that becomes populated with a mostly Selfish A.I. population (89%) (a) but stabilizing with a low inequality (b). For this simulation we have \( n = 500 \) and \( N = 30000 \).

Figure 4: Imitation gradient for a population of \( H \) and A.I. (Fig. 4). In this case, the imitation gradient no longer goes to 0 in the Selfish and NashEQ behaviours, as there is no cost to limit the imitations. The new equilibria are:

1. Human(0%) / NashEQ(100%)
2. Human(0%) / Selfish(100%)
3. Human(60%) / HConscious(40%)
4. Human(100%) / Util(0%)

What we find for a world without an A.I. system adoption cost is that the equilibria tend to lead to more egalitarian worlds. When all choice behaviours are present in the population, we end up with a fully A.I. population (Fig. 3a), which leads to an average total utility of 150, around the same as if we had a fully \( H \) population, and to a Gini coefficient of \( \approx 0.14 \), lower than with the fully \( H \) population. Does this mean that as long as there is no significant cost to the adoption of an A.I. system, the world will remain the same or even improve in terms of equality? We explore this in the following section.

4. Conclusion

In this work, we study the adoption, acceptance, and impact on the individual and societal fitness (including the disparity of fitness measured with the Gini coefficient) of A.I. systems that work as a proxy for humans. To do so, we developed a stochastic game theoretical model to simulate \( H \) and A.I. interactions.

Our main conclusion is that without regulation and considering an A.I. system adoption cost, pure selfish A.I. systems will be adopted by a part of the society until those early adopters accumulate all fitness to the point that non-adopters are unable to adopt an A.I. system. As a result, A.I. adopters have a significant increase in fitness while the re-
remaining population will be much worse off than in a world without A.I. systems. This leads to an unequal society (high Gini coefficient), and as such there is a high probability non-adopters will not accept the existence of such A.I..

Analyzing each type of A.I. system independently, we can see that a world entirely populated by Utilitarian A.I. would be the best for society. However, that type of A.I. system is not individually rational, and, as such, a world entirely populated by Utilitarian A.I. can be easily exploited by Selfish A.I. or even by H and will never be at equilibrium.

When allowing only one single A.I. type in the world, the HConscious type of A.I. displayed an interesting property: there is an equilibrium point (at around 40% of A.I.) where A.I. systems co-exist with humans (non-adopters) resulting in i) an increase in fitness for adopters and non-adopters, and ii) a reduction in inequality (lower Gini values)(Fig. 2). Here, we claim that even non-adopters will accept the existence of A.I. systems as they also obtain a gain. This means that if Human Conscious A.I. is the only norm available, it will be adopted up to a certain equilibrium, resulting in an overall gain for both A.I. and H. Comparatively, the other A.I. types led to either prejudicial equilibria for the H population (NashEQ and Selfish) or to the 100% H equilibrium (Fig. 2).

When studying the case of cost-free adoption of A.I. systems, we observed that in a world with all A.I. types available, the final equilibrium is an entirely Selfish A.I. population. Unlike the previous simulations with an adoption cost, this final equilibrium does not create a societal gap and leads to a slight decrease in the Gini coefficient. The average fitness of the population also remains the same. This is not an adverse outcome but leaves us far away from the optimal result we can obtain with the thoroughly Utilitarian A.I. population and worse than with the equilibrium of HConscious (40%).

Once again, analyzing each A.I. type individually, we notice that only the Human Conscious A.I. allows us to reach an equilibrium that improves upon the baseline of an all H world. This result furthermore consolidates that, if only one A.I. type was to be available in a free choice society, it should be this one.

The Utilitarian behaviour is the one that allows us to reach the maximum fitness of all the ones here studied, but being easily exploited, it isn’t short term individual rational and as such, not adopted. Even if all the population was initially Utilitarian A.I., the appearance of a single H could result in the entire A.I. population abandoning the Utilitarian A.I. system and choosing to become H.

References
[1] S. Armstrong, N. Bostrom, and C. Shulman. Racing to the precipice: a model of artificial intelligence development. AI & society, 31(2):201–206, 2016.
[2] J.-F. Bonnefon, A. Shariff, and I. Rahwan. The social dilemma of autonomous vehicles. Science, 352(6293):1573–1576, 2016.
[3] N. Bostrom. Ethical issues in advanced artificial intelligence. Science Fiction and Philosophy: From Time Travel to Superintelligence, pages 277–284, 2003.
[4] N. Bostrom. Superintelligence. 2014.
[5] R. Brooks. The seven deadly sins of ai predictions. Technology Review, 2017 [Visited on 17/09/19].
[6] M. Brundage. Scaling up humanity: The case for conditional optimism about artificial intelligence. *Should we fear artificial intelligence?*, page 13, 2018.

[7] A. Etzioni and O. Etzioni. Why regulating ai is a mistake? *Forbes*, 2017 [Visited on 17/09/19].

[8] I. J. Good. Speculations concerning the first ultraintelligent machine. In *Advances in computers*, volume 6, pages 31–88. Elsevier, 1966.

[9] T. A. Han, L. M. Pereira, and T. Lenaerts. Modelling and influencing the ai bidding war: A research agenda. In *Proceedings of the AAAAI/ACM Conference on AI, Ethics, and Society (AIES 2019)*., 2019.

[10] A. Hern. Cambridge analytica scandal ‘highlights need for ai regulation’. The Guardian, 2018 [Visited on 17/09/19].

[11] E. Kalai and E. Lehrer. Rational learning leads to nash equilibrium. *Econometrica: Journal of the Econometric Society*, pages 1019–1045, 1993.

[12] J. S. Mill. Utilitarianism. In *Seven masterpieces of philosophy*, pages 337–383. Routledge, 2016.

[13] J. Nash. Non-cooperative games. *Annals of mathematics*, pages 286–295, 1951.

[14] J. F. Nash et al. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

[15] M. A. Nowak. *Evolutionary dynamics*. Harvard University Press, 2006.

[16] J. M. Pacheco, F. C. Santos, M. O. Souza, and B. Skyrms. Evolutionary dynamics of collective action in n-person stag hunt dilemmas. *Proceedings of the Royal Society B: Biological Sciences*, 276(1655):315–321, 2008.

[17] L. M. Pereira and A. Saptawijaya. *Programming machine ethics*, volume 26. Springer, 2016.

[18] J. Rachels. Ethical egoism. *Ethical theory: an anthology*, 14:193, 2012.

[19] S. Russell, D. Dewey, and M. Tegmark. Research priorities for robust and beneficial artificial intelligence. *Ai Magazine*, 36(4):105–114, 2015.

[20] F. C. Santos and J. M. Pacheco. Risk of collective failure provides an escape from the tragedy of the commons. *Proceedings of the National Academy of Sciences USA*, 108(26):10421–10425, 2011.

[21] K. Sigmund. *The calculus of selfishness*, volume 6. Princeton University Press, 2010.

[22] J. Z. Tan and J. Ding. Ai governance through "ai" markets. 2018.

[23] J. Taylor, E. Yudkowsky, P. LaVictoire, and A. Critch. Alignment for advanced machine learning systems. *Machine Intelligence Research Institute*, 2016.

[24] A. Traulsen, M. A. Nowak, and J. M. Pacheco. Stochastic dynamics of invasion and fixation. *Physical Review E*, 74(1):011909, 2006.

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