The Building and Assurance of Agent-Based Models: An Example and Challenge to the Field

Professor David Midgley
INSEAD
Boulevard de Constance
77305 Fontainebleau
FRANCE
david.midgley@insead.edu

Professor Robert Marks
AGSM
UNSW, Sydney 2052
AUSTRALIA
bobm@agsm.edu.au

&
Dinesh Kunchamwar
School of Computing
National University of Singapore
SINGAPORE
kdinesh@gmail.com

Acknowledgements
The support of the INSEAD R&D fund and the AGSM is gratefully acknowledged, as are helpful comments from participants at the *Marketing Science* conference, Rotterdam, June 2004, the *EUMAS Agent-Based Modeling* workshop, Barcelona, December 2004, the *UCLA Human Complex Systems* workshop, Lake Arrowhead, May 2005 and workshop on *Agent-Based Models of Market Dynamics and Consumer Behaviour*, University of Surrey, January 2006. We would also like to thank Peter McBurney for his help on the issue of software verification. Marks would like to acknowledge the hospitality of the INSEAD Marketing group.

Abstract

The assurance — the verification and validation — of agent-based models is difficult, because of the heterogeneity of the agents, and the possibility of the emergence of new patterns of macro behaviour as a result of the interactions of agents at the micro level. We use an agent-based model of the complex interactions among consumers, retailers, and manufacturers to explore issues of model assurance. Our explorations indicate two challenges for the agent-based models field. The first challenge is to address the critical issue of software verification. The second challenge is to overcome the many methodological challenges that exist in empirically validating these models, some of which we will outline in our paper. We will also propose a method based on the Genetic Algorithm to address both these challenges, but our experiments, and the lack of good data for many kinds of agents, suggest a minimalist approach to building and assuring agent-based models in general.

1 Introduction

1.1 The purpose of the paper

Agent-based (AB) models are a relatively new and important approach to representing and exploring phenomena of heterogeneous agents interacting. Such phenomena occur in many social sciences, including economics, business, and marketing. Taking a disaggregate perspective to the various agents of which such human systems are composed, and utilizing the power of modern object-oriented programming languages, AB models have the potential to be more sophisticated, subtle and faithful to the complexity of such phenomena than do more traditional modeling approaches such as econometrics or game theory or indeed older approaches to simulation such as system dynamics. Recognizing this, researchers in many fields have begun to develop, implement and publish many interesting AB models in scholarly journals and on the web. These developments are encouraging to researchers like us, who believe that many human systems are complex, non-linear, and exhibit emergent behaviour, and are thus poorly modeled by existing approaches. Our fear is, however, that the potential of AB models may not be fully realized unless the critical tasks of assuring — verifying and validating — such models are given considerably more attention in the scholarly literature, and the many current hurdles to achieving adequate assurance overcome by improved methodologies. We further believe that AB modeling will not be recognized as an important scientific method unless those who develop them begin to pay greater attention to assurance and in turn are supported by methodologists who develop improved methods to achieve these ends.

The fact that these issues do not attract enough attention is easily demonstrated. In a recent issue of this journal, Tay and Lusch (2005) clearly demonstrate the power of AB models to address problems that other methods cannot—in this case competitive market strategy in complex and ill-defined environments—and outline an interesting and valuable model addressing this topic. But these authors declare their goal to be simply establishing face validity for their model, and for further validation refer ‘readers who are interested in rigorous analysis of simulation models’ to a citation that pre-dates AB models. We demonstrate in our paper that this step is not so easily taken—many of the issues in the rigorous analysis of ABMS are yet to be fully understood. Indeed, even the 'simple' step of establishing face validity may itself be a significant challenge, requiring the development of new methodologies. These issues of validation are also wider than the scholarly literature. An AB model addressing an important policy issue (the deterrence effects of tax audits) developed by a US government department remains as yet un-validated (Bloomquist, 2004). In their paper entitled 'why are economists skeptical about agent-based simulations,' Leombruni and Richiardi (2005) demonstrate that AB models have not yet been accepted in many top journals—partly because lack of validation makes criticism that 'simulations do not prove anything' difficult to refute.

The purpose of our paper is to examine the issues in verifying (a vital step, not yet discussed in the AB models literature) and validating AB models and to propose some methodologies that may be helpful in achieving these goals. We do this using as an example a model we developed that addresses an important and not well-understood issue in business. This example allows us to draw some important conclusions about AB model assurance and to set out the challenges that these conclusions present to the field of AB modeling. The two critical challenges we shall propose are (1) to value simplicity more than theoretical sophistication in model specification, and (2) to incorporate assurance methodologies into model development from the start.

1.2 Structure of the paper

The second section describes our example AB model—the *Supermarket ABM*—and its implementation in RePast. We present the philosophy behind the design of this model and the detailed specifications of the three types of agent. The third section discusses what we believe should be the first step in the assurance of AB models, namely software verification. The fourth section looks at the issues involved in what we believe should the second step, validation of the model. The fifth section proposes an approach to assurance based on the ideas of Miller (1998) and the Genetic Algorithm and which we illustrate with our results from assuring the *Supermarket ABM*. These results led us to change our perspective on modeling AB models and to rethink how one might specify and test such models. The sixth and final section of the paper sets out our conclusions and the challenges for the field, together with a simple 4-step process that we believe should be taken in building and assuring any AB model.

2 An Example: the Supermarket ABM

2.1 The research problem

Understanding the complex interactions among consumers, retailers, and manufacturers that lead to market and economic outcomes such as consumer satisfaction, and retailer and manufacturer profits is an important issue in business, but it is not well understood. We are undertaking a research program exploring this phenomenon, and our specific focus is that of non-durable products sold in supermarkets.

Individual aspects of this problem have been discussed in many literatures and from many perspectives. For example, the field of marketing has a long tradition of modeling the impact of marketing actions on the sales and market share of products (e.g. Cooper and Nakanishi, 1993; Hanssens, Parsons and Schultz, 2003). Similarly, game theorists have addressed the interaction between consumers and product manufacturers (e.g. Carpenter, Cooper, Hanssens and Midgley, 1988) and between manufacturers and distributors (e.g. Iyer, 1998).

We do not, however, believe the complete system has been adequately modeled to date. Here we define 'complete system' to mean a set of consumers purchasing a category of products (for example, shampoo), the competing retailers that make these products (amongst many others) available to consumers and the competing manufacturers that supply these products to the retailers and promote them through advertising and store displays. If we require realism in our specification of each agent within the system, and if our objective is to model the multiple-period interactions of interest to managers as well as to scholars, then modeling the complete system is indeed a very difficult problem. We might observe thousands of consumers buying the category as a part of their weekly shopping trip, several retailers vying for their custom not only for this category of product but many other categories besides, and several manufacturers promoting their own brands in the focal category. All—consumers, retailers, and manufacturers—can be viewed as goal-oriented agents who learn and adapt their patterns of behaviour over multiple interactions.

Existing approaches to understanding the supermarket setting have used analytical equations to represent part of this system (for example, in the study of the impact of store promotions on consumers). The rest of the system is, however, often viewed as exogenous to these models (which, for example, provide no explanation of why retailers chose promotions in reaction to the past choices of consumers) and so these models remain incomplete. Even where game theory has been employed to model the interactions among different types of agent, this has often been for 'one-period' or 'two-period' games rather than the interactions over multiple periods that characterize this setting. While the existing literature has undoubtedly added greatly to our understanding, it might not capture the richness of agent interactions in this setting or the longer-term dynamics of the market-place. As a consequence, our knowledge of these interactions and dynamics remains incomplete and the normative prescriptions we make from such a partial view of this system may well be incorrect.

Given all of the above, we believe there are three strong arguments for considering AB modeling techniques as a way to gain a more complete, integrated and dynamic understanding of the supermarket setting. First, it is easier to incorporate our existing knowledge about the nature of human-decision-making processes into AB models than it is into analytical equations (for example, decision-makers using elimination by aspects in their choices, or consumers paying selective attention to store displays). AB models allow a flexibility of representation that is not found in more traditional approaches. Second, we believe that individual consumers, retailers, and manufacturers have differing decision-making processes and behaviors. For consumers, we naturally think of different market segments, but we would also not expect Carrefour to make decisions in the same way as does Tesco, or Proctor and Gamble the same way as does Unilever (for reasons of history, organization, costs, etc.). Incorporating heterogeneity in existing econometric approaches, while not impossible, usually results in clumsy simplifications or in equations whose solution is intractable. In contrast, heterogeneity is the essence of AB models. Third, many current approaches are 'top down', imposing analytical structures on markets that are useful to the researcher. In contrast, historical markets are built 'bottom-up' from the actions of independent agents of differing types. By imposing structures, rather than allowing interactions, we might be artificially constraining the system in ways which we do not understand and which might not reflect the historical dynamics or behaviour of the system. AB models potentially allow us to overcome this limitation.

The objectives of our current work are thus to use a bottom-up approach to modeling the supermarket setting, in particular the ideas and techniques of AB modeling. In taking this approach we shall build on the existing literature to specify the decision-making and interactions of the three types of agents (consumers, retailers, and manufacturers). In accordance with our opening remarks we shall, however, also seek to *assure* (to verify and validate) this *Supermarket ABM*.

2.2 The basic modeling philosophy

The basic philosophy of our model is one of memory and decision rules. An agent has memory of what worked for it in the past and rules for deciding which new opportunities to consider and how to evaluate them against known alternatives. This basic philosophy applies to all three types of agent, although the retailer and manufacturer agents are concerned with profits, whereas the consumer agent is concerned with consumption satisfaction. The retailer and manufacturer agents are also conceptualized as having larger memory and more systematic decision-making than does the consumer agent. Similarly the retailer and manufacturer agents are fully informed of each other's proposals through their close interaction, whereas the consumer agent may only become aware of new offers through advertising or in-store promotion. Finally, following industry practice the retailer and manufacturer agents operate on quarterly planning periods, whereas consumer agents operate in a weekly time frame. We now describe each type of agent in more detail.

2.3 The three types of agent

The *consumer agent* becomes aware of brand attributes (two features, plus price) in two ways. First, when the agent sees advertising (with a probability that depends on the level of advertising of the brand relative to its competitors). Observing advertising also reduces the agent's uncertainty on the advertised attribute. Second, when the agent observes an in-store promotion on visiting the store during a week in which there is such a promotion (with a simple probability of observing the promotion in the store). Observing a promotion also reduces uncertainty on the price attribute. The probability that the consumer will go shopping in any week is modeled as a Poisson process with an individual-specific parameter.

The consumer agent is assumed to make screening decisions about which brand to put into their consideration set using a lexicographic rule, and decisions about how to choose a brand in this set using a compensatory rule. To be added to the set, a brand that the agent recently became aware of must be better than any brand already in the consideration set on the most important attribute to that agent (and on the second most important attribute if there is a tie between two brands on the first, etc.). 'Better' implies having more of the attribute than any existing brand by an increment which represents the cognitive cost of expanding the consideration set.

At the point of purchase the agent becomes certain of the actual prices of those brands in their consideration set for that week, and applies a compensatory rule to choose which to buy. The rule is applied by computing an overall score for each brand (the sum of the agent's beliefs about attribute levels weighted by their importance to that agent), a score which is corrected for risk when the agent is uncertain about one or more of the non-price attributes. The brand purchased is the one with the greatest risk-adjusted score. Once a brand has been purchased, the agent becomes aware of its true attribute levels, uncertainty drops to zero, and the score is recomputed. Provided this score is above the agent's individual threshold level of satisfaction, the brand is retained for consideration on the next purchase occasion; otherwise it is dropped as 'unsatisfactory'.

In contrast, the focus of the *retailer agent* is on those store promotions that make the greatest total category profit—reflecting the general interests of retailers, which are obviously different from those of manufacturers selling brands within the category. The retailer agent retains a memory of previously successful promotions, including attributes of the promotion itself (discount off normal price and whether an aisle display was used) as well as the total category profit generated. This memory is updated each quarter so that it retains the best promotions.

For the retailer agent consideration is simple. They choose a certain number of weeks on which to run promotions in the next quarter (a state variable they can change from quarter to quarter) and have a fixed policy that only one brand can promote in any one week (which follows historical examples). Next, they are aware of all the promotions being offered by manufacturers for the upcoming quarter. They consider all proposals systematically, and choose a specific proposal for action in two steps. First, the proposed promotions are compared with those in memory brand by brand, establishing which are most similar on promotional attributes and then ascribing the category profit achieved previously to the new promotion. Second, the agent simply chooses the number of promotions that earn it the most profit, without considering which brand is associated with that promotion. In weeks where the retailer does not schedule a promotion for a brand, the normal price of that brand applies and in some weeks all brands are offered at their normal prices.

The focus of the *manufacturer agent* is also to make profits, in this case for their brand, but their world is more complex than the retailer's. First, the manufacturer agent can choose to change their wholesale price and their weekly advertising level from quarter to quarter. They can also choose which attribute to emphasize in their advertising and what to say about that attribute (e.g. the level they wish to communicate). Second, they need to remember two separate classes of events, corresponding to normal and promotional periods in the retail store. For normal periods, the agent's memory includes the previously most profitable settings of price and advertising. For promotional periods, the agent remembers the previously most profitable promotions (including discount, aisle display and brand profit). Third, the manufacturer agent needs to make promotional offers to the retailer for the next quarter. They do this by first asking the retailer how many promotions will be scheduled for that quarter. They then offer the equivalent number of their most profitable promotions to the retailer. The manufacturer agent will not, however, be awarded all the promotions they request, owing to competition from other manufacturers.

For demonstration purposes the *Supermarket ABM* has been implemented in RePast and successfully run with one retailer, five manufacturers, and 1000 consumers for many simulated weeks of interaction. With these settings the model has the 37 parameters, which are shown in Table 1. These parameters should not be confused with the variables and contents of the memories used by the agents in their decisions and interactions. Rather they are (1) global constants which define items such as the size of memory, allowable price changes, mark-ups, etc. or (2) means and variances of distributions used to generate heterogeneity across agents for items such as attribute importance weights, risk propensity or the attributes of different brands.

[Table 1 about here]

2.4 The realism of the model

Our AB model is built from two sources: the literature, especially that on consumer behaviour, and industry knowledge. The resulting model is realistic—at least to some degree of face validity—but is evidently complex in overall structure. Yet the model has limitations. For just a few examples, we would note that in reality (1) manufacturers take explicit account of the actions of their competitors (here they do not—rather competition is indirectly inferred from results), (2) retailers and manufacturers negotiate over prices and promotions (here they simply accept/reject offers), and (3) consumers forget advertising (here they do not). So while the model is complex, it is not fully based on either the literature or industry knowledge—it remains a considerable simplification of both. This point is important because, as we shall argue later, the trade-off between realism and simplicity is a difficult one to judge. We would also argue that this is not a particular feature of the way we built our AB model. Anyone building an AB model for the supermarket setting is likely to have to make similar trade-offs and arrive at AB models that are complex but not fully representative of whatever theoretical literature or practical knowledge the modeler might bring to bear on the problem. Moreover, this is also likely to apply to any AB model, marketing or otherwise, that attempts to model with systems with three or more sets of interacting agents. To develop this point in a more general direction we now turn to the issues involved in assuring AB model such as these.

3 Verification of the software implementation of the AB model.

3.1 The step before validation

As a consequence of our work, we have come to realize that there is an important step prior to the validation of an AB model. This precursor step is the *verification* of the software—put simply, that the software correctly implements the conceptual model the researcher intended. We believe that this step is largely ignored in the AB model literature and must be given more attention if the field is to progress. Moreover, this is not a trivial step or issue. In developing an AB model the conceptual ideas of the researcher have to be translated into specific programming code, with many choices as to how the details of these ideas are implemented. Often this may involve several academics, research assistants, and computer programmers working together and bringing different skills and perspectives to the project. This process is exactly analogous to software development in general, and so we believe there is much for AB modelers to learn from the literature on that topic.

The importance of verification and validation, and the distinction between them, has been much debated in the software development literature. To quote an early and influential text, one has to first demonstrate that one is 'solving the equations right' before moving on to demonstrate that one has 'solved the right equations' (Boehm, 1981). Another frequently quoted paper is that of O'Keefe *et al.* (1987, p82), who argue that verification is 'substantiating that a system correctly implements its specifications', whereas validation is 'substantiating that the system performs with an acceptable level of accuracy.' Gonzalez and Barr (2000) point out, however, that that many writers remain confused between the two steps and argue that better definitions are needed. They review trends in this debate and put forward the following definition of verification for the intelligent systems field '…the process of ensuring that the intelligent system (1) conforms to specifications, and (2) that its knowledge base is consistent and complete within itself' (2000, p412). 'Conforms to specifications' they mean simply as the process of having a documented specification and checking that the system conforms to this. We note that the written specifications for many AB models are not readily available, nor is there much discussion of the checking process. 'Consistent and complete' they mean as a software system which follows from the researcher's assumptions and which is free from internal errors. Internal errors could be conflicts, redundancies, or circularities, which might lead to unreachable code, cycles, and other forms of non-termination, pathological interactions between elements, dead-end modules, unneeded elements and missing links (all of which are common problems in programming). Note that this type of error is not the same as a "bug" (meaning an observed failure to execute)—verification sets a higher standard than simple debugging. It is unclear how many AB models have been checked for such problems of internal consistency, since seldom do AB model research papers mention such checks.

The natural question at this point is how one goes about verifying an AB model. For analytical models, it is possible to verify that the equations have been correctly solved. The proof is normally presented in the paper and open to inspection by reviewers and readers. For very simple programs, it is also possible for reviewers or readers to inspect the code in a technical appendix. Verification is not, however, an easy task for complex software such as an AB model, which may require several hundred lines of code, and which moreover is often also embedded within a development platform such as RePast.

Indeed, whether software can be completely verified is topic of controversy in the software literature, especially in the long-running debate over whether it is useful to have formal proofs of the correctness of any program (Glass, 2002). More recently, while some writers still think it impossible to completely verify large, complex systems with many parameters (Kelton et al 2001, Shervais et al 2003), others argue that there are several methods and practical steps that can be applied which, while they may not completely verify the software, can go a considerable way towards that goal. We shall focus on the latter here.

3.2 Verification methods

One practical step that is advocated by many writers is an "inspection process" akin to quality assurance in manufacturing—an idea identified as one of the major turning points in improved software development (Goldberg 1999), and originally developed by Fagan (1976) . The key component of this process is a small number of experts with defined roles who go over the code in a structured manner using techniques such as 'paraphrasing' (verbalizing the meaning of each line of code at a higher level than the source text). Research has consistently shown that external review reduces errors significantly, even though there are conflicting findings in the literature on the best way to organize this process (Glass, 1999). Inspections have become a common feature of software development. This can be contrasted with AB modeling, where code is often not externally reviewed either in development or before publication. This, of course, requires that the code be available for reviewers and readers, if not in an appendix, then at least at an open web page. We note that the possibility of external review of code, and the verification that results from this, is a key reason for the commercial popularity and technical success of open-source software.

Beyond simple inspections, there are a variety of more formal methods that can be applied to code verification. These include source code analysis (manual, tool-based or automated), automatic theoretic verification, deriving automata from the program and finite state verification (Holzmann 2000; Hailpern and Santhanam 2002; Cobleigh, Clarke and Osterweil 2002). Schreiber (2002) suggests a role for extreme bounds testing—does the model continue to make sense at the margins?—and sensitivity analyses—to which parameters or combinations of parameters is the model especially sensitive and is this consistent with the specification? Later we shall add another method to this list—automated non-linear testing (Miller, 1998)—since it has contributed much to the development of our own approach and also incorporates several aspects of the preceding discussion. Our purpose here is not, however, to detail all these methods, but to demonstrate that there is much that can be done to verify complex software, and to contrast this with the little that is actually done in AB modeling. We would accept, of course, that many AB models are not safety-critical or commercially important in the sense that some of the software systems for which these formal verification methods were developed are—for example, air-traffic control or banking systems. Increasingly, however, such computer models might be used in decisions that result in litigation, such as merger decisions or liability cases,: it is therefore worthwhile making some attempt to verify the software implementing our models.

4 Validating the AB Model

If verification is solving the equations 'right', then validation is showing that one has solved the 'right equations.' And most writers see the proof that the 'right equations' have been solved by reference to some empirical reality or test. For example, in the intelligent systems field, Gonzalez and Barr (2000) see validation as 'the process of ensuring that the output of the intelligent system is equivalent to those of human experts when given the same inputs.' In the simulation field, Dijkum and Kuijk (1999) also see empirical testing as key when they ask (echoing the Turing test) 'can human beings discriminate between the outcomes of a computer model and the outcomes of the real system the computer is modeling?' And in the view of one profession, validation is 'the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model' (AIAA 1998). Thus the central approach to the 'validation' of AB models will most likely continue the scientific tradition of empirical validation and testing. We say 'will most likely' not only because validation has only recently become a major topic of concern within the AB modeling field, but also because validation of AB models will involve many novel methodological challenges such as to make the direction of the field as yet unclear. Here we shall focus on the empirical approach because we believe it is the central challenge facing the field and because it is the focus of our work (and we shall not discuss other useful approaches such as 'docking,' Axelrod 2003).

Among these novel challenges are Moss and Edmonds' (2005) conclusion that for AB models there are at least two stages of empirical validation, corresponding to the (at least) two levels at which AB models exhibit behaviour: the micro and the macro. The first stage is the micro-validation of the behaviour of the individual agents in the model, which they suggest might be done by reference to data on individual behaviour. An example of this kind of micro-validation (which might also be called agent calibration) is the work by Garcia, Rummel and Hauser (2006) on the wine industry. They calibrated the preferences of their consumer agents from conjoint analyses based on surveys of actual consumers. These agents then interacted with agents representing profit-maximizing wine producers. The second stage is macro-validation of the model's aggregate or emergent behaviour when individual agents interact, which Moss and Edmond suggest can be done by reference to aggregate time series. Moreover, since the interactions at the micro level may result in the emergence of novel behaviour at the macro level, there is an element of surprise in this behaviour, which, with the possibility of leptokurtosis and clustered volatility, may be highly non-Gaussian and difficult to verify using standard statistical methods. As Moss and Edmonds note, at the macro level only qualitative validation judgments might be possible as a consequence. A similar point has been made by McKelvey and Andriani (2005), who note that analyses of such data must take account of extreme events and infinite variance.

In similar vein, LeBaron (2006) suggests three steps to empirical validation. First, attempt to replicate difficult empirical features: for example, does the model fit facts not otherwise explained? Second, put the parameters under evolutionary control, when the AB model is using evolutionary processes (such as the Genetic Algorithm) in order to search the parameter space for better combinations of values. Third, use the results from laboratory experiments with human subjects to validate features of the model.

The two-stage approach of Moss and Edmonds appropriately reflects the complexity of AB models, where interaction among agents at the micro level emerges as behaviour exhibited at the macro level. There could, however, be more than two levels: for instance, the individual, the family, the neighborhood, the city, the state, and the nation. Since the forms of interaction could depend on the level, this is not just a scheme of aggregation or categorization. That is, the macro level behaviour is not simple superposition of the micro behaviour of the agents, bur arises through their interactions. As Bar-Yam (2003) notes, there may exist a class of AB models where the emerged, macro behaviour is insensitive to variation in, say, the initial conditions of the simulation of each agent in the AB model. Ideally, we should like to identify such equivalence classes of models, but the essence of emergence means that the problem of equivalence class identification is complex, with no simple way to predict emergent behaviour from initial conditions, apart from actually simulating the model and observing the macro behaviour.

5 A Common Methodology for Destructive Verification and Empirical Validation

5.1 Miller's approach extended

Both Schreiber (2002) and Midgley and Marks (2004) have suggested that a possible approach to the assurance of AB models might lie in Miller's Automated Non-Linear Testing System. Miller (1998) demonstrated the use of optimization algorithms (such as the Genetic Algorithm) to 'break' the target model. This is done by searching for a set of reasonable perturbations to the model's parameters that produce an extreme deviation from the original prediction of the model. That is, the objective function in the optimization is specified to reward lack of fit. As Miller points out, by careful choice of objective functions, one can test different aspects of the model. Miller applied this approach to the World3 model of the Club of Rome (Meadows, 1974) and was able to show that small changes to just a few parameters resulted in significantly differing results from those originally published. As he notes '…the occurrence of such events does not necessarily imply a faulty model—good models must be responsive to their parameters. Nonetheless, they do indicate the potential for extreme errors, as well as suggest structural areas of the model that might require further investigation and refinement' (1998, p829). In essence, Miller proposed an automated system for the destructive testing of complex models—combining ideas of extreme bounds, sensitivity analysis and robustness. If we apply this system to an AB model that has been documented and 'inspected,' and we are not able to produce extreme, excessively sensitive or excessively insensitive or non-robust behaviour across a range of relevant objective functions, then we would be more confident in the verification of our model.

Schreiber and Midgley and Marks also point out, however, that if an AB model is embedded in an optimization algorithm, then it is equally possible to fit that AB model to empirical data using the same algorithm. In this case the objective function would be specified to reward closeness of fit. Furthermore, building on the ideas of (Bleuler *et al*, 200?) if we separate the required software into three distinct modules — namely, AB model, objective function specification, and optimization algorithm — and have these communicate via a common data interface, we can envisage a flexible platform for advanced AB model assurance. Bleuler *et al* point out the advantages of separating the software into distinct modules. These advantages stem from a separation of the problem-specific code (e.g. the AB model) from algorithm-specific (e.g. choice of optimization algorithm) and the flexibility of choice and development that result. Here we extend these ideas a little further and note that there are also advantages in having the objective function as a separate module—especially when, first, we wish to use this for assurance, and, second, we may wish to optimize against multiple objectives (as will be discussed later). No such platform currently exists in a complete form but we shall illustrate our application of some of these ideas in the following section and we hope these go part way toward this perhaps ambitious goal.

5.2 An Illustration of the Common Method Applied to the Supermarket ABM

We have implemented some elements of the common method by embedding the Supermarket ABM in a Genetic Algorithm (GA) optimizer (using the readily available JGAP code). This is illustrated in Figure 1.

[Figure 1 about here]

We chose the GA because of its robust optimization properties—especially given the non-linear nature of our AB model—and because of our prior familiarity with the GA. We should point out, however, that with separation of modules proposed above, the researcher has complete flexibility to choose the optimization algorithm best suited to their problem (e.g. hill-climbing, tabu search, etc.).

5.3 Destructive verification

Here our goal is to produce extreme, excessively sensitive or excessively insensitive or even non-robust behaviour from the AB model. To that end we tried various objective functions such as: maximize the market share of one producer at the expense of all others, equalize the market shares of producers, maximize the retailer's profit, maximize the manufacturers' profits, maximize the satisfaction of consumers. Many other possibilities can be envisaged here. We should also note that we did these exercises independently; later we will raise the issue of whether verification and validation should both be seen as the simultaneous optimization over several criteria.

Following Miller the search space was constrained to reasonable perturbations from our original parameter settings (even good models might fail with extreme perturbations). To reduce the search space, we selected 16 parameters of particular interest from the 37 parameters in the model. These parameters were perturbed by the optimizer within +/- 20% bounds of their original values and only allowing integer increments on the perturbations to further reduce search time. The GA was run with an adequate population size and for sufficient generations to obtain convergence (25 and 50, respectively). An example of the sort of output obtained from these exercises would be a set of parameter perturbations that resulted in a market share of 82% for one producer. While this outcome is not completely implausible, it is of concern because we are modeling an oligopoly with fairly equal competitors. And this concern is heightened when inspection revealed that only four of the 16 parameters had been pushed even close to the 20% bounds. This suggests that the Supermarket ABM is overly sensitive to some combinations of some parameters. Equally, we found that it is much less sensitive to other combinations. This imbalance suggests flaws in either the conceptual specification or the software implementation of this specification. An example of the former could be the inclusion of a parameter at the micro-level that does not have a significant influence on macro-level behaviour. An example of the latter could be a parameter whose expression is wrongly implemented in the code. These example dampen sensitivity, equally we can imagine flaws that magnify it.

A number of other critical issues arose during these verification exercises. First, we found code that was not invoked during the runs and needed to be examined for possible deletion. Second, we found code that was incorrect and needed to be modified. Third, we found an issue with the use of random distributions to generate the requisite degree of agent heterogeneity (particularly for consumer agents). This introduces considerable noise into the optimization process, often such as to make it difficult to get convergence. As an expedient, we drastically reduced the variance of these distributions. This allowed the GA to converge but obviously sacrifices many of the benefits of the disaggregate approach.

5.4 Empirical validation

We can also use the optimizer to fit the model to historical data. For example, we could simply change the objective function to be the fit between one or more of the AB model's output time series and historical time series data on market share, sales or profits, and we would then seek to maximize this. But we have not done this for the *Supermarket ABM* for three reasons. First, as discussed above, in the verification stage it became apparent that the model is either wrongly specified or our software implementation flawed. Second, adequate data on a total retail system—which might include several manufacturers and retailers as well as many consumers—are not readily available; this itself raises important issues about the design of AB models. Third, the very process of implementing the common method raises an important philosophical trade-off between the simplicity and realism of our AB models that we believe deserves more debate. This trade-off is the first of two important issues that we shall return to in our conclusions; here we shall briefly touch on data availability, as this leads to the second important issue for our conclusions.

Excellent data on consumer purchasing patterns exist and are often integrated with data on the advertising and store promotions that consumers have been exposed to prior to purchase. Indeed, such data exist not only in aggregate form but also from individual household purchasing panels, making it possible to calibrate our consumer agents at the micro-level, as suggested by Moss and Edmonds (2005), as well at the macro-level through aggregate sales and share data.

Data on the retailers and manufacturers are, however, much harder to find, particularly data on costs and profits, but, more critically, data on decision-making. The problem is not only access to confidential data, but also what is not observed or recorded. For example, we do not observe the total set of promotional offers that producers make to retailers, simply the ones the retailers accept and implement and that therefore appear in the consumer panel data. And since these negotiations are often not well documented, we must either to get access to firm meetings or possibly design other calibration methods, such as choice experiments with representative retail managers. Similarly, we do not observe all the decision-making inputs to the producer, simply the resulting actions as implemented through agreements with the retailers, which again may make micro-validation (at the decision-making level) difficult. Contrast this with the availability of data for the consumer. There we have data on the total marketing environment faced by the consumer when making a decision. The prospects for micro-validating consumer agents definitely look very promising, but the prospects for micro-validating retailer and producer agents appear more challenging from the perspective of both methodology and access.

This suggests that, at least in marketing, AB models may need hybrid approaches to validation, in which only some micro agent types are validated. Absent the possibility of validation of all micro types, macro validation (even if only qualitative in nature) bears a heavier burden in validation of the entire AB model. Given the difficulties of validation at the emergent, macro level discussed above, those seeking greater confidence in the AB model via traditional econometric statistics might be disappointed.

Nor should those who seek to validate AB models under-estimate the issues involved in obtaining data on other variables, such as costs and profits. Both are much more commercially sensitive information to firms than are sales or market share, and are also subject to complex measurement issues (cost allocation procedures, aggregation of accounting entities, etc.) especially at the product category level that marketing scholars and managers primarily focus on.

Another central issue, and one on which we have not seen much debate in the AB modeling literature, is that of the initial conditions/parameter values and their relationship to empirical validation. In building the *Supermarket ABM* we face this dilemma because we are very unlikely to obtain data from the beginning of the commercial history of any chosen market. Rather, any data we obtain will relate to a particular window in time, for example, weekly scanner panel data for the last three years. In general, any available historical data will capture a system of interactions among reasonably savvy actors — consumers, retailers, and producers — that have already been through a process of learning about each other's patterns of interaction before the available time series begins. In one sense this is convenient, since an older system will exhibit less variance than one in which most agents are still experimenting and learning through trial and error. But if our AB model includes the possibility of learning, as it does, how should we initialize it? At any point (perhaps never observed historically), and then allow our agents to learn? Or at a historically observed point, which may provide less opportunity for our agents to learn? Especially if we seek robustness in our model to rare events in its environment, absence of sufficient model learning might result in a "brittle" model (Holland 1983), vulnerable to an unprecedented event. We need to think carefully about how incorporate this separation between history and the observed data into our model, in particular how to specify parameters relating to initial conditions and parameters relating to observed behaviour. This problem is analogous to the market-share modeling of price and advertising impacts where brand preferences developed before the data window might be modeled as the intercept term in an econometric model. What are the 'intercepts' in our AB model, and how might these be either estimated from the given data or calibrated from independent sources?

The major conclusion from our efforts to develop an empirical validation methodology is that we need to be much more influenced by the type and nature of the data we can plausibly obtain before we begin to specify our AB models, rather than developing from theory and then seeking appropriate data to fit the demands of this theory. We shall return to this second important issue in our conclusions.

Last, there are conceptual issues in fitting AB models to historical data concerning how the differing degrees of fit to various output variables are combined and/or weighted. AB models can generate a multiplicity of outputs at different levels of analysis and observation windows. For example, how would we value the degree of fit to the individual behaviour of a consumer agent as compared to the fit to the aggregate producer market share? We might have excellent fit to the aggregate output, while simultaneously having counter-balancing poor fits to different segments of agents. This issue is well-known in econometric modeling of consumer data, but with AB models we have a much broader canvas to consider. For example, we may have excellent fit to consumer data but poor fit to retailer or producer data. How do we value that model as compared to one with good fit to the retailer data but poor fit to the consumer data? While we shall not develop these ideas in this paper, we believe the recent literature on evolutionary multi-objective optimization may be of relevance here (Coello et al 2002), and, in particular, concepts relating to the optimization of chaotic systems with conflicting criteria (Rodriguez-Vasquez and Fleming, 2005).

6 Conclusions and next steps

Working on the construction and verification of the *Supermarket ABM* enabled us to reach one conclusion regarding the trade-off between realism and simplicity when we specify ABMs. Thinking about the data realistically available for validation enabled us to reach a second conclusion also shaping the specification of these models.

6.1 Realism versus simplicity

Our experiences with verification have taught us that it is very difficult to verify even moderately complex models. And if one cannot verify one's model, it is not clear that one should be attempting to validate it. We came into this project with the traditional science mindset of building on the extant literature and a deep knowledge of the context. It is possible that this is the wrong approach. Any developed literature tends to emphasize nuances and sophistications leading to complexities in the model. Deep knowledge of the context tends to further add to this complexity. As a result we end up with a model with many parameters, distributional assumptions, and complex interactions and housekeeping. The resulting search space for verification is large indeed and the possibility of building adequate confidence in the basic workings of the model is not that high.

In contrast, we now think the emphasis should be on minimalism. For example, what are the one or two key aspects of consumer behaviour that will explain 80% of the variance in purchases? Equally, what is the simplest decision-making model for a retailer faced with competing promotional offers? And so forth, with the over-riding goal of building the simplest model that will capture a substantial part of the actual phenomena.

This might be seen as an appeal to Occam's Razor. Parsimony is highly valued in science and most researchers building a model with deep theoretical and unobservable constructs would indeed seek good fit to the data with as few parameters as possible (Simon and Wallach, 1999). However, for AB models we think this goes further than simple parsimony. AB models are inherently complex in and of themselves because of the interactions between the various classes of agents and the emergent behaviour that results. Thus we would argue for *minimalism* rather than parsimony, especially in terms of the number of parameters that need to be verified or calibrated. Too many parameters make it very difficult or almost impossible to assure these models.

Note this is a substantial challenge. It is easy to build minimalist models; it is far less easy to build ones that capture a substantial part of the actual phenomena. Here we might echo Einstein, 'make everything as simple as possible, but not simpler.' And as a field we need to develop norms as to what is 'as simple as possible.'

6.2 Models should be built with validation data in mind

In our particular example it has become clear that the most reliable data will relate to individual consumer purchasing behaviour. This is simply because more commercial investment goes into collecting those data. Therefore in this area lies the best opportunity for micro-validation. In contrast, the retailer and manufacturer models will always be harder to micro-validate. This suggests to us a changed modeling approach, whereby the consumer agent is built essentially bottom-up from the data. The more assumption- and parameter-based modeling might then refer to the other types of agents, who could be calibrated by the fitting exercise. This would also reduce the number of parameters—and thereby considerably facilitate model assurance.

Although we have not yet fully articulated this idea, we do think that the nature of available data should play a greater role in the formulation of AB models than it does in the current literature. This is not to say it should be the only determinant: theory needs also to be evident and indeed may suggest the need for new measures. We believe, however, that all models should be built with validation more clearly in mind.

6.3 Next steps

The next steps for us in building our *Supermarket ABM* are clear. We have already specified a second version with simpler agent decision-making and interactions and half the number of parameters. Consumer researchers or behavioral decision theorists may well be uncomfortable with these simplifications. But we believe they will allow us to assure this model. Once that step has been achieved we can move forward to address sophistications that advanced theory or the validation itself suggests are worthwhile. But we will do this with a minimalist approach, only adding complexity where the payoff in improved validation is compelling. And we have thought more about the sorts of data that will be available for this validation, thinking that has influenced the redesign of the model from the bottom-up. We are currently working on obtaining these data and once they are to hand we will complete the final specifications of the model. From that time we want to follow a 5-step process that we would also recommend to the AB models field as a whole, namely:

1. Publish the detailed specifications of the model on the web;
2. Enlist the help of a small number of programming experts to inspect and correct the code implementation of this specification;
3. Subject the code to destructive testing using the GA with the aim of establishing parameter sensitivity and identifying pathologies arriving from agent interaction;
4. Empirically validate the model against real data, using the GA to;
    a. Calibrate the model parameters on half the data;
    b. Validate by testing the fit of the calibrated model on the remainder; and
5. Compare this model with other models in the nested manner recommended by methodologists (do even simpler models explain the phenomena as well, if theory suggests any sophistications do these explain the phenomena substantially better, etc.)

We recognize this is a broad and ambitious agenda, but it is an agenda we believe must be addressed if AB models are to achieve their evident potential.

References

AIAA Guide for the verification and validation of computational fluid dynamics simulations. AIAA Standards Series. Reston, VA: AIAA 1998.

Axelrod Robert. Advancing the art of simulation in the social sciences. Japanese Journal for Management Information Systems 2003, 12(3).

Bar-Yam Y. Unifying principles in complex systems, in Converging Technologies for Improving Human Performance: Nanotechnology, Biotechnology, Information Technology and Cognitive Science, M.C. Roco and W.S. Bainbridge, editors, New York, NY: Springer, 2003: pp. 380 - 409.

Bleuler Stefan. Laumanns Marco. Thiele Lothar. Zitzler Eckart. PISA — a platform and programming language independent interface for search algorithms. Mimeo. http://www.tik.ee.ethz.ch/pisa/

Bloomquist Kim M. Multi-agent based simulation of the deterrent effects of taxpayer audits. 97th Annual Conference of the National Tax Association, Minneapolis, MN, November 2004.

Boehm Barry. Software Engineering Economics. New York, NY: Prentice-Hall, 1981.

Carpenter Greg. Cooper Lee. Hassens Dominique. Midgley David. Modeling asymmetric competition. Marketing Science 1988, 7(4): pp. 393-412.

Cobleigh J.M. Clarke L.A. and Osterweil L.J. FLAVERS: a finite state verification technique for software systems. IBM Systems Journal 2002, 41(1): pp. 140-165.

Coello C.A.C. VanVeldhuizen D.A. Lamont G. Evolutionary Algorithms for Solving Multi-Objective Problems. Boston, MA: Kluwer 2002.

Cooper, Lee and Masao Nakanishi. Market share analysis. Boston, MA: Kluwer, 1993.

Dijkum Cor van. Kuijk Etzel van. Validation of simulation models in the social sciences. In: Dijkum Cor van. DeTombe Dorien. Kuijk Etzel van, editors. Validation of simulation models Amsterdam: SISWO 1999: pp. 7-29.

Fagan Michael. Design and code inspections to reduce errors in program development. IBM Systems Journal 1976, 15(3): pp. 182-211.

Garcia Rosanna. Rummel Paul W. Hauser John. Co-opetition for the diffusion of resistant innovations: a case study in the global wine industry using and agent-based model. Agent-based models of market dynamics and consumer behavior. Institute of Advanced Studies, University of Surrey, Guildford January 2006.

Glass Robert L. Inspections – some surprising findings. Communications of the ACM 1999, 42(4): pp. 17-19.

Glass Robert L. The proof of correctness wars. Communications of the ACM 2002, 45(8): pp. 19-21.

Goldberg R. Turning points in software development. IBM Systems Journal1999, 38(2&3): pp. 25-229.

Gonzalez Avelino J. Barr Valerie. Validation and verification of intelligent systems—what are they and how are they different? Journal of Experimental & Theoretical Artificial Intelligence 2000, 12: pp. 407-420.

Hailpern B. Santhanam P. Software debugging, testing and verification. IBM Systems Journal 2002, 41(1): pp. 4-12.

Hanssens Dominique. Parsons Leonard. Schultz Randall. Market response models: econometric and time series analysis. Boston, MA: Kluwer, 2003. Second Edition.

Holland, John H. Escaping brittleness. In Proceedings Second International Workshop on Machine Learning, pp. 92-95, 1983.

Holzmann Gerard J. Software verification at Bell Labs: one line of development. Bell Labs Technical Journal 2000, January-March. pp. 35-45.

Iyer, Ganesh. Coordinating channels under price and non-price competition. Marketing Science 1998, 17(4): pp. 338-355.

JGAP. http://jgap.sourceforge.net

Kelton David. Sadowski Randall. Sadowski Deborah. Simulation with Arena, New York, NY: McGraw-Hill, 2001. Second edition.

LeBaron Blake. Agent-based computational finance in Handbook of Computational Economics, Volume 2, edited by Leigh Tesfatsion and Kenneth L. Judd, Amsterdam: Elsevier Science, forthcoming 2006.

Leombruni Roberto. Richiardi Matteo. Why are economists skeptical about agent-based simulations? Physica A, 2005 (355): pp. 103-109.

McKelvey Bill. Andriani P. Why Gaussian statistics are mostly wrong for strategic organization? Strategic Organization 2005, 31(2): pp. 219-228.

Meadows D.L. Behrens W. W. Meadows D.H. Naill R.F. Randers J. and Zahn E.K.O. Dynamics of Growth in a Finite World. Cambridge, MA: Wright-Allen 1974.

Midgley David F. Marks Robert. The interaction among consumers, retailers and manufacturers: an agent-based model. Marketing Science Conference, Eramus University, Rotterdam, June 2004.

Miller John. Active nonlinear tests (ANTs) of complex simulations models. Management Science 1998, 44(6):820-830.

Moss Scott. Edmonds Bruce. Sociology and simulation: statistical and qualitative cross-validation. American Journal of Sociology 2005, 110(4): pp. 1095-1131.

O'Keefe R.M. Balci O. Smith E.P. Validating expert system performance. IEEE Expert Systems 1987(Winter): 81-86.

Rodríguez-Vázquez, K. Fleming P.J. Evolution of mathematical models of chaotic systems based on multi-objective genetic programming. Knowledge and Information Systems 2005, (8): pp. 235-256.

Schreiber Darren. Validating agent-based models: from metaphysics to applications. Annual conference of the Midwestern Political Science Association, Chicago, IL: April 2002.

Shervais Stephen. Wakeland Wayne. Raffo David. Evolutionary verification and validation of software process simulation models. Mimeo. http://prosim.pdx.edu/prosim2004/abstract/wakeland_ext_abs.pdf

Simon Herbet A. Wallach Dieter. Cognitive modeling in perspective. Kognitionswissenschaft 1999, 8(1): pp. 1-4.

Tay Nicholas S.P. Lusch Robert F. A preliminary test of Hunt's General Theory of Competition: using artificial adaptive agents to study complex and ill-defined environments. Journal of Business Research 2005, 58: pp. 1155-1168.

Table 1

Parameters of the model

	Number of parameters
Consumer agent	
Global constants	
Number of weeks for advertising awareness calculations	1
Size of consideration set	1
Mean and variance of distributions	
Attribute importance	6
Cognitive costs	2
Risk adjustment	2
Satisfaction threshold	2
Inter-purchase time Poisson lambda	2
Chance of observing a store promotion	2
Retailer agent	
Global constants	
Number of best promotions remembered	1
Slotting fee	1
Quarterly increment/decrement to mark-up	1
Range of mark-ups allowed	2
Manufacturer agent	
Global constants	
Quarterly increment/decrement to wholesale price	1
Range of advertising levels allowed	2
Quarterly increment/decrement to advertising level	1
Size of normal memory	1
Size of promotional memory	1
Fixed and variable costs	2
Mean and variance of distributions	
Product attributes	6

Figure 1

An approach to verifying & validating agent-based models