JASSS

Uri Wilensky and William Rand (2007)

Making Models Match: Replicating an Agent-Based Model

For information about citing this article, click here

Abstract

Scientists have increasingly employed computer models in their work. Recent years have seen a proliferation of agent-based models in the natural and social sciences. But with the exception of a few "classic" models, most of these models have never been replicated by anyone but the original developer. As replication is a critical component of the scientific method and a core practice of scientists, we argue herein for an increased practice of replication in the agent-based modeling community, and for widespread discussion of the issues surrounding replication. We begin by clarifying the concept of replication as it applies to ABM. Furthermore we argue that replication may have even greater benefits when applied to computational models than when applied to physical experiments. Replication of computational models affects model verification and validation and fosters shared understanding about modeling decisions. To facilitate replication, we must create standards for both how to replicate models and how to evaluate the replication. In this paper, we present a case study of our own attempt to replicate a classic agent-based model. We begin by describing an agent-based model from political science that was developed by Axelrod and Hammond. We then detail our effort to replicate that model and the challenges that arose in recreating the model and in determining if the replication was successful. We conclude this paper by discussing issues for (1) researchers attempting to replicate models and (2) researchers developing models in order to facilitate the replication of their results.

Keywords:

Introduction

1.1   One of the foundational components of the scientific method is the idea of replication (Popper 1959; Latour and Woolgar 1979). Under this conception, in order for an experiment to be considered acceptable by the scientific community the scientists who originally performed the experiment must publish the details of how the experiment was conducted. This description of the experiment is then read by another group of scientists who carry out the experiment themselves. They then ascertain whether the results of the new experiment are similar enough to the original experiment to state that the experiment has been replicated. This process confirms the fact that the experiment was not dependent on any local conditions, and that the written description of the experiment is satisfactory enough to record the knowledge gained in the permanent record.

1.2   Agent-based modeling (ABM) is a new form of scientific experimentation. If ABM is to become an integral part of scientific practice then it must develop standards of practice similar to those that have been developed for other experimental methodologies[1]. Thousands of agent-based models (ABMs) have been published in the last few decades, but, with some notable exceptions, very few of these models have been publicly replicated (Epstein and Axtell 1996; Axelrod 1997a). We argue that such replications should become a more common practice within the ABM community.

1.3   We will further argue that replication may be even more important within the realm of computational models than it is within the realm of physical experiments. Replicating a physical experiment proves that the original experiment was not a one-time event, and makes the results and model embodied by that experiment available to the replicater as a tool in their own research. Replicating a computational model has these benefits as well, but replication of a computational model also increases our confidence in the model verification, leads to a reexamination of the original validation of the model, and at the same time, facilitates a common language and understanding among modelers.

1.4   Despite these benefits, replication of ABMs occurs infrequently. Part of the reason for this is that not a lot has been written about replication attempts. Knowledge of how to replicate, and how to validate the results of a replication is not widespread within the ABM community and is not typically included in courses on ABM methodology. This in turn impedes others from carrying out replications. Building up a body of cases of replication would enable researchers to take a step back and extract general principles regarding the replication process. This paper embraces this agenda. We begin with a brief review of the history of scientific replication, followed by some working definitions. We continue by describing the rationale for and benefits of scientific replication. Then, we summarize the relevant literature with respect to replication of ABMs. From there we move on to describe a particular case study: our attempt to replicate the Ethnocentrism model created by Axelrod and Hammond (2003). We then extract some general principles from our experience for both model builders and model replicaters. We conclude by noting future directions of research for the methodology of model replication.

History and Discussion of Scientific Replication

2.1   Though discussion of the scientific method and replication of experiments goes back to the time of the ancient Greek philosophers, Karl Popper best enunciated this idea as a fundamental part of the scientific method[2]:

> Only when certain events recur in accordance with rules or regularities, as in the case of repeatable experiments, can our observations be tested—in principle—by anyone. We do not take even our own observations seriously, or accept them as scientific observations, until we have repeated and tested them. Only by such

repetitions can we convince ourselves that we are not dealing with a mere isolated 'coincidence', but with events which, on account of their regularity and reproducibility, are in principle inter-subjectively testable. (Popper 1959)

2.2 In this statement Popper explicates one of the benefits of replication, that by replicating an experiment it can be concluded that the experiment is not a 'coincidence.' He also places emphasis on the fact that a scientist should not even believe their own observations until they are replicated. Observations, he declares, cannot be utilized within the scientific method until they are replicated.

2.3 Some sociologists of natural science, most notably Collins (1985), argue that though many scientists claim that replication is a fundamental part of the scientific endeavor, natural scientists do not often participate in the replication process. Other reports have claimed that replication is carried out but simply not published (Giles 2006). These claims about natural science do not impact the claims made in this paper, since our argument is premised on the fact that replication is beneficial to the computational modeling community independent of its benefit to natural scientists and that, in fact, replication is even more important for computational models than for physical experiments.

2.4 In Collins's further critique of the scientific method he argues that when replication has been performed in the past it has been a rhetorical act and not a scientific one, i.e. that its most common use has been to discredit and marginalize radical experiments (Collins 1985). In some sense true replication may not even be possible. The scientist can overcome Heraclitus' age-old objection that one cannot step in the same river twice by assuming that there is some independent standard to judge replication, but the determination of this independent standard requires experiments which must also be replicated, and thus we wind up in an infinite regress (Taylor 1996). This is part of what informs Medawar's discussion that the scientific process is described fallaciously (Medawar 1991). With these objections in mind, if we are nonetheless to proceed with the practice of model replication, it is incumbent upon us to agree on some standards for replication[3].

2.5 Before we embark on the main thrust of the paper it will be useful to establish some definitions. Terms such as *model, conceptual model,* and *replication* are often used loosely, and different papers use the terms with different connotations. For our purposes in this paper, we will establish the following definitions. By the term *model* we mean a simplified representation of a real-world process or object[4]. An *agent-based model* (ABM) is a model that utilizes, for its basic ontological units, numerous autonomous, heterogeneous agents that follow simple rules. By *conceptual model* we refer to some description, often textual, of a real-world process or object that is not executable and thus has some ambiguities with regards to how to map inputs to outputs[5]. An *implementation* of a conceptual model (also known as an *implemented model*, or an *operationalization*) is a formalization of that model into a computational format such that the model can be given input and generates output. The implemented models that we will be discussing in this paper consist of source code coupled with its executable compiled code. Even in these executable models there are many ambiguities that can exist because of underspecifications in processes like floating point arithmetic (Belding 2000; Izquierdo and Polhill 2006; Polhill, Izquierdo and Gotts 2005; Polhill, Izquierdo and Gotts 2006). However, despite these caveats, for pragmatic purposes, we will consider executable models to be canonical implementations. In this paper we are primarily concerned with agent-based implementations of models. However, many of the issues that we will discuss apply equally well to other forms of implemented models.

2.6 Though many conceptions of replication may exist, for the purposes of this paper, we will define *replication* as the implementation (*replicated model*) by one scientist or group of scientists (*model replicaters*) of a conceptual model described and already implemented

(*original model*) by a scientist or group of scientists at a previous time (*model builders*). The implementation of the replicated model must differ in some way from the original model, and, per our definition, the implementation of the replicated model must be executable, not another formal conceptual model. Since replication refers to the creation of a new implementation of a conceptual model based on the previous results of an implementation, the terms *original model* and *replicated model* always refer to implemented models. Moreover, since this paper is concerned with model replication, any named reference to a researcher's model is a reference to their model implementation (e.g., Axelrod-Hammond model, Wilensky-Rand model).

2.7   An original model and an associated replicated model can differ across at least six dimensions: (1) time, (2) hardware, (3) languages, (4) toolkits, (5) algorithms and (6) authors. This list is ordered based upon how likely the replication effort is to produce different results from the original model. Of course more than one of these dimensions can be varied at once. We will describe each of these dimensions in turn.

2.8
        Time: A model can be replicated by the same individual on the same hardware and in the same toolkit and language environment but rewritten at a different time. This is the least likely to produce significantly different results but, if it does, that would indicate that the published specification is inadequate, since even the original researcher could not recreate the model from the original conceptual model. This is the only dimension of replication that will always be varied.

2.9
        Hardware: The model could be replicated by the same individual but on different hardware. At the minimum, by a change in hardware we mean that the implemented model was run on a different machine. However more interesting results may be obtained by replicating the model on a different hardware platform. Regardless, in these days of hardware independent languages, neither of these changes should provide significantly different results, but, if the results are different, investigations (often technical) are warranted and could point to, for example, the model being susceptible to small changes in the order of events.

2.10
        Languages: The model could be replicated in a different computer language. By a *computer language* we refer to the code entered by the user in order to create the implemented model. For example, Java, Fortran, Objective-C and NetLogo are all different languages. Often the syntax and semantics of a language (e.g., procedural versus functional languages) have a significant effect on how the researcher translates the conceptual model into an actual implementation, and thus replication in a new language can show differences between the conceptual model and the implementation. Even apparently minor details in language and algorithmic specifications, like the details of floating point arithmetic and differences between implementations of protocols can cause differences in replicated models (Izquierdo and Polhill 2006; Polhill and Izquierdo 2005; Polhill, Izquierdo and Gotts 2005; Polhill, Izquierdo and Gotts 2006). For a model to be widely accepted as part of scientific knowledge it should be robust to such changes.

2.11
        Toolkits: The model could be replicated in a different modeling toolkit that is based in the same computer language. A toolkit, in this sense, is a set of program libraries written in a particular language for the purpose of aiding the development of a model. For instance, Repast (Collier, Howe and North 2003), Ascape (Parker 2000) and Mason (Luke et al. 2004) are distinctly different toolkits though they are all written in Java. NetLogo (Wilensky 1999) is an interesting example because, though NetLogo is written in Java, the user does not write Java code to develop models. Therefore we classify NetLogo as both a toolkit and a language. With

many different modeling toolkits available for use, the results of replicating a model in a different toolkit can often not only illuminate issues with the conceptual model but also with the toolkits themselves.

2.12
Algorithms: The model could be replicated using different algorithms. For example, there are many ways to implement search algorithms (e.g., breadth-first, depth-first), or to update a large number of objects (e.g., in order of object creation, in a random order chosen at the beginning of the run, in a random order every time). In fact, a replicated model may simply carry out the steps of a model in a different order than the original model. All of these differences can potentially create differences in the results.

2.13
Authors: Individuals different from the original researcher can replicate the model. This is a strong test of a model's replicability. If another researcher can take a formal description of the model and recreate the same results, then we have good evidence that the model is accurately described and the results are robust to changes in the dimensions of replication that have been altered.

2.14   In our replication effort that we will detail below, the original model was replicated by different authors, in a different language, within a different toolkit, with some different algorithms, on different hardware, at a different time. In other words it varied every dimension of replication that we have described.

2.15   A *successful replication* is one in which the replicaters are able to establish that the replicated model creates outputs sufficiently similar to the outputs of the original model. The criterion by which the replication will be judged successful or not is called the *replication standard* (*RS*). Different replication standards exist for the level of similarity between model outputs. Axtell et al. (1996) examined this question of standards. They developed three categories of replication standards for a replication experiment. The first category of replication standards, "numerical identity" is difficult to establish as it entails showing that the original and replicated model produce the exact same numerical results. One of the reasons this is difficult is that it has been shown that running the same program on the same machine with the same parameters does not guarantee "numerical identity" (Belding 2000). The second category of replication standards is "distributional equivalence." Here the goal is to show that the two implemented models are sufficiently statistically similar to each other. To meet this RS researchers often show statistical indistinguishability, i.e. that given the current data there is no proof that the models are not distributionally equivalent[6] (Axtell et al. 1996; Edmonds and Hales 2003). The final category of replication standards is "relational alignment." Relational alignment exists if the results of the two implemented models show qualitatively similar relationships between input and output variables, e.g., if you increase input variable *x* in both models then if output variable *y* increases in the first model it should also increase in the second model.

2.16   After deciding on the category of RS for a replication effort, it is important to define the particular RS more concretely. Within the three categories of replication standards described above there are many specific replication standards that could be defined. ABMs usually produce large amounts of data, much of which is usually irrelevant to the actual modeling goal. Thus replication of outputs like *x* and *y* coordinates, time stamps, or particular random numbers is probably irrelevant to showing that a replicated model is a correct replication. Usually, one must choose appropriate functions on a subset of the output variables to be the measures for replication.

2.17   After the particular measures have been chosen it is also necessary to choose how often the

results will be compared. One specific RS is to simply match a particular set of outputs at the end of the run. Another more detailed RS would be to match a set of values at various intermediate points throughout the run. Finally, one could attempt to match all the outputs throughout the run demonstrating equivalence in the evolution of the outputs over time. This last RS is perhaps most in the ABM spirit, that is, concerned less with equilibrium and more with dynamics over time. As Epstein asserts in his seminal 1999 paper, "If you didn't grow it you didn't explain its emergence. (Epstein 1999)"

2.18 For the replication effort reported in this paper, we decided on a RS category of distributional equivalence between the replicated and original models. To determine distributional equivalence we must show that output data from the replicated model are sufficiently statistically similar to the output data from the original model. We must first establish focal measures in order to fully specify the RS for this replication effort. Instead of trying to show distributional equivalence for every output variable, we choose a few measures and demonstrate distributional equivalence for those metrics alone. For this project, we identified three such measures, which are described later in this paper. Note that the measures themselves must also be reproduced in the replicated model, which means determining if they too have been reproduced successfully. The problem of measure replication was an issue in the replication effort described below, and will be discussed in more depth later[7]. However in order to prevent an infinite regress of validations, we proceed here with the assumption that a determination of the successful replication of the measures can be achieved by comparing the replicated measures to the conceptual description of the measures.

The Benefits of Replication

3.1 A successful replication of a physical experiment advances scientific knowledge because it demonstrates that the experiment's results can be repeatedly generated and thus the original results were not an exceptional case. As a "side benefit" of a replication, the knowledge and data embodied by that experiment is available to be utilized by the model replicater as a tool to advance his/her own research agenda. That is the replicater now has a working version of the original experiment that gives him/her a tool that they can use to further explore the phenomenon beyond the original publication. Replication of a computational model has these benefits as well. A successful replication of a computational model demonstrates that the results of the original model were not an exceptional occurrence.

3.2 In addition to the above benefits, the process of replicating a computational model can contribute to the scientific community in many other ways. As we describe below, replication of computational models specifically contributes to the processes of model verification, model validation, and in developing a shared understanding of modeling. By *model verification* we mean the process of determining whether an implemented model corresponds to the target conceptual model. Model verification is equivalent to making sure that the implemented model is "correct", which in experimental science is sometimes referred to as making sure that the operationalization[8] of the model is "correct." By *model validation* we mean the process of determining whether the implemented model corresponds to and explains some phenomenon in the real world[9]. By *shared understanding* we mean the creation of a set of terms, idioms and best practices that can be utilized by model authors to communicate about their models.

3.3 To illustrate what we mean by verification and validation, let us explore a simple example. Suppose we wanted to build an agent-based model of a flock of birds. We would first write down a base level description of how the birds operate at the individual level. This description might look something like: "Each bird tries to align its flight direction with the birds around it and get closer to the birds nearby, but not too close." This textual description is an example of a conceptual model. We would then take the conceptual model and create a computer program

that embodies it. In most agent-based modeling languages, the program would consist of three main sections of code, known as "procedures". The three procedures might be named "align", "cohere" and "separate." Verification is the process of ensuring that the individual components in the model correctly implement the conceptual model. In other words verification is making sure that the behavior of the computational model corresponds to the textual description. To verify the NetLogo flocking model, we would need to ascertain that the NetLogo procedures caused the simulated birds to align, cohere and separate as described in the conceptual model. Once such a correspondence has been established, some model authors might be satisfied, but they would not have shown that their computational model corresponds in any way to reality; to accomplish this task they would have to undertake a *validation* effort. As a first step in validation, we might check if the computer visualization of the flock indeed looks like a flock. If it does, we have a kind of face validity. But to get beyond face validity, we would need to measure some properties of real flocks, such as angular accelerations, average relative velocities and size of subgroups within the flock and compare them to the corresponding measures within the model. Toner and his colleagues (2005) report on similar measures used to describe flocks — measures that can be observed in computational models, equational models, and real world data and compared between them[10]. If we have chosen appropriate measures and we find a correspondence, our confidence in the validity of the model increases.

3.4   Replication supports the model verification process because if two distinct implementations of a conceptual model are capable of producing the same results, that lends support to the hypothesis that the implemented model correctly implements the conceptual model. During the model replication process, if differences between the original model and the replicated model are discovered, it may be that the replicated model needs to be fixed. It may also be the case that the original model is not a verified, implementation of the conceptual model.

3.5   Replication supports model validation because validation is a process that determines a correspondence between the outputs from an implemented model and real-world measures. If the replicated model produces different outputs than the original model then that raises questions as to which outputs correspond more to real world data. If the replicated model's outputs are closer to the real world data that lends support to the validity of the replicated model as compared with the original model. More importantly, model replication raises questions about the details of the original modeling decisions and how they correspond to the real world. These questions help clarify whether there is sufficient correspondence between the original model and the real world. Replication forces the model replicater to examine the face validity of the original model by re-evaluating the original mapping between the real world and the conceptual model, since the replicater must re-implement those same concepts. Most model replicaters are not simply blindly following directions, but instead they are themselves researchers and have a vested interest in understanding what the model means in terms of its explanatory power for the phenomenon that they are investigating. As a result, they must consider what explanatory power the model has during the replication process. For instance, a political scientist, replicating the ethnocentrism model described below, must consider the correspondence between the model and real ethnocentric practices. By doing so they become engaged in the validation process, by way of replication.

3.6   We believe that by developing a suite of "best practices" (Jones 2000) with respect to replication, the field of ABM as a whole will advance. The replication process helps us develop a language for describing the modeling process. Creating a culture of replication would foster a shared understanding of the modeling process in the ABM community. In much the same way that statisticians share an understanding of what is meant by "mean" and "standard deviation" and when to apply various statistical tests, over time, replication of ABM experiments will help us define terms of art such as "time-step", "shuffled list" and "vision cone" and lead us to classify ABM rules so they can be matched up with patterns of data.

3.7   These additional benefits may make replication even more important in developing a scientific basis for agent-based modeling than it is for physical experiments. In the next section we will discuss the history of scientific replication and define the process of replication with respect to agent-based models.

Prior Research on Replication of Agent-Based Models

4.1   Since ABM is a relatively new methodology most research has aimed at showing off the power of ABM and not at developing core guidelines for how to utilize ABM. However there are exceptions. One of the earliest attempts was in some ways more ambitious than replication. Axtell, Axelrod, Epstein and Cohen (1996) specifically examined how to "dock" two different ABMs. Docking involves taking models that originally had different goals and showing that they can produce similar outcomes. In this research, they examined two different models, Epstein and Axtell's Sugarscape (1996), and Axelrod's Cultural Model (1997b) and showed that they both led to similar generalizations about the world. In some ways this is similar to what Grimm et al. (2005) call "pattern-oriented modeling", where the attempt is to show that different models can reproduce similar patterns using contrasting alternative theories. Bigbee, Cioffi-Revilla and Luke (2005) also replicated the Sugarscape model in a new ABM toolkit (MASON). Anderson and Fischer (1986) docked a Monte Carlo version of the Garbage Can model with the original Cohen, March, and Olsen (1972) model. North and Macal (2002) did a docking of the beer game. The original beer game was implemented as a systems dynamics model (Forrester 1961), but North and Macal implemented it in a functional programming environment (Mathematica) and two agent-based platforms (Swarm and Repast). Similarly, Edwards et al. (2003) compared an individual-based model to a mean field aggregate model. Densmore (2004) who was also involved in the North and Macal work has documented the use of docking as a benefit to researchers and mentions additional docking experiments. Moss (2000) has developed what he calls a "canonical environment" to facilitate the docking of social simulation models.

4.2   Cohen, Axelrod and Riolo (1998) replicated eight "classic" ABMs. Though they never published the results, a description of their efforts can be found on their website. Hales, Rouchier and Edmonds (2003) first held a workshop on model-to-model analysis in 2003 and this workshop continues to be held on a regular basis. Edmonds and Hales (2003) have separately published their own replication effort. Edmonds and Hales's replication uncovered some unknown exceptions to a well-published research result. Edmonds and Hale (2005) have also published on replication and its relation to theoretical experiments. Fogel, Chellapilla, and Angeline (1999) illustrate a case of replication that questioned the validity of a previous model; in replicating Arthur's (1994) El Farol model he found a potentially conflicting result. Rouchier (2003) replicated another economic-based multi-agent model of speculation. While Galan and Izquierdo (2005) replicated Axelrod's "Evolutionary Approach to Norms" model (1986). Cioffi-Revilla and Gotts (2003) went so far as to compare two models in very disparate research domains (i.e., military conflict and land-use / land-change). In the following we describe our attempt at a replication of an agent-based model.

The Case Study

5.1   We begin the story of our replication effort by explaining the original conceptual model. We then describe the original implemented model, and follow with a recounting of several attempts to improve the replication of the original model. We have deliberately employed a narrative structure while discussing this replication. This narrative structure is meant to explain clearly the process of replication and foster a conversation about how replication occurs. Given the critiques of Collins and others with respect to the inaccuracy of accounts of scientific replication, we have attempted to present our experience to the reader as closely as

we could, describing events as they unfolded.

The Conceptual Model

5.2   At Northwestern University's 2003 conference on Complex Systems, Axelrod presented a conceptual description of an ABM that explored the evolution of ethnocentrism. This conceptual model has also been published in several other formats (Axelrod and Hammond 2003; Hammond and Axelrod 2005a; Hammond and Axelrod 2005b). Below we present the exact description of the model from the 2003 Axelrod and Hammond paper, which is the nearest written description to the conceptual model presented at the 2003 Northwestern conference:

> The model makes three assumptions. First, each interaction is a Prisoner's Dilemma of a single move, thereby eliminating the possibility of direct reciprocity. Second, interaction is local, and so is the competition for scarce resources including space for offspring. Third, the traits for group membership and behavioral strategy are typically passed on to offspring, by means of genetics, culture, or (most plausibly) both.
>
> The model is very simple. An individual agent has three traits. The first trait is a tag that specifies its group membership as one of four colors. The second and third traits specify the agent's strategy. The second trait specifies whether the agent cooperates or defects when meeting someone of its own color. The third trait specifies whether the agent cooperates or defects when meeting an agent of a different color. In this model, an ethnocentric strategy is one that cooperates with an agent of ones own color, and defects with others. Thus only one of the four possible strategies is ethnocentric. The other strategies are cooperate with everyone, defect with everyone, and cooperate only with agents of a different color. Since the tags and strategies are not linked, the model allows for the possibility of "cheaters" who can be free riders in the group whose tag they carry.
>
> The simulation begins with an empty space of 50x50 sites. The space has wrap around borders so that each site has exactly four neighboring sites. Each time period consists of four stages: immigration, interaction, reproduction, and death.
>
>   1. An immigrant with random traits enters at a random empty site.
>   2. Each agent receives a initial value of 12% as its Potential To Reproduce (PTR). Each pair of adjacent agents interacts in a one-move Prisoner's Dilemma in which each chooses whether or not to help the other. Giving help has a cost, namely a decrease in the agent's PTR by 1%. Receiving help has a benefit, namely an increase in the agent's PTR by 3%.
>   3. Each agent is chosen in a random order and given a chance to reproduce with probability equal to its PTR. Reproduction is asexual and consists of creating an offspring in an adjacent empty site, if there is one. An offspring receives the traits of its parent, with a mutation rate of 0.5% per trait.
>   4. Each agent has a 10% chance of dying, making room for future offspring.

(Axelrod and Hammond 2003)

The results of the original implementation of this model as well as the replicated version are discussed below.

The Original Implemented Model

5.3   The results of the Ethnocentrism model that were presented at the NICO conference were from an implementation by Hammond in coordination with Axelrod utilizing the Ascape ABM toolkit (Parker 2000). In the original version of the Ethnocentrism model, "immigration", "interact", "birth" and "death" rules are added to an implemented model. These rules correspond roughly to the descriptions of events in the conceptual model above.

5.4   The most important general result of this model was to show that "ethnocentric" behavior arose under a wide variety of circumstances. In this case ethnocentric behavior was demonstrated by having a high percentage of ethnocentric genotypes and a large amount of behavior consistent with ethnocentrism. Moreover it was shown that cooperation levels remained high, indicating that many individuals were meeting and cooperating with individuals of their own type (ethnicity). Axelrod and Hammond described these qualitative results numerically through a number of behavioral measures. Three of the most important such measures are:

     a. **Cooperation (COOP)**: This is the percentage (out of all the interactions that took place in the final 100 time steps of the run) in which an agent cooperated instead of defecting.

     b. **Ethnocentric Genotypes (CD_GENO)**: This is the percentage of agents (out of all agents that existed in the final 100 time steps) that had the ethnocentric strategy.

     c. **Behavior Consistent with Ethnocentrism (CONSIS_E)**: This is the percentage out of all interactions in the final 100 time steps where an individual cooperated with an individual of the same type (ethnicity) or defected against an individual of another type (ethnicity).

5.5   The percentage of ethnocentric genotypes in the final 100 time steps of a 2000 time step run (CD_GENO) was 76% with a low standard deviation. In addition 74% of all interactions in the final 100 time steps resulted in a cooperation event (COOP) and the percentage of interactions that were consistent with ethnocentrism in the final 100 time steps was 88% (CONSIS_E). After developing the original model implementation, Axelrod and Hammond wrote several papers on the subject and published the results widely. They also made available on a website the original source code from the implemented model and all of the data that they had collected for their publications. A screenshot of this model can be seen in Figure 1.
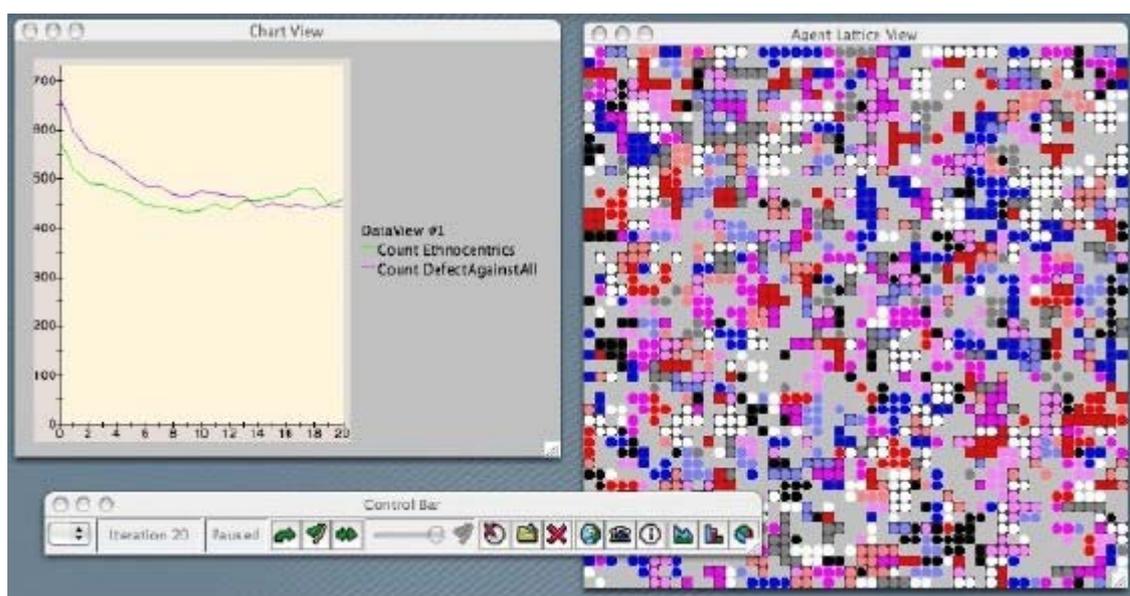

Figure 1. Screenshot of the original Axelrod-Hammond model

The Replication Experience and its Validation

5.6   Wilensky first replicated the Axelrod-Hammond model in the NetLogo language (Wilensky

1999) on the basis of the talk that Axelrod gave at Northwestern University in 2003. Because of the ease of coding within the NetLogo environment, Wilensky was able to replicate this model during Axelrod's talk and Wilensky showed this first version to Axelrod soon afterward.

5.7 Axelrod and Wilensky were pleased that the replicated model seemed to confirm Axelrod's results. But they did not have time at the conference to thoroughly compare the models. Wilensky realized that Axelrod's description of the conceptual model had some ambiguities in it. Looking at Axelrod and Hammond's description quoted above we see the following text: "Each agent receives a initial value of 12% as its Potential To Reproduce (PTR). " This would seem to suggest the potential to reproduce is reset to its base level of 12% at each "time period". However, in the original verbal description it was not clear when this happened. Did it occur after a reproduction event? Did it occur at the beginning of a model step? Questions like these were clarified via email communications between Wilensky and Axelrod. A "final" version of the replicated model was then sent to Axelrod to examine because Wilensky wanted to publish the replicated model in the NetLogo Models Library. This version of the model is available online at http://ccl.northwestern.edu/ethnocentrism/wilensky/. At this point Wilensky thought that his replication correctly captured the rules of Axelrod's conceptual model but he asked Axelrod to determine if the replication was indeed correct. Axelrod gave the Wilensky model to a graduate student who ran the same experiments with the Wilensky implementation that were originally run with the Axelrod-Hammond implementation and then compared the results to see whether the results of the replicated model matched the original model. In order to conduct this test, the RS had to be further specified. Wilensky, Rand and Hammond selected the three important measures described earlier as the focal measures of the replication and a goal of achieving statistical equivalence for those three measures. The Axelrod and Hammond results consisted of ten runs of their model, and they reported the averages and standard deviations across these ten runs. Wilensky and Rand decided to conduct the same number of runs, though the small number of data points meant that a goal of significant statistical equivalence would be difficult to achieve.

Table 1: A comparison of the original model results with the first replication results (averaged over 10 runs)

	Axelrod/Hammond		Wilensky		t-values
	Avg.	Std. Dev.	Avg.	Std. Dev.	
Ethnocentric Consistent Actions	88.47%	1.64%	88.09%	1.10%	0.609
Cooperative Actions	74.15%	1.55%	79.65%	2.22%	-6.424
Ethnocentric Genotypes	76.31%	3.02%	69.14%	4.59%	4.127

5.8 Axelrod and his student discovered that there were differences between the two model implementations that indicated that statistical alignment had not been achieved. They recorded these results in an unpublished report, which they sent to Wilensky. These differences can be observed in Table 1. The t-values were calculated using a standard t-test assuming that the population distributions have equal means and variances. Note that here we are interested in proving that the two sample distributions are drawn from the same population distribution, not different population distributions, thus a low magnitude t-value is good, because it indicates the probability that two samples with the given means and standard deviations could be drawn from the same population distribution. The table shows that the Wilensky model resulted in fewer individuals employing ethnocentric strategies and more cooperation than in the Axelrod-Hammond model and that these differences are statistically significant. Given the relative simplicity of the ethnocentrism conceptual model, Wilensky and Rand were surprised that the replication did not succeed. Wilensky asked Rand to investigate the differences between the

two model implementations with the goals of (a) understanding the mechanisms that led to the divergence in the results, (b) determining which (or whether both) of the two model implementations were externally valid[11], and (c) modifying the Wilensky model to achieve distributional alignment with the Axelrod-Hammond model.

5.9   Since two different toolkits and languages were used, there were differences in the style of the two models. Ascape is a java library created by Parker (2000) at The Brookings Institution to support the development of ABMs. It was developed as an outgrowth of the Sugarscape project (Epstein and Axtell 1996). In Ascape the primary agent type is a landscape, which is called a "scape." Scapes control and manipulate other agents that operate within their boundaries. By adding "rules" to the scapes the user can specify particular processes that other agents in the landscape carry out. NetLogo, on the other hand, was created by Wilensky of Northwestern University, and utilizes a different paradigm of thinking about agents. In NetLogo there are two primary agent types, "turtles" and "patches." Both of these agents can be asked to carry out tasks by the user but neither of these agent types is more primary than the other. However despite this fundamental difference in the two modeling toolkits, NetLogo can emulate the Ascape rules paradigm by giving rules as a task to all "turtles." Though the two implementations of the ethnocentrism model were written in different computer languages it was not clear whether the languages were the source of the observed mismatch, or whether the differences stemmed from different interpretations of the conceptual model.

5.10 In beginning the comparison analysis, Rand noticed that the basic method of agent interaction appeared to be different in the Wilensky model than in the descriptions of the Axelrod-Hammond model. Wilensky had implemented the interaction as a two-way simultaneous prisoner's dilemma where each agent played against each neighbor and it was determined immediately whether both agents would cooperate or defect. This method of interaction was similar to previous computational implementations of the prisoner's dilemma, but was different than the way the interaction was described in Axelrod and Hammond's ethnocentrism papers. This appeared to be a clear difference between the two model implementations, and during conversations between Rand and Hammond they both agreed that this aspect of the Wilensky model needed to be modified in order to have the Wilensky model replicate the Axelrod-Hammond model.

5.11 For example, Figure 2 illustrates three agents in the ethnocentrism model, the middle agent will cooperate with the agent above it and defect against the agent below it, and the other two agents will reciprocate. Let us assume that the top agent goes first, then the middle one, and finally the bottom one. In the original Axelrod-Hammond model this was carried out by having the top agent, since it is cooperating, give the agent below it .01 of its probability to reproduce (PTR) and instantly increasing the middle agent's PTR by .03. The middle agent then would go, and would decrease its PTR by .01 and increase the top agent's PTR by .03. Since the middle agent was defecting against the bottom agent it would do nothing. Finally the bottom agent would execute and do nothing since the only agent it was next to was the middle one and that agent it was defecting against.
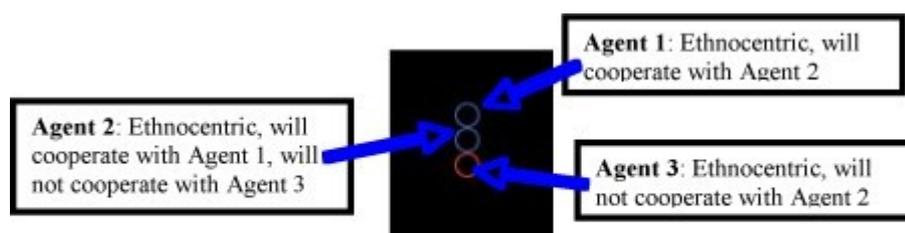

Figure 2. Three agents in the ethnocentrism model

5.12 However, in the original Wilensky model, the top agent would play a simultaneous PD with the middle agent. This would mean that both would announce their strategies simultaneously. Since both were cooperating, both would increase their PTR by .02 immediately. Then the middle agent would activate, since it had already interacted with the top agent it would not do so again, but it would interact with the bottom agent. In this case both agents would announce a "defect" strategy and their PTR's would not change. Finally the bottom agent would activate, but since it had already interacted with the middle agent and had no other neighbors it would do nothing.

5.13 Rand modified the Wilensky model to utilize the interaction method of the Axelrod-Hammond model. However, after these changes, the new implemented model still produced different results from the Axelrod-Hammond model. In fact, as Rand and Wilensky discussed these different results, they realized that the new Wilensky-Rand implementation of the model produced the same outputs as the original Wilensky version, despite the apparent difference in the method of interaction. After some analysis, they were able to prove that whether agents acted independently and at different times from each other or in parallel and at the same time, the results were identical. As long as each agent interacted with each of its extant neighbors once, and the potential to reproduce was updated correctly then the net effect of all interactions was the same.

5.14 If you look back at Figure 2 above, it becomes clear that the PTR of each agent is different at different parts during the interaction phase. For instance, in the Axelrod-Hammond interaction phase, after the first agent finishes executing the PTR's of the agents are from top to bottom: .11, .15, and .12. After the second agent, they are .14, .14, .12, and after the third agent: .14, .14, and .12. In the original Wilensky interaction phase, the PTR's would be .14, .14, and .12 immediately after the first agent acts. In general the eventual results were always the same even if during the interaction phase the values were different at different times.

5.15 Rand then went back to investigate the results of Axelrod's graduate student. In order to determine if the Wilensky model was a successful replication of the Axelrod-Hammond model, Axelrod's graduate student had introduced new measures into the Wilensky model that corresponded with measures in the Axelrod-Hammond model. As mentioned in **Section III**, part of replication involves replicating the original outputs or results generated by the model. These measures are often independent of the model mechanisms. Wilensky had not implemented all of the measures in the original Axelrod-Hammond model and thus Axelrod's graduate student had had to introduce some of these measures into the Wilensky model. Rand corresponded with Axelrod and was able to obtain a copy of the Wilensky model including his graduate student's additional code for new measures that she had introduced and utilized in determining the success of the replication. While examining this version of the Wilensky model as modified by Axelrod's graduate student, Rand realized that the student had added several variables to the model implementation in order to produce exactly the same measures as the original Axelrod-Hammond model. Upon examination of some of these measures it became clear that the graduate student had misinterpreted, based upon the variable names, some of the results that were already being calculated by the Wilensky model. In particular, the graduate student had assumed that measures of genotypes were measures of interactions. For instance the student interpreted "cc-count" as the number of cooperate-cooperate events that had occurred, not as the number of altruistic agents that existed, which was its actual definition. Essentially the measures that the graduate student was utilizing in the comparison were not the same as the measures in the original Axelrod-Hammond models. When the graduate student thought she was counting the number of times agents cooperated, she was actually counting the number of altruistic agents. This confusion is understandable given the differences in the implementation of the interaction phase described above. Since these measures were the basis for the claim that the Wilensky model exhibited more cooperation and less ethnocentrism than the Axelrod-Hammond model, Rand thought it possible that this was the cause of the observed differences.

5.16   Thus Rand modified the Wilensky-Rand model to output the correct measures. This new
model implementation did generate different results than the version of the Wilensky model
modified by Axelrod's graduate student. This version of the model is available online at
http://ccl.northwestern.edu/ethnocentrism/corrected/. However the new Wilensky-Rand model
was still statistically distinguishable from the Axelrod-Hammond. In particular, the standard
deviation of the measures of interest was much higher in the corrected Wilensky-Rand model
than it was in the Axelrod-Hammond model. In fact these new results, if anything, were more
different from the Axelrod-Hammond model than the original replication. Even the first
measure (Ethnocentric Consistent Actions) which was previously indistinguishable from the
original results is now statistically distinguishable. These results are illustrated in Table 2.
There was still more work to do to find sources of difference.

Table 2: Corrected Replication Results averaged over 10 runs (Model online at:
http://ccl.northwestern.edu/ethnocentrism/corrected/)

	Axelrod-Hammond		**Wilensky-Rand (Corrected)**		t-values
	Avg.	Std. Dev.	Avg.	Std. Dev.	
Ethnocentric Consistent Actions	88.47%	1.64%	86.97%	2.38%	6.017
Cooperative Actions	74.15%	1.55%	80.01%	0.83%	-10.540
Ethnocentric Genotypes	76.31%	3.02%	67.78%	5.81%	4.119

5.17   To determine why the results from the model implementations were still different, Wilensky
and Rand again examined the source code for the models. They realized that the Axelrod-
Hammond model utilized a different order of events than did their model. In the Axelrod-
Hammond model the order was immigrate, interact, birth, death. In the Wilensky-Rand model
the order was immigrate, birth, death, interact. This appeared, at first, to be an unimportant
difference between the two implemented models, but as Wilensky and Rand discussed this
difference in ordering, they realized that there might be an effect on the reproduction rate of
the population as a whole. With the birth event happening after immigration but before the
interaction event, new immigrants would not have a chance to raise their reproduction rate
before the birth event resulting in fewer births and thus a smaller population. They conducted
experiments that confirmed their realization; the Wilensky model resulted in fewer individuals
in the population than in the original model. Fewer individuals in a population mean that there
are more chances for improbable effects to dominate the population. As a result, the variance
around the means of the target measures was higher than it should be. They reasoned that
aligning the order of events of the two models would increase the number of individuals in the
Wilensky-Rand model and would likely bring it into statistical agreement with the Axelrod-
Hammond model.

5.18   Thus the Wilensky-Rand model was modified once again, and comparisons were made
between the two model implementations. The variance of the measures of interest in the
Wilensky-Rand model was decreased, but it was still higher than it was in the Axelrod-
Hammond model. There were yet more sources of differences to find.

5.19   Rand met with Hammond in order to try to uncover more model differences. In their
conversation, Hammond noticed another difference between the two implementations. In the
Axelrod-Hammond model the list of agents was shuffled[12] before the reproduction event
took place at each time step. In the Wilensky-Rand model the list was in an unshuffled but
arbitrary order. This arbitrary order was established at the beginning of the model run and was

not varied. When the list is unshuffled some agents are preferentially selected over others. Since the space for reproduction is a scarce resource, the preferentially selected agents have a greater chance of reproducing. This causes a bias throughout a run of the Wilensky-Rand model toward these preferentially selected agents. If these preferred agents were, for example, a particular genotype then that would systematically bias the model results. For instance, if their was a concentration of altruistic agents in the upper left corner of the world and they always got a chance to reproduce before any of the agents in the bottom right, then they would not have to compete with the agents in the bottom right for space and would come to dominate the population. One of the dimensions of replication we described is algorithms. The Wilensky-Rand model used a different algorithm (sorted) than the original Axelrod-Hammond (shuffled) model. Before deciding to align the algorithms it is fruitful to consider whether the model should be robust to such a difference. In this case, actions in the real world are usually not taken in a sorted fashion with one person always acting before another. Thus, on the basis of face validity, there is no reason why the model should be robust to this difference, and the algorithms should be aligned in the replication.

5.20 Wilensky and Rand therefore modified their model to account for the difference in agent ordering that Hammond had pointed out. When the modified model was run and results generated, this final version of the Wilensky-Rand model produced results that were statistically similar to the Axelrod-Hammond model on the three main measures, which was the RS agreed upon by both Wilensky-Rand and Axelrod-Hammond. The t-values were also much better across all three measures than any of the previous replication attempts. Despite the improvements, the second measure (Cooperative Actions) was still statistically distinguishable at the 95% confidence level. Nonetheless, we decided that since the main claims of the original paper were based on the persistence of ethnocentric behavior and not on the amount of cooperation, the statistical differences in this measure could be ignored for our purposes though the differences in this measure warrant further investigation in the future[13]. These results can be seen in Table 3 and the final model is illustrated in Figure 3. This model is available online at http://ccl.northwestern.edu/ethnocentrism/final/.
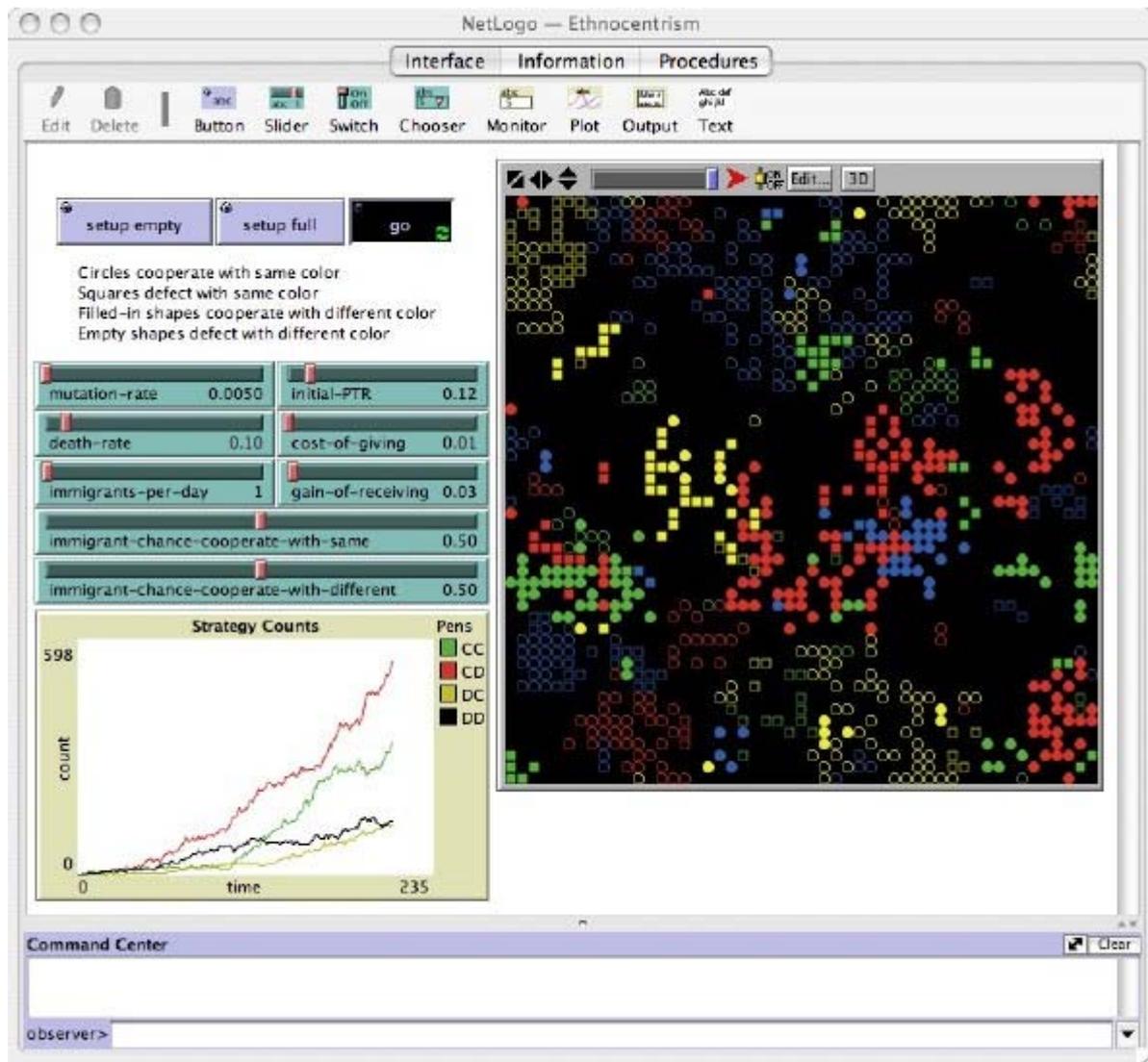
Figure 3. The final model

Table 3: Final Replication Results averaged over 10 runs (Model online at http://ccl.northwestern.edu/ethnocentrism/final/)

	Axelrod-Hammond		Wilensky-Rand (Final)		t-values
	Avg.	Std. Dev.	Avg.	Std. Dev.	
Ethnocentric Consistent Actions	88.47%	1.64%	88.71%	1.97%	-0.296
Cooperative Actions	74.15%	1.55%	76.77%	2.50%	-2.817
Ethnocentric Genotypes	76.31%	3.02%	74.73%	4.03%	0.992

Issues for Model Replicaters

6.1 The replication experience we have recounted herein highlighted several issues that we would like to see discussed by scientists interested in model replication. Foremost, it is important to think about the replication standard—what criteria will be used to determine whether a

replication has been achieved. Typically, a scientist working on a physical experiment will not attempt to exactly reproduce numerical results produced by another scientist. Instead, the RS is to reproduce to the level of precision necessary to establish the hypothesized regularity. This means that the RS itself changes depending upon the question being asked. As discussed earlier, Axtell et al. (1996) listed three general categories of replication standards: "numerical identity", "distributional equivalence", and "relational alignment." These are examples of how different experiments can achieve different levels of replication. Taking care to specify the RS in advance facilitates the replication effort.

6.2 A second issue is the level of detail in the original paper describing the conceptual model, and the ensuing level of communication between the model developer and model replicater. Due to page limits and the desire to limit technical detail in a paper, research papers are usually quite concise and thus every word may have meaning for model replicaters. It may be necessary to contact the original implementers of the model[14]. Many times they will quickly be able to clear up misunderstandings about the original model. Sometimes, the researchers who developed the conceptual model may not have implemented the model, having delegated that job to a programmer. In that case, it may be important to talk with both the author and implementer, as there is an increased chance that, unbeknownst to the author, the original implementation was not a veridical implementation of the conceptual model (i.e. the implemented model could never have been verified).

6.3 On the other hand it is beneficial to delay contacting the original authors until after a first attempt to recreate the original model has been made, since part of the goal of reproducing scientific results is to make sure that the published papers detail the process well enough to preserve the results. Thus by constructing the replicated model from the original paper first, the model replicater serves a valuable service by noting which parts of the conceptual model are not sufficiently described in the original paper. As discussed above, these difference have an impact on the verification of the model because it may indicate that the conceptual model is not detailed enough to be correctly verified. Moreover, it is possible that the differences between the published conceptual model and an implementation can be scientifically interesting and result in new discoveries. These discoveries could affect the validation of the original model because it may be shown, for example, that the replicated version of the model produces output that better corresponds to measures of the real world.

6.4 As part of this process it may be necessary to become familiar with the toolkit in which the original model was written. Taking the time to learn the toolkit can result in a better understanding of how the original model operates. NetLogo is a different toolkit than Ascape. All ABM toolkits have a metaphor or central concept that they utilize to structure their primitives. Becoming familiar and working with this concept will often help the replicater understand some of the "black magic" of the original model they are attempting to reproduce. On the other hand, it may also be beneficial to deliberately implement a strategy that is counter to the paradigm of the original model. By replicating a model in a new language or toolkit and ignoring the biases imposed on the original model by its chosen language or toolkit, differences between the conceptual and implemented models may be easier to observe.

6.5 To facilitate the model replication process it is usually necessary to obtain the source code of the original model. This will enable the model replicater to examine the source in detail and even do line-by-line comparisons with the replicated model. Often this will illuminate discrepancies in the two model implementations that are not obvious from the written descriptions. In addition, in most cases the published results of the original model do not completely explore the parameter space of results that the original model can produce. By obtaining a copy of the original model, it is possible to explore unpublished parts of the parameter space and determine if the two model implementations reproduce similar results. Exploring beyond the published space may significantly alter the conception of what can be

learned from the model results about the real world. This was the case in at least two replication attempts that have been previously attempted (Fogel, Chellapilla and Angeline 1999; Edmonds and Hales 2003). In both of these cases the validation of the original model was called into question by the replication of that model. In these cases, the replication process illuminated differences in the results that were not expected from the description of the conceptual model and were not obtained from the original model. As a result, the replicaters published a different view of the world that was implied by the results of the replicated model.

6.6   Though exposure to the source code and the original model is important eventually, if done too early in the replication process, it may result in "groupthink" whereby the replicater unconsciously adopts some of the practices of the original model developer and does not maintain the independence necessary to replicate the original model, but instead essentially "copies" the original model (Janis 1982). The above considerations represent important tradeoffs and merit careful consideration of the level and timing of communication between the original model authors and the model replicaters.

6.7   In Table 4 we present a list of items that replicaters should include in publications of replication. Alongside these items, we describe some possible ways to specify these items and we list the specifications that we chose while preparing this paper. We have also made supplementary materials from our replication attempts available at http://ccl.northwestern.edu/ethnocentrism/. These issues and possible answers are not meant to be a complete list but to be the start of a conversation about such social norms.

Table 4: Details To Be Included In Published Replications

Categories of Replication Standards:

               Numerical Identity, Distributional Equivalence, Relational Alignment

               Distributional Equivalence

Focal Measures:

               Identify Particular measures used to meet goal

               COOP, CD_GENO, CONSIS_E

Level of Communication:

               None, Brief Email Contact, Rich Discussion and Personal Meetings

               Personal Meetings between replicaters and authors

Familiarity with Language / Toolkit of Original Model:

               None, Surface Understanding, Have Built Other Models in this language / toolkit

               Surface familiarity with Ascape, deep familiarity with Java

Examination of Source Code:

               None, Referred to for particular questions, Studied in-depth

               Mainly examined when particular questions existed about implementation details

Exposure to Original Implemented Model:	
	None, Run, Re-ran original experiments, Ran experiments other than original ones
	Ran the model a few times but just to get a feel for the interface
Exploration of Parameter Space:	
	Only examined results from original paper, Examined other areas of the parameter space
	Only examined results from original paper

Issues for Model Authors

7.1   Model authors can also make a model replicater's job easier by considering a few issues when building new ABMs. First of all, the section of a research paper where the conceptual model is described needs to be well specified. Details of the specification may not appear to be important to the model developer but are important if a model replication is to be successful. It is important to carefully consider the level of detail of articulation of the conceptual model. For example, is it sufficient to describe the model using text alone? Or expand into a pseudo-code description of the model? Or publish the full source code of the model? Even the complete source code for the model may not suffice. For discovering differences between some replications, replicaters may require pseudo-code or source code of the modeling toolkit in which the model was authored. That could in turn necessitate a description of the machine the model was run on, and so on, and so on. Publishing the complete source code of the original model may facilitate the replication process but it may also have costs. In practice, scientists make some assessment of the balance between the advancement of scientific knowledge and their own professional advancement. Once the complete source for a model is made public, for instance, it gives competing researchers quick and easy access to the author's research methods, which may entail the professional cost of enabling another scientist to leap ahead of the original model developer. Thus, determining a standard of publication is necessary for ABM to move ahead as a methodology. It is our recommendation that, at the very least, "pseudo-code" of how the model was implemented be made publicly available[15]. In the longer run, we would expect the field to converge on a common standard form or language for model publication.

7.2   Another issue is to what extent the model developer presents a sensitivity analysis of the results generated by their model. In some cases it may be clear that there are small modifications to the original implementation of the model that drastically affect the results. It is exactly these sensitive differences that model authors need to publish. Even if they do not as yet have explanations for the sensitivity, it is important to point these out as directions for future research. It is also important to consider how the details of the model correspond to the process that the model is attempting to recreate. For instance, in the model presented here, all agents give rise to new agents before any agents die in that time step. In the real world birth and death events for different individuals are interleaved. Does this simplification alter the interpretation of the results of the model? These simplifications must be robust. If the original model makes a different simplification choice than the replicated model and either choice appears to be irrelevant in the real world, then the results should match. If they do not, then the simplification choice needs to be investigated in more depth. In the case of the Axelrod-Hammond model, different simplifications regarding the order of events did have an effect on some of the measures, but this effect did not alter the basic claim that ethnocentrism arises

under a wide variety of circumstances. Nonetheless these differences merit further investigation. This is one important way in which the process of replication can shed light on model validity, or making sure that the model has appropriate correspondence to reality.

7.3   As noted earlier, the model author may not be the original implementer. Such a division of labor can be efficient but it has significant costs. In essence, translating a conceptual model that someone else designed into an implemented model is quite similar to the process of replication. Knowing as we do how hard it is to do a faithful replication; there is a danger that when the model author and implementer are different, the implementation will not be veridical (i.e., that if / when the model undergoes the verification process, it will turn out to be an incorrect implementation of the conceptual model). This will in turn make further replication attempts even harder to do. Our recommended resolution of this trade-off is to suggest that model authors implement their own models using "low-threshold" (Papert 1980; Tisue & Wilensky 2004) languages and toolkits. Such languages and toolkits are designed to be simple enough so that model authors need not be general-purpose programmers, yet can faithfully implement their models. These design affordances reduce the costs and inefficiencies of having a single person be both the model designer and implementer and enable model authors to gain the benefit of more likely verity. An additional benefit of this approach is that the author can more freely experiment with alternative implementations and thus uncover and resolve threats to model validity.

7.4   In Table 5, we present these issues in a concise form, as a list of items that model authors should consider when making available their results. We also present possible ways those issues could be addressed, and how Axelrod and Hammond addressed them[16]. These issues and possible responses are not meant to be a complete list but rather to be the start of a conversation about such social norms.

7.5   In order to truly facilitate the model replication process, it is advisable for model authors to examine their conceptual models through the lens of a potential model replicater. It is only by going through the replication process that a researcher can understand how to adequately describe their conceptual model for other researchers. If model authors consider whether their conceptual model descriptions are detailed enough that a model replicater could replicate the model from those descriptions published models would undoubtedly be more replicable. Moreover, by establishing a norm for model authors to also engage in replication of models, we will accumulate a body of cases of model replications and the field will gain a better understanding of best practices for model replication.

Table 5: Details To Be Included In Published Models

Level of Detail of Conceptual Model:

Textual Description, Pseudo-code
Textual Description

Specification of Details of the Model:

Order of events, Random vs. Non-random Activation
Order of events was specified in the paper form, method of activation was not clear

Model Authorship / Implementation:

Who designed the model, who implemented the

model, and how to contact them

Axelrod designed the original model, Hammond implemented the Ascape version

The model was further refined by both, Email addresses for both were provided

Availability of Model:

Results beyond that in the paper available, Binary available, Source Code available

Results beyond paper and source code available on website

Sensitivity Analysis:

None, Few key parameters varied, All parameters varied, Design of Experiment Analysis

In the paper a few parameters were varied, but many were varied on the website

Conclusion

8.1   As we have shown in this paper, model replication is not as straightforward a process as it may seem. There are many decisions and considerations that must be carefully examined by both model replicaters and model authors. However, despite and indeed because of this lack of straightforwardness, model replication is a critical component of the scientific process. Although there is a widespread attitude that treats model replication as an afterthought, a detail that is necessary to fulfill a scientific obligation, we have argued that it is part and parcel of the validation of scientific knowledge. It shows that the original knowledge was not contingent on particular conditions, illuminates what is contingent and what necessary in a model and provides additional tools to the modeling community to further explicate these conditions of validity. We have argued that computational model replication may be even more beneficial to the scientific community than is the replication of physical experiments. Since computational model replication can affect the model verification process, it can alter our view of conceptual models. In addition, the effect of model replication on the validation process directly increases our knowledge about the real world. Finally, the replication of computational models fosters a shared understanding of modeling concepts and practices.

8.2   In the replication experiment detailed herein, differences were discovered between the Wilensky-Rand model and the Axelrod-Hammond model, and the replicated model had to be modified to produce the original results. Moreover, the process that was required to determine that the replication was successful was complicated and involved unforeseen problems. By accumulating "best practices" and "patterns" of replication, the ABM community can start to place agent-based modeling on a firmer footing. The ABM community has been engaged in a conversation about building standards for ABM. This paper contributes to that conversation and furthers the establishment of such standards and guidelines. We have presented a discussion of some of the issues that need to be resolved within the realm of model replication, and have made initial recommendations that we hope further the conversation. However, it is only ongoing consideration of these issues by both model replicaters and model authors that will improve ABM practice and facilitate the widespread adoption of ABM as a standard methodology.

🌐 Acknowledgements

🌐 Notes

[1] Please note, throughout this paper we refer to ABM both as a methodology and as a field or discipline. We see ABM as both, a methodology and a nascent field.

[2] This particular Popper quotation is also used as a definition of replication in Collins ([1985](#)).

[3] In response to these cautions and to the sociological work by Collins, Medawar, and Latour, we chose to write this paper in a narrative style. In a paper specifically about the *process* of doing science we believe it is important to describe our process as accurately as possible.

[4] The choice of how to simplify, what to foreground and what to background, is at the heart of the modeling process.

[5] Though conceptual models usually take the form of written descriptions, they can take other forms. For instance, they could be diagrams, images, aural descriptions, or even pseudo-code. It should be noted that our definition of conceptual models includes both informal models, as well as non-executable formal models like UML diagrams or flowcharts.

[6] It should be noted that it might be impossible to conclusively prove that two models are distributionally equivalent due to the problem of induction and the stochastic nature of these models.

[7] The philosophical problems concerning the validity of replication mentioned above are not just theoretical, but are ubiquitous in the practice of replication.

[8] This term comes from experimental social science. Another related term from measurement theory is reliability. A necessary condition for a model to be verified is that it is reliable. A reliable model is one the produces the same results over time. However, there are many dimensions to reliability, for a more complete discussion see Carmines and Zeller ([1979](#)).

[9] Many distinct definitions of validation have been proposed by philosophers of science (see[Kleindorfer et al. 1998](#)), but the one given here should suffice for our discussions.

[10] For additional examples of validation of agent-based models with real world data see Grimm et al.'s ([2005](#)) paper on pattern-oriented modeling.

[11] A full discussion of validation is beyond the scope of this paper, but validation can occur either at the level of macro-results or micro-rules (Wilensky and Reisman 2006). In this case Wilensky was asking Rand to consider the validity of the micro-rules by comparing the models' micro-rules to those observed in reality.

[12] By *shuffled*, we mean that the order of the list was rearranged randomly each time the list was iterated. By *unshuffled*, we mean that the order of the list was the same each time it was iterated.

[13] We had a limited number of data points (10) for the Axelrod-Hammond model. The Wilensky-Rand model results bear out even when averaged over 100 runs.

[14] This is a strategy that may have limited applicability. However, as ABM is still relatively young, the majority of original model implementers are still alive and accessible.

[15] Grimm et al. (2006) have recently explored this issue with respect to ecological modeling.

[16] As described in Table 4, Axelrod and Hammond made available a large amount of data on their website, including the full source code of the model and detailed accounts of experiments that they had run but not published. This far exceeds the average amount of information made publicly available by agent-based model authors. However, despite these efforts it still required a considerable amount of effort to perform this replication.

References

ANDERSON, P A and Fischer G W (1986). A Monte Carlo Model of a Garbage Can Decision Process. In March, J G and Weisinger-Baylon, R (Eds.) *Ambiguity and Command*, Pitman.

ARTHUR, W B (1994). Inductive Reasoning and Bounded Rationality. *The American Economic Review* 84(2): 406-411.

AXELROD, R (1986). An evolutionary approach to norms. *The American Political Science Review,* 80(4), 1095-1111.

AXELROD, R (1997a). "Advancing the Art of Simulation in the Social Sciences". In Conte R, Hegelsmann R and Terna P (Eds.) *Simulating Social Phenomena*, Berlin, Springer-Verlag: 21-40.

AXELROD, R (1997b). The Dissemination of Culture: A Model with Local Convergence and Global Polarization. *The Journal of Conflict Resolution* 41(2): 203-26.

AXELROD, R and Hammond, R A (2003). *The Evolution of Ethnocentric Behavior*. Paper presented at the Midwest Political Science Convention, Chicago, IL.

AXTELL, R, Axelrod, R, Epstein, J M and Cohen, M D (1996). Aligning Simulation Models: A Case Study and Results. *Computational and Mathematical Organization Theory* 1 123-141.

BELDING, T C (2000). Numerical Replication of Computer Simulations: Some Pitfalls and How To Avoid them, University of Michigan's Center for the Study of Complex Systems, Technical Report.

BIGBEE, G, Cioffi-Revilla, C and Luke, S (2005). Replication of Sugarscape using MASON. Paper presented at the European Social Simulation Association, Koblenz, Germany.

CARMINES, E G and Zeller, R A (1979). *Reliability and Validity Assessment*. London, Sage Publications.

CIOFFI-REVILLA, C and N Gotts (2003). Comparative analysis of agent-based social simulations: GeoSim and FEARLUS models. *Journal of Artificial Societies and Social Simulation,* 6(4)10 http://jasss.soc.surrey.ac.uk/6/4/10.html.

COHEN, M, Axelrod, R and Riolo, R (1998). CAR Project: Replication of Eight "Social Science" Simulation Models, http://www.cscs.umich.edu/Software/CAR-replications.html.

COHEN, M D, March, J G and Olsen, J P (1972). A garbage can model of organizational choice. *Administrative Science Quarterly,* 17(1), 1-25.

COLLIER, N, Howe, T and North, M (2003). Onward and Upward: The Transition to Repast 2.0. Paper presented at the First Annual North American Association for Computational Social and Organizational Science Conference, Pittsburgh, PA.

COLLINS, H M (1985). *Changing Order: Replication and Induction in Scientific Practice.* London, SAGE Publications.

DENSMORE, O. (2004). Open Source Research, A Quiet Revolution. *http://backspaces.net/research/opensource/OpenSourceResearch.html*.

EDMONDS, B and Hales, D (2003). Replication, Replication and Replication: Some Hard Lessons from Model Alignment. *Journal of Artificial Societies and Social Simulation* 6(4)11 http://jasss.soc.surrey.ac.uk/6/4/11.html.

EDMONDS, B and Hales, D (2005). Computational simulation as theoretical experiment. *Journal of Mathematical Sociology,* 29, 1-24.

EDWARDS, M, Huet, S, Goreaud, F and 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)9 http://jasss.soc.surrey.ac.uk/6/4/9.html.

EPSTEIN, J and Axtell, R (1996). *Growing Artificial Societies: Social Science from the Bottom Up*. Cambridge, MA, MIT Press.

EPSTEIN, J (1999). Agent-based computational models and generative social science, *Complexity 4(5): 41-60.*

FOGEL, D B, Chellapilla, K and Angeline, P J (1999). Inductive Reasoning and Bounded Rationality Reconsidered. *IEEE Transactions on Evolutionary Computation* 3(2): 142-146.

FORRESTER, J W (1961). *Industrial Dynamics*. Cambridge, MA: MIT Press

GALAN, J M and Izquierdo, L R (2005). Appearances Can Be Deceiving: Lessons Learned Re-Implementing Axelrod's 'Evolutionary Approach to Norms'. *Journal of Artificial Societies and Social Simulation* 8(3)2 http://jasss.soc.surrey.ac.uk/8/3/2.html.

GILES, J (2006). The trouble with replication. *Nature* 442(7101): 344-7.

GRIMM, V, Revilla, E et al. (2005). Pattern-Oriented Modeling of Agent-Based Complex Systems: Lessons from Ecology. *Science* 310: 987-991.

GRIMM, V, Berger, U, et al. (2006). A standard protocol for describing individual-based and agent-based models. *Ecological Modelling* 198: 115-126.

HALES, D, Rouchier, J and Edmonds, B (2003). Model-to-Model Analysis. *Journal of Artificial Societies and Social Simulation* 6(4)5 http://jasss.soc.surrey.ac.uk/6/4/5.html.

HAMMOND, R A and Axelrod, R (2005a). Evolution of Contingent Altruism When Cooperation is Expensive. *Theoretical Population Biology*: In Press.

HAMMOND,, R A and Axelrod, R (2005b). The evolution of ethnocentrism, University of Michigan, Technical Report.

IZQUIERDO, L R, and Polhill, J G (2006). Is your model susceptible to floating point errors? *Journal of Artificial Societies and Social Simulation*, 9(4)4 http://jasss.soc.surrey.ac.uk/9/4/4.html.

JANIS, I L (1982). *Groupthink: psychological studies of policy decisions and fiascoes*. Boston, Houghton Mifflin.

JONES, C (2000). *Software assessments, benchmarks, and best practices*. Boston, MA, Addison-Wesley Longman Publishing Co., Inc.

KLEINDORFER, G B, O'Neill, L and Ganeshan, R (1998). Validation in simulation: Various positions in the philosophy of science. *Management Science,* 44(8), 1087-1099.

LATOUR, B and Woolgar, S (1979). *Laboratory Life: The Social Construction of Scientific Facts*. Beverly Hills, Sage Publications.

LUKE, S, Cioffi-Revilla, C et al. (2004). *MASON: A Multi-Agent Simulation Environment*. Paper presented at the 2004 SwarmFest Workshop, Ann Arbor, MI.

MEDAWAR, P B (1991). *The Threat and the Glory: Reflections on Science and Scientists*. Oxford, Oxford University Press.

MOSS, S (2000). Canonical tasks, environments and models for social simulation. *Computational and Mathematical Organization Theory,* 6(3), 249-275.

NORTH, M J and Macal, C M (2002). The Beer Dock: Three and a Half Implementations of the Beer Distribution Game. Paper presented at Swarmfest.

PAPERT, S (1980). *Mindstorms: Children, Computers, and Powerful Ideas.* New York: Basic Books.

PARKER, M (2000). *Ascape* [computer software], The Brookings Institution.

POLHILL, J G, and Izquierdo, L R (2005). Lessons learned from converting the artificial stock market to interval arithmetic. *Journal of Artificial Societies and Social Simulation,* 8(2)2 http://jasss.soc.surrey.ac.uk/8/2/2.html.

POLHILL, J G, Izquierdo, L R, and Gotts, N M (2005). The ghost in the model (and other effects of floating point arithmetic). *Journal of Artificial Societies and Social Simulation,* 8(1)

5 http://jasss.soc.surrey.ac.uk/8/1/5.html.

POLHILL, J G, Izquierdo, L R, & Gotts, N M (2006). What every agent-based modeller should know about floating point arithmetic. *Environmental Modelling and Software,* 21(3), 283-209.

POPPER, K R (1959). *The Logic of Scientific Discovery*. New York, Harper & Row.

ROUCHIER, J (2003). Re-implementation of a multi-agent model aimed at sustaining experimental economic research: The case of simulations with emerging speculation. *Journal of Artificial Societies and Social Simulation* 6(4)7 http://jasss.soc.surrey.ac.uk/6/4/7.html.

TAYLOR, C A (1996). *Defining Science: A Rhetoric of Demarcation*. Madison, Wisconsin, The University of Wisconsin Press.

TISUE, S and Wilensky, U (2004). NetLogo: A simple environment for modeling complexity. Paper presented at the International Conference on Complex Systems (ICCS 2004), Boston, MA, May 16-21, 2004.

TONER, J, Yuhai, T and Sriram, R (2005) Hydrodynamics and phases of flocks, *Annals of Physics* 318(1):170-244.

WILENSKY, U (1999). *NetLogo* [computer software]. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

WILENSKY, U and Reisman, K (2006). Thinking like a wolf, a sheep or a firefly: Learning biology through constructing and testing computational theories. *Cognition & Instruction,* 24 (2), 171-209.

Return to Contents of this issue