FOUNDATIONS OF COMPUTING AND DECISION SCIENCES
Vol. 45 (2020) No. 1
ISSN 0867-6356
DOI: 10.2478/fcds-2020-0002 e-ISSN 2300-3405

Morality, protection, security and gain: lessons from a minimalistic, economically inspired multi-agent model

Maciej Komosinski Tomasz Żok [*]

Abstract. In this work, we introduce a simple multi-agent simulation model with two roles of agents that correspond to moral and immoral attitudes. The model is given explicitly by a set of mathematical equations with continuous variables and is characterized by four parameters: morality, protection, and two efficiency parameters. Agents are free to adjust their roles to maximize individual gains. The model is analyzed theoretically to find conditions for its stability, i.e., the fractions of agents of both roles that lead to an equilibrium in their gains. A multi-agent simulation is also developed to verify the dynamics of the model for all values of morality and protection parameters, and to identify potential discrepancies with the theoretical analysis.

Keywords: morality, social norm, equilibrium, mathematical modeling, simulation

1. Introduction

The story of humans trying to address problems and questions related to morality and ethics is a long one. From Aristotle to Kant, the endeavor continued to formulate definitions of morality, identify moral behavior, distinguish right from wrong, decide what criteria should be used to evaluate one's actions, and establish one's responsibility for such actions. Discoveries in evolutionary biology motivated new kinds of questions, such as whether altruism could emerge spontaneously and what could be the benefits of moral behavior in terms of the chances of survival and the access to resources in the environment [3, 29, 10]. Finally, psychological and sociological experiments allowed to perform quantitative and qualitative measurements of human actions, and perform statistical analyses of various factors to estimate the degree to which they influence human actions that have a moral dimension [7, 8].

[*] Poznan University of Technology, Institute of Computing Science, Piotrowo 2, 60-965 Poznan, Poland. Email: {maciej.komosinski, tomasz.zok}@cs.put.poznan.pl.

Recently, investigations regarding morality gained additional motivation due to the rise of artificial intelligence and the increase in the autonomy of artificial agents. There is a large body of research discussing what properties should artificial agents possess in order to behave ethically and be aligned with human values [1, 11, 26, 24, 2, 27, 21, 13, 9, 6]. Apart from philosophical considerations, interest grew in modeling moral behaviors [14, 12, 18, 20] and applying formal models to describe actions of abstract agents [23] using mathematical equations or formal logic. Such models often involve a population of agents, described either as a set of discrete, individual entities, or as a set of numbers that only capture global properties or statistics of such agents (their total number, average income, etc.). To model the dynamics of populations of agents, tools used traditionally in biology, in the analysis of discrete dynamical systems and their evolutionary stability – such as differential equations and game theory – are particularly appropriate [16].

Such formal models and descriptions of agents and their behavior facilitate the simulation of working systems that implement these models, so instead of static, off-line, numerical analysis of descriptive models, agents take some form of virtual existence and perform interactions in simulated environments. This allows for the development of agent-based social simulations (ABSSs) that help investigate social dilemmas and experiment with potential scenarios. Such applications let researchers analyze the behavior of the complex system as a whole [19, 4, 5, 22, 12, 15, 28, 25], including its evolution, stability, and even the inheritance and mutation of agents' traits. In order to study such phenomena, some authors [1, 5, 26, 27] mention artificial life [17] as the appropriate or relevant approach.

Motivated by the advantages that multi-agent simulations can offer and by the scarcity of such models that concern morality – even though quantitative, empirical studies concerning the morality of human behavior exist – in this work, we follow this line of research and create a minimalistic, economically inspired multi-agent model that exhibits non-trivial behavior.

- The minimalism of the model is due to the fact that there are only two roles of agents ("workers" and "thieves") that reflect moral and immoral attitudes, respectively. There are four parameters that will be denoted with capital letters throughout this article. Two of them have a sociological interpretation that influences the attractiveness of each role ($MORALITY$ and $PROTECTION$), and two other parameters describe the simulated reality ($EFFICIENCY\_OF\_WORK$ and $EFFICIENCY\_OF\_THEFT$). Despite the fact that it is very easy to model complex, diverse phenomena using simulation and multi-agent systems, we keep the model as simple as possible and refrain from introducing additional agent roles and parameters in this initial formulation.

- The economic aspect of the model is that all gains in the model are generated by "workers" only and that the total income is dependent on their number. This is in contrast to popular game-theoretic models where the payoff matrix can be arbitrarily adjusted and is usually a modification or extension of the standard iterated prisoner's dilemma.

- The non-trivial behavior of the model results from the fact that there are interesting tradeoffs in agent behavior dependent on the values of parameters and on the proportion of agent roles, and that the model is in part non-linear.

Given these properties of the model, we want to explore how, precisely, the parameters influence the choice of roles of agents, what is the nature of all edge cases, and how stable they are. This is possible because the simplicity of the proposed model allows for its extensive theoretical analysis and makes it easy to interpret. We verify the behavior of the model by implementing a multi-agent simulation, performing a series of computational experiments, and comparing their outcomes with the theoretical analyses. The detailed description of the model follows below.

1.1. The model and its interpretation

Consider a set of agents (a population of individuals) where each agent can choose and take one of the two roles: either to gain resources by their own work, or by stealing from other agents. To capture these characteristics of the two behaviors, we will call the two subsets of agents *workers* and *thieves*. While these particular names are used to appeal to human intuition, other names could also be used with a similar interpretation, such as organisms and parasites, the honest and the fraudsters, etc.

In every moment, each agent can change their role. We assume that the choice of roles is rational and is based solely on the estimated gains for each role. Both estimated gains depend on the current state of the population (i.e., the number of agents in each subset) and on the four global parameters of the model: the efficiency of work, the efficiency of theft, morality, and protection.

Since the set of all agents consists of two subsets, let us denote by $workers$ and $thieves$ fractions of each role in the set, so

$$workers + thieves = 1$$

The gain of a worker is calculated in the following way:

- The global parameter $PROTECTION \in [0, 1]$ determines how much of their gain each worker allocates to safeguard against attacks of thieves.
- The global parameter $EFFICIENCY\_OF\_WORK \in [0, 1]$ determines how efficient is the work – i.e., what fraction of the true worth of work is obtained by the worker.
- As a result, the potential gain of each worker is

$$potentialgainworker = (1 - PROTECTION) \cdot EFFICIENCY\_OF\_WORK$$

- A part of this potential gain is stolen by thieves; the stolen amount depends on the global $EFFICIENCY\_OF\_THEFT$ parameter, and it is inversely proportional to the protection expenditure made by each worker:

$$tosteal = potentialgainworker \cdot (1 - PROTECTION) \cdot EFFICIENCY\_OF\_THEFT$$

Note that this means that *tosteal* is proportional to $(1 - PROTECTION)^2$. This is because in the model we assume that the amount used earlier to protect worker's gain, or some part of this amount (e.g. a reinforced door), must be destroyed by a thief or requires some investment or effort to break, or additional risk to be taken – so this part of worker's gain decreases the effective value of stolen goods.

- The final gain of a worker is therefore

$$gainworker = potentialgainworker - tosteal \tag{1}$$

and if there are no thieves in the set of agents, then $gainworker = potentialgainworker$.

For thieves, their gain is calculated as follows:

- Since we don't differentiate between individual agents within a given role, all thieves get the same fraction of gains that can be stolen from workers (note that in this model, thieves do not rob other thieves), so each thief expects

$$gainthief = \frac{workers \cdot tosteal}{thieves} \tag{2}$$

If there are no thieves in the population, *gainthief* is undefined.

- The global $MORALITY \in [0, 1]$ parameter determines the decrease in attractiveness (value) of gains that were stolen. The gain for a thief that takes into account the *MORALITY* parameter is

$$potentialgainthief = \frac{workers \cdot (tosteal - MORALITY)}{thieves + t_1} \tag{3}$$

where t_1 is the fraction of the population that corresponds to one agent (the agent that considers becoming a thief). In the discrete case where the total number of agents is known, $t_1 = \frac{1}{number\_of\_agents}$. In the continuous case, t_1 is an infinitely small value. When every agent in the population is a thief, calculating *potentialgainthief* is not necessary because there are no agents who are workers and could consider becoming a thief.

As mentioned earlier, each agent decides which role they want to assume in the population by comparing *gainworker* and *potentialgainthief*. Since agents (rationally) want to maximize their gains, they choose the role that yields the higher gain. When estimating the gain due to being a thief, an agent considers *potentialgainthief*, i.e., the gain of a thief reduced by *MORALITY*, but when an agent actually becomes a thief, they earn *gainthief*.

The total gain in the population is therefore $totalgain = workers \cdot gainworker + thieves \cdot gainthief$ as long as $thieves > 0$, otherwise there are no thieves and $totalgain = workers \cdot potentialgainworker = potentialgainworker$. The average gain of an agent is *totalgain* divided by the total number of agents in the population, but since the numbers of agent roles are normalized, the average (weighted by the fraction of both roles) gain of an agent in the population has the same value as the total gain of the population: $averagegain = totalgain$.

One alternative interpretation of agents (instead of workers and thieves) would be robots that can either gather energy ("energy harvesters" instead of "workers") or convert energy (transform it into some other form, store it, or sell – let's call them "transformers" instead of "thieves"). In this interpretation, gains can be measured in energy units instead of representing a monetary value. If there is excessive demand for energy from transformers, some agents change their role and become energy harvesters instead of converting energy. If there is too much energy available, harvesters change role and become transformers to convert or store energy. In this alternative interpretation, *EFFICIENCY_OF_WORK* would reflect how efficient are energy harvesters, *EFFICIENCY_OF_THEFT* would be how efficient is energy transformation, *MORALITY* would be the threshold of profitability that would convince a harvester to become a transformer, and *PROTECTION* would be the fraction of energy used by energy harvesters and energy transformers for their own functioning. With appropriately adjusted parameter values, such a system would tend to self-stabilize, gather and convert energy without the need for manual control of the roles of these robots.

2. Dynamics of the model: the analytical approach

The primary question regarding the model is how the state of the population depends on the four global parameters: the efficiency of work, the efficiency of theft, morality, and protection, assuming that their values are known and constant. Let us determine stable states, i.e., states where agents will not want to change their roles. There are two kinds of stable states: *uniform* when all agents take on the same role, and *mixed* when both roles coexist in the population.

No role changes will occur in the population when no agent wants to change their role, i.e., gains for both roles will be equal:

$$gainworker = potentialgainthief$$

The left side of this equation depends, as (1) defines it, on the efficiency of work, the efficiency of theft, and protection. The right side depends on the same quantities, and additionally, on the *MORALITY* parameter and the proportion of thieves (3). Assuming that the values of the four global parameters are constant, both sides of the equation depend only on one variable – the proportion of thieves. Therefore, the requirement for the lack of role changes in the model can be formulated as the equilibrium equation

$$gainworker - potentialgainthief(thieves) = 0 \tag{4}$$

Solving this equation for $thieves \in [0, 1]$ shows that there are four possible outcomes depending on all four global parameters:

A. There is a single solution.

B. There are infinitely many solutions.

C. There are no solutions, and the left side of (4) is always positive.

D. There are no solutions, and the left side of (4) is always negative.

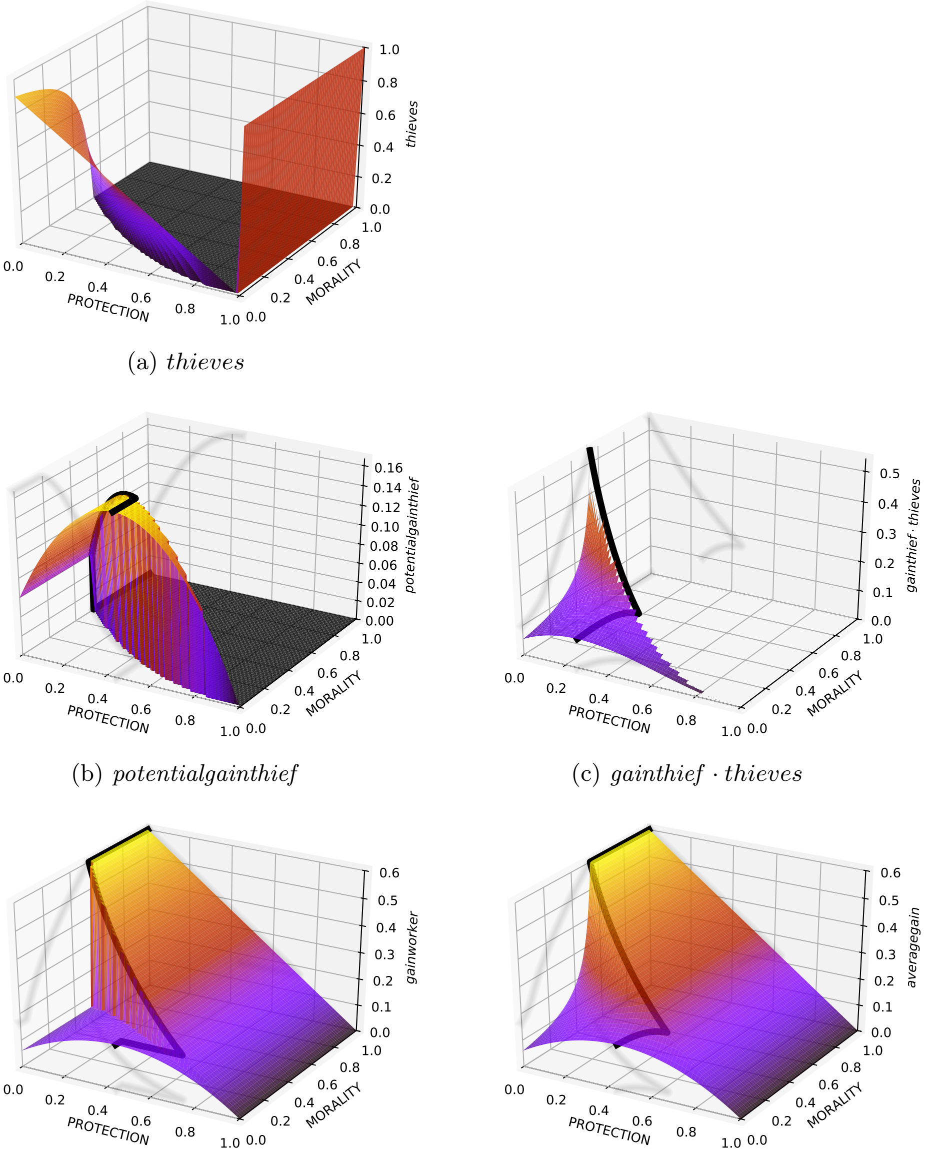
In the case of A, the solution will describe a mixed equilibrium state. In B, no role will be preferred in the population independently from the fraction of workers and thieves. This occurs in edge cases when $MORALITY = 0$ and either $EFFICIENCY\_OF\_WORK = 0$ or $PROTECTION = 1$. This combination of parameter values results in $gainworker = 0$ and $potentialgainthief = 0$, so agents do not change their initial roles. For cases C and D, the population will reach a uniform stable state, with $thieves = 0$ for C and $thieves = 1$ for D. However, a careful analysis revealed that in order for D to be true, $MORALITY$ should be negative, which is outside of the valid range for this parameter. With $MORALITY \geq 0$, the only other way for D to occur is to have a system with thieves gaining even when they constitute the entire population. This is against the definition of our model, where thieves can only gain by stealing from workers. On the other hand, a uniform stable state for C (no thieves) is possible – in particular, for specific values of $PROTECTION$ and $MORALITY$, as discussed later.

Let us now discuss the influence of the two efficiency parameters (work and theft) on the behavior of agents in the population. If we assume $EFFICIENCY\_OF\_WORK = 0$, then the gains of both workers and thieves are zero. If $EFFICIENCY\_OF\_THEFT = 0$, then, independently of the values of $PROTECTION$ and $MORALITY$, there will be no thieves in the population. We can, therefore, say that for the specific cases where any of the efficiencies of activities performed by the two roles are zero, the model becomes degenerate. The remaining cases where the efficiencies are positive lead to an interesting and non-trivial dynamics of the model. In the following analyses, we assumed $EFFICIENCY\_OF\_WORK = 0.6$ and $EFFICIENCY\_OF\_THEFT = 0.9$. Other values of these parameters yield different gains of both roles of agents and different stable states, but shapes of the functions presented below remain similar, and without losing generality, one can discuss and examine the consequences of this particular set of two efficiency values.

The equilibrium equation (4) depends only on the proportion of $thieves$. Similarly, $gainthief$ and the most interesting variable – $averagegain$ – depend only on $thieves$. Therefore, identifying a solution to the equilibrium equation (4) (which would determine a mixed stable state), or demonstrating the lack of such a solution (which would determine a uniform stable state) unambiguously describes all the properties of the model given constant global parameter values. This allows one to describe the dynamics of the system using plots shown in Fig. 1.

2.1. The discontinuity for maximal protection

The general discussion of the characteristics of the model will be provided in Sect. 4; below, we focus on specific boundary conditions that need to be examined in detail.



(a) *thieves*

(b) *potentialgainthief*

(c) *gainthief* · *thieves*

(d) *gainworker*

(e) *averagegain*

Figure 1: Analysis of the model and the dependence of its various quantities on $PROTECTION$ and $MORALITY$. The range of $[0, 1]$ for $PROTECTION$ and $MORALITY$ is divided uniformly into 63 intervals, so each plot presents $64 \cdot 64$ values. In Fig. 1c some data points are missing because *gainthief* is undefined when $thieves = 0$. The thick black lines in Figs. 1b-1e and their projections indicate the highest gain achievable for each $MORALITY$ value given the optimal $PROTECTION$.

Fig. 1a demonstrates a clear discontinuity for $PROTECTION = 1$. This is where the value of *thieves*, independently from $MORALITY$, reaches the global maximum of 1. This exceptional case causes discontinuity, because for $PROTECTION < 1$, the proportion of *thieves* is equal or close to zero. This difference is due to the fact that spending the entire gain of a worker on security ($PROTECTION = 1$) results in their lack of gain ($gainworker = 0$) and, as a consequence, the lack of gains that can be stolen ($tosteal = 0$). In this case, solving (4) requires finding the zero of the function *potentialgainthief* (3), which occurs if and only if (i) $workers = 0$ and $thieves = 1$, or (ii) $MORALITY = 0$.

The (i) case is a stable uniform state. Note that we earlier excluded states where $thieves = 1$, because thieves cannot gain anything when they are the only agents in the population. This current conclusion does not contradict previous considerations, because, in the currently considered specific situation, nobody gains at all – including workers – and this is what causes the discontinuity visible in Fig. 1a. The (ii) case with $PROTECTION = 1$ and $MORALITY = 0$ is a quasi-stable state described in the beginning of this section. In such a case, agents will not want to change their roles in the population, so the analytical result of $thieves = 1$ for this case that is shown in the plot is just one specific value from the infinite set of values that are all solutions to (4).

2.2. The role of morality

Solving the equilibrium equation (4) for *thieves* allows one to reveal interesting aspects of the model. It turns out that $thieves = 0$ if either

$$MORALITY > EFFICIENCY\_OF\_WORK \cdot EFFICIENCY\_OF\_THEFT \tag{5}$$

or

$$PROTECTION > 1 - \sqrt{\frac{MORALITY}{EFFICIENCY\_OF\_WORK \cdot EFFICIENCY\_OF\_THEFT}}. \tag{6}$$

Tracking the thick black line in Fig. 1d for $PROTECTION = 0$ confirms that it represents a range of $MORALITY \in (0.54, 1]$, as predicted by (5). The other inequality (6) can be used for the remaining values of $MORALITY$ to better understand the model dynamics. For example, $MORALITY = 0.2$ allows one to predict that there will be no thieves if only $PROTECTION \gtrsim 0.392$; this can be verified by looking at Fig. 1a and noticing that at $PROTECTION \approx 0.4$ and $MORALITY \approx 0.2$, the value of *thieves* starts to be positive.

These two inequalities demonstrate the unique role of the $MORALITY$ parameter. Given the values of $EFFICIENCY\_OF\_WORK$ and $EFFICIENCY\_OF\_THEFT$, one can predict whether a system can stabilize in a uniform state of $thieves = 0$ with zero expenses on protection, or – if not possible – what is the minimal value of $PROTECTION$ to eliminate all thieves.

Additionally, (6) clearly shows that thieves are inevitable only in a system with $MORALITY = 0$ (the only exception here is maximal $PROTECTION = 1$). Any positive value of $MORALITY$ guarantees that a system without thieves is possible given sufficient, non-extreme protection.

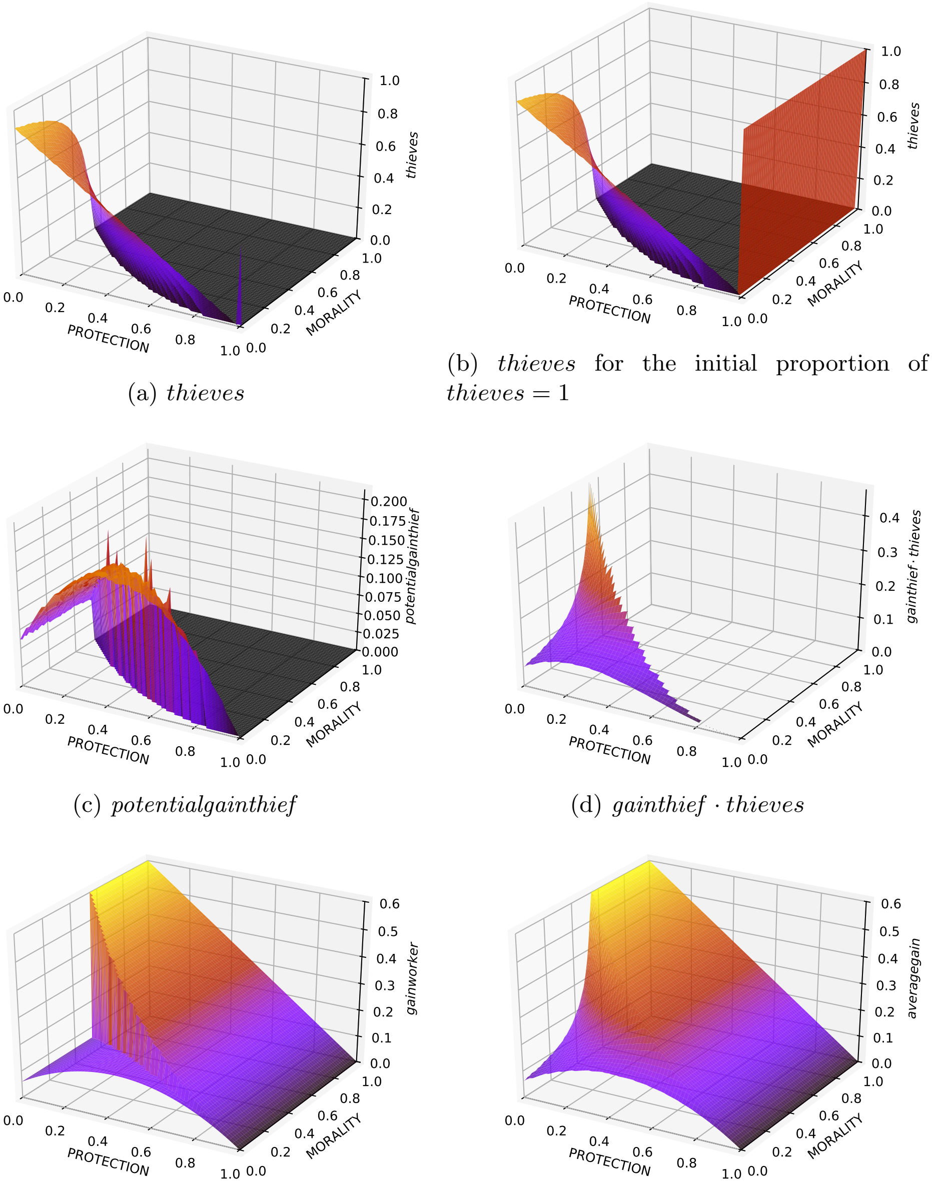
3. Dynamics of the model: multi-agent simulation

In order to verify the behavior of the model in a more realistic setup, a multi-agent computer simulation was implemented and performed. In a simulated two-dimensional world, there is a constant number of agents each located at some (x_i, y_i) coordinates, and each agent can freely change its state (being either a worker or a thief). Agents can change roles in random intervals of time. The expected length of these intervals is a global parameter and can be adjusted; the shorter the intervals, the faster the convergence of the population. Agent choices are rational – each agent wants to maximize their gains based on perfect and accurate information about gains of both roles available to each agent. Following the equations that describe the model, gains of both roles depend on the number of worker agents, the number of thief agents, and the four global parameters of the model. When some agent becomes a thief, they locate the nearest worker and move to that location in order to commit a theft. Apart from visualization, these spatial properties do not influence gains or decisions of agents, because the model assumes a homogeneous distribution of resources in the simulated world. Nevertheless, contrary to the analysis performed in the previous section, in this simulation we deal with individual, discrete agents, so the variables that describe the state of the system are no longer continuous.

The primary goal of the investigation of the multi-agent simulation is to identify the proportion of *thieves* such that the equilibrium condition (4) would hold. In the simulation, the number of *thieves* in the population is initially set to some value, and then it can change. The fundamental question now is whether the system will ultimately converge to the same state for *any* initial number of thieves, and if it will, whether this state will be consistent with the state that resulted from the theoretical analyses described in the previous section.

Since the four global parameters of the model and fractions of both agent roles in the population are expressed as real numbers that are within the range $[0, 1]$, and the definition of the model is mostly based on addition and multiplication, particular attention was paid to boundary conditions, which may cause some variables to become zero or to grow to positive or negative infinity.

The results of the multi-agent simulations are shown in Fig. 2. Compared to theoretical analyses from the previous section, there is only one clearly visible difference, which is the dependency of the proportion of *thieves* on $PROTECTION$ and $MORALITY$, but only in the specific case when $PROTECTION = 1$. Note however that the plots based on the multi-agent simulation visualize a completely different underlying process – here, the numbers that are visualized correspond to virtual entities, and the system needed some time to stabilize and converge to a stable state (possibly, with micro-fluctuations where one agent would oscillate between the two roles), while in the previous section, they were the numerical solutions to mathematical equations. To obtain data shown in Fig. 2, the simulation was repeated independently for



(a) *thieves*

(b) *thieves* for the initial proportion of $thieves = 1$

(c) *potentialgainthief*

(d) $gainthief \cdot thieves$

(e) *gainworker*

(f) *averagegain*

Figure 2: The dependence of various quantities of the model on *PROTECTION* and *MORALITY* resulting from the multi-agent simulation. In Fig. 2d some data points are missing because *gainthief* is undefined when $thieves = 0$.

$65 \cdot 65$ combinations of different values of *PROTECTION* and *MORALITY*, and each simulation was run until the proportion of agent roles in the simulated environment converged to a stable value, as Fig. 3 illustrates.

A more careful comparison of Figs. 1 and 2 reveals a subtle difference – plots based on numerical analysis are more smooth, while plots based on the results of the simulation have limited resolution in vertical axes because they are based on individual, indivisible agents. For plots presented in Fig. 2, the simulation consisted of 125 agents; the lower is the number of agents employed in the simulation, the more pronounced is the difference between both approaches.

We will now investigate the discrepancy in the proportion of *thieves* for *PROTECTION* $= 1$ in more detail, leaving the general discussion of the characteristics of the model to Sect. 4.

3.1. The discontinuity for maximal protection

The particular case where protection is maximal was discussed in detail in Sect. 2.1, where this situation was analyzed theoretically. Let us now compare earlier theoretical results with the outcomes of the multi-agent simulation. When workers spend everything on protection (*PROTECTION* $= 1$), in consequence they profit nothing (*workergain* $= 0$). However, the situation of potential thieves is not so straightforward. Workers might be indifferent and look at the prospect of becoming a thief as

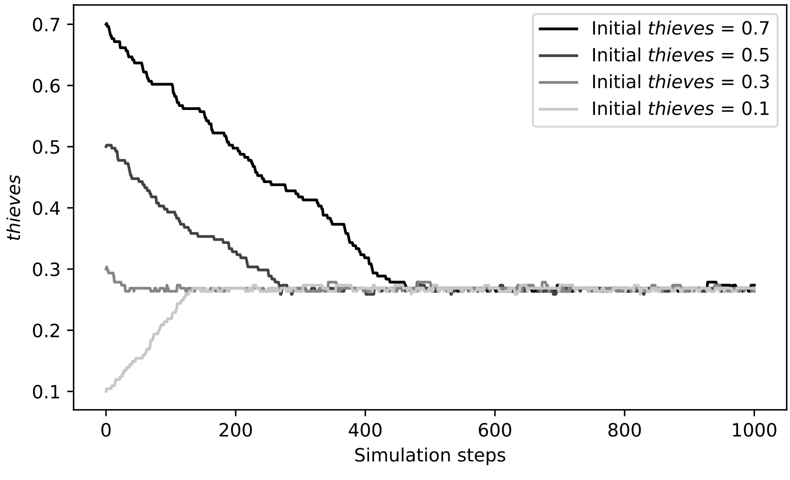


Figure 3: Dynamics of changes in the proportion of *thieves* in time for multi-agent simulation with 200 agents, depending on the initial proportion of thieves in the population, for *PROTECTION* $= 0.1$ and *MORALITY* $= 0.4$. Minor oscillations are due to a few agents changing their role in the same way in a single simulation step; this may happen since the expected gains of both roles are calculated once in every simulation step.

neither better nor worse than their current state. Alternatively, due to the morals guiding workers, they might see the thief profession as even worse than having zero income. These two cases are reflected by scenarios (I) $\mathit{potentialgainthief} = 0$ and (II) $\mathit{potentialgainthief} < 0$, respectively.

Scenario (I) occurs when either (Ia) $MORALITY = 0$ or (Ib) $thieves = 1$, and in this scenario (I) neither agent sees any benefit in changing their role. Therefore, the initial population of agents does not change over the course of the simulation. Case (Ia) occurs because agents cannot choose their role based on incentives, and they lack any moral guidelines to follow. An example of (Ia) is evident in Fig. 2a as a sudden peak of 0.5 (which was the initial fraction of *thieves* in the simulation) for $PROTECTION = 1$ and $MORALITY = 0$. On the other hand, (Ib) corresponds to a hypothetical system in which everyone is a thief. They rob nobody and have zero gain, but at the same time, they reject the idea of becoming workers. It is important to notice that the decision to become a thief requires morality sufficiently low, but the opposite decision – to become a worker – is not influenced by morality. Thieves become workers only if it clearly benefits them, and in a system where all gains are spent on protection, this is never the case. To demonstrate (Ib), a system with initial $thieves = 1$ was simulated and Fig. 2b clearly shows the final state with $thieves = 1$ for $PROTECTION = 1$ and any value of $MORALITY$.

Scenario (II) occurs when $MORALITY > 0$ and $thieves < 1$. This scenario corresponds to a population with at least a few workers who gain nothing. Nevertheless, the idea of becoming a thief is rejected because, for agents, it is even worse than having zero profits (as the potential gain of a thief is negative). In consequence, the final state of the system has $thieves = 0$. This scenario can be seen in Fig. 2a, where initially *thieves* was $0.5 < 1$ on the whole line where $PROTECTION = 1$ and $MORALITY > 0$.

This analysis of the multi-agent simulation shows that $PROTECTION = 1$ constitutes the only case where the final number of thieves in the system depends on the initial number of thieves. In particular, when initially $thieves = 1$, the results of the simulation are fully consistent with theoretical analyses, as Fig. 2b demonstrates.

4. Discussion and conclusions

Sects. 2 and 3 discussed the behavior of the model in detail, paying particular attention to boundary cases. While such corner cases are technically interesting and needed to be examined, in realistic scenarios the system will operate outside of its extremes – individuals will not spend their entire income on protecting against thieves, not everybody will be perfectly moral or immoral, the efficiency of work and theft will be neither 0% or 100%, etc. Let us investigate the most general and practical conclusions that the model reveals within its typical range of operation.

Ignoring the extremes of $PROTECTION = 1$, the most obvious observation from Figs. 1a, 2a and 2b (these plots present consistent characteristics) is that for some combinations of $PROTECTION$ and $MORALITY$, there are no thieves in the population. The role of a thief is unattractive either because the morality of each agent is too high to become a thief, or because the amount of goods that can be stolen (which is inversely proportional to the workers' expenditure for protection) is too low – or because of the non-linear interaction of the two discouraging factors.

When there are no thieves, the population consists only of workers, so the average gain of a worker and the average gain of the population, and also the total gain of the population (due to the normalization of the number of agents) are all the same, as demonstrated in Figs. 1d and 1e, and also 2e and 2f. When high morality of agents is sufficient to discourage them from becoming thieves, expenditures for protection are just losses – reflected in a linear decrease in *gainworker* and *averagegain*. The highest, optimal gain is achieved with no spending on protection needed thanks to sufficiently high morality, as the thick black line in Fig. 1e demonstrates.

When the morality drops and agents are tempted to become thieves, the highest gains (albeit lower than the optimum ones) can be achieved by progressively increasing expenditures for protection (see the middle section of the thick black dropping line in Fig. 1e). This allows the population to stay free from thieves at the cost of spending resources on protection.

Finally, when the morality further decreases and agents are not discouraged from becoming thieves, some of them actually become thieves (Figs. 1a, 2a and 2b) and gain by stealing from workers, because the prospect of becoming a thief yields higher gains than being a worker (compare Figs. 1b and 1d, and analogously Figs. 2c and 2e). This is when workers should start spending slightly less on protection (even though morality drops) – for workers, the cost of protecting against thieves is higher than the cost of being robbed. For extremely low morality, decreasing expenditures for protection ensures the highest gain of workers and also the highest gain of thieves, and therefore the highest gain of the entire population – see the lowest section of the thick black line in Fig. 1e. This is because for low morality, moderate protection is unable to discourage agents from becoming thieves, yet it decreases gains that can be achieved by both workers and thieves.

As mentioned in Sect. 2, the two efficiency parameters of the model (the efficiency of work and the efficiency of theft) influence the output parameters such as the number and the gains of workers and thieves, but they do it in a gradual way. The most interesting and complex behavior of the system results from the intermediate values of the two efficiency parameters. The extreme values (0.0 or 1.0) of these parameters simplify the model – for example, for maximal efficiency of work (1.0) and minimal efficiency of theft (0.0), no agent wants to become a thief independently of the values of $PROTECTION$ and $MORALITY$, and gains of workers (and therefore gains of the entire population) are inversely linearly dependent on $PROTECTION$ only.

Superrational agents would choose to be workers even for non-extreme values of the efficiency of work and the efficiency of theft. Independently from the level of $MORALITY$, this would allow such agents to entirely avoid spending on *PROTECTION*, thus ensuring maximal gains equal to $EFFICIENCY\_OF\_WORK$. In this regard, maximum morality in the model (or morality high enough to make stealing unattractive for specific efficiency parameter values and no protection) yields the same effects as superrational behavior.

5. Summary

This work introduced a minimalistic, economically inspired multi-agent model that featured two parameters with sociological interpretation, namely "morality" and "protection". We have explored how the parameters of the model influence the choice of roles of agents, what is the nature of all edge cases, and how stable they are. We have confirmed that the multi-agent simulation produced consistent results with the theoretical analysis of the model; the only difference was the case of maximal protection, where the results of the multi-agent simulation depended on the initial proportion of agent roles in the population. This behavior was consistent with the goals of the model, and in this particular regard, the discrepancy demonstrated the advantage of the simulation over the analytic, continuous approach. The simulation is closer to reality because it models individual agents and their behaviors, while the analysis deals only with aggregated or averaged characteristics of the population. The multi-agent simulation can also model temporal dynamics of the population along with instabilities or oscillations (illustrated in Fig. 3) that are due to the discrete nature of agents.

One of the conclusions from this particular model is that a highly moral population leads to the highest total and individual gains without the need of protection expenditures, and for less moral agents, some amount of protection is required to achieve optimal gains. Whether the conclusions should influence decisions made by humans depends on how accurately this model reflects reality; it was, however, constructed primarily with simplicity in mind. It was intended to be interpretable and easy to investigate analytically, including the ability to visualize the results for all combinations of protection and morality parameter values. Considering this as a starting point, we avoided the use of complex, non-linear functions, yet the model can be easily extended and enriched with additional functional dependencies and parameters. The values of the global parameters, the number of agent roles, their behaviors, and interactions can be further adjusted to be consistent with psychological and sociological findings and with the nature of human behaviors.

The primary areas for improvement and further development are: turning $MORALITY$ and $PROTECTION$ into probability distributions (or, in case of the multi-agent model, individual properties of agents that can differ) instead of global parameters, making the influence of $MORALITY$ and $PROTECTION$ more realistic by introducing complex non-linear relationships that follow empirical results [7, 8], performing sensitivity analyses, and expanding the influence of morality – currently, the value of $MORALITY$ deters thieves from stealing amounts lower than this value (3) so it acts like a temptation threshold, while the opposite mechanism would also be justified, likely requiring the introduction of another parameter. Another extension of the model would be to consider the topology of the environment and restrict information flow that influences agent decision making and interactions to their neighborhood, and to make such local interactions stochastic.

After calibration of the model and verification of its consistency with empirical results based on studies of human behavior, such simulations can be used for the prediction of possible moral choices and the evaluation of their outcomes before decisions are actually made. To facilitate further developments, the sources of the implementation are published both in java and javascript, and the working program itself is available for experimentation at `http://en.alife.pl/morality`.

References

[1] Allen C., Varner G., and Zinser J. Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(3):251–261, 2000.

[2] Anderson M. and Anderson S. L. Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4):15, 2007.

[3] Ayala F. J. The biological roots of morality. *Biology and Philosophy*, 2(3):235–252, 1987.

[4] Bazzan A. L. C., Bordini R. H., and Campbell J. A. Agents with moral sentiments in an iterated prisoner's dilemma exercise. Technical report, 1997.

[5] Bazzan A. L. C., Bordini R. H., and Campbell J. A. Evolution of agents with moral sentiments in an iterated prisoner's dilemma exercise. In Parsons S., Gmytrasiewicz P., and Wooldridge M., editors, *Game Theory and Decision Theory in Agent-Based Systems*, pages 43–64. Springer, 2002. URL: `https://doi.org/10.1007/978-1-4615-1107-6_3`, `doi:10.1007/978-1-4615-1107-6_3`.

[6] Belloni A., Berger A., Besson V., Boissier O., Bonnet G., Bourgne G., Chardel P. A., Cotton J.-P., Evreux N., Ganascia J.-G., et al. Towards a framework to deal with ethical conflicts in autonomous agents and multi-agent systems. In *CEPE 2014 Well-Being, Flourishing, and ICTs*, pages paper–8, 2014.

[7] Birnbaum M. H. Morality judgments: Tests of an averaging model. *Journal of Experimental Psychology*, 93(1):35, 1972.

[8] Chiu C.-y., Dweck C. S., Tong J. Y.-y., and Fu J. H.-y. Implicit theories and conceptions of morality. *Journal of Personality and Social Psychology*, 73(5):923, 1997.

[9] Coelho H., da Rocha Costa A. C., and Trigo P. On agent interactions governed by morality. In *Interdisciplinary Applications of Agent-Based Social Simulation and Modeling*, pages 20–35. IGI Global, 2014.

[10] DeScioli P. and Kurzban R. Mysteries of morality. *Cognition*, 112(2):281–299, 2009.

[11] Floridi L. and Sanders J. W. On the morality of artificial agents. *Minds and Machines*, 14(3):349–379, Aug 2004. URL: `https://doi.org/10.1023/B:MIND.0000035461.63578.9d`, `doi:10.1023/B:MIND.0000035461.63578.9d`.

[12] Gotts N. M., Polhill J. G., and Law A. N. R. Agent-based simulation in the study of social dilemmas. *Artificial Intelligence Review*, 19(1):3–92, 2003.

[13] Gunkel D. J., Bryson J. J., and Torrance S. The machine question: AI, ethics and moral responsibility, 2012.

[14] Harsanyi J. C. Can the maximin principle serve as a basis for morality? A critique of John Rawls's theory. *American political science review*, 69(2):594–606, 1975.

[15] Hill R. P. and Watkins A. A simulation of moral behavior within marketing exchange relationships. *Journal of the Academy of Marketing Science*, 35(3):417–429, 2007. URL: `https://doi.org/10.1007/s11747-007-0025-5`, `doi:10.1007/s11747-007-0025-5`.

[16] Hofbauer J. and Sigmund K. *Evolutionary games and population dynamics*. Cambridge University Press, 1998.

[17] Komosinski M. and Adamatzky A., editors. *Artificial Life Models in Software*. Springer, London, 2nd edition, 2009. URL: `http://www.springer.com/978-1-84882-284-9`, `doi:10.1007/978-1-84882-285-6`.

[18] Kuhn S. T. Reflections on ethics and game theory. *Synthese*, 141(1):1–44, 2004.

[19] May R. M. and Leonard W. J. Nonlinear aspects of competition between three species. *SIAM journal on applied mathematics*, 29(2):243–253, 1975.

[20] McLaren B. M. Computational models of ethical reasoning: Challenges, initial steps, and future directions. *IEEE intelligent systems*, (4):29–37, 2006.

[21] Moor J. Four kinds of ethical robots. *Philosophy Now*, 72:12–14, 2009.

[22] Nawa N. E., Shimohara K., and Katai O. Does diversity lead to morality? On the evolution of strategies in a 3-agent alternating-offers bargaining model. In *Workshop on Evolutionary Computation and Multi-Agent Systems (ECOMAS) at the Genetic and Evolutionary Computation Conference (GECCO-2001)*, pages 317–320, 2001.

[23] Rahwan I. *Interest-based negotiation in multi-agent systems*. PhD thesis, University of Melbourne, Department of Information Systems, 2004.

[24] Robbins R. and Hall D. Decision support for individuals, groups, and organizations: Ethics and values in the context of complex problem solving. In *AMCIS 2007 Proceedings*, page 329, 2007.

[25] Saptawijaya A. and Pereira L. M. Towards modeling morality computationally with logic programming. In *International Symposium on Practical Aspects of Declarative Languages*, pages 104–119. Springer, 2014.

[26] Sullins J. P. When is a robot a moral agent? *IRIE: International Review of Information Ethics*, 2006.

[27] Wallach W., Allen C., and Smit I. Machine morality: bottom-up and top-down approaches for modelling human moral faculties. *AI & Society*, 22(4):565–582, 2008.

[28] Wiegel V. and van den Berg J. Combining moral theory, modal logic and MAS to create well-behaving artificial agents. *International Journal of Social Robotics*, 1(3):233–242, 2009.

[29] Wilson E. O. The biological basis of morality. *The Atlantic Monthly*, 281(4):53–70, 1998.

Received 07.11.2018, Accepted 07.02.2020