Scientometrics (2014) 99:663–688
DOI 10.1007/s11192-014-1239-1

Mechanism change in a simulation of peer review: from junk support to elitism

Mario Paolucci · Francisco Grimaldo

Received: 14 June 2013 / Published online: 16 February 2014

Abstract Peer review works as the hinge of the scientific process, mediating between research and the awareness/acceptance of its results. While it might seem obvious that science would regulate itself scientifically, the consensus on peer review is eroding; a deeper understanding of its workings and potential alternatives is sorely needed. Employing a theoretical approach supported by agent-based simulation, we examined computational models of peer review, performing what we propose to call *redesign*, that is, the replication of simulations using different *mechanisms*. Here, we show that we are able to obtain the high sensitivity to rational cheating that is present in literature. In addition, we also show how this result appears to be fragile against small variations in mechanisms. Therefore, we argue that exploration of the parameter space is not enough if we want to support theoretical statements with simulation, and that exploration at the level of mechanisms is needed. These findings also support prudence in the application of simulation results based on single mechanisms, and endorse the use of complex agent platforms that encourage experimentation of diverse mechanisms.

Keywords Peer review · Agent-based simulation · Mechanism change · Rational cheating · BDI approach · Restrained cheaters

M. Paolucci (✉)
Institute of Cognitive Sciences and Technologies, Italian National Research Council, Via Palestro 32, 00185 Rome, Italy
e-mail: mario.paolucci@gmail.com

F. Grimaldo
Departament d'Informàtica, Universitat de València, Av. de la Universitat, s/n, Burjassot, 46100 Valencia, Spain
e-mail: francisco.grimaldo@uv.es

Introduction

Peer review at the center of the research process

Science has developed as a specific human activity, with its own rules and procedures, since the birth of the scientific method. The method itself brought upon unique and abundant rewards, giving science a special place, distinct from other areas of human thinking. This had been epitomized in the 1959 “two cultures” lecture, whose relevance is testified by countless reprints (see Snow 2012), presenting the natural sciences and the humanities as conflicting opposites, evidenced by their misunderstandings and the resulting animosity in the world of academia.[1]

Famous studies on science (the first example that comes to mind is that of Kuhn 1996) have been carried out with the tools of sociology and statistics. In the meantime science, co-evolving with society, has developed and refined a set of procedures, mechanisms and traditions. The most prominent about them is the mechanism of paper selection by evaluation from colleagues and associates—peers, from which the name peer review—, which is so ingrained in everyday research process to make scientists forget its historical, immanent nature (Spier 2002).

Nowadays, simulation techniques, and especially social simulations, have been proposed as a new method to study and understand societal constructs. Time is ripe to apply them to science in general and to peer review in particular. Simulations of peer review are starting to blossom, with several groups working on them, while the whole field of scientific publications is undergoing remarkable changes due, for instance, to the diffusion of non-reviewed shared works,[2] and to the open access policies being enacted after the 2012 approval of the Finch report from the UK government. These are only symptoms of the widespread changes brought about by the information revolution, which has transformed general access and dissemination of content, and specifically access and dissemination of science. In the future, the paper as we know it might be superseded by new forms of scientific communication (Marcus and Oransky 2011). The measures we use to evaluate scientific works will also keep changing. Consider for instance the impact factor, not yet fully superseded by the *h*-index, which in turn is likely to be substituted by alternatives as the $P_{top\ 10\ \%}$ index (Bornmann 2013), or complemented by usage metrics. The simplest of the latter, paper downloads, has already been shown to exhibit unique patterns; reading and citing, apparently, are not the same (Bollen et al. 2009).

In the meantime, collective filtering (e.g. reddit) and communal editing (e.g. Wikipedia) have found a way to operate without the need of the authority figures that constitute the backbone of peer review. Academia, on the contrary, still shows a traditional structure for the management of paper acceptance, maintaining a “peer evaluation” mechanism that is the focus of the present work. We note in passing that the peer review mechanism is only one of the many aspects of collective scientific work, which has been little studied as such (see for a recent exception Antonijevic et al. 2012); it is only one of a number of intersected feedback loops, serving the purpose of quality assurance between others.

Peer review, as a generic mechanism for quality assurance, is contextualized by a multiplicity of journals, conferences and workshops with different quality requests, a kind of multilevel selection mechanism where the pressure to improve quality is sustained

[1] These, that we simplify as opposite fields, have had their share of mixing and cross-fertilization; consider (Cohen,1933), and the digital humanities works starting with Roberto Busa in the 1940s.

[2] Consider for example http://arxiv.org/.

through the individualization of a container—for now we will consider only journals as representative of this—and the consequent pressure to defend its public image (see Camussone et al. 2010, for a similar analysis).

Journals as aggregates of papers, then, play two other important roles: first, as one of the most important (although indirect) elements in deciding researchers' career, at least for the current generation of researchers, that are often evaluated on the basis of their publications on selected journals. Second, journals play a role in the economy of science: their economic sustainability—or profit, as shown by the recent campaign on subscription prices[3]—depends on quality too, but just indirectly; profits would be warranted as well in a self-referential system, based not on quality but on co-option.

At the core of the above intersected feedback loops lies the mechanism of peer evaluation, that is based on custom and tradition, leveraging on shared values and culture. The operation of this mechanism has been the target of much criticism, including accusations of poor reliability, low fairness and lack of predictive validity (Bornmann and Daniel 2005)—even if this calculation often lacks a good term of comparison. Declaring a benchmark poor/low with respect to perfect efficiency is an idealistic perspective, while a meaningful comparison should target other realistic social systems. All the same, we must mention that some studies have demonstrated the unreliability of the journal peer review process, in which the levels of inter-reviewer agreement are often low, and decisions can be based on procedural details. An example is reported by Bornmann and Daniel (2009), who show how late reviews—that is, reviews that come after the editors have made a decision—, would have changed the evaluation result in more than 20 % of the cases. Although a high level of agreement among the reviewers is usually seen as an indicator of the high quality of the process, many scientists see disagreement as a way of evaluating a contribution from a number of different perspectives, thus allowing decision makers to base their decisions on much broader information (Eckberg 1991; Kostoff 1995). Yet, with the current lack of widely accepted models, the scientific community has not hitherto reached a consensus on the value of disagreement (Lamont and Huutoniemi 2011; Grimaldo and Paolucci 2013).

Bias, instead, is agreed to be much more dangerous than disagreement, because of its directional nature. Several sources of bias that can compromise the fairness of the peer review process have been pointed out in the literature (Hojat et al. 2003; Jefferson and Godlee 2003). These have been divided into sources closely related to the research (e.g. reputation of the scientific institution an applicant belongs to), and sources irrelevant to the research (e.g. author's gender or nationality); in both cases the bias can be positive or negative. Fairness and lack of bias, indeed, is paramount for the acceptance of the peer review process from both the community and the stakeholders, especially with regard to grant evaluation, as testified by empirical surveys.

Finally, the validity itself of judgments in peer review has often been questioned against other measures of evaluation. For instance, evaluations from peer review show little or no predictive validity for the number of citations (Schultz 2010; Ragone et al. 2013). Also, anecdotal evidence of peer review failures abound in the literature and in the infosphere; since mentioning them contributes to a basic cognitive bias—readers are bound to remember the occasional failures more vividly than any aggregated data—we will not delve on them, striking as they might be.[4] Some of the most hair-rising recent failures (Wicherts 2011) recently raised the attention to the emerging issue of research integrity,

[3] See http://thecostofknowledge.com/.

[4] The curious reader may start from the obvious place—Wikipedia http://en.wikipedia.org/wiki/Peer_review_failure, retrieved on 10 Jan 2013 - for an amusing tale about trapezia.

that converged in documents such as the Singapore statement.[5] These documents highlight the hazard of cheating in science, by exploiting the loopholes of peer review. This behavior goes under the name of rational cheating.

Rational cheating is defined in (Callahan 2004) as a powerful force, often augmented by rational incentives for unethical misconduct. Factors like large rewards for winners, perception of diffused cheating, limited or non-existing punishments for breaking the rules put pressure on researchers to work against the integrity and fairness required from themselves. Rational cheating is currently believed to be an important factor, potentially contributing to the failure of peer review and with that, to the tainting of science by cronyism. In this work, we aim to contribute to the understanding of rational cheating in peer review. Before moving on, however, we point out how the peer review mechanism has never been proved—or, for the matter, disproved—to work better than chance. Indeed, at present, little empirical evidence is available to support the use of editorial peer review as a mechanism to ensure quality; but the lack of evidence on efficacy and effectiveness cannot be interpreted as evidence of their absence (Jefferson et al. 2002).

This might sound strange, given the central position of peer review in the scientific research process as it is today. Why this lack of attention for such a crucial component? The answer, besides the habituation component, is probably to be found in the nature of the peer review as a complex system, based on internal feedback, whose rules are determined by tradition in a closed community. In this paper, we aim to contribute to an undergoing collective effort to understand peer review, in order to improve it, preparing the ground for its evolution by reform.

Simulation of peer review

To model the dynamics of science, a broad range of quantitative techniques have been proposed, such as: stochastic and statistical models, system-dynamics approaches, agent-based simulations, game theoretical models, and complex-network models (see Scharnhorst et al. 2012, for a recent overview).

In accordance with the aim and with the tool—that is, the type of modeling being used—a number of conceptualizations of science have been proposed that: explain statistical regularities (Egghe and Rousseau 1990); model the spreading of ideas (Goffman 1966), scientific paradigms (Sterman 1985) or fields (Bruckner et al. 1990); interrelate publishing, referencing, and the emergence of new topics (Gilbert 1997); and study the evolution of co-author and paper-citation networks (Börner 2010). In this paper we focus on a specific aspect of science: the peer review process applied to assess the quality of papers produced by scientists, aimed to the publication of textual documents.

The peer review process can be generally conceptualized as a social judgment process of individuals in a small group (Lamont 2009). Aside from the selection of manuscripts for their publication in journals, the most common contemporary application of peer review in scientific research is for the selection of fellowship and grant applications (Bornmann 2011). In the peer review process, reviewers sought by the selection committee (e.g. the editor in a journal or the chair in a conference) normally provide a written review and an overall publication recommendation.

Simulations and analysis of actual data on the review process are the ingredients that have been proposed to improve our understanding of this complex system (Squazzoni and Takács 2011). In this work, with an agent-based approach, we develop a computational

[5] http://www.singaporestatement.org/.

model as an heuristic device to represent, discuss and compare theoretical statements and their consequences. Instead of using the classic "data-model-validation" cycle, we take advantage of the social simulation approach that, following Axelrod (1997), could be conceptualized as "model-data-interpretation." In this approach, we use the computer to draw conclusions from theoretical statements that, being inserted in a complex system, have consequences that are not immediately predictable.

Agent-based modeling and simulation has been proposed as an alternative to the traditional equation-based approach for computational modeling of social systems, allowing the representation of heterogeneous individuals, and a focus on (algorithmic) process representation as opposed to state representation (Payette 2011). Research agendas and manifestos have appeared defending that agent-based models should become part of the policy-maker's toolbox, as they enable us to challenge some idealizations and to capture a kind of complexity that is not easily tackled using analytical models (Scharnhorst et al. 2012; Conte and Paolucci 2011; Paolucci et al. 2013).

An already classic example of the agent-based simulation approach to the study of science is presented in (Gilbert 1997), where the author succeeds into finding a small number of simple, local assumptions (the model), with the power to generate computational results (the data), which, at the aggregate level, show some interesting characteristics of the target phenomenon—namely, the specialty structure of science and the power law distribution of citations among authors (the interpretation).

A simulation with simple agents is performed in Thurner and Hanel (2011), where the authors propose an optimizing view of the reviewer for his or her own advantage. To this purpose, they define a submission/review process that can be exploited by a *rational cheating* (Callahan 2004) strategy in which the cheaters, acting as reviewers, reject papers whose quality would be better than their own. In that model, the score range for review is very limited (accept or reject) and in case of disagreement (not unlikely because they allow only two reviewers per paper), the result is completely random. They find out that a small number of rational cheaters quickly reduces the process to random selection. The same model is expanded by Roebber and Schultz (2011), focusing not on peer review of papers but on funding requests. Only a limited amount of funding is available, and the main focus is to find conditions in which a flooding strategy is ineffective. The quantity of cheaters, differently from this study and from Thurner and Hanel (2011), is not explored as an independent variable. The main result obtained is a strong dependence of results from the mechanism chosen (e.g. number of reviews, unanimity, etc.). In Grimaldo et al. (2012), the authors introduce a larger set of scores and use three reviewers for paper; they analyze the effect of several left-skewed distributions of reviewing skill on the quality of the review process. They also use a disagreement control method for programme committee update in order to improve the quality of papers as resulting from the review process (Grimaldo and Paolucci 2013).

A similar approach has shown that there is a quantitative, model-based way to select among candidate peer-review systems. Allesina (2012) uses agent-based modeling to quantitatively study the effects of different alternatives on metrics such as speed of publication, quality control, reviewers' effort and authors' impact. As a proof-of-concept, it contrasts an implementation of the classical peer review system adopted by most journals, in which authors decide the journal for their submissions, with a variation in which editors can reject manuscripts without review and with a radically different system in which journals bid on manuscripts for publication. Then, it shows that even small modifications to the system can have large effects on these metrics, thus highlighting the complexity of the model.

Other researchers have considered reviewer effort as an important factor and have studied its impact on referee reliability. Their results emphasize the importance of homogeneity of the scientific community and equal distribution of the reviewing effort. They have also shown that offering material rewards to referees tends to decrease the quality and efficiency of the reviewing process, since these might undermine moral motives which guide referees' behavior (Squazzoni et al. 2013).

Agent-based simulation and mechanisms

Agent-based simulation, even if its application to this class of problems (modeling science, and peer review in particular) is starting to spread, is still a controversial approach. As its practice grows, its armor begins to show the chinks. These include a tendency to build ad-hoc models (Conte and Paolucci 2011), the inability to do proper sensitivity analysis for changes in the process as opposed to changes in parameters, the risk of overfitting, and oversimplification. All these issues are amplified by the self-referentiality of the simulation community, and from the difficulty to actually define what is a result in silico.

To frame the problem, we can consider the practice of simulation as a discipline under pressure from two opposing forces. The first is the tendency to simplify to the extreme, which can produce models that are, at best, only incremental in complexity with respect to an equation-based model; at worst, models that could actually be easily converted into equation-based ones, applying game theoretical approaches or mathematical techniques as the master equation of mean field theory (see Helbing 2010, for a review). The second force is the pressure for cognitive modeling, coming from researchers interested in sociology, social psychology, and from system designers interested in cognitively-inspired architectures. They have an interest in devising complex architectures that, while plausible at the micro level as inspired by complete representation of cognitive artefacts and folk psychology, are prone to overfitting and difficult to verify.

Our view is that the first force is destructive for the agent-based field, as it is simply inclined towards its reinclusion in the sphere of equation-based treatment. This is not to be considered harmful, especially as it would cause an expansion of the technical and explanatory power of the equation-based field, but could lead to failure by oversimplification. Moreover, discretization often produces results that are different, for example, from those produced by mean field approximation (see Edwards et al. 2003), even for simple models. The path of simplification is attractive and has long been promoted by researchers in the field with the Keep it simple, stupid! (*KISS*) slogan (Axelrod 1997, p. 5). We see a twofold problem here. First, a *KISS* approach contrasts with a descriptive approach, as argued in Edmonds and Moss (2005). But more than losing part of the descriptive approach, a simplification mindset tends to drive out from simulations the part that characterizes them most: the ability to play with mechanisms (Bunge 2004), in the form of processes or algorithms, instead of playing just with distributions and parameters. We are hardly suited to propose a solution of this tension between the simple and the complicated in the field of simulation. We can only suggest that the focus on *mechanismic* (Bunge 2004) explanation could compensate the attraction towards simple models, and propose—also thanks to tools that support that focus—an example of how different results could be, if the employed mechanisms are different.

A *mechanism* is originally defined as a process in a concrete (physical, but also social) system. Focusing on the mechanism implies paying attention on how the components of a system interact and evolve. A mechanism in a system is called an *essential* mechanism when it is peculiar to the functioning of that system, in contrast to both mechanisms that are common to other systems, and mechanisms that do not contribute to system definition, that is, that could be modified or replaced without losing the specificity of the system. As the modelist will immediately recognize, an essential mechanism is the one that should be preserved through the process of abstraction. A mechanism, as Bunge contends, is essential for understanding; it is essential for control.

In the approach that we propose, modeling for agent-based simulation is a matter of finding out, by abstraction and by conjecture, both the variables and the processes that characterize the individual agents. Thus, agent-based simulation is the ideal ground for experimentation with mechanisms, that we are going to perform in the rest of this paper. A warning, however, must be issued on our algorithmic interpretation of mechanisms, something that Bunge would not appreciate: he states that "algorithms are ... not natural and lawful processes ... can only imitate, never replicate, in silico some of what goes on in vivo" (Bunge 2004, p. 205). Nevertheless, in the context of the modeling activity, we maintain that our proposal improves on what could be seen as just an algorithmic variation, if only for the reason that what we propose is indeed algorithmic variations, but in the context of a modeling approach that leans on individual-oriented, plausible micro foundations.

A replication of peer review simulation

The tension between simple models, that lay open to mathematical modeling of some sort, and richer models, that reflect a larger part of the reality they want to describe but do not lend themselves to mathematical formalization, and thus lose in generality, characterizes the simulative approach from its inception. In this work, we wish to add another dimension to this tension. We show how a simple model, whose results we reproduce, is stable enough to be replicated with a completely different approach; however, when we add freedom in the mechanisms, rather than on parameters only, the results may change qualitatively.

Specifically, we compare results from a simple simulation of peer review to results obtained from a belief–desire–intention (BDI)-based one. The BDI model of rational agency has been widely relevant in the context of artificial intelligence, and particularly in the study of multi-agent systems (MASs). This is because the model has strong philosophical assumptions such as: the intentional stance (Dennett 1987), the theory of plans, intentions and practical reasoning (Bratman 1999) and the speech acts theory (Searle 1979). These three notions of intentionality (Lyons 1997) provide suitable tools to describe agents at an appropriate level of abstraction and, at the level of design, they invite to experiment different mechanisms.

The simple-agent simulation we take as our starting point is the one proposed in Thurner and Hanel (2011). After replicating it with our BDI model, we explore some algorithmic variations that are inspired by some of the models examined above in the state of the art. In Fig. 1 we present the sources that will influence our model in the rest of the work in a schematic form.

Replication is considered at the same time a neglected and indispensable research activity for the progress of social simulation (Wilensky and Rand 2007). This is because models and their implementations cannot be trusted in the same way as we trust mathematical laws. On the contrary, trust in models can only be built by showing convergence of different approaches and implementations. Replication is the best solution to obtain *accumulative* scientific findings (Squazzoni 2012, chap. 4.2).

Here, we pursue a kind of qualitative matching that goes beyond both replication and docking (i.e. the attempt to produce similar results from models developed for different purposes, Wilensky and Rand (2007)). Once qualitative congruence is obtained, a successful replication should demonstrate that the results of an original simulation are not driven by the particularities of the algorithms chosen. Our work shows that the original result is replicable but fragile.

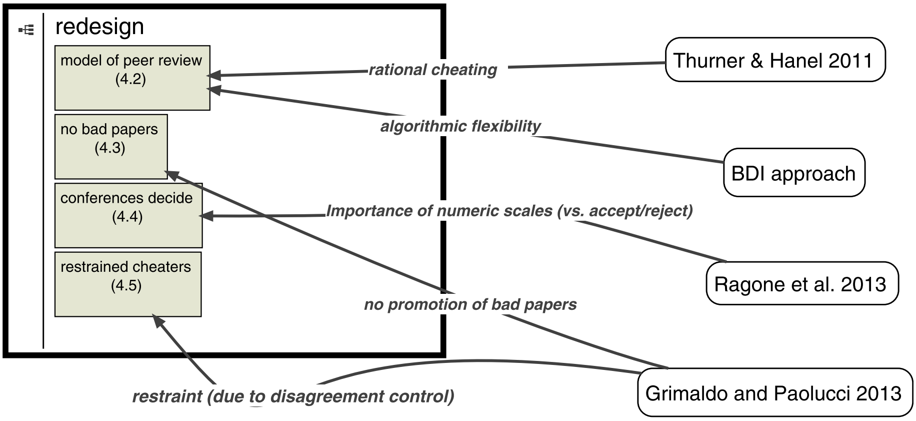
redesign
model of peer review (4.2)
no bad papers (4.3)
conferences decide (4.4)
restrained cheaters (4.5)
rational cheating
algorithmic flexibility
Importance of numeric scales (vs. accept/reject)
no promotion of bad papers
restraint (due to disagreement control)
Thurner & Hanel 2011
BDI approach
Ragone et al. 2013
Grimaldo and Paolucci 2013

Fig. 1 Influences on our peer review model. The different configurations inside the redesign box report the section number where the relative experiment will be discussed

The rest of the paper is organized as follows: the next section presents a new agent-based model of peer review that allows to flexibly exchange the mechanisms performed by the entities involved in this process; "The model in operation" section shows the model in execution and describes the metrics we obtain from it regarding the number and quality of accepted papers; in "Results and comparison" section we present a qualitative replication of the peer review simulation by Thurner and Hanel (2011) and we analyze how the original results appear to be fragile against small changes in mechanisms. "Discussion" section compares the findings obtained in the different scenarios. Finally, "Conclusion" section summarizes the general lessons learned from the proposed redesign and identifies directions for future work.

The model

In this section, we define the entities involved in the peer review process and propose a new agent-based model to describe its functioning.

Peer review entities

The conceptualization of the peer review process presented in this paper identifies the following three key entities: the *paper*, the *scientist* and the *conference*. The *paper* entity is the basic unit of evaluation and it refers to any item subject to evaluation through a peer review process, including papers, project proposals and grant applications. In the present work, we focus on reviewer bias in peer review, and in particular on the bias caused by rational cheating. The simplest way to represent it is to attribute an intrinsic *quality* value for each paper, that readers (and more specifically, reviewers) can access; however, factors other than quality may contribute to evaluation, constituting a bias on the reviewers' side. We are aware that we are compressing on a single value the multifaceted and different areas—Bornmann et al. (2008) dealt with relevance, presentation, methodology and results—that compose scientific merit. Though, we consider this approximation suitable for the present purposes, leaving for future expansions the consideration of multidimensional factors such as topic, technical quality and novelty as they are used, for example, by Allesina (2012).

Scientist entities write papers, submit them to conferences (as we define them below) and review papers written by others. Regarding paper creation, the quality value of a paper will depend on the scientific and writing skills of the authors. Scientists will also be characterized by the reviewing strategies adopted by them during the evaluation process, such as the competitor eliminating strategy used by rational cheaters in Thurner and Hanel (2011), which we will use as a source of bias in the present model.

The *conference* entity refers to any evaluation process using a peer review approach. Hence, it covers most journal or conference selection processes as well as the project or grant evaluations conducted by funding agencies. Every paper submitted to a conference is evaluated by a certain number of scientists that are part of its programme committee (*PC*). Although work dealing with the predictive validity of peer review has questioned the validity of judgments prior to publication (Smith 2006), it has also pointed out the lack of alternatives; in our simulation, by assuming a single a priori value for paper quality, we defend the validity of the ex-ante evaluation of the potential impacts of a paper, as opposed to the ex-post process of counting citations for papers (see Jayasinghe et al. 2003, for a more detailed discussion).

The conference is where all the process comes together and where a number of questions arise, such as: what effects do different reviewing strategies produce on the quality of accepted papers? The peer review model detailed below is meant to tackle this kind of questions by concretizing the different issues introduced for the general entities presented above.

Proposed model

The proposed model represents peer review by a tuple $\langle S, C, P \rangle$, where S is the set of *scientists* playing both the role of authors that write papers and the role of reviewers that participate in the programme committee of a set C of *conferences*. *Papers* (P) produced by scientists have an associated value representing their intrinsic value, and receive a review value from each reviewer. These values are expressed as integers in an N-values ordered scale, from strong reject (value 1) to strong accept scores (value N); we choose such a discrete scale to mirror the ones traditionally used in reviewing forms (i.e., from full reject to full accept through a sequence of intermediate evaluations).

Conferences $c \in C$ are represented by a tuple $\langle PC, rp, pr, av, uw \rangle$. To perform the reviewing process, conferences employ a subset of scientists $PC \subseteq S$, initially chosen at random, as their programme committee, whose size is determined on the base of three parameters: the amount of received papers in the specific year, the number of required reviews per paper rp, and the number of reviews asked per PC member pr (see algorithm 1 for details). Then, those papers whose average review value is greater than the minimum acceptance value av are accepted. In accordance with Thurner and Hanel (2011), to model the adaptation of a conference request to the level of papers submitted to it, this threshold paper quality is updated as shown in Eq. 1, where uw is an update weight and $avgQuality_{year\text{-}1}$ indicates the average quality of the papers accepted by the conference in the previous edition.

$$av_{year} = (1 - uw) * av_{year-1} + uw * avgQuality_{year-1} \tag{1}$$

In algorithm 1 we show the pseudocode executed when celebrating a conference. First, function `CallForPapers` in line 3 broadcasts the conference call for papers and receives papers submitted during a fixed period of time. Then, the function `AdjustPC` in line 4, if the number of reviewers in the initial PC is insufficient in face of the received contributions, enlarges it temporarily with randomly chosen scientists. Subsequently, the *for* statement starting in line 5 initiates the evaluation process: the function

AskForReviews returns the reviews from *rp* reviewers, distinct from the author and randomly chosen from the PC; after calculating the average of those values (see function ComputeAvgReview), lines 8–2 deal with acceptance. Ties, that is, papers with an average review value exactly equal to the acceptance value are randomly accepted or rejected. Lines 22 and 23 respectively compute the average quality of accepted papers as well as the new acceptance value for the next year as for Eq. 1. Finally, accept and reject notifications are sent to the authors by the procedures NotifyAccepts and NotifyRejects.

Algorithm 1 Pseudocode to celebrate a conference.

```
Input: Celebration Year ($year$), Conference Acceptance Value ($av$), Scientists ($S$), Reviews
    Per Paper ($rp$), Papers Per Reviewer ($pr$)
 1: $AccPapers \leftarrow \phi$
 2: $RejPapers \leftarrow \phi$
 3: $RcvPapers \leftarrow CallForPapers(year, av)$
 4: $PC \leftarrow AdjustPC(S, \lceil |RcvPapers| * rp/pr \rceil)$
 5: for all $p$ such that $p \in RcvPapers$ do
 6:     $Reviews \leftarrow AskForReviews(p, rp, PC)$
 7:     $avgReviewValue \leftarrow ComputeAvgReview(Reviews)$
 8:     if $avgReviewValue > av$ then
 9:         $AccPapers \leftarrow AccPapers \cup \{p\}$
10:     else
11:         if $avgReviewValue < av$ then
12:             $RejPapers \leftarrow RejPapers \cup \{p\}$
13:         else
14:             if $Random() \geq 0.5$ then
15:                 $AccPapers \leftarrow AccPapers \cup \{p\}$
16:             else
17:                 $RejPapers \leftarrow RejPapers \cup \{p\}$
18:             end if
19:         end if
20:     end if
21: end for
22: $avgQuality \leftarrow CalculateAvgQuality(AccPapers)$
23: $newAv \leftarrow (1 - uw) * av + uw * avgQuality$
24: $NotifyAccepts(AccPapers)$
25: $NotifyRejects(RejPapers)$
```

Every scientist $s \in S$ is represented by a tuple $\langle ap, aq, as, rs \rangle$. Regarding paper production, each scientist has an associated author productivity *ap*, representing the number of papers uniformly written per year. Produced papers are of the form $p = \langle a, iv \rangle$, being $a \in S$ the author of the paper and $iv \in \{1, .., N\}$ the intrinsic value (quality) of the paper. This intrinsic value is calculated considering the author quality $aq \in \{1, .., N\}$ and the author skill value $as \in [0,1]$. Whereas *aq* represents the standard author quality, *as* represents his/her reliability. Hence, while scientists as authors normally (that is, with probability *as*) write papers of value *aq*, once in a while (with probability $1 - as$) they produce something different. In that case, the intrinsic value is a random integer in the full $\{1, .., N\}$ scale, so that all agents can occasionally produce papers of any value. Once a paper is available, it is submitted to the first succeeding conference announcing its call for papers. After submission, the review process takes place. In the following, we present two possible scenarios for evaluation. In the first one, agents tell the value of the reviewed paper to the conference, where the decision takes place. In the second one, on the contrary, the reviewers decide for a “yes or no” evaluation, and the conference simply counts the votes. Let’s now see the two scenarios and detail what cheating means in each of them.

Valued review scenario

As a reviewer, each scientist follows a reviewing strategy (*rs*) that drives its evaluation behavior. We simulate two reviewing strategies, namely *Correct* and *Rational Cheating*. The *Correct* strategy corresponds to fair behavior, simply reporting the actual quality of the paper (i.e. its internal value *iv*) when an evaluation is requested.

The *Rational Cheating* strategy, on the contrary, lies about those papers whose intrinsic value is greater or equal than the quality of the reviewer when he behaves as an author. Thus, it attempts to clear the way for its own papers—by preventing better papers to appear.

In a scenario in which reviewers communicate a quality value representing their evaluation of the paper, cheating will be performed by tweaking this value. The cheating strategy that we implemented is presented in the pseudocode of algorithm 2. The desired cheating value is calculated so that the final average review value is (just) under the acceptance value of the conference, provided the rest of the reviewers give the *Correct* score (see line 4). That is, the cheating value is first calculated as the highest integer that satisfies Eq. 2.

$$\frac{cheatingValue + (rp - 1) * iv}{rp} < av \tag{2}$$

In addition, for increased plausibility, the rational cheater agent ensures that the returned review value will neither be greater than the actual quality of the paper (i.e. its internal value *iv*) nor too far from the actual value. This requires an additional parameter *rd*, referring to the maximum permitted distance between the intrinsic value of the paper and the review value communicated by the reviewer (see line 6). Thus, the *rd* parameter allows defining scientists that put a limit on their unfair behavior (e.g. to avoid being detected).

Algorithm 2 Pseudocode of the *Rational cheating* reviewing strategy in the valued review scenario.

```
Input: Paper Intrinsic Value ($iv$), Maximum Quality Value ($N$), Reviewer's Author Quality
    ($aq$), Conference Acceptance Value ($av$), Maximum Review Distance ($rd$)
Output: Review Value for the paper ($reviewValue$)
 1: if $iv < aq$ then
 2:     $reviewValue \leftarrow iv$
 3: else
 4:     $cheatingValue \leftarrow \lceil rp * av - (rp - 1) * iv - 1 \rceil$
 5:     $maxValue \leftarrow iv$
 6:     $minValue \leftarrow max(iv - rd, 1)$
 7:     if $cheatingValue > maxValue$ then
 8:         $reviewValue \leftarrow maxValue$
 9:     else
10:         if $cheatingValue < minValue$ then
11:             $reviewValue \leftarrow minValue$
12:         else
13:             $reviewValue \leftarrow cheatingValue$
14:         end if
15:     end if
16: end if
```

Accept/reject review scenario

The model can easily be tuned to evaluate a scenario in which reviewers only communicate a recommendation to accept or to reject. This amounts to using a review scale defined by two values. This is the scenario presented also by Thurner and Hanel (2011).

In the Accept/Reject scenario, reviewers give a review value of 2 when they want a paper to be accepted and a review value of 1 when they want it to be rejected. To this purpose, they must take into consideration the current acceptance value (*av*) of the conference (while in the previous scenario, where they were expected to communicate only the value, they could ignore it—unless they were cheating).

Accordingly, conferences have been modified to accept those papers whose average review value is greater than the mid-range 1.5, whereas those with an average review value exactly equal to 1.5 are randomly accepted or rejected. With respect to scientists, algorithm 3 shows the pseudocode executed when reviewing a paper using the *Correct* strategy. The pseudocode used for reviewing a paper using the *Rational cheating* strategy is shown in algorithm 4. Here, reviewers give the accept recommendation (i.e. 2) only if the quality of the paper is within a minimum quality value (*qmin*) and the reviewer's author quality (*aq*), otherwise they answer with the reject value (i.e. 1). Note that the minimum quality value (*qmin*) is different from the conference acceptance value (*av*) as it is meant only to define an acceptance range [*qmin*, *aq*].[6]

Algorithm 3 Pseudocode of the *Correct* reviewing strategy in the Accept/Reject scenario.

Input: Paper Intrinsic Value (*iv*), Conference Acceptance Value (*av*)
Output: Review Value for the paper (*reviewValue*)

```
if $iv < av$ then
    $reviewValue \leftarrow 1$
else
    $reviewValue \leftarrow 2$
end if
```

Algorithm 4 Pseudocode of the *Rational cheating* reviewing strategy in the Accept/Reject scenario

Input: Paper Intrinsic Value (*iv*), Reviewer's Author Quality (*aq*), Minimum Quality Value (*qmin*)
Output: Review Value for the paper (*reviewValue*)

```
if $(qmin \leq iv)$ AND $(iv < aq)$ then
    $reviewValue \leftarrow 2$
else
    $reviewValue \leftarrow 1$
end if
```

The model in operation

The proposed peer review model has been implemented as a MAS over Jason (Bordini et al. 2007), which allows the definition of BDI agents using an extended version of AgentSpeak(L) (Rao 1996). AgentSpeak(L) is an abstract programming language based on

[6] Rational cheaters in Thurner and Hanel (2011) also use an acceptance range of [90, *RevQlty*], being 100 the average paper value.

a restricted first order logic dealing with events and actions, and represents an elegant framework for programming BDI agents. In turn, Jason is an interpreter written in Java, which conforms to the language AgentSpeak(L). The main idea we need to understand about the BDI agency model is that we can talk about computer programs as if they have a mental state. When it comes to our peer review model, we are talking about an implementation with computational analogies of beliefs, desires and intentions (e.g. the available papers, the desire to publish them and the reviewing strategies a referee could follow).

Thus, our MAS represents both scientists and conferences as agents interacting in a common environment. The environment handles the clock system and maintains the agents' belief bases. As every agent lives in its own thread, the system runs in a (simulated) continuous time. Thus, agents can concurrently react to the passage of time by triggering different plans such as that of writing new papers or celebrating a new edition of a conference. Communication between conferences and scientists takes place within these celebrations: conferences broadcast their call for papers, which cause scientists to decide whether to submit their available papers; reviewers from the PC are asked for reviews of papers; and authors are notified about the acceptance or rejection of candidate papers.

The implemented MAS is highly configurable; the number and characteristics of both conferences and scientists can be independently set, following different statistical distributions (e.g. uniform, normal, beta...). Thus, the MAS can be configured to run different experiment settings and evaluate the effects of the parameters in the proposed peer review model.

The results shown in the rest of the paper have been obtained by running a set of simulations involving 1,000 scientists that try to publish on ten equivalent conferences per year, in a time frame of 40 years. Experiments have been replicated five times for each value of the parameter space. Scientists write two papers uniformly distributed over the year ($ap = 2$), so that the overall production amounts to 2,000 papers per year. Paper intrinsic values are expressed as integers in a 10-values ordered scale ($N = 10$) and author qualities (aq) follow a (stretched and discretized) Beta distribution with $\alpha = \beta = 5$. The beta distribution is the obvious choice for a statistic in a fixed interval as the one we are using—the alternative being a normal distribution with cut tails, much less flexible, for example, in terms of central value. We chose this shape, a bell shaped curve with mean 5.5 and symmetrically distributed between one and ten, in the hypothesis that average papers are more common than either excellent or bogus papers. Author skills (as) follow a uniform distribution in [0.5,1], that we consider a moderate level of noise in the production of papers. In turn, conferences use an initial set of 200 reviewers as their PC, where each member reviews 2 papers at maximum ($pr = 2$) and each paper gets two reviews ($rp = 2$). Acceptance values are initially set to the mean quality value ($av = 5.5$) and are adjusted using a 10 % of update weight ($uw = 0.1$) as explained in Eq. 1.

For each run, we measure the number and quality of accepted papers over time. In the example shown in Fig. 2, in which we show a sample run produced by the valued scenario (see "Valued review scenario" section) with 10 % of the agents following the rational cheating strategy, the quality of accepted papers, that starts rather low, grows in time until it reaches a plateau in 2030, around which it oscillates.[7]

[7] The code for our implementation is available at http://www.openabm.org/model/4025.

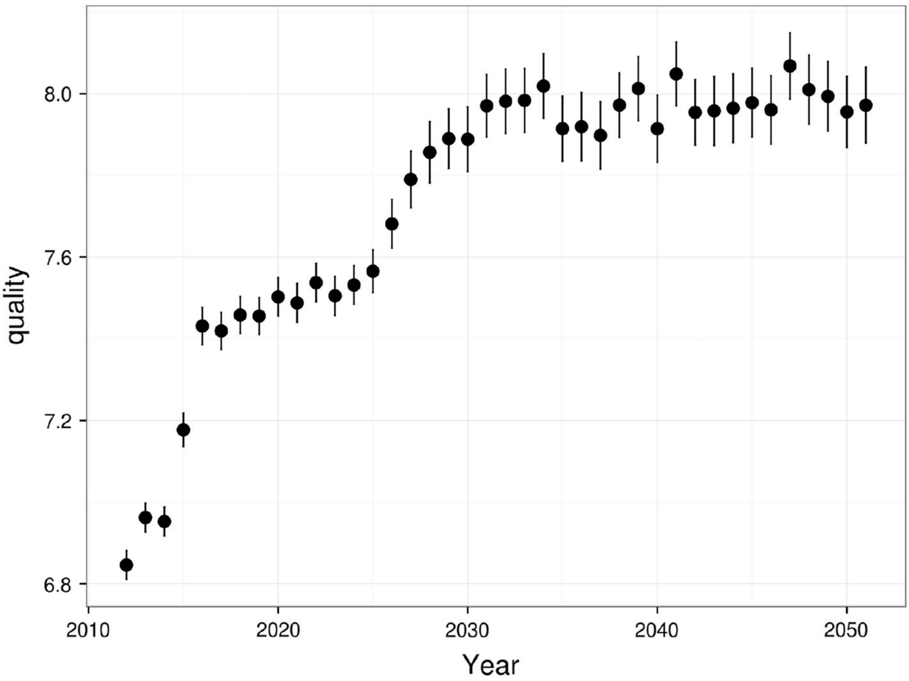

Fig. 2 Evolution in time of accepted papers quality with a fixed amount (10 %) of rational cheaters. Average quality with *error bars*

Results and comparison

Redesign

We now apply our model to replicate the results presented by Thurner and Hanel (2011). For the sake of brevity, we will call it the TH-HA model from the initials of the authors. Our purpose is to discover if our model is capable of reproducing qualitatively the results obtained, without employing the same processes and algorithms. While different algorithms are mentioned in Wilensky and Rand (2007) as one of the potential dimensions on which replications differ from originals, they had in mind technical details as search algorithms or creation order. We use this dimension in a wider sense, more similar to Bunge's mechanisms (Bunge 2004): what we are going to explore are alternative recipes for peer review as inspired from real world observation, an thus involved in the analogic relation between model and object. Thus, instead of simply trying to reproduce numerically the results from the TH-HA model, we will adapt the parameters of our model (as presented in "Proposed model" section to the target of replication, but we will maintain as much as possible the logical flow and processing of our model.

Exploring parameter spaces is already a complicated task, and rightfully enough, there is substantial concern in the agent-based simulation community on its application, as testified, between others, from the recommendations included in the recipe for social simulation presented by Gilbert and Troitzsch (2005), which include validation and sensitivity analysis, to the emphasis given to replication (Edmonds and Hales 2003).

The approach proposed here is different; we are going to perform a kind of validation that doesn't just explore the space of parameters but compares different mechanisms as models of the same target phenomenon. Thus, we perform a simulation that aims to find comparable indication through different mechanisms; we try to align two different models, hoping to see which conclusions are reinforced, and which ones obtain different indications. More than a model replication, we could call this process model *redesign*.

In the specific, we have two models that start from different assumptions and use different techniques; while our model is inspired by cognitive, descriptive modeling, and makes use of a multi-threaded, BDI-based MAS (Bordini et al. 2007), the model by Thurner and Hanel (2011) applies techniques from the physics-inspired simulation world. Having being developed independently, the two models have in common only the target—the real world equivalent—but might differ in fundamental choices regarding what is retained and what is abstracted away.

We believe that this kind of confirmation is actually stronger than a simple replication, and, as we will see in the following, also a good way to point out which results are dependent on the specificity of the mechanisms chosen. The approach helps also to avoid ad-hoc modeling, thus reinforcing the value of generative explanations (Conte and Paolucci 2011).

Replication of the TH-HA model

In our attempt to replicate qualitatively the results from Thurner and Hanel (2011), we start with the configuration that resembles more closely the one proposed in the original paper. Thus, we run experiments varying the percentage of rational cheaters from zero to 90 %; each paper gets two reviews ($rp = 2$), and the reviewers give simply accept/reject scores (following the scenario described in "Accept/Reject review scenario" section), decided on the basis of their (always accurate) evaluation of the papers. As mentioned above, papers' intrinsic qualities take values in a scale of integers from 1 to 10 included ($N = 10$). The only source of noise here is the variable quality of papers (depending from the quality of authors in the way explained above in "Proposed model" section). As rational cheaters in Thurner and Hanel (2011) accepted a few low quality papers,[8] also here they accept papers that have a quality between 4 ($qmin = 4$) and their own author quality. We expect, in accordance with the original paper, to see a marked decrease of quality with the growth of cheaters—perhaps even worse than completely random acceptance, that in our scale is expressed by a *5.5* quality of accepted papers.

A set of categories for replication are suggested in Wilensky and Rand (2007), and we report the choices made in this work in Table 1.

We let the system evolve for 40 years, and present the averaged results for the last ten years (averaging through the last five, or just for the last year shows the same pattern; generally, as from Fig. 2, the simulation reaches a stable state in the last 10 years). The results, as shown in Fig. 3, confirm the expectation that the presence of rational cheaters would cause an initial steep drop in quality of nearly one point, followed by what looks like a descent, convex first then linear, to essentially random quality, reached with 80 % of rational cheaters.

We consider this result as a successful qualitative reproduction of Thurner and Hanel (2011). In the ideal situation described by our model with the parameters above, a small

[8] In th-ha a paper is accepted by a rational cheater when the quality is between 90 andthe quality of the author, while the minimum accept value in the initial turn should havebeen 100 (and it grows thereafter).

Table 1 Replication standards

Categories of replication standards	Approach chosen
Focal measures	Average quality of accepted papers
Level of communication	Brief email contact (we asked for confirmation of what revealed to be a typo in one of the formulas; the authors answered immediately)
Familiarity with language / toolkit of original model	None (no toolkit was specified in the target work)
Examination of source code	None
Exposure to original implemented model	None
Exploration of parameter (and mechanism) space	We expanded the exploration to different mechanisms, instead of to parameters.

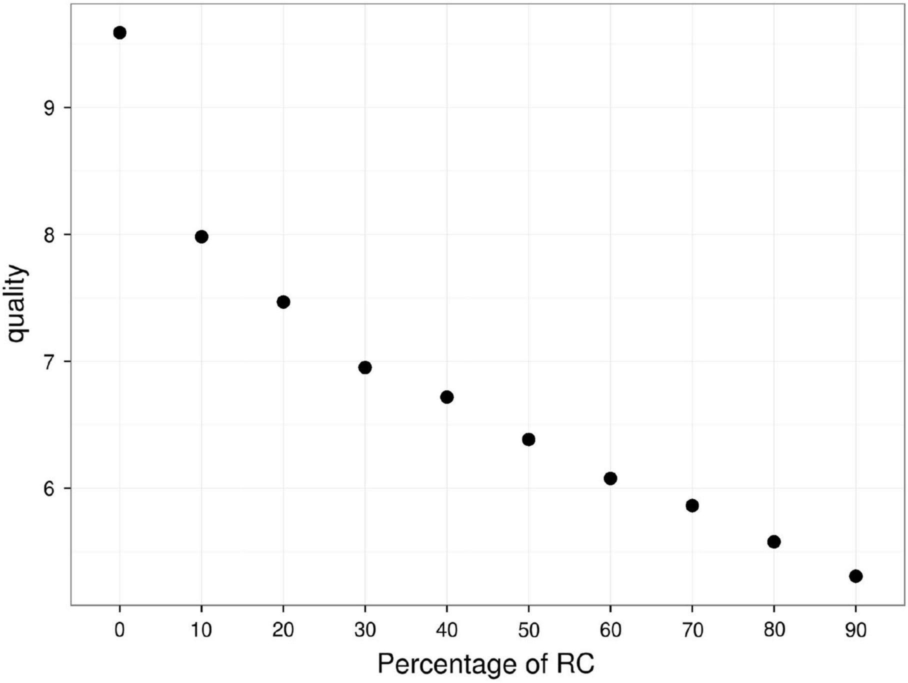

Fig. 3 Replication scenario; average quality (with *error bars*, barely visible in the plot) of accepted papers by percentage of rational cheaters averaged on the last ten years of simulation. Ten percent of rational cheaters cause a steep drop in quality. Results confirm TH-HA qualitatively

quantity of rational cheaters is able to substantially hamper the performance of the system as a whole. The recommendation that would follow, then, is to keep a tight watch against rational cheating in order to suppress it at its inception. But is this indication stable with respect to variations in the parameters - and, what is more important, with respect to variations in the mechanisms we have borrowed from Thurner and Hanel (2011)?

This second question concerns us the most. To answer it, we perform a few modifications of the mechanisms underlying our model: first, the rational cheaters do not push low-level papers ("No bad papers" section); then, the reviewers, instead of just giving a reject/accept evaluation, send out the actual quality score resulting from their review, leaving to the conference the task of averaging them and deciding on the result (see "Conferences decide" section). Both these choices, as we will argue, are supported by plausibility claims. Finally, inspired by the changes observed in the results, we implement another mechanism in which rational cheaters restrict themselves from sending implausible evaluations; results are shown in "Restrained cheaters" section.

No bad papers

In this experimental setting, we employ again the accept/reject scenario with a difference in the mechanism of rational cheaters; they only accept papers between the average value of *5.5* ($qmin = 5.5$) and their own quality as authors. As a consequence, low-quality (4 and 5 on our scale) papers are not inserted in the system, and thus we expect an increase in quality, either by translating the curve up or by changing its shape.

Results (again for the last 10 years in a 40 years simulation) are shown in Fig. 4 by percentage of rational cheaters. With the removal of low quality papers, also the initial marked sensitivity disappears, making the response of the system to the injection of rational cheaters nearly linear. Also the quality of papers remains higher than the average quality even in the worst case, arriving just below seven. The initial sensitivity of the

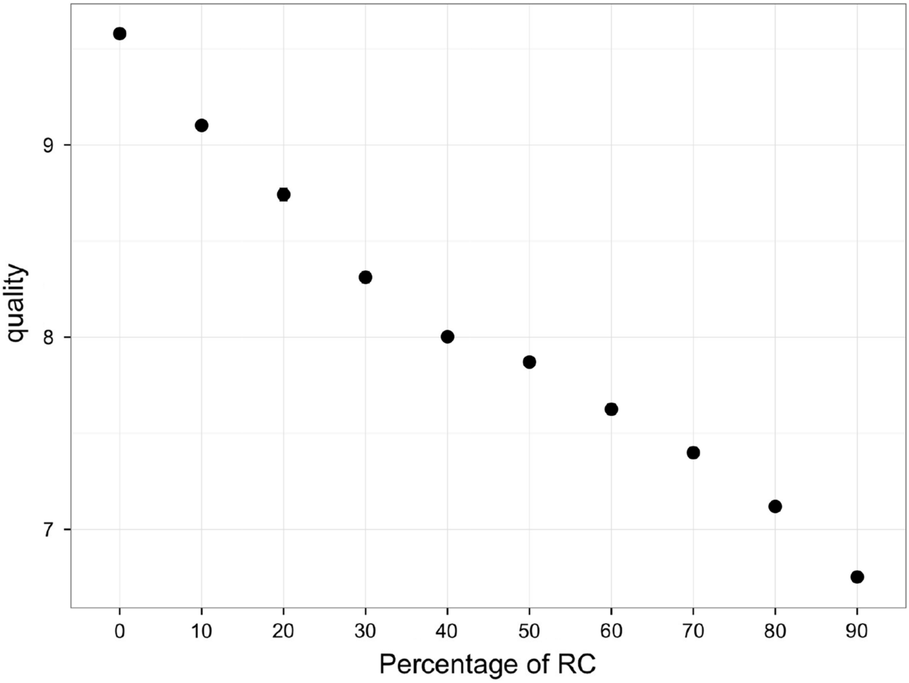

Fig. 4 No bad papers scenario; average quality (*error bars* are not visible at this scale) of accepted papers by percentage of rational cheaters, averaged on the last 10 years of simulation. The decrease in quality is approximately linear. Quality remains higher than in Fig. 3

model to rational cheaters disappears; thus, by comparing this configuration with the previous one (Fig. 3), the indication is that pushing bad papers has a critical role in bringing the system to failure with few rational cheaters. In other words, the model indicates that rational cheaters are fatal for the functioning of peer review only if, in addition to being hostile to papers better than their own, they also promote low-quality papers. If they do not, they remain detrimental, but much less dramatically.

Conferences decide

Let us now proceed to make further changes to the acceptance mechanism. In this section, instead of modifying the preferences of rational cheaters, we modify the review process and give more responsibility to the conference. Following the approach presented by Thurner and Hanel (2011), we have until now modeled reviewers as giving an accept/reject judgment. Being just two reviewers, they end out frequently in ties. Not surprisingly (but somehow conveniently), ties are decided randomly—a mechanism that is due to amplify the effect of rational cheating, because couples of reviewers that include a rational cheater will often end out in ties on good papers.

Ties, however, are naturally eliminated by one of two mechanisms—having three reviewers, which requires additional resources, or stretching the evaluation scale. Why stretching the scale helps? Because when the reviewers pass on to the conference the actual value of their evaluation, a number between 1 and 10 ($N = 10$), the chances of a review converging to the center (average review, *av* value equal to *5.5*) are made negligible. Moreover, extended scales of values are customarily employed in conferences, workshops, and journals, thus making this representation a more accurate micro description (Moss and Edmonds 2005). This algorithm has been described in the model section (see "Valued review scenario" section) with the name of *valued review scenario*. We implement it here and we run a set of simulations comparable to the replication ones.

Results, presented in Fig. 5, put in evidence that the original result indeed was dependent from these random ties. Communicating directly the evaluation value makes the rejection of excellent papers much more difficult, because the cheater will have to throw in unbelievable scores. The performance of the simulated peer review system remains extremely good until the number of rational cheating extends to half of the population or more.

The shape of the curve obtained is concave and not convex, indicating a low sensitivity to the entry of rational cheaters. When conferences decide, the system tolerates rational cheaters, without significant quality decrease, upto about 30 %; then the decrease continues regularly down to 90 %, in a concave shape instead of the convex one seen in the replication. Moreover, the quality of papers remains over the middle point even in the worst case, arriving just below *6.5* for *90 %* rational cheaters. The sharp initial sensitivity of the model to small numbers of rational cheaters completely disappears; thus, by comparing this configuration with the replication (Fig. 3), the indication is that the random allocation of uncertain papers, just as the promotion of bad papers, has a critical role in bringing the system to failure with few rational cheaters.

Restrained cheaters

Up to this point, we have shown how the reaction curve—that is, the curve that summarizes the quality value as more and more rational cheaters enter the system—has a shape that depends on the applied mechanism. While in the replication case this shape qualitatively

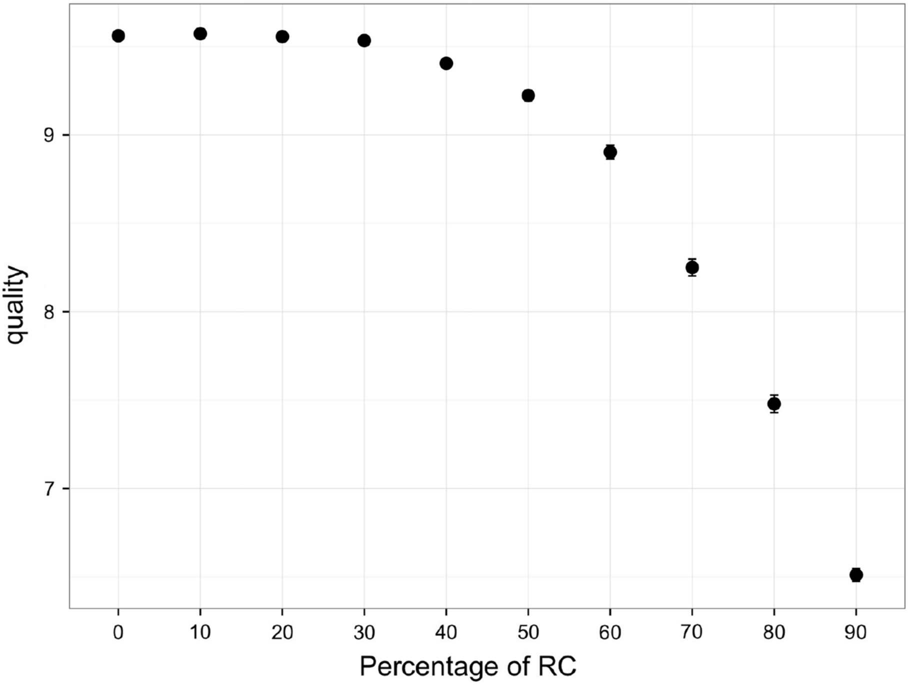

Fig. 5 Conferences decide. Average quality (*error bars* are barely visible at this scale) of accepted papers by percentage of rational cheaters, calculated on the last 10 years. The system tolerates with little reaction upto 30 % or rational cheaters. Peer review performs excellently until the ratio of rational cheaters exceeds 60 %

confirmed the results of the TH-HA model, we found out that two plausible modifications of that mechanism invert the curvature of that shape and remove the strong initial sensitivity; in other words, small algorithmic changes did cause a qualitatively different result.

In this section, we present results from another variation. Until now, we had not taken advantage of the *rd* parameter that controls, so to say, the self restraint of rational cheaters, preventing them from attributing to papers a score that is too distant from the actual one (previous settings deactivated this restraint by setting $rd = 10$). However, issuing reviews that are bound to be in disagreement with other supposedly non-cheating reviewers is risky; while a certain amount of disagreement is unavoidable and perhaps even healthy, giving widely diverging values puts the reviewer to the risk of being detected (Grimaldo and Paolucci 2013). Now we run a set of simulations with the same scenario as the last one ("Conferences decide" section) but we activate this "restraint" mechanism; here we show results obtained for $rd = 5$. In Fig. 6, a surprising result awaits: the trend inverts at starts, rational cheaters causing an slight increase instead of a decrease for the overall quality of the system. How is this possible? Simple enough: in a setting where rational cheaters show restraint, the papers that get accepted are only the excellent ones. The strategy of rational cheaters retorts against them, ending in an *elitist* situation - the mechanism of acceptation locks up so much that normal papers cannot get through, while the very best ones can. Rational agents, designed for blocking papers that are "too good", end up instead in promoting excellence.

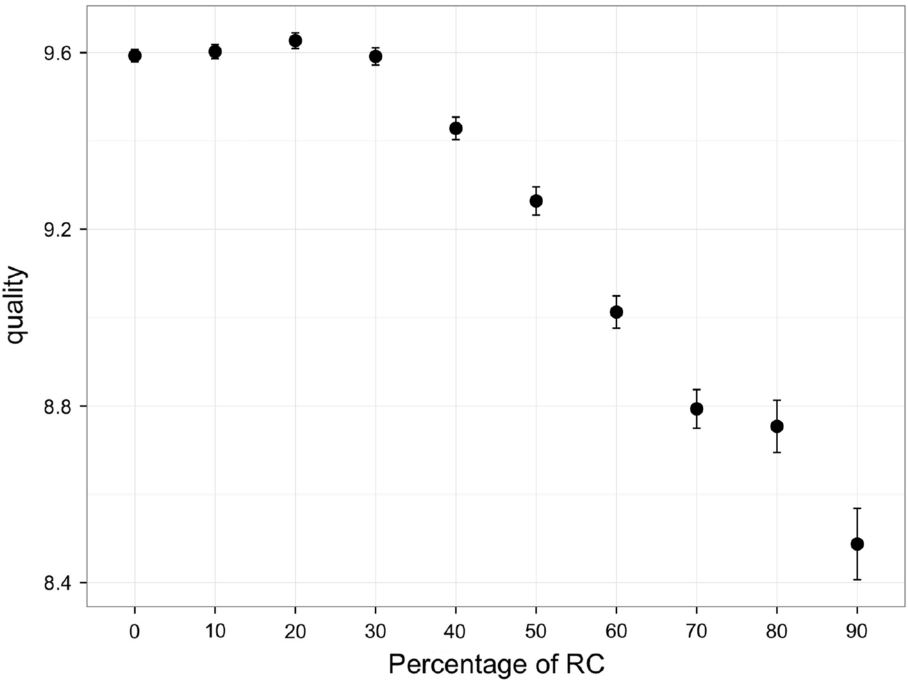

Fig. 6 Results from restrained rational cheaters. Average quality with *error bars* of accepted papers by percentage of rational cheaters, calculated on the last ten years. Surprisingly, quality increases initially with the number of cheaters—the restraint allowing the very best papers to pass, and only those, causing an *elitist effect*

As an example, consider a rational cheater whose author skill is just over the average—*6* in our scores. This reviewer will reject all the papers with a score better than his, by downgrading them upto the limit imposed by *rd*. If this cheater receives a paper of value 7, he will attribute it a value of *3*; being that in the *rd* range, the paper will be rejected with an average score of *5*. If, on the contrary, the received paper has quality 9, the rational cheater would go for a score of 2, but that would conflict with the restraint; thus, it will converge to the lowest restrained evaluation available—a score of 4, which allows the paper to be accepted unless the other evaluator is also a cheater. This initially unforeseen mechanism is the cause of the elitist effect that emerges from our simulation.

Also note that we allow conferences to exist even when they accept only a minimal amount of papers, and this favors the elitist effect. While it is not obvious how this can be compared to TH-HA, which does not have a concept of distinct conferences, the filtering mechanism is being helped by not having a minimum quantity of papers to be accepted accept regardless of their quality. However, we checked that even setting a reasonable minimum number of papers per conference does not change the result in a qualitative way.

Have we, as the authors, been cheating too in finding this mechanism? We let this judgment to the reader; while it is true that this last setting has devised to keep cheating under control, it is also true that the change in the mechanism with respect to the *conferences decide* one is minimal and not implausible. However, this elitist effect should come with a reduction of the number of papers accepted overall. Let's consider this issue next, comparing the number of accepted papers along the four scenarios that we examined.

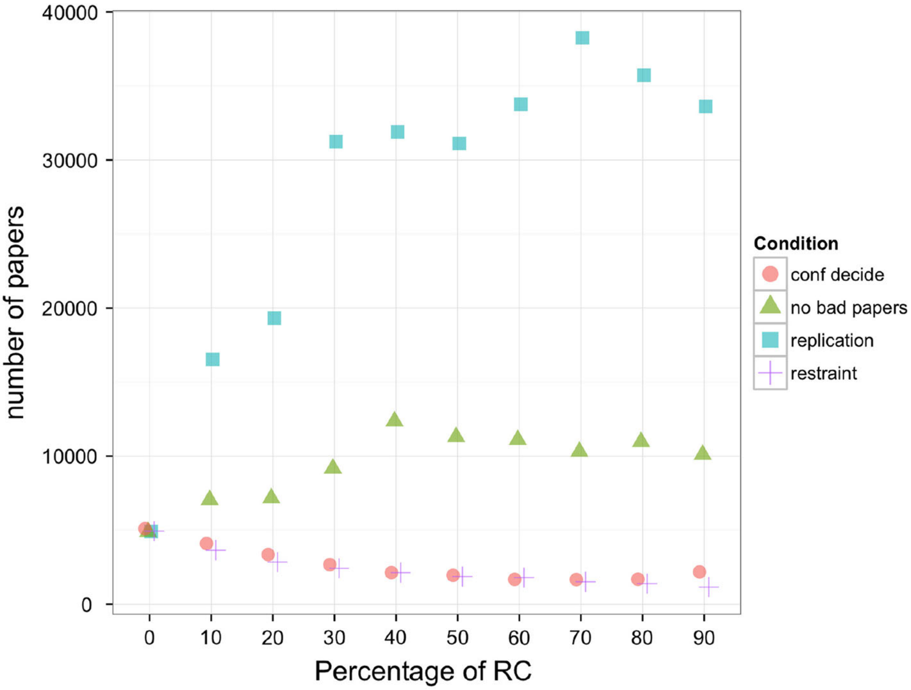

Fig. 7 Number of published papers for different mechanisms. Scenarios that perform better (*restraint* and *conf decide*) allow only few papers to be published. The *no bad papers* scenario is in a middle position; allowing bad papers in in the *replication* scenario increases massively the number of published papers

Number of accepted papers

In the discussion so far, we focused on quality only. What about quantity, that is, how many papers are accepted in each of the different scenarios? Obviously enough: scenarios that accept more papers are bound to exhibit lower quality. In Fig. 7 we present a summary, showing the number of accepted papers by condition. In the *replication* scenario, an initial explosion in the number of accepted papers corresponds to the sudden drop in quality; the scenario where rational cheaters do not push *bad papers* also increases the number of accepted papers, but only moderately when compared with the previous one. Both the *conferences decide* scenario and the *restraint* one, on the contrary, decrease the number of accepted papers with the increase of the rational cheaters ration; this maps, respectively, on the slight decrease and slight increase that we see in Figs. 5 and 6. Regrettably, we do not have indications on the quantity of papers accepted in the TH-HA original formulation, and thus we cannot confirm or disconfirm the validity of our replication from this point of view.

There are two general lessons to be learned here. The simpler one is that averaged quality hides much—if we add measures considering also the number of accepted papers, then it shows immediately how the elitist effect is connected to a decrease in the number of accepted papers. Moreover, the quantity of accepted papers seems to be a good indicator for the quality of peer review with respect to different mechanisms; the quantity of accepted papers is also easier to measure than the quality, the latter being assessed only a posteriori by the number of citations, which is, however, sometimes subject to fashion and fads.

Discussion

The above presented results give indications in two main directions, concerning the object of study, and the method applied; peer review, and agent-based modeling.

For peer review, we have confirmed that unfair play (here, rational cheating), in a first approximation, is likely to impact the performance of the review process rather heavily. However, by modifying the mechanisms employed (see a comparison for all scenarios in Fig. 8), we have also been able to lessen this impact, and even to reverse it when we come to a specific scenario (see "Restrained cheaters" section). This evidence points out to a more complex role of cheating in an evaluation system, ranging from sabotaging effect, to a sort of "useful idiot" role. In turn, this suggests a multi-level approach for containment of cheaters. Indeed, cheating behavior could be directly addressed at the individual level with enforced norms, and/or it could be made harmless by using the opportune mechanism at the collective level.

Concerning its applicability to concrete situations, we are aware that the model presents some important weaknesses; first of all, the attribution of a single quality value to papers, which in this version of the model is directly accessible, without noise, to the reviewers. We nevertheless believe that these weaknesses are justified as first approximations; in

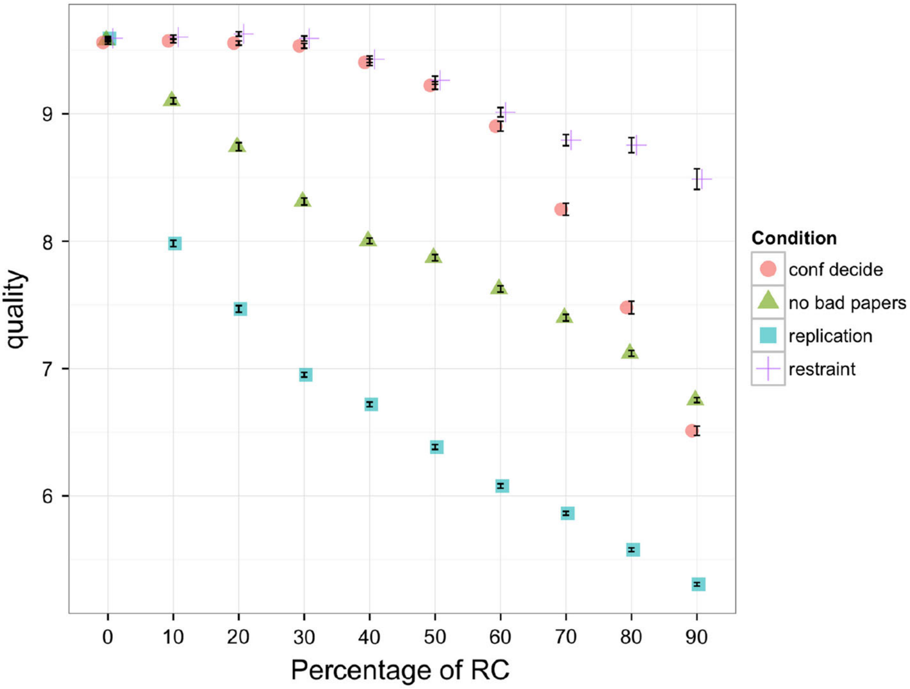

Fig. 8 Average quality of papers for different mechanisms. Only the replication scenario drops sharply for a small amount of rational cheaters

addition, they have been required for comparability with the TH-HA model that inspired us. Removing some of these weaknesses, for example by carrying on validation with data collected from a journal or conference, would be an interesting topic for future studies. However, notwithstanding the extensive informatization of the peer review process, actual data are still hard to obtain due to privacy issues and to uncertain copyright status. Peer review text and evaluations remain, curiously, one of the few online activities that are performed, for the most part, without explicit acceptance of copyright/ownership transfer clauses. As a consequence, published data for validation appears limited and highly aggregated (see Grimaldo and Paolucci 2013 and Ragone et al. 2013 for a discussion).

Should we be reassured and conclude that cheating is manageable? Hardly so. When reading the results under the light of an evolutionary approach, it is not easy to say which situation—between the one in Fig. 3 and in Fig. 5, for example—is the most dangerous. While in the first one the sharp decrease is easily detectable, the second case gives "generous" rational cheating a chance to invade the system being undetected until a large amount of reviewers are turned into rational cheaters, reminding of the cooperative invasions of TFT populations by genetic drift (Nowak and Sigmund 1992). The study of this phenomenon requires an evolutionary simulation, perhaps under group selection, and will be the object of future work.

For what concerns the applied methodology, that is, agent-based modeling, and in particular a BDI approach, we found it to be crucial for letting us focus on the mechanisms as defined above (see "Agent-based simulation and mechanisms" section). Mechanism-based redesign motivated us to explore the variations presented above, pointing out the fragility of the initial result against this class of changes.

In our exploration, we did not check only the mechanisms that we have illustrated in this paper. The results that we have presented are a representative sample of a much larger set of experiments that we have been running, introducing variations in the mechanisms; in general, results follow one of the presented patterns. Other mechanisms that we have tested but not reported for matters of space include using three reviewers ($rp = 3$) or requiring unanimous consensus from reviewers before accepting a paper. Using three reviewers invalidates substantially the Thurner and Hanel (2011) mechanism that applies dice rolling when the result is uncertain (with two yes/no reviews, uncertain results are bound to happen frequently) and that was one of the causes of the sharp drop in Fig. 3. In this paper, this already happens for the scenario where conferences decide ("Conferences decide" section, results in Figs. 5, 6). Requiring unanimity does not change the results qualitatively.

This begs the general questions: is mechanism exploration useful? Is it necessary? Is it feasible? In the case that we present, analyzing the results obtained with different mechanisms has been certainly useful. More in general, we believe it to be necessary for all the simulations aimed to describe society. Indeed, society consists for large part of evolved and/or agreed upon mechanisms. Thus, in order to get a better simulative understanding of society, in the spirit of recent attempts as the FuturICT initiative (Helbing et al. 2012; Paolucci et al. 2013), exploration of alternative mechanisms should necessarily be performed. The last question is perhaps more difficult to answer. Running consistent exploration of the parameter space gives agent-based simulation a hard time - and parameters are, after all, just numbers: what about the much wider space of possible algorithms?

A tentative answer, which is all we can offer here, would touch the difference between all possible mechanisms and the likely ones, perhaps leaning on a micro level plausibility argument. In perspective, simulations that explore different plausible mechanisms, perhaps supported by crowdsourcing mechanisms (Paolucci 2012).

Conclusions

In this paper, we replicated results from Thurner and Hanel (2011) by using a different approach: we employed independent agents (Wooldridge 2009; Bordini et al. 2007), a different structure of multiple conferences , and different quality distributions: we call this approach *redesign*. Notwithstanding those differences, we obtained a clear qualitative replication of the original results.

Once the replication is established, our approach to modeling (with explicit, "intelligent" agents as opposed to spin-like agents) naturally made us want to test the solidity of the result with respect to the employed *mechanisms*. With respect to mechanism change, the result showed surprisingly fragile. Simple, plausible changes in the mechanisms showed that peer review can withstand a substantial amount of cheaters, causing just a graceful decline in total quality. By not favoring the publication of low-quality papers, peer review becomes more robust and less random. By moving from an accept/reject review to a numerical score, hence accepting those papers whose average review value is greater than the acceptance value of the conference, the initial drop disappears as well. Remarkably, a further change that enables a plausible restraint mechanism for cheaters results in an inversion of the tendency, from decrease to increase, generating an unexpected *elitist* effect (see Fig. 8 for a summary of these results).

Our conclusion is then twofold. First, peer review and rational cheating show in our model a complex interaction: depending on the mechanisms employed, it can cause a quality collapse, a graceful decay, or even a slight quality increase (*elitist effect*).

Secondly, we point out mechanism exploration as a key challenge for agent-based modeling and simulation. Especially for social simulation, models should always control for mechanism effect, at least for those mechanisms that appear to be plausible at the micro level, in the description of the agent processes.

Acknowledgments The authors are indebted to Rosaria Conte, Luis Gustavo Nardin, and Federico Cecconi for their advice and encouragement. We would like to offer special thanks to Federica Mattei and Mindy Thoresen-Sertich for their help with the language. The support of two anonymous reviewers greatly improved the clarity of the paper. The publication of this work was partially supported by the PRISMA project (PON04a2 A), within the Italian National Program for Research and Innovation (Programma Operativo Nazionale Ricerca e Competitività 2007–2013), and by the FuturICT coordination action (FP7/2007–2013) under grant agreement no. 284709. It was also jointly supported by the Spanish MICINN and the European Commission FEDER funds, under grant TIN2009-14475-C04.

References

Allesina, S. (2012). 'Modeling peer review: An agent-based approach'. *Ideas in Ecology and Evolution 5*(2), 27–35

Antonijevic, S., Dormans, S., & Wyatt, S. (2012). Working in virtual knowledge: Affective labor in scholarly collaboration. In Wouters P., Beaulieu A., Scharnhorst A., Wyatt S., (eds.), *Virtual knowledge—experimenting in the humanities and the social sciences*. Cambridge: MIT press.

Axelrod, R. (1997). *The complexity of cooperation: Agent-based models of competition and collaboration*, 1st printing edn. Princeton: Princeton University Press.

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009) A principal component analysis of 39 scientific impact measures. *PloS One 4*(6), e6022+.

Bordini, R. H., Hübner, J. F. & Wooldridge, M. (2007). Programming multi-agent systems in AgentSpeak using Jason. Chichester: Wiley.

Börner, K. (2010). *Atlas of science: Visualizing what we know*. Cambridge, Mass: MIT Press.

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science & Technology 45*(1), 197–245.

Bornmann, L. (2013). A better alternative to the h index. *Journal of Informetrics 7*(1), 100+, doi:10.1016/j.joi.2012.09.004

Bornmann, L., & Daniel, H.-D. (2005). Selection of research fellowship recipients by committee peer review. Reliability, fairness and predictive validity of Board of Trustees' decisions. *Scientometrics 63*(2), 297–320.

Bornmann, L. & Daniel, H.-D. (2009). The luck of the referee draw: the effect of exchanging reviews. *Learned Publishing 22*(2), 117–125.

Bornmann, L., Nast, I., & Daniel, H.-D. (2008). Do editors and referees look for signs of scientific misconduct when reviewing manuscripts? A quantitative content analysis of studies that examined review criteria and reasons for accepting and rejecting manuscripts for publication. *Scientometrics 77*(3), 415–432.

Bratman, M. E. (1999). *Intention, plans, and practical reason*. Cambridge: Cambridge University Press.

Bruckner, E., Ebeling, W. & Scharnhorst, A. (1990). The application of evolution models in scientometrics. *Scientometrics 18*, 21–41.

Bunge, M. (2004). How does it work?: The search for explanatory mechanisms. *Philosophy of the Social Sciences 34*(2), 182–210.

Callahan, D. (2004). Rational cheating: Everyone's doing It. *Journal of Forensic Accounting*. pp. 575+.

Camussone, P., Cuel, R. & Ponte, D. (2010). ICT and Innovative Review Models: Implications For The Scientific Publishing Industry. In 'Proceedings of: WOA 2010, Bologna, 16–18 giugno 2010', pp. 1–14.

Cohen, M. R. (1933). Scientific method. In Seligman E. R. A., Johnson A., (eds.), *Encyclopeadia of the social sciences*. New York: MacMillan and Co., pp. 389–386.

Conte, R. & Paolucci, M. (2011). On Agent Based Modelling and Computational Social Science. Social Science Research Network Working Paper Series.

Dennett, D. C. (1987). *The intentional stance (Bradford Books)*. reprint edn, Cambridge: The MIT Press.

Eckberg, D. L. (1991). When nonreliability of reviews indicates solid science. *Behavioral and Brain Sciences 14*, 145–146.

Edmonds, B. & Hales, D. (2003). Replication, replication and replication: Some hard lessons from model alignment. *Journal of Artificial Societies and Social Simulation 6*(4).

Edmonds, B. & Moss, S. (2005). *From KISS to KIDS - An 'Anti-simplistic' Modelling Approach*, Vol. 3415 of Lecture Notes in Computer Science, Berlin: Springer, pp. 130–144.

Edwards, M., Huet, S., Goreaud, F. & Deffuant, G. (2003). Comparing an individual-based model of behaviour diffusion with its mean field aggregate approximation. *Journal of Artificial Societies and Social Simulation 6*(4).

Egghe, L. & Rousseau, R. (1990). *Introduction to informetrics: quantitative methods in library, documentation and information science*. Amsterdam: Elsevier Science Publishers.

Gilbert, N. (1997). A simulation of the structure of academic science. *Sociological Research 2*(2), 1–25.

Gilbert, N. & Troitzsch, K. G. (2005). *Simulation for the Social Scientist*, 2nd edition. Buckingham: Open University Press.

Goffman, W. (1966). Mathematical approach to the spread of scientific ideas—the history of mast cell research. *Nature 212*(5061), 449–452.

Grimaldo, F. & Paolucci, M. (2013). A simulation of disagreement for control of rational cheating in peer review. *Advances in Complex Systems* pp. 1350004+.

Grimaldo, F., Paolucci, M. & Conte, R. (2012). Agent simulation of peer review: The PR-1 model. In Villatoro, D., Sabater-Mir, J., & Sichman, J.S., (eds.), *Multi-agent-based simulation XII*, Vol. 7124 of lecture notes in computer science chapter 1. Springer Berlin / Heidelberg, Berlin: Heidelberg, pp. 1–14.

Helbing, D. (2010). *Quantitative sociodynamics: stochastic methods and models of social interaction processes*. Berlin: Springer.

Helbing, D., Bishop, S., Conte, R., Lukowicz, P. & McCarthy, J. B. (2012). FuturICT: Participatory computing to understand and manage our complex world in a more sustainable and resilient way. *European Physical Journal 214*(1), 11–39.

Hojat, M., Gonnella, J. & Caelleigh, A. (2003). Impartial judgment by the "Gatekeepers" of science: Fallibility and accountability in the peer review process. *Advances in Health Sciences Education 8*(1), 75–96.

Jayasinghe, U. W., Marsh, H. W. & Bond, N. (2003). A multilevel cross-classified modelling approach to peer review of grant proposals: The effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society - Series A - Statistics in Society 166*, 279–300.

Jefferson, T., Alderson, P., Wager, E. & Davidoff, F. (2002). Effects of Editorial Peer Review: A Systematic Review. *JAMA 287*(21), 2784–2786.

Jefferson, T. & Godlee, F. (2003). *Peer Review in Health Sciences*. London: Wiley.

Kostoff, R. N. (1995). Federal research impact assessmentaxioms, approaches, applications. *Scientometrics 2*(34), 163–206.

Kuhn, T. S. (1996). *The structure of scientific revolutions*, 3rd edn, Chicago: University of Chicago Press.

Lamont, M. (2009). *How Professors Think: Inside the Curious World of Academic Judgment*. Cambridge: Harvard University Press.

Lamont, M. & Huutoniemi, K. (2011). Opening the black box of evaluation: How quality is recognized by peer review panels. *Bulletin SAGW 2*, 47–49.

Lyons, W. (1997). *Approaches to intentionality*. Oxford: Oxford University Press.

Marcus, A. & Oransky, I. (2011). Science publishing: The paper is not sacred. *Nature 480*(7378), 449–450.

Moss, S. & Edmonds, B. (2005). Sociology and simulation: Statistical and qualitative cross-validation. *American Journal of Sociology 110*, 1095–1131.

Nowak, M. A. & Sigmund, K. (1992). Tit for tat in heterogeneous populations. *Nature 355*, 250–253.

Paolucci, M. (2012). Two scenarios for Crowdsourcing Simulation. In Paglieri, F., Tummolini, L., Falcone, R. & Micel, M., (eds), *The goals of cognition: Essays in honour of Cristiano Castelfranchi*. London: College Publications.

Paolucci, M., Kossman, D., Conte, R., Lukowicz, P., Argyrakis, P., Blandford, A., Bonelli, G., Anderson, S., Freitas, S., Edmonds, B., Gilbert, N., Gross, M., Kohlhammer, J., Koumoutsakos, P., Krause, A., Linnér, B. O., Slusallek, P., Sorkine, O., Sumner, R. W. & Helbing, D. (2013). Towards a living earth simulator. *The European Physical Journal Special Topics 214*(1), 77–108.

Payette, N. (2011). For an integrated approach to agent-based modeling of science. *Journal of Artificial Societies and Social Simulation 14*(4), 9.

Ragone, A., Mirylenka, K., Casati, F. & Marchese, M. (2013). On peer review in computer science: analysis of its effectiveness and suggestions for improvement. Scientometrics pp. 1–40.

Rao, A. S. (1996). AgentSpeak(L): BDI agents speak out in a logical computable language, in 'Proc. of MAAMAW'96', number 1038 in 'LNAI', Heidelberg: Springer, pp. 42–55.

Roebber, P. J. & Schultz, D. M. (2011). Peer review, program officers and science funding. PLoS One 6(4), e18680+.

Scharnhorst, A., Börner, K. & van den Besselaar, P., eds (2012). *Models of Science Dynamics: Encounters Between Complexity Theory and Information Sciences*. Berlin: Springer.

Schultz, D. M. (2010). Are three heads better than two? how the number of reviewers and editor behavior affect the rejection rate. *Scientometrics 84*(2), 277–292.

Searle, J. (1979). *Expression and meaning: Studies in the theory of speech acts*. Cambridge: Cambridge University Press.

Smith, R. (2006). Peer review: a flawed process at the heart of science and journals. *JRSM 99*(4), 178–182.

Snow, C. P. (2012). *The two cultures*. Cambridge: Cambridge University Press.

Spier, R. (2002). The history of the peer-review process. *Trends in Biotechnology 20*(8), 357–358.

Squazzoni, F. (2012). *Agent-Based Computational Sociology*. Chichester: Wiley

Squazzoni, F., Bravo, G. & Takács, K. (2013) Does incentive provision increase the quality of peer review? an experimental study. *Research Policy 42*(1), 287 – 294.

Squazzoni, F. & Takács, K. (2011). Social simulation that 'peers into peer review. *Journal of Artificial Societies and Social Simulation 14*(4), 3.

Sterman, J. D. (1985). The growth of knowledge: Testing a theory of scientific revolutions with a formal model. *Technological Forecasting and Social Change 28*(2), 93 – 122.

Thurner, S. & Hanel, R. (2011). Peer-review in a world with rational scientists: Toward selection of the average. *European Physical Journal B-Condensed Matter 84*(4), 707.

Wicherts, J. M. (2011). Psychology must learn a lesson from fraud case. *Nature 480*(7375), 7.

Wilensky, U. & Rand, W. (2007). Making models match: Replicating an agent-based model. *Journal of Artificial Societies and Social Simulation 10*(4), 2.

Wooldridge, M. (2009). *An introduction to MultiAgent systems*, 2nd edn. Chichester : Wiley.