Fecha de recepción 11/04/2021
Fecha de aceptación: 22/08/2021
Pp. 60–99

Using Agent-based Models (ABM) for Ethical Prescriptions

Raúl González Fabre
Universidad Pontificia Comillas - Madrid
rgfabre@comillas.edu

Abstract

We propose to use Agent-based models (ABM) for estimating trend consequences of following concrete ethical prescriptions in impersonal social relationships. We discuss the potential and limitations of this methodology and its utility for the many sources of ethical prescriptions in our society--among them, University courses in professional and corporate Ethics, CSR, Sustainability, and similar ones.

We follow an Aristotelian conception of Ethics, where the goodness of a prescription results both from its relation to moral principles and from its expected consequences. ABM allow to enrich the knowledge about expected consequences in defined contexts with a generality, similar to the one intended by the prescription. Thus, ABM may contribute to the discernment of ethical prescriptions.

For demonstration, we develop a basic example of ABM for the study of commutative justice in a stylized market. We derive some hypotheses from that model, and propose several ways for further confirmation or refusal of those hypotheses.

Keywords: Agent-based models, ethical prescriptions, Business Ethics, Ethics and Economics.

RAÚL GONZÁLEZ FABRE

Uso de modelos basados en agentes (ABM) para prescripciones éticas

Resumen

Proponemos aquí usar modelos basados en el agente (ABM) para estimar las consecuencias tendenciales de seguir prescripciones éticas concretas en relaciones sociales impersonales. Discutimos el potencial y las limitaciones de esta metodología, y su utilidad para las muchas fuentes de prescripciones morales en nuestra sociedad--entre ellas, los cursos universitarios de Ética profesional y de la empresa, RSC, sostenibilidad y semejantes.

Seguimos una concepción aristotélica de la Ética, donde la bondad de una prescripción resulta tanto de su relación con principios morales como de sus consecuencias esperadas. Los ABM permiten enriquecer el conocimiento sobre las consecuencias esperadas en contextos definidos con generalidad semejante a la pretendida por la prescripción. De esta manera pueden contribuir al discernimiento de las prescripciones éticas.

Como ejemplo, desarrollamos un ejemplo básico de ABM para el estudio de la justicia conmutativa en un mercado estilizado. Derivamos algunas hipótesis de ese modelo, y proponemos varios caminos para la ulterior confirmación o refutación de esas hipótesis.

Palabras clave: Modelos basados en el agente, prescripciones éticas, Ética de los Negocios, Ética y Economía.

RAÚL GONZÁLEZ FABRE

Utilisation des modèles basés sur les agents (ABM) pour les prescriptions éthiques

Résumé

Nous proposons d'utiliser des modèles à base d'agents (ABM) pour estimer les conséquences tendancielles du respect de prescriptions éthiques concrètes dans les relations sociales impersonnelles. Nous discutons du potentiel et des limites de cette méthodologie et de son utilité pour les nombreuses sources de prescriptions éthiques dans notre société - parmi lesquelles, les cours universitaires en éthique professionnelle et d'entreprise, RSE, durabilité, etc.

Nous suivons une conception aristotélicienne de l'Éthique, où la bonté d'une prescription résulte à la fois de son rapport aux principes et de ses conséquences attendues. L'ABM permet d'enrichir la connaissance des conséquences attendues dans des contextes définis avec une généralité proche de celui visé par la prescription. Ainsi, ils peuvent contribuer à améliorer le discernement des prescriptions éthiques.

Pour démonstration, nous développons un exemple de base d'ABM pour l'étude de la justice commutative dans un marché stylisé. Nous dérivons quelques hypothèses de ces ABM et proposons plusieurs manières de confirmer ou de refuser davantage ces hypothèses.

Mots clés : Modèles à base d'agents, prescriptions éthiques, éthique des affaires, éthique et économie.

RAÚL GONZÁLEZ FABRE

Uso de modelos baseados em agentes (ABM) para prescrições éticas

Resumo

Propomos aqui usar modelos baseados em agentes (ABM) para estimar as consequências da tendência de seguir prescrições éticas concretas em relações sociais impessoais. Discutimos o potencial e as limitações desta metodologia e a sua utilidade para as diversas fontes de prescrições morais na nossa sociedade — nomeadamente, cursos universitários de Ética Profissional e da Empresa, RSC, Sustentabilidade e afins.

Seguimos uma concepção aristotélica da ética, onde a bondade de uma prescrição resulta tanto da sua relação com os princípios morais quanto das suas consequências esperadas. Os ABM permitem enriquecer o conhecimento sobre as consequências esperadas em contextos definidos com uma generalidade semelhante à pretendida pela prescrição. Deste modo podem contribuir para o discernimento das prescrições éticas.

Como exemplo, desenvolvemos um exemplo básico de ABM para o estudo da justiça comutativa num mercado estilizado. Derivamos algumas hipóteses a partir deste modelo e propomos vários caminhos para confirmar ou refutar essas hipóteses.

Palavras-chave: Modelos baseados no agente, prescrições éticas, Ética dos Negócios, Ética e Economia.

I. Introduction

In our pluralistic society many sources propose prescriptions for decision making. Families, educational institutions, NGO, churches, political movements, private companies, the State, speakers in mass media, agents in social networks... provide prescriptions with the intention of guiding all kinds of decisions by others.

Among the prescriptions with which we obsequy each other, the ethical ones may be characterized by their going beyond the receiver's immediate convenience, in extension and/or in time. There may be no conflict, if what is prescribed happens to coincide with the individual's immediate convenience, but generally such convenience is not the support for arguing an ethical prescription.

Here we are interested only in ethical prescriptions about how to decide in impersonal relationships. In those relations we interact with people we need not know personally in any detail--often we do not even know how many they are. Markets, citizenship, bureaucracies public and private... are typical loci of impersonal interactions. They are different from relations with relatives, friends, neighbors, close coworkers and the like, where we decide taking very much into account the concrete personality and story of our counterpart.

Ethical prescriptions about impersonal relations must be somehow motivated, explicitly or implicitly, given that their issuer is telling some other to decide differently from the latter's immediate convenience. Among the several possible motivations--think of authority or emotion, for example--we focus here on rationality. When resourcing to rationality as a motivator, as it is usually done in the courses of Ethics in University grades, rational justification of our proposed ethical prescription becomes central.

Decisions always take place in particular contexts. Ethical prescriptions are often more general; in consequence, they must identify their conditions of application. Some may be universal ('never do A') or restricted to broader or narrower contexts ('if the case can be understood under B, do A').

Justifying rationally an ethical prescription implies to identify at least two possibilities eligible by the agent in the stylized situation for which the normative statement pretends applicability. Then we must explain rationally why one precise possibility must be chosen (and/or why the others must not).

Following the general Aristotelian line of thought, we hold that justifying a prescription requires to consider both its relation to ethical principles--for example, which values it realizes and how--and its expected consequences--for the agent undertaking that option, for other agents affected and for the society as a whole. This means that neither of the two elements--principles or expected consequences--is by itself enough for ethical justification, but a certain ('prudential') balance between them must be searched for a normative statement to be proposed as good.

It can be thought that while the relation of a certain prescription to ethical principles can be studied in general, its expected consequences can only be estimated in fully particular situations. This is not wrong, but it tends to split the ethical process in two, divorcing Ethics from consequences in a quite Kantian fashion. Common sentences like 'Business Ethics is very nice but does not apply to real life', sometimes heard from corporate people, express that divorce.

Some kind of 'reflective equilibrium' between the general and the particular is needed to bridge the gap. We propose here using Agent-based models (ABM) to make the gap smaller--though not to eliminate it completely: learning from consequences in concrete situations will always be necessary. The study of the consequences of ethically different ways of action in quite general simulated setups can help to make ethical prescriptions more credible as good in the Aristotelian sense.

In this article we propose a methodology for doing so.

II. A certain use of ABM

Agent-based models allow to explore expected consequences of an ethical prescription related to impersonal relations, or at least to start it.

The basic components of an ABM are three:

a. *agents* understood as units of decision that interact attempting to maximize their respective target functions;
b. within a common *environment*--fixed or with a predictable evolution. The environment includes the agents' rules of interaction and the topology of their possible connections--random, geographical, any other type of network, pre-established or generated along the simulation...;
c. along a *time* scale often discrete, so that the starting point at step t+1 is the ending point at step t.

The programmer defines fully these three elements, but the results of the model are not intentionally programmed. They emerge from the computer interaction of the agents in that environment along time, thus allowing for events macro that are not an expected result of the original programming.

The application of ABM for discussing an ethical prescription is easy:

- The general situation for which our ethical prescription is proposed can be modeled as the environment of the simulation.
- ABM facilitates to program heterogeneous agents. In this use, we are interested on agents that differ only in their following our prescription or not--using in consequence some other decision algorithm. That difference will allow to classify them in ethical types. Varying the initial proportions of those ethical types we can study the influence of the prescription over some results of the simulated system.
- Selection and mutation mechanisms can be programmed within the environment, so allowing for studying variations in the proportion of the different ethical types in the population. The selection should happen over a fitness variable comparable across agents, thus not completely subjective. Ethical type and convenience are subjective concepts but money possessed by each agent, for example, is not.

In consequence, ABM can be programmed where the environment and the agents are all the same, except for one detail: the ethical type of the agents with regard to our prescription. Such ABM would allow to explore the consequences of the interaction along time between different initial proportions of those ethical profiles, in the stylized environment. After some initial steps, the system stabilizes, and some trend hypotheses can be extracted regarding:

- Consequences for the agents that follow our ethical prescription: do they tend to 'invade' the system or rather to be expelled from it, or a stable equilibrium between the ethical types is reached? Are they more or less successful in the stylized environment than the agents with other ethical types? Do the agents tend to cluster according to ethical types?
- Consequences of the presence of agents following the ethical prescription under study, for the general results of the model. For example: do more or less agents fail when more 'ethical' agents are present? Are there more or less transactions? Does the system results in bigger or smaller inequality, as measured over the variable used for selection?

This is the basic use of ABM we are proposing for contributing to the rational evaluation of ethical prescriptions. In section 5 we present an example, but before it is important to discuss briefly some epistemological points.

III. Epistemological considerations

Let us start with a brief recapitulation of a couple of points already presented in section 1.

First, expected consequences is not all that matters in the rational evaluation of an ethical proposal. However, there are lines of thought that think otherwise (Consequentialisms, of which the most relevant is Utilitarianism). And there is still another line that holds that consequences have no relevance at all in the rational justification of an ethical prescription (Kantian). Quite in an Aristotelian fashion, here we stand in a middle point, so to speak, where estimated consequences are important for justifying the ethical goodness of a prescription but not the only relevant consideration.

Second, the main idea of using ABM for estimating consequences is to shorten somehow the gap between the generality of most ethical prescriptions--for example, the ones proposed in the formation in professional Ethics at the University--and the concrete situations where they may apply.

That gap is never fully bridged with ABM. In fact, it is never bridged, not even with a very detailed knowledge of particular situations, because the future may be somehow structurally different from the past and unexpected consequences may happen. Any model, even the mental model of the particular situation that a good professional forms, is susceptible to miss relevant aspects that reveal themselves only after acting in the situation, not before, while deciding how to act. But ethical prescriptions make sense precisely as decision guides, that is, before decision.

The learning of particularities through some kind of 'reflective equilibrium' is always necessary.

A first step in bridging that gap happens because ABM allows for ethical heterogeneity in the population of agents. Most social theories of a Neoclassical or Austrian flavor suppose ethical homogeneity through some kind of 'representative agent', that for mathematical solvability ends up being an egoistic one. The conclusions of such theories are taught from the secondary school onwards in basic courses of social sciences. As a result, many people estimate consequences of ethical prescriptions using those theories as hermeneutic principles, without being aware of the ethical suppositions there embedded.

Incorporating moral heterogeneity, ABM can enlighten ethical prescriptions through an exploration of expected consequences in worlds ethically plural. But their reach and limits must be clearly stated.

Regarding its reach, the ABM methodology presents great flexibility. Particularly, it needs not use derivable functions because no mathematical analysis takes place. Some important consequences are:

- The agents may decide in a lexicographic way, that is, first on one criterion, then on another only among the possibilities that complied with the first, etc. This is adequate for ethical prescriptions, which often propose criteria for decision making that must have absolute priority over other criteria. For example, let us suppose--as in our example in section 5--that a first criterion is fairness in the transaction and a second is maximization of the agent's gains. An agent that follows such ethical prescription will maximize her gain only within the transactions that fulfill the criterion of fairness. Fairness has then lexicographic priority over gains.
- The lexicographic character may extend not only to the way some agents interact but also to their choice of counterpart for the next interaction. Morality often implies that agents of

certain ethical types give priority to pairing with other agents that have exhibited in the past the same morality (understood as the same decision-making pattern). Previous experience is the most effective way of signaling to others what kind of agent one is, thus what can be expected of oneself in an interaction. If there is some influence of morality over the choice of counterparts, pairing cannot be merely random. Systems of reputation arise that can be easily modeled in ABM.

- Finally, the possibility of non-derivable functions finds an application also in evolutionary mechanisms. For example, in ABM, selection can be easily programmed using thresholds--which imply discontinuous functions--instead of direct proportionality. This allows to enrich the concept of 'competitive success', so important for evaluating consequences in many social contexts. On top of the usual understanding of that concept as the ability to win in each round of the competition, a second idea can be proposed consisting in remaining above the threshold along the simulation time: competitive sustainability, so to speak. We would be modelling then a 'maximum competitive success'--triumph in the competition--and a minimum one--survival in the competition. Any competitive trajectory in between would be 'successful' to a degree. Obviously, some kind of competitive success is necessary for ethical prescriptions in impersonal relations to stand: an ethical prescription that generally leads the agents following it to be excluded from the social competition in question, may not qualify as a good one.

ABM has some secondary additional advantages born out of its flexibility. For example:

- ABM can be designed with a degree of abstraction similar to the ethical prescription whose consequences they intend to estimate. As mentioned in section 2, most prescriptions are not referred to entirely particular situations, but pretend some general application. They exhibit some degree of abstraction, that can be easily modeled in ABM.
- ABM allow for studying whatever consequences we desire in the system where the ethical prescription is applicable. Consequences for the agents acting on that prescription,

proportions of agents of each ethical type programmed, impact over the results of the whole system (its efficiency, inequality, or whatever other aspect we may consider relevant).

So far the epistemological advantages of ABM for our purpose. Their epistemological limits are also very relevant. They all come from the fact that ABM are not only very flexible but, at the same time and for the same reason, also very specific. ABM are not solved analytically but through numeric computation. That allows for modeling complications in the environment, the agents and the time that wouldn't result in equations analytically solvable, but requires:

- to assign numbers to all parameters and independent variables[1] that define the starting point of the model.
- to establish precise algorithms for deriving values of all dependent variables along time, from parameters, independent variables and previous values of the dependent ones.

Doing that amounts to defining fully the structure of the model, with no unknowns left. As the results of the model depend, we may suppose, on that structure, they can be varied in a non-obvious manner simply by changing relevant details in the design and starting point of the simulation. Apart from malicious changes on the side of the programmer to obtain her desired results, also mistakes in the code and artifacts generated by the programming language used are possible and difficult to detect. All of them would limit strongly the scientific value of any conclusions obtained from ABM.

There are several ways to face this challenge, but they are not totally satisfactory even taken together:

[1] We call 'parameters' to starting values fixed across the different runs of an ABM, while 'independent variables' are different in different runs, though invariable within each run. The objective of our use of ABM is to study the influence of independent variables with some ethical meaning over the chosen dependent variables--the model results--; hence that several runs of the model with different values of the independent variables are needed.

- publishing the full program of the simulation may allow evaluation by other scholars patient enough to read the code. Unluckily there is not an agreed standard for building ABM--most proposals refer merely to how to document them--which makes evaluation more difficult.
- programming the same ABM in several independent languages, to avoid artifacts derived from the computer language used.
- favoring the replication of interesting ABM by other scholars, so that they can check the results obtained.
- using values taken from empirical studies for parameters and independent variables, or at least plausible values.
- introducing some relevant randomness in the operation of the agents, so running the simulation as many times as random seeds are programmed, to explore whether the conclusions are robust to that randomness.

Given our purpose and the above limitations, we must be very careful in the interpretation of ABM results. They must not be taken as conclusions but as hypotheses: probable consequences of following a certain ethical prescription in the general situation modeled. That implies a provisional character, pending examination of their robustness via more simulations and also via 'reflective equilibrium' with real cases that can be understood as concretions of that general situation.

On the other hand, such hypotheses must be presented in terms of trends--in which direction an independent variable seems to influence a dependent one--, not of concrete values. Sensitivity studies may be useful to see the impact of the proportion of agents following an ethical prescription over different dependent variables of interest, as compared with other independent variables incorporated in the model.

With all these precautions, it is not the same to be completely unaware of possible consequences, or to have merely an intuition regarding them, possibly based on Neoclassical or Austrian suppositions, as having explored those consequences in a more systemic way for ethically heterogeneous social setups. ABM can help with this, and so contribute to improving the quality of ethical prescriptions for impersonal domains.

IV. Related work

Edmonds *et al.*[2] emphasize the importance of the purpose of computational models for evaluating them and affirm rightly the need for declaring that purpose first thing when presenting a model. The present article proposes a very specific and, to our knowledge, quite new purpose for ABM: discussing ethical prescriptions in generalized setups with moral type variability of agents.

That is by no means the first time that computational models have been focused on moral phenomena. Computer simulations based on the decisions of individual agents--'tournaments', ABM, its close relative cellular automata, dynamical replicators--have been aimed at explaining determinate aspects of moral dynamics in societies since the 70's[3]. Veit has called this approach the "Explaining Morality Program"[4].

The underlying concept is shared with the Libertarian tradition: social order, which often requires most agents to behave not-immediately maximizing their convenience, may be born out of individual decisions without a central authority imposing such 'morality'. Almost

[2] Bruce Edmonds *et al.*: "Different Modelling Purposes", *Journal of Artificial Societies and Social Simulation*, 22:(3):6, June (2019).

[3] Rainer Hegselmann, "Moral dynamics", in Robert A. Meyers, (ed.): *Encyclopedia of complexity and systems science*, (New York: Springer, 2009): 5677-5692.

[4] Walter Veit, "Modelling morality", in Ángel Nepomuceno Fernández, et al. (eds.): *Model-based reasoning in science and technology: inferential models for logic, language, cognition and computation*, (Cham: Springer, 2019): 83.

textually, that's the beginning question of the seminal work of Axelrod: "Under what conditions will cooperation emerge in a world of egoists without central authority?"[5]

The second basic conceptual tenet in the standard use of computational models for explanation of moral dynamics is evolution[6], understood in a quasi-biological sense: variation, selection and propagation of certain traits among the agents' population. For example, Danielson starts asking: "What sort of agents, amoral or moral, do better playing a series of representative games?"[7]

A more recent approach, also explanatory, drops the assumption of immediately self-interested agents. It intends to explain the emergence of morality as a personal phenomenon inside the agents themselves, driven by external success at the social games considered. That was already proposed by Hegselmann at the end of his article[8]. Internal architectures of the agent like BDI or PESC[9] have been proposed for this purpose.

Another line of research intends to use computational models to discuss theories from moral philosophy. Ruvinsky proposes to use computational simulations for "assessing the utility of an ethical theory or system"[10], what she calls "computational ethics". Wiegel[11] uses ABM in order to decide between act and rule Utilitarianisms, while Lasquety-Reyes[12] considers rightly that ABM may be specially suited for virtue ethics, which is by definition also agent-based.

[5] Robert Axelrod, *The evolution of cooperation*, (New York: Basic Books, 1984): 3.

[6] Steven Mascaro, *et al.*, *Evolving Ethics: The New Science of Good and Evil*, (Exeter: Imprint Academic, 2010).

[7] Peter Danielson, *Artificial morality: virtuous robots for virtual games*, (London. New York: Routledge, 1992): 17.

[8] Peter Danielson, *Artificial morality: virtuous robots for virtual games*, (London. New York: Routledge, 1992): 17.

[9] Jeremiah Lasquety-Reyes: "Computer Simulations of Ethics: the Applicability of Agent-Based Modeling for Ethical Theories", pp. 18-28, *European Journal of Formal Sciences and Engineering* 1(18): 10, May-August 2018.

[10] Alicia I. Ruvinsky, "Computational Ethics", in Mary Quigley (ed.), *Encyclopedia of information ethics and security*, (Hershey: Information Science Reference, 2008): 80.

[11] Vincent Wiegel, *SophoLab: Experimental Computational Philosophy*, (Maastricht: Technische Universiteit Delft, 2007)

[12] Jeremiah Lasquety-Reyes, "Towards Computer Simulations of Virtue Ethics". *Open Philosophy* 2: September (2019): 399-413.

How does our research fit into this landscape? Let us consider some aspects:

First, our basic unit of enquiry is not a moral theory but merely a behavioral prescription, which may be arrived at from several different theories--and by different social prescribers. That leaves out of our purpose the last, rather incipient use of ABM for discussing ethical theories. Mutatis mutandis, however, an important purpose enunciated by Ruvinsky coincides with our intent. She says: "By simulating various ethical theories within an agent society, researchers can study the effects of individual computationally moral actions within the society on the ultimate ethical equilibrium of the society"[13]. That is precisely done in our example (see 5.3, Figure 1 below), only that not for different theories but only for different prescriptions, whatever their theoretical roots.

Second, we are not interested in modeling the 'inside' of the moral agent. We intend to examine the trend consequences of acting according to a moral prescription exogenously assumed by a portion of the agents. The internal motives of the agent for acting according to such prescription are not studied; we simply assume that it does. In consequence, agents are not adaptive, though the system where they interact changes with their evolutionary success or failure. The closest similarity we have found is with Müller[14]. He also uses an extrinsic approach to ethical options of customers, in order to study the changes, it produces in the innovation process of the companies.

Finally, some of the conclusions of the 'Explaining Morality Program' may be useful for our purpose--which is very different from that Program's, to the extent that the basic suppositions of a certain model are supposed to represent well enough the conditions under

[13] Jeremiah Lasquety-Reyes, "Towards Computer Simulations of Virtue Ethics". *Open Philosophy* 2: September (2019): 399-413, :79.

[14] Matthias Müller, *An agent-based model of heterogeneous demand*, Wiesbaden, Springer, 2017, ch. 6.

which an ethical prescription under study intends applicability. That may not be the case if explanatory models simplify too much, for example supposing homogeneous agents, random pairing, non-accumulative fitness or non-lexicographical decision making. For our purpose, the opposite is essential to study moral dynamics in impersonal relations: ethical heterogeneity of the agents, reputation-guided pairing, accumulative fitness--wealth, power or the like--, and lexicographical decision making by at least some agents. It is difficult to identify trends in the consequences of holding a certain ethical prescription if those four elements are not kept. ABM is particularly congenial to keeping all of them, more than Replicator Dynamics models for example.

Our model could be qualified as 'predictive' to a certain extent in the sense proposed by Edmonds *et al.*[15]. It intends to predict trend consequences "unknown to the modeller", consisting of "a pattern or a relationship" that, if the model has been properly validated, aspires to be "reliable" with respect to its declared parameters and algorithmic design.

The other conditions posed by Edmonds *et al.*[16] for a predictive model with real scientific value are not that much fulfilled, as we may see in 3. Each simulation like the one exemplified in 5 is only an "illustration", not based on empirical data. Its reliability depends on the approach chosen by the "modeller"[17] capturing well enough the basic moral dynamics where the agents follow different prescriptions. The results of any such model cannot "be unambiguously checked to see if it holds"[18] but they can be tested through successive extensions, other simulations and dialogue with experts. Only after such procedures, the hypotheses coming out of these simulations would have "a useful degree of accuracy". In our case, that usefulness means robustness enough to enrich the general 'estimation of consequences' of the prescription under

[15] Matthias Müller, *An agent-based model of heterogeneous demand*, (Wiesbaden: Springer, 2017): ch. 6, #2.5.

[16] Matthias Müller, *An agent-based model of heterogeneous demand*, (Wiesbaden: Springer, 2017): ch. 6, #2.5.

[17] José M. Galán, *et al.*: "Checking Simulations: Detecting and Avoiding Errors and Artefacts": in Bruce Edmonds and Ruth Meyer, *Simulating Social Complexity. A Handbook*, (Cham: Springer, 2017): 119-140.

[18] Bruce Edmonds and Ruth Meyer, *Simulating Social Complexity. A Handbook*, (Cham: Springer, 2017): #2.5.

study, and thus contribute to evaluate its goodness in the general contexts modeled.

V. Example

In this section we demonstrate with a quite simple example the use of ABM for an ethical prescription in market Ethics. Due to space constraints we cannot document fully the corresponding ABM, but the Matlab program can be found in internet.

V.1. An ethical prescription

A classical prescription in economic Ethics can be stated: 'In any market transaction between two agents, the parties must share the benefits of the transaction in proportion to their contribution to producing them'. This follows the general idea of fairness as proportionality to contribution, already proposed by Aristotle[19] and developed as theories of 'commutative justice' and 'just price' by the Scholastics in the Middle Ages and Early Modernity.

The mentioned agents, parties in a market transaction, could act otherwise in principle. For example, the benefits could be distributed in proportion to the relative negotiating power of the parties, not their contribution, so that more powerful agents get more than their 'fair share' in the transaction. Let us call these agents 'type HE' (*Homo Œconomicus*) while the agents acting according to the above prescription of economic fairness are 'type FA' (Fair Agents).

Choosing one or other position in the negotiation is an ethical option of each agent. If you are the most powerful agent in a transaction, acting according to our moral prescription

[19] Aristotle, *Nicomachean Ethics*, Trad. R. Crisp, Cambridge, U.K.; (New York: Cambridge University Press, 2000): bk V.

implies that you are not going to extract as much as possible of it but some less, in order to keep the fairness of the whole business. Fairness--a form of ethical goodness adequate to markets--is put to your election against maximum gain. You have to choose.

The positive relation of our prescription to principles--fairness in this case--is the object of the theories of commutative justice and just price. More arguments can be piled up to support the idea, for example of a Kantian nature, based on the Golden Rule, etc.

But that ethical prescription is not totally justified by its strong relation to moral principles. Some consideration of consequences must also be done before we can propose it to economic agents as ethically good. If we can rationally expect that following that prescription results in ruin and exclusion from the market for any agent so acting, the prescription can hardly be proposed as good in general. The same if very negative consequences can be rationally expected for the market or for society. So a certain exploration of expected consequences is also necessary for holding our prescription as ethically good.

V.2. ABM: design

To explore the consequences of that prescription, let us pose an ABM of a stylized market with a high number of agents (*nAgents*=1000), that are going to interact along *nTics*=3000 cycles.

The interaction between agents consists in four basic steps:

- Pairing, let us suppose *agent i* with *agent j*.
- They produce jointly a surplus.
- That surplus is negotiated and distributed between the two agents.

- Each agent calculates the reputation of the other agent in its eyes. Reputation is accumulative, but previous values are degraded with each new transaction between the same agents, multiplying it by the parameter *repDegTime*.

Each agent is characterized by two internal variables:

- The resources in its possession (*money*), that changes along the simulation, spending part or all of it in transactions and adding what it gets from those transactions. The value of money at the beginning of each cycle (*money1*) stands for the power of negotiation of the agent.
- Its moral character (*moralChar*), that may be 1 (type HE) or 2 (type FA). It is fixed for each agent, except for mutation.

The initial values of those internal variables are totally independent. In the first tick agents are not statistically richer or poorer (*money1*) because of their type (*moralChar*). Any differences will result from the successive transactions in the simulation.

money is randomly distributed according to a bounded Pareto distribution	α =1.16; min value=1; max value=30.
moralChar is 2 for a certain proportion *pFA0* of the agents, different for each simulation, and 1 for the rest	The cases considered are *pFA0*={0.0, 0.2, 0.4, 0.6, 0.8, 1.0}. The first value (*pFA0*=0.0) implies that at the start all agents are type HE, while the last one (*pFA0*=1.0) means that all agents are type FA.

Table 1: Basic agent internal variables

Of these two, only the second is an independent variable that changes from one simulation to another. The first is actually a parameter common to all simulations. There are several other parameters:

Number of agents	*nAgents*=1000
Proportion exploration/exploitation in the choices for pairing	*explore*=0.2
Maximum number of interactions an agent can start in a tick	*nPar*=3
Maximum number of other agents each agent will consider as possible partners for pairing	*nCan*=6
Surplus created per unit of investment in a transaction	*p*=0.1
Decay rate of previous reputation of other agents with new transactions with the same agent	*repDegTime*=0.9
The minimum money necessary for an agent to remain in the simulation	*survCond*=1
The probability of mutation in moralChar for each agent at each tick	*mutProb*=0.0001

Table 2: Parameters

The general dynamics of the model is, for each tick:

1. Normalization of money to the same average of the initial distribution: *money*. This is made to avoid any inflationary effect along simulated time as a result of introducing in the system surpluses of past transactions.
2. Pairing of agents. In 20% of the cases (*explore*=0.2) the agents choose randomly a partner among agents that have no previous reputation in their eyes. In the remaining 80% of the

cases, each agent selects a maximum of *nCan* other agents to examine, and pairs with the one with the highest reputation in its eyes provided that it has some remaining *money* to invest.

3. Production between the agents paired. Supposing that *agent i* and *agent j* have paired for a transaction, having $money_i$ and $money_j$ respectively, each invests $min[money_i, money_j]$ in the transaction. The surplus produced is $p*2*min[money_i, money_j]$.
4. Distribution of the surplus produced between the agents that generated it. See Table 3 below for the rules of distribution.
5. Update of reputations of agents in the eyes of their partners. The way of doing it depends of the moral type of the agent (see below).
6. Selection via exclusion-replacement. Agents with *money<survCond* 'fail' in the market and are replaced with agents with money assigned with the same distribution used at the beginning of the simulation (bounded Pareto with α=1.16; min value=1; max value=30).

   moralChar is randomly copied from one of the remaining agents (those that didn't 'fail'). This implies that if agents of a certain moral type (*moralChar*) tend to fail more often than the average, that type's proportion in the simulation will decrease.
7. Random mutation. Agents may mutate their moral profile with low probability (*mutProb*), from FA to HE or vice-versa. This is to prevent the system from getting stuck in *pFA=0* or *pFA=1*. Even if all agents have the same moral profile, some small amount of the opposite profile will always enter the simulation through mutation.

The moral type of an agent influences the steps 4 (distribution), 5-2 (reputation-pairing) above:

Step 4. Let us call *mi* the *money1* of *agent i* at the beginning of the step and *mj* the *money1* of *agent j*. The payments per unit of jointly produced surplus depend on the moral type of agents involved in the transaction:

	agent j: type HE	agent j: type FA
agent i: type HE	agent i gets: *mi/(mi+mj)* agent j gets: *mj/(mi+mj)*	agent i gets: *max[1/2, mi/(mi+mj)]* agent j gets: *min[1/2, mj/(mi+mj)]*
agent j: type FA	agent i gets: *min[1/2, mi/(mi+mj)]* agent j gets: *max[1/2, mj/(mi+mj)]*	agent i gets: *1/2* agent j gets: *1/2*

Table 3: Payments

This implies that agents type HE get as much of the joint surplus as their relative power allows (*money1*), when facing another agent HE or an agent FA with less power. On the other hand, agents type FA get 1/2 (their fair share of the surplus, given that both agents contributed the same to producing it) when facing another agent FA or an agent HE with less power. The functions *max* and *min* are not derivable. They used here to express the lexicographic preference of fairness over gain for agents type FA.

Steps 5-2. For agents type HE, the reputation of any other agent is merely the surplus they got in transactions with it, per unit of their own investment. In consequence, they choose partners from whom they obtained the most in the past.

For agents type FA this criterion is second to the type of their potential partner. They prefer partners type FA, and within that, those from whom they got maximum return per unit of investment. The lexicographical preference for fairness applies thus also in the choice of partners.

No agent can know the type of other agent, unless it has experienced that type actually. No mind reading is possible, nor there is any other signaling previous to transaction. If such experience has not happened (there were no previous transactions between the agents), agents type FA suppose that they are meeting another agent type FA.

Summarizing, the basic tenets of our initial ABM are:

It represents a stylized market where agents have different wealth and power of negotiation (*money*), and belong to one of two different moral types (*moralChar*), one corresponding to the full use of that power for getting maximum gains (type HE) and the other to restricting the use of that power within the limits of fairness (type FA). Making variable the initial proportion of each (*pFA0*), we can study the influence of assuming the moral prescription that agents FA follow, in a 'world' where they are neither all nor only one.

The wealth/power variable (*money*) establishes an 'objective' dynamics in the ABM, on which selection depends entirely. That dynamics is accumulative: agents with more wealth/power in a step of simulation tend to get more in the following step if they are type HE. We may expect that the big tend to become bigger and the small, smaller till they 'fail' in the market and leave it to be replaced by a 'medium' one. In consequence, the variable wealth/power plays the role of evolutionary fitness.

The moral type of each agent (a 'subjective' aspect) influences the way in which it acts in that 'objective' dynamics, both regarding the distribution of surplus and the pairing of agents for transaction. The last aspect requires to incorporate some reputational dynamics in the ABM.

In principle, we can expect that agents type HE will get (statistically) more wealth/power from each transaction than agents type FA, but at the same time they will have (statistically) worse reputation in the eyes of other agents, and thus less likelihood to be chosen as partners.

Knowledge is local and limited. No agent has knowledge of the system as a whole, nor of how other agents are doing. All agents in each transaction choose partner among a maximum of *nCan*=6 other agents (over a total of *nAgents*=1000), often (*explore*=0.2) based on their own experience with those agents, expressed as reputations of one agent in the eyes of another. Reputations are not communicated among agents; so that moral types and performances are a private issue. Even if it has remaining money, no agent can make more than *nPar*=3 transactions in each simulation step.

V.3. ABM: conclusions

The model is used for six simulations for the different values of the moral type (*moralChar*) in initial proportions: *pFA0*={0.0, 0.2, 0.4, 0.6, 0.8, 1.0}. Given that the initial distribution of *money* and the pairing of the agents have a random component, we run each case 24 times for different random seeds and take as result the average (the median would also be a reasonable central measure for this purpose, giving similar trend results).

We are interested in three kinds of results:

- Proportion of the agents following the moral prescription under study (proportion of agents type FA: *pFA*).
- How those agents do in the environment modeled as compared to agents of moral type HE.

- How their presence affects relevant values of the whole system.

Let us see some hypotheses of those three kinds than can be derived from the results of our ABM:

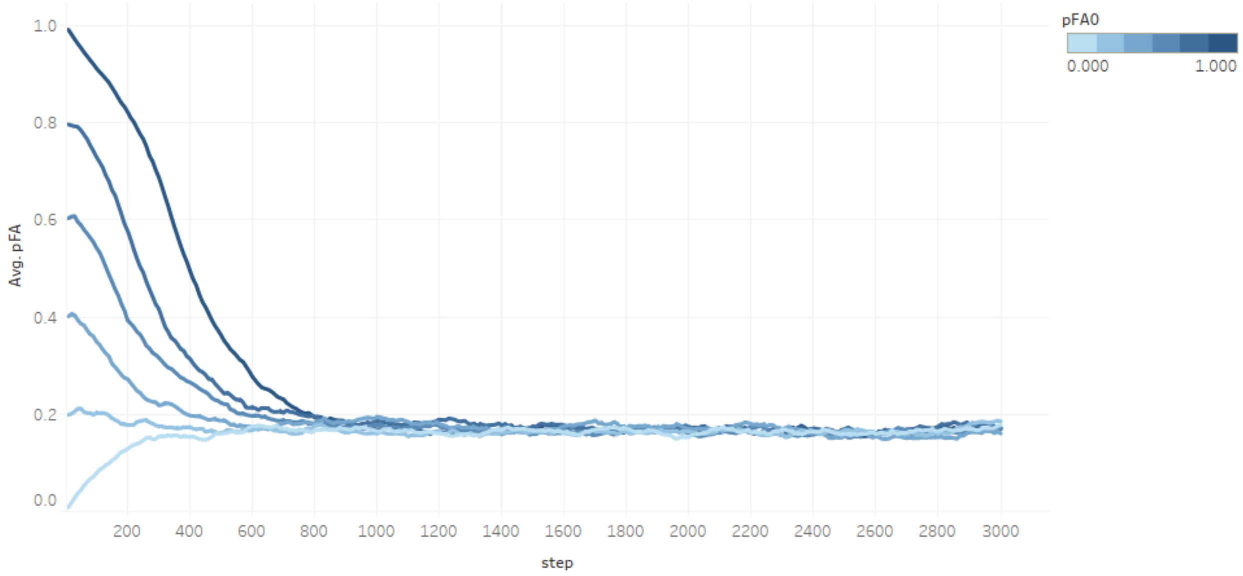


Figure 1: Proportion of agents type FA (following the moral prescription) along simulation time

Figure 1 allows to make a first hypothesis: whatever the starting point in the proportion of agents type FA (*pFA0*), the system tends to converge to the same stable proportion *pFA*, which is not null (as if all agents where pure maximizers in each decision, type HE), nor one (as if all agents ended up following the moral prescription of fairness we study, type FA). A system with mixed moral types seems sustainable in time.

One the other hand, that equilibrium *pFA* must be a consequence of structural factors, given that the only non-random independent variable is *pFA0*, and this does not influence the simulation after step 900 (aprox.). In the last book of *E.N.*, Aristotle[20] suggests that whoever has

[20] Aristotle, *Nicomachean Ethics*, Trad. R. Crisp, Cambridge, U.K.; (New York: Cambridge University Press,

a moral project for her society, must become a teacher or a legislator. If our hypothesis stands, being a politician--someone acting to change social structures--seems more effective for the moralization of society. Teaching ethics makes more of an indirect sense--forming supporters of structural changes--than a direct one--getting agents to behave according to our moral prescription. Maybe we should be teaching at the University more of Politics to accompany and make effective the professional Ethics, so fashionable after the crisis.

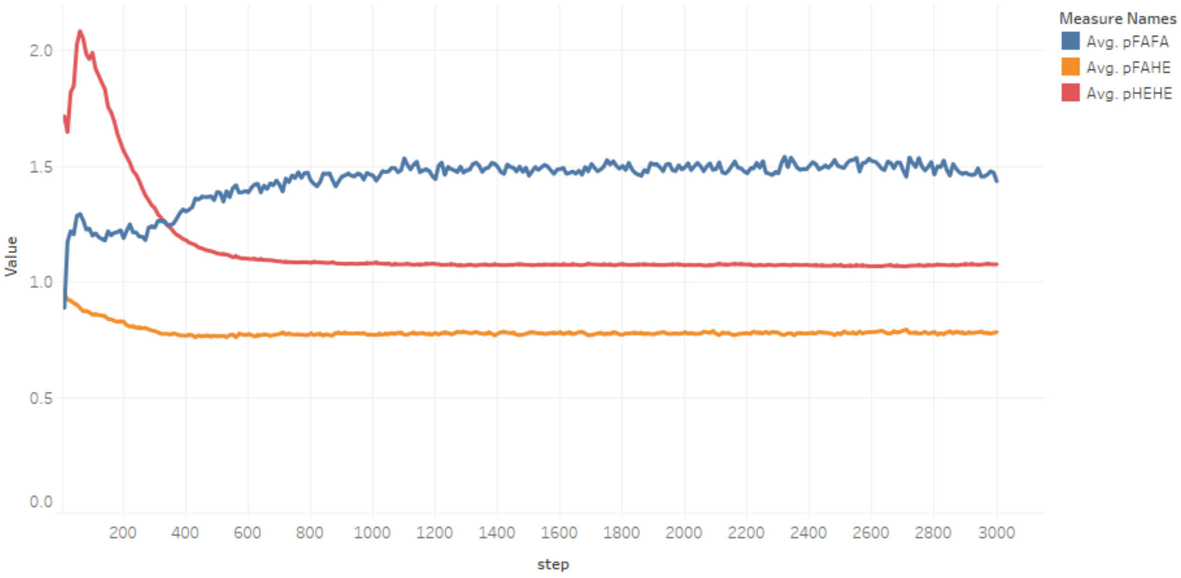


Figure 2: Proportion of transactions between agents of the same moral type (pFAFA and pHEHE) and mixed type (pFAHE) along simulation time

Figure 2 shows the moral types involved in each transaction actually happened. Values over 1 mean that a certain transaction happens more likely than the average, and values under 1 the opposite.

It can be seen that transactions with mixed types involved (*pFAHE*) are less frequent

2000): bk V.

proportionally than they should be with random pairing, while transactions with the same moral type on both sides (*pFAFA* and *pHEHE*, for the two types FA and HE, respectively) happen more often than random. The corresponding hypothesis is that our reputation system tends to create clusters (in a loose sense), particularly between agents that follow the moral prescription under study (agents type FA).

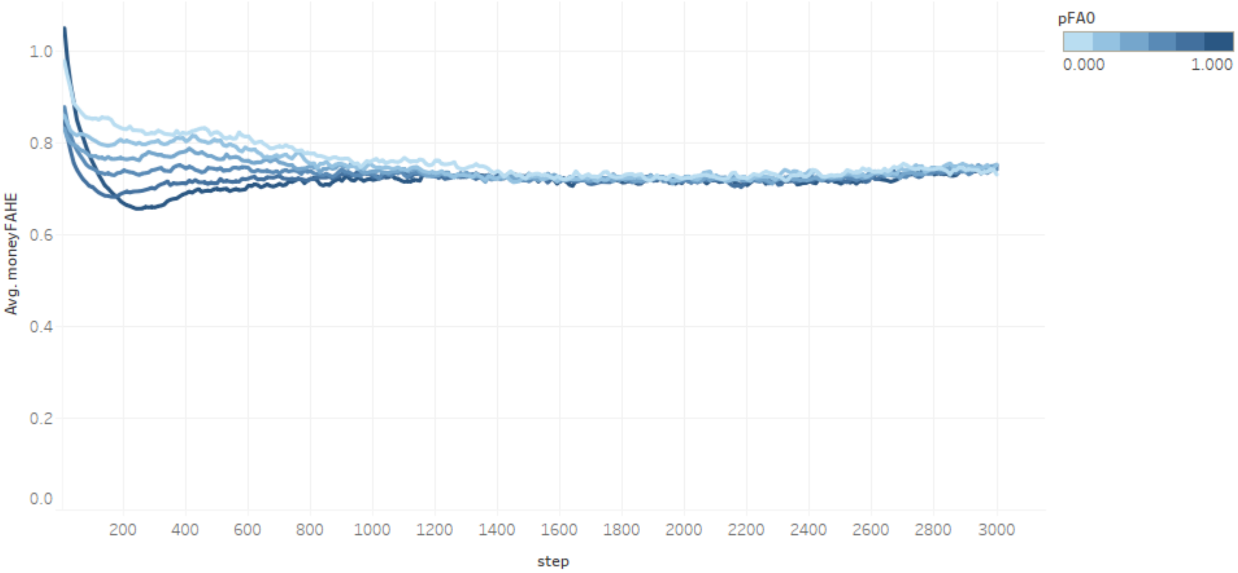


Figure 3: Average money of agents type FA divided by average money of agents type HE, along simulation time

The agents acting according to our moral prescription (type FA) do in average worse than those who act in a purely maximizer fashion (type HE). If they were performing equally, the quotient of their average money would stabilize around 1, but it actually does around 0.72. In consequence, following this moral prescription costs money to the agent.

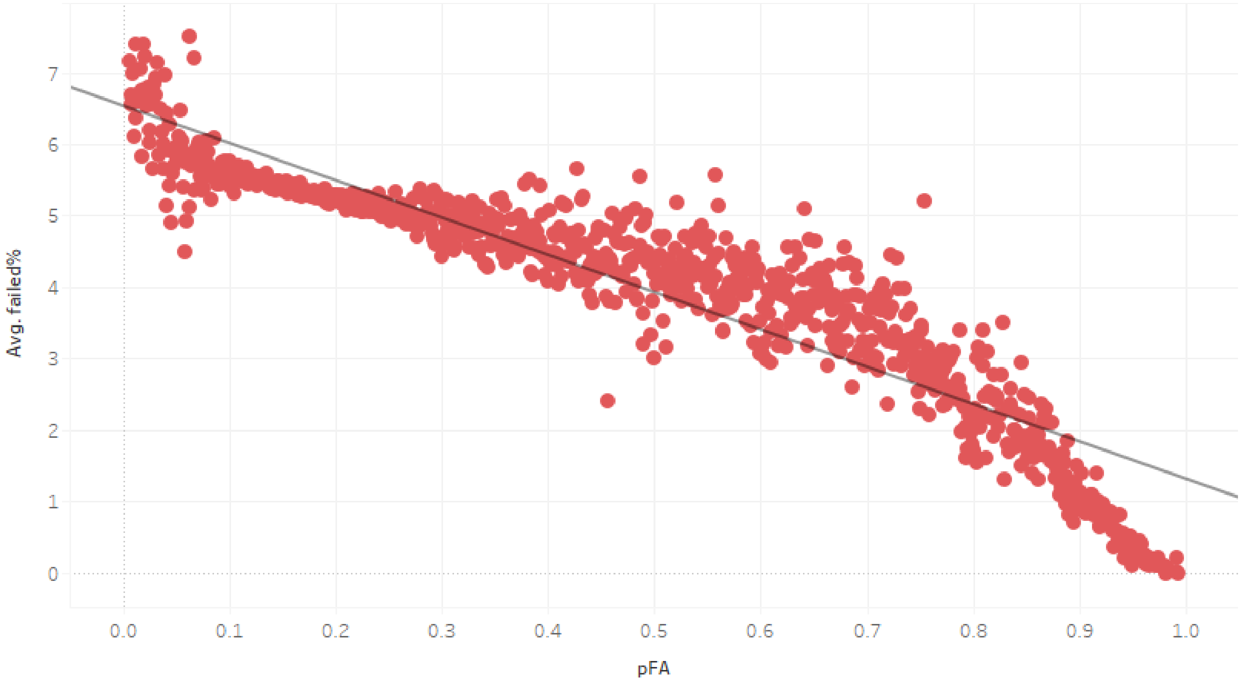


Figure 4: Percentage of agents that fail in the market vs actual proportion of agents type FA

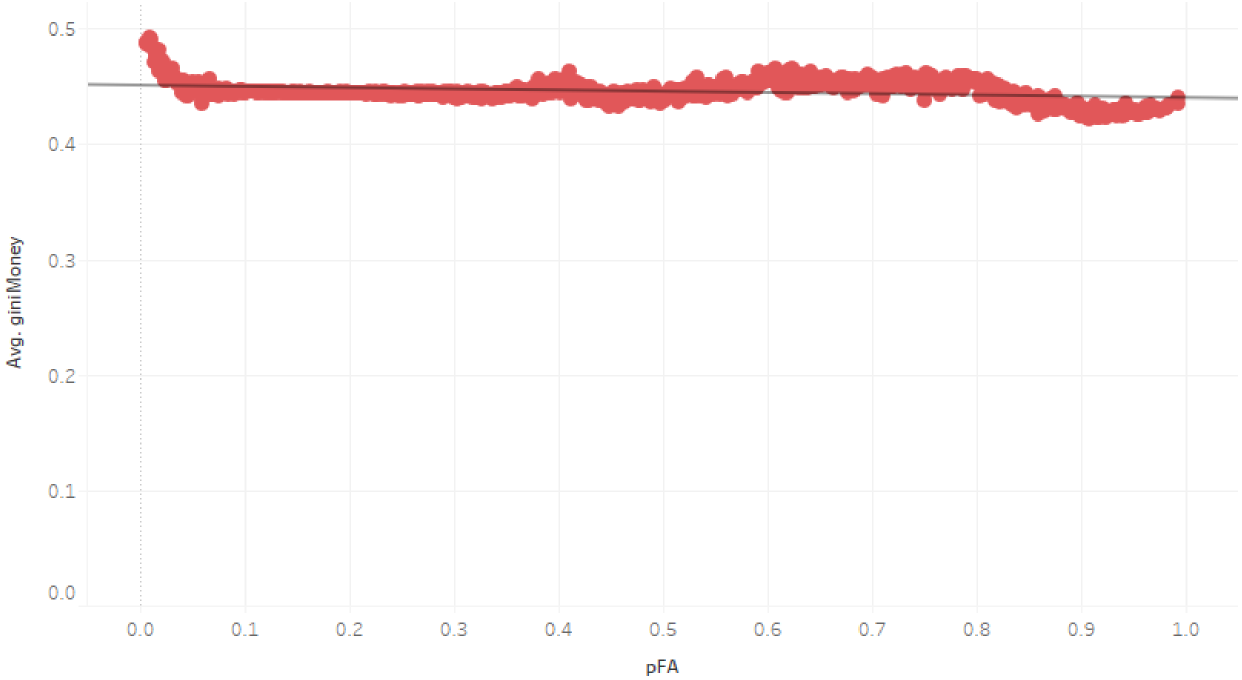


Figure 5: Gini index of wealth of the agents vs actual proportion of agents type FA

Figures 4 and 5 suggest that increasing the number of people that act according to the moral prescription under study diminishes the number of agents of any type that fail in the system, but does not change the inequality in the system--as measured by the Gini index of wealth (*money*).

It's easy to see that agents type FA do not take advantage of their superior power--if they have it--when distributing shares with the agents that are close to fail. So less of the latter actually fall below the survival threshold.

Regarding inequality, however, that effect seems to be compensated by the fact that the new agents replacing the failed ones have in average medium wealth, thus reducing the Gini index.

V.4. ABM: some extensions

As we have seen in our previous example, using a basic ABM some hypotheses can be formulated about the consequences of following a certain moral prescription. That is merely an starting point for investigating the point. Given that social structures are continuously changing, the hypotheses are unlikely to ever become proven theses. But some of them can be integrated into moral discernment, if they show to be robust, holding in a number of different circumstances. On the contrary, if those hypotheses show to be highly fragile to structural changes, maybe we have to correct them--or our previous intuitions-- specifying under which circumstances they stand, or simply recognizing that no connection can be made.

A good example of the latter can be found in Figure 4. Intuitively we may think that acting more fairly in a stylized market--as opposed of using all power to maximize gains--would diminish inequality. Figure 4 shows that not to be the case. More exploration can be done to see whether it would be under other (and which) conditions.

We can systematize the directions in which such exploration can proceed, with some examples:

a. *Variations of the basic model*

The first obvious modification can be introduced by converting parameters into independent variables.

Our model contains a number of parameters that are common to all simulations (see Table 2). Let us convert the first (*nAgents*) into a variable that can take the values: {250, 500, 1000, 2000}. All other parameters remain the same. We graph in Figure 6 the equivalent to Figure 1:

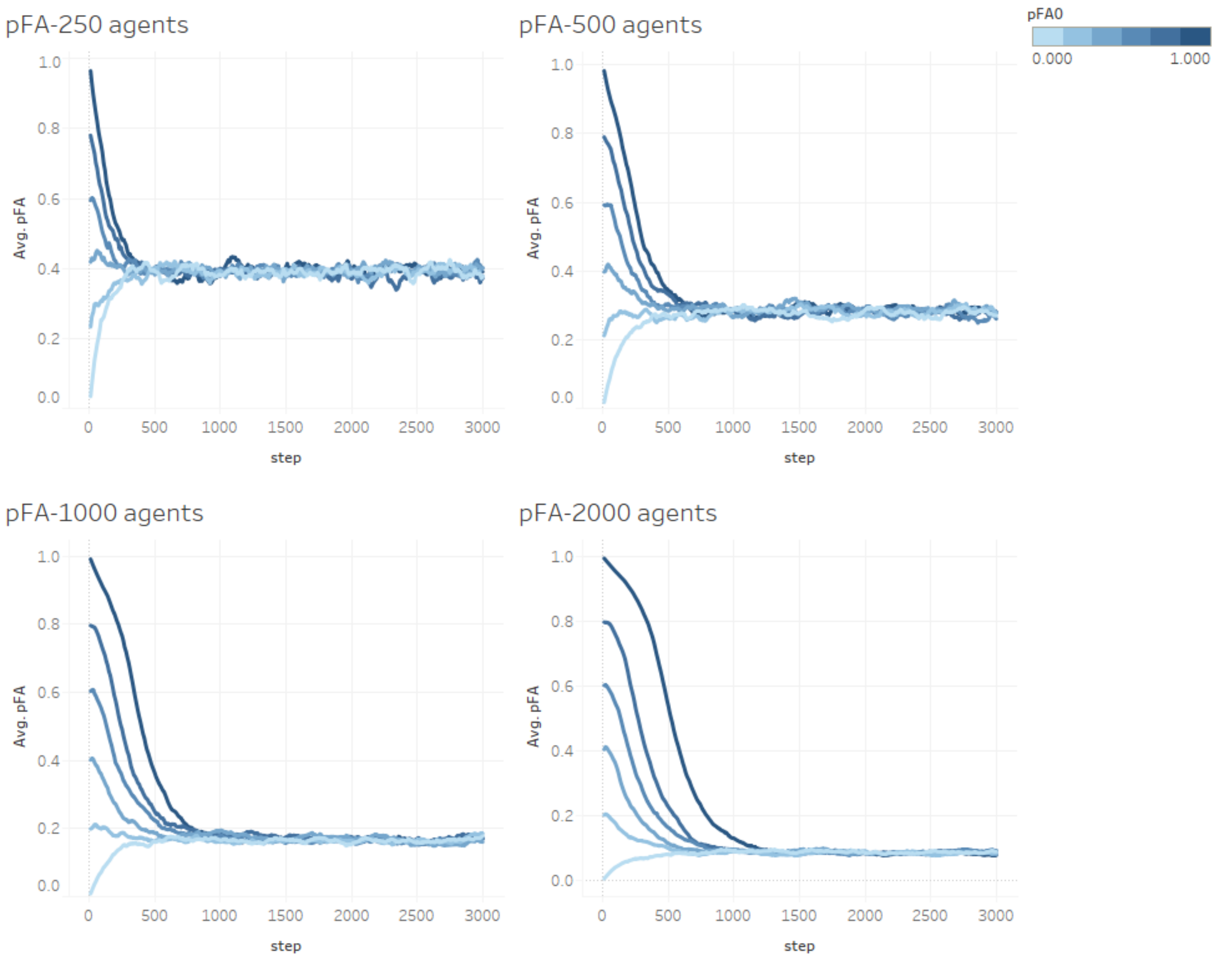


Figure 6: Proportion of agents type FA along simulation time, for different values of the number of agents

We see that our first hypothesis regarding the convergence of *pFA* to a single value for all *pFA0*, still holds. However the value in question tends to diminish when the number of agents in the simulation increases. Given the structure of our simulation, it is not difficult to guess that the reason is the different weight of reputation. With less agents in the market, each one forms relatively soon a reputation of most of the others, while with more agents that will not happen. An obvious conclusion--this is, a new hypothesis--can be formulated: in markets with many agents and a reputation mechanism based only on immediate experience, the moral type FA is less sustainable.

Modifying parameters in the same model may take a more complex form. Let us go back to *nAgents*=1000, and consider a new possibility: that the initial distribution of money is precisely equal, not following the bounded Pareto distribution described in Table 1. The average is the same in both cases, but now all agents receive exactly that average.

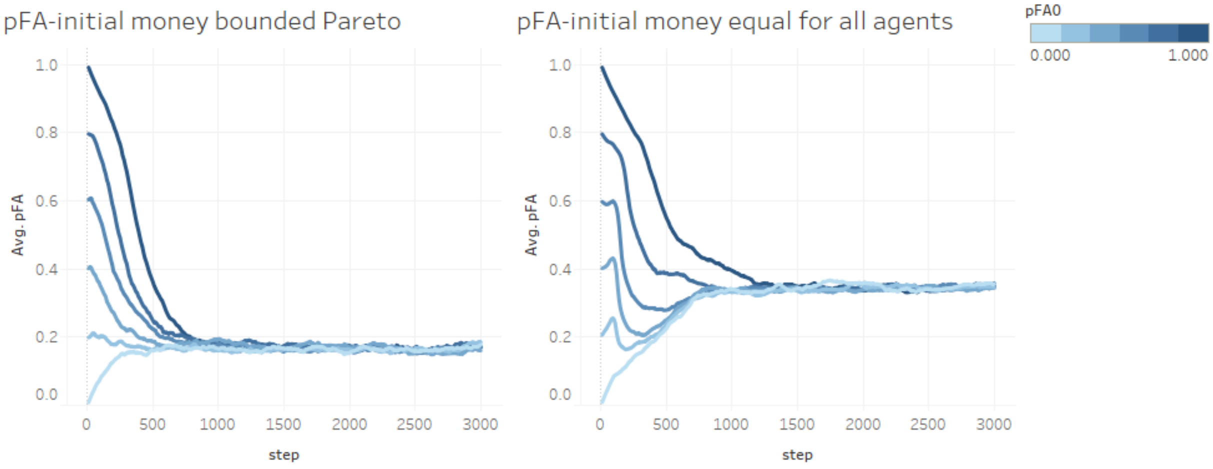


Figure 7: Proportion of agents type FA along simulation time, for different initial distributions of money

The hypothesis of convergence of the proportion of agents following our moral prescription on a single value holds. But with no starting inequality, that value is higher: 0.35 vs 0.17 in our example. That suggests an additional non-obvious hypothesis: that the average 'moral quality' in which the market stabilizes is inversely related to its initial inequality.

Another easy modification consists in posing more questions to the same model. Keeping *nAgents=1000*, let us ask about the relative ages of the agents belonging to each moral type considered:

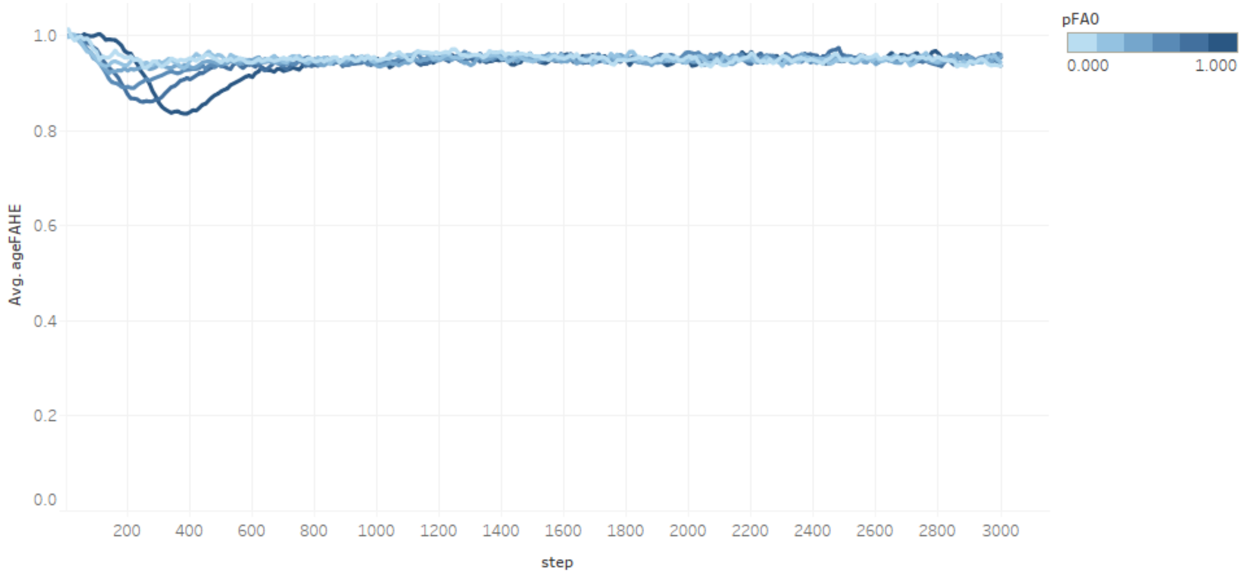


Figure 8: Average age of agents type FA divided by average age of agents type HE, along simulation time

The 'age' here considered is the number of cycles in the model, from incorporation to failure, if this happens. All agents start with 0 and increase it by 1 every step they remain alive. When an agent fails and it's replaced, the new agent starts again with *age*=0.

Figure 8 shows that acting according to our moral prescription has a cost in survival within the model. If there were none, the quotient would be 1, but it is slightly lower. Comparing with Figure 3, we can see that the cost in *age* seems smaller than the cost in *money*. This is reasonable because the selection condition has a threshold functional form (not a proportional one), and we have seen in Figure 2 that there is some clustering among agents type FA and in Figure 4 that agents type FA tend to diminish the rate of failure in the simulation. All this makes sense from an economic point of view.

Figure 9 confirms that agents type FA fail slightly more often than agents type HE.

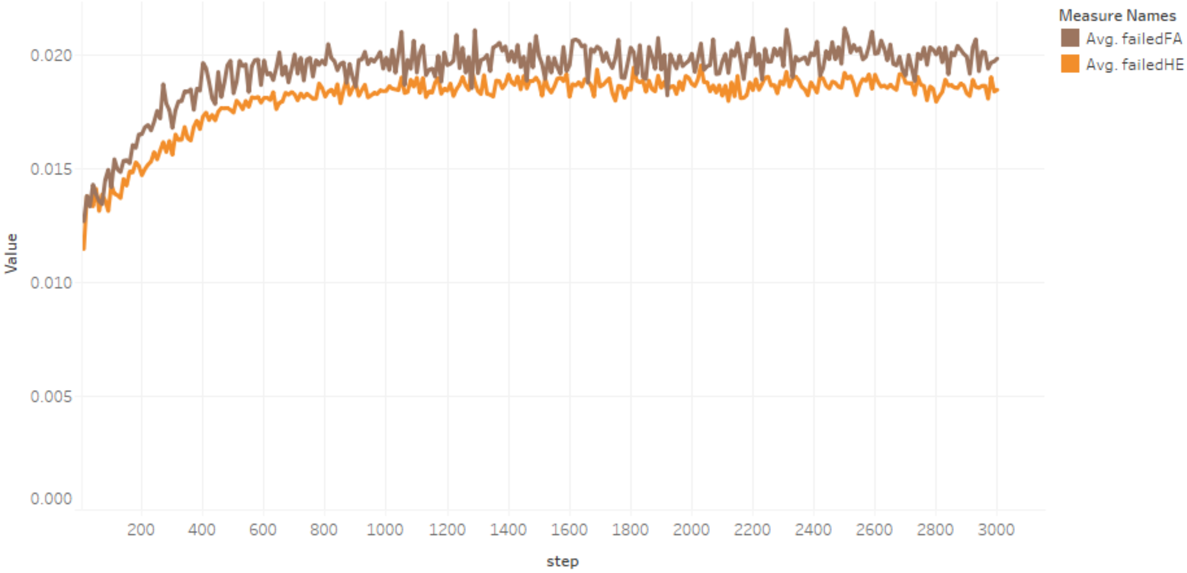


Figure 9: Proportion of agents failed according to type (FA and HE) along simulation time

More complex variations can be introduced, for example changing the algorithms used for pairing, forming reputations or failing and selection. In each case, the intention may be twofold: on the one hand, to check whether previous hypotheses still hold; on the other hand, to formulate new hypotheses, which in turn may lead to new variations of the model.

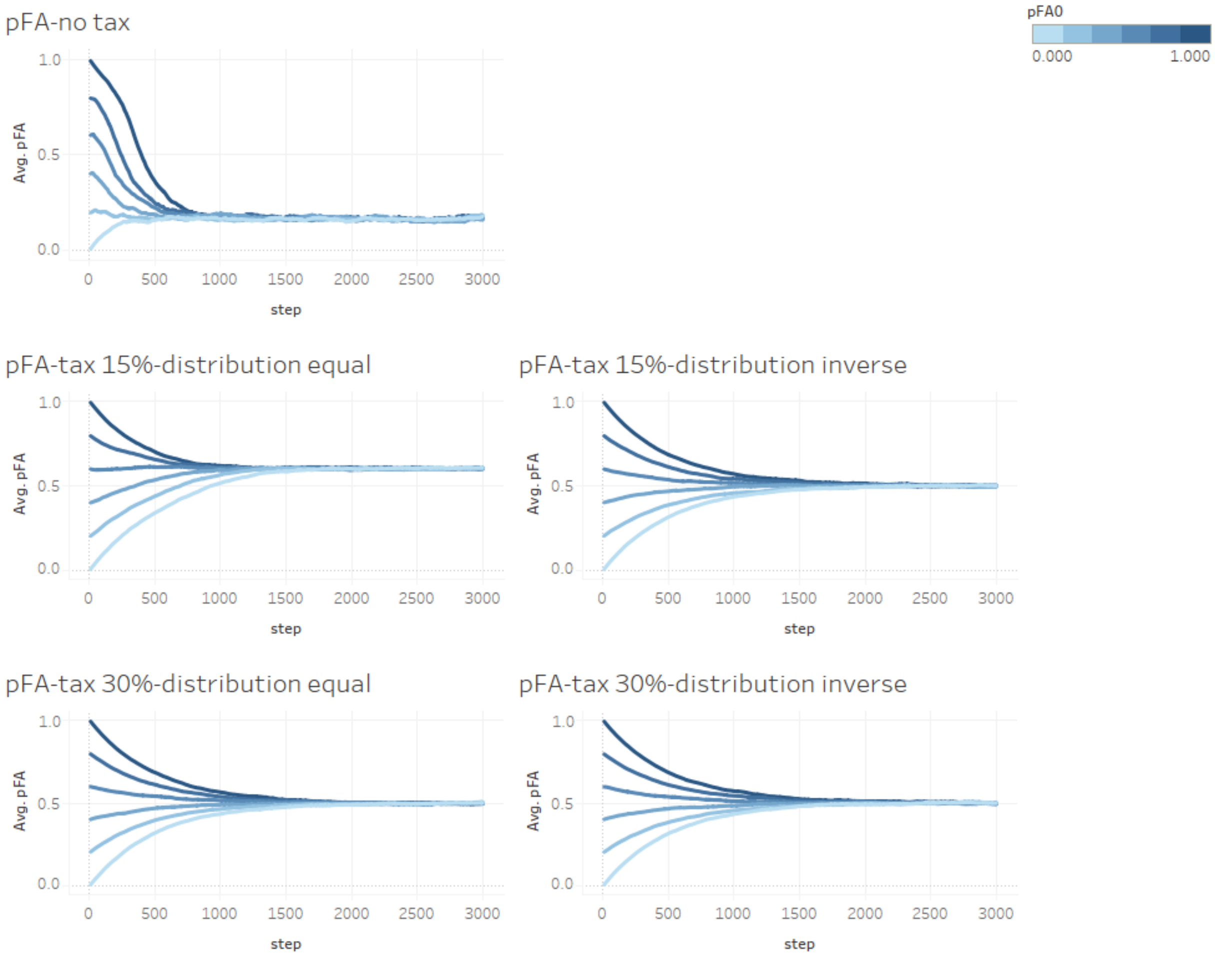
b. Adding new mechanisms to the basic model.

On top of those small variations of the model, there is always the possibility of enriching it via deeper changes to add relevant aspects that have not been considered yet. There are many possibilities; for example:

- In our basic model the number of agents is fixed. When an agent fails, it is replaced with another. Separating failure and entrance in two different algorithms would allow for a more realistic stylization of the market.
- The basic model is also technologically static. The function of production is always the same. But agents in a market may invest in future production capacity, detracting that money from their capacity for present transactions. The idea of fairness as appropriation of the surplus proportion that one has contributed to produce, would still be relevant, though the calculation of a fair distribution would not be the same: Table 3 would have to be modified.
- In the basic model all agents have capacity to consider the same number of candidates for transaction (*nCan*) and to initiate the same number of transactions (*nPar*). Those values could change from agent to agent via, for example, a publicity expenditure detracted from the agent's resources for transactions.
- The basic model takes only direct experience of each agent as a source of reputation of the other agents in its eyes. Interpersonal or even public mechanisms of reputation building could be added.
- No taxes are considered in the basic model. Some could be introduced, to examine its impact over the 'ethical quality' of the market as measured through *pF0*. Will a certain tax increase or decrease the evolutionary stable proportion of agents following our prescription?

As it is quick to describe, we use this last line as an example of possible model development. Let us introduce an indirect tax over the surplus and distribute the collected amount among all agents. We shall consider four cases, combining: the tax rates: {15%, 30%}; and the way of distributing the money collected: equally to all agents, inversely to the respective money (money1) they own at the beginning of the cycle ('richer' agents receive less, 'poorer' agents receive more). In all other aspects the model is the same described in 5.2.

Let us look only at a couple of results:



pFA-no tax

pFA-tax 15%-distribution equal

pFA-tax 15%-distribution inverse

pFA-tax 30%-distribution equal

pFA-tax 30%-distribution inverse

Figure 10: Proportion of agents type FA along simulation time, for different tax rates and distributions

Figure 10 shows that the confluence in an evolutionary stable value of *pFA* holds after introduction of the indirect tax. But the value itself: (1) increases considerably with the first 15% of tax rate and then remains approximately the same, while (2) it is not much affected by the distribution of the money collected by the tax.

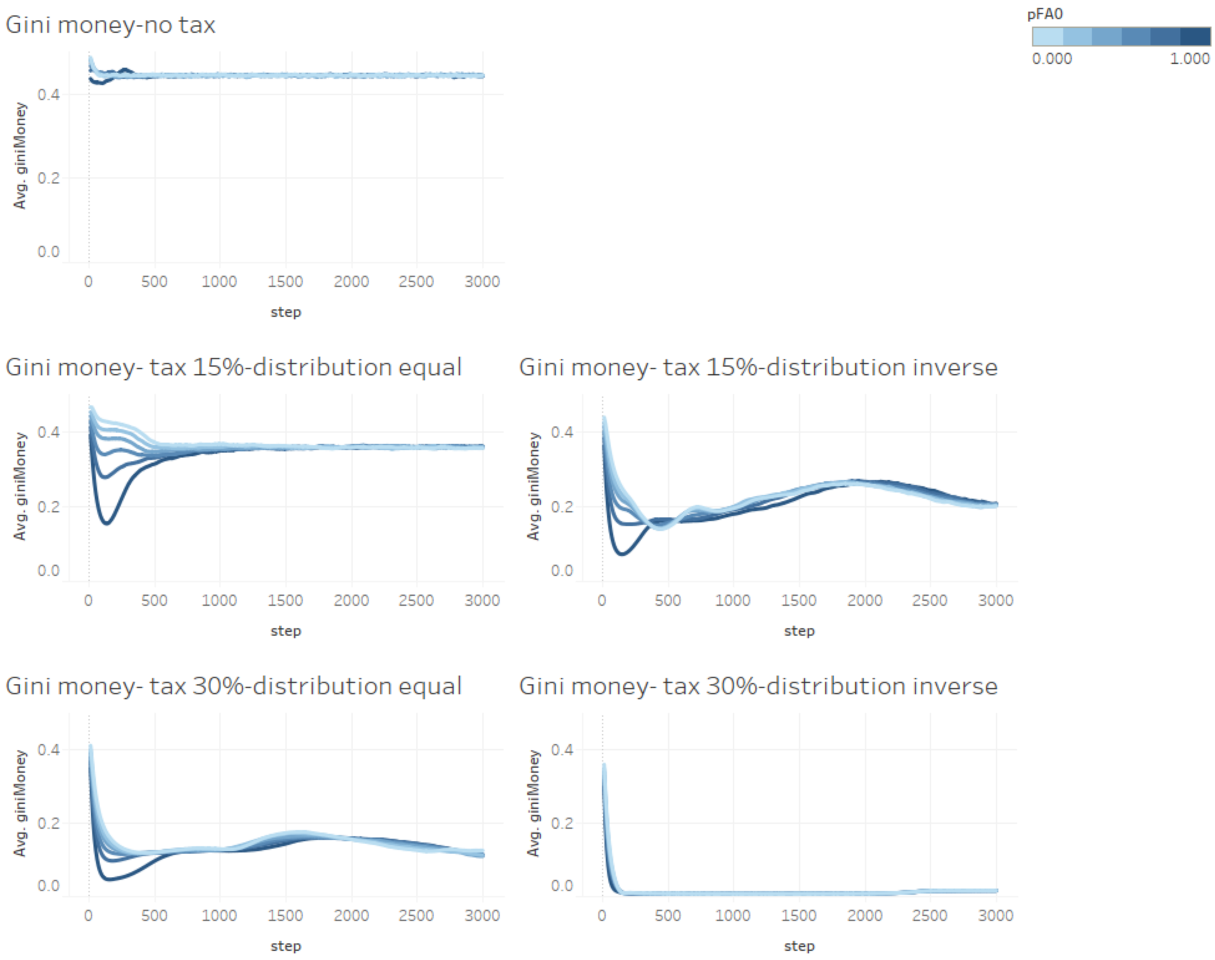


Figure 11: Gini index of agents' money along simulation time, for different tax rates and distributions

Figure 11 shows what could be expected: inequality in the model decreases with the increase in the tax rate and with the inverse distribution among the agents.

We said above that teaching Politics along with Ethics would be necessary in the social and administrative sciences of our Universities. Figures 10 and 11 support the idea. An indirect tax, which is a political measure, has strong consequences not only for a social property of the system (inequality) but also for its average ethical quality (*pFA*). Concerns about the ethical patterns in society can be realized at least partially through political means.

c. Via models for which some analogy can be established

Published models of social systems may provide some confirmation of hypotheses obtained from our own ABM. The analogy--and thus the corresponding confirmation--must be understood loosely and cautiously, of course. But provided that the conditions mentioned in 4 are fulfilled, it may be established. Those conditions were: ethical heterogeneity of the agents; reputation-guided pairing; accumulative fitness; and lexicographical decision making by at least some agents.

The examination may be useful not only to confirm or dis-confirm our post-model hypotheses, but also as clues to replicate their model in ABM methodology, and to identify new aspects of interest that could be incorporated in our models--relevant parameters and independent variables, algorithms, outputs...

d. Extra-model reflective equilibrium and expert consultation

Finally, direct experience and dialogue with persons involved in areas that can be understood under the stylized environment of the ABM model, will help to check our hypotheses and, again, discover aspects of the situation that are sufficiently general as to be incorporated in the ABM.

VI. Conclusions

In a pluralistic society many sources of ethical prescription coexist. It is realistic to assume that ethical prescriptions do not necessarily result from the system where they are applied, nor they are followed by all agents in that system. Different ethical ways of acting in markets, for example, may be proposed from outside the market system, and may be differently adopted by some market agents but no others.

Proposing computer simulations to estimate trend consequences of prescriptions may help institutions proposing such prescriptions to make them 'good' in an Aristotelian sense: better balanced between the principles they realize and the consequences both personal and social that can be expected from following them.

In this article we have proposed to use ABM for estimating those consequences in the case of impersonal relationships. ABM's methodological flexibility allows to model both the essential elements of moral life and the contexts where the prescription under study is proposed to be applied, with similar generality.

We have mentioned the epistemological limitations of ABM for this purpose. We conclude that ABM allows to formulate hypotheses that can reinforce or weaken the original prescription by showing different trend effects of following it. With successive simulations we can study the conditions under which those hypotheses stand or fail. Only those hypotheses that exhibit robustness to changes plausible in the general context where the prescription intends to be valid, can be incorporated into moral discernment.

And finally, we have demonstrated this particular use of ABM with an example consisting of a prescription of commutative justice in market relationships. We proposed a basic ABM of a stylized market, identified some hypotheses from its results and proposed several lines for extension of the model and progressive confirmation or refusal of those hypotheses.