# Stepping on Earth: A Roadmap for Data-driven Agent-Based Modelling

Samer Hassan[1,2], Luis Antunes[3], Juan Pavon[1], and Nigel Gilbert[2]

[1] GRASIA: Grupo de Agentes Software, Ingeniería y Aplicaciones, Departamento de Ingeniería del Software e Inteligencia Artificial, Universidad Complutense de Madrid, Madrid, 28040, Spain {samer, jpavon}@fdi.ucm.es

[2] CRESS: Centre for Research in Social Simulation, Department of Sociology, University of Surrey, Guildford, Surrey, GU2 7XH, United Kingdom n.gilbert@surrey.ac.uk

[3] GUESS: Group of Studies in Social Simulation, LabMAg, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal xarax@di.fc.ul.pt

**Abstract.** In the last few years an increasing number of more complicated Agent-Based Models have been designed with the aim of approaching reality more closely, usually by introducing more and more empirical data. In this paper we review different approaches that could be adopted: KISS, KIDS and a new one that lies between both. We then propose a new logic of simulation driven by empirical data. Because the emphasis on collected data affects the model's design and initialisation, as well as its validation, we provide guidelines for data injection, flow diagrams of well-defined stages, and suggest the application of some Artificial Intelligence technologies.

**Key words:** agent-based modelling, data-driven, deepening, empirical social simulation

## 1 Introduction

In a previous paper [9], we have argued in favour of using data to illuminate and inform several steps of multi-agent-based simulations for the study of complex social phenomena. In this paper we explain the benefits of data use in simulations, as well as providing methods, techniques, and tools to perform such injection. The paper will focus on the main points to illustrate our approach. We address the classical approaches to Social Simulation in section 2, and several initiatives that foster the exploration character of simulations in section 3. We outline a methodological roadmap by detailing tools, techniques and technologies to handle data (section 4), and finally propose a complete picture of this 'new logic of simulation' to be used in data-driven exploratory social simulations (section 5). Finally, we discuss some of its difficulties and implications.

## 2 The Classical View

### 2.1 The Logic of Simulation Diagram

Agent -based modelling is founded on a methodology that has been described as a “logic of simulation” [6]. This logic, shown in diagrammatic form in Fig. 1, is a representation of the classical scientific experimentation applied to the simulation. The Target is the observed phenomenon. As a result of a process of Abstraction, a Model, a simplification of this phenomenon, can be obtained. This Model, in this case an Agent-Based Model, can be simulated to obtain results, the Simulation data. A process of Data gathering (qualitative, quantitative, or both) can be used to extract the Collected data from the Target. The comparison of this data and the simulation output allows a process of validation. If there is structural similarity between them, the ABM is validated and considered a good representation of the phenomenon. If there is not, the model should be modified and the simulation repeated until the output fits the gathered data.

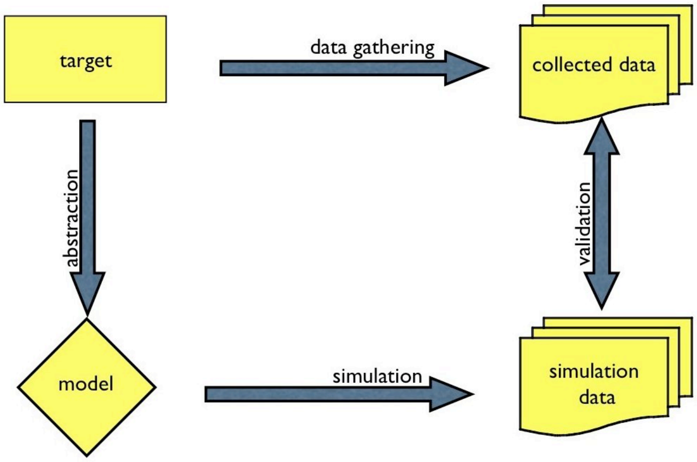

**Fig. 1.** Diagram showing the classical logic of simulation (after [6]).

### 2.2 KISS Paradigm

Abstraction is a process of generalization by removing some elements from an observable phenomenon (or from a model) in order to build a model that can be handled. The amount of information kept in this process is an issue for debate, but it usually depends on the purpose of the model.

Agent-based social simulation usually follows the KISS principle (“Keep It Simple, Stupid”) proposed by Axelrod [2] following the “Occam’s razor” argument. The justification to use it relies on the importance and practicality of simplicity in modelling. Simplicity is helpful for transmitting the model to the scientific community, promoting understanding and extensibility. Besides, making an abstract and simple model is often supported with the argument that such models are more general, and therefore have possible applications in many real cases. Another reason for its spreading is that building a simple model is, simply, easier than building a complex one. And furthermore, it is not only the design: it is easier to implement, analyse and check [3].

## 3 Breaking the Rules: Several Initiatives

### 3.1 Data-driven Modelling

Most ABM follow the KISS principle, and therefore try to be rather abstract and generic. As the empirical data is specific to one site and time, in order to do that they often decide to use standard distributions in several steps of the design: configuring the initial conditions of simulations, distributing objects spatially, determining exogenous factors or aspects of the agents' behaviour [9].

However, an increasing number of ABM are appearing, specially in recent years, which follow different approaches that try to be more realistic by getting closer to the target. Increasing complexity of the models, against the KISS paradigm, is strongly linked to a more intense use of the data available. This view implies breaking with the modelling "for the sake of simplicity" and can even slightly modify the classical logic of simulation. In this section two alternatives to KISS are presented, while in the subsection 5.1 an alternative logic is proposed.

### 3.2 KIDS

The KIDS approach, formulated by Edmonds and Moss [3], opposes KISS and promotes the principle, "Keep It Descriptive, Stupid". This alternative has achieved some notoriety in the ABM community (e.g. [10, 4])because, while they consider KISS attractive and understandable, they do not find it realistic or useful. KIDS asks modellers to begin with the most similar model to the target, in spite of its complexity. Only afterwards, should the model be analyzed to see which parts could be simplified while preserving the behaviour. With further simplifications, a KIDS ABM could be used in several contexts while being sure that it has good foundations.

However, the authors admit that "Neither the KISS nor the KIDS approach will always be the best one, and complex mixtures of the two will be frequently appropriate. "

### 3.3 A New Perspective: Deepening KISS

Whereas in KISS the models are designed as simple as possible and only made more complex when difficulties are met; and in KIDS the idea is to start with a model that is descriptive in face of evidence and made progressively more simple and abstract as more evidence and understanding allows it; there is a third way that we coin "deepening". This deepening phase is a part of a more comprehensive methodology described in [1]. The idea is to start from something close to a KISS model, but following Sloman's prescription of a 'broad but shallow' design [12]. Then, through the use of evidence and especially data, as we prescribe in [9], a collection of models can be developed and explored, allowing for the designer to follow the KIDS prescription without really aiming at more simplicity of abstraction. Once the design space (of agents, societies and experiments) has been reasonably explored, the best features of each model in the collection can be used to design a stronger model, and the process iterated.

This deepening principle allows for models to be made as complex as necessary, but no more than the designer wants, for the sake of control of the model and adaptiveness to the research questions posed. It is the exploration of the models themselves that will inspire further deepening, or allow the process to stabilise and other features to be addressed.

Deepening and its underlying methodology allow for the iterative refinement and exploration of all the objects in the undertaken scientific questions. Hypotheses, theories, conjectures, programs, models, simulations are all situated in complex design spaces, which, together with the modeller (and even stake-holder, see participatory simulation [7]), are explored to find the best combinations to allow an in-depth understanding of the target phenomenon.

The ultimate aim is not to provide a model that answers all the stake-holders' questions, but rather to provide an exploratory environment that allows them to make a more informed decision by knowing their problem in depth. In a subsequent section we will see how data injection can help do a better job at this.

## 4 Methodological Roadmap

### 4.1 Why Data

While programming and running simulations, there are a lot of necessary simplifications. Usually, the removal of arbitrary assumptions will be made with the help of data. But raw data are hard to use in a simulation, and often modellers have to resort to techniques that filter away unnecessary data complexity.

One technique is to use statistical measures to provide parameters to probability distributions. For instance, 'let's assume that salaries follow a Gamma distribution.' This abstract description of a given quantity spares us form delving into real data, but it is by no means more general than a sensible use of data. The only way we can consider the use of random distributions more adapted to the problem at hand, and possibly more general to encompass other similar problems, is precisely by the use of several collections of data, carefully tested with statistical techniques against the distributions we are advocating. Each of these techniques will allow for a quantified error (confidence of fit to the distribution). Also, those distributions are usually static, so not particularly adapted to the dynamical nature of computer simulations.

Not knowing usually what the correct statistical distribution is, it is probably better to use one or more empirical distributions. Or, a typical set of data that could be followed could be preferable. The problem is that 'typical' is hard to define formally. The statistical methods aim to define that notion. Another fundamental problem with probability distributions is that they are good to describe static overall behaviours, especially from an *a posteriori* perspective. They have many more problems in providing the emphreasons that may cause individual behaviour.

The power of random distributions to fit well (and quantifiably) a collection of data, and their mathematical elegance, give us no reasons why models based on them can be more general than the scope of the data collection they are based on. That generality can only be achieved by the use of even more data, and more statistical fits.

Since we advocate the use of data not only through statistical models, but in other phases of the simulation development, we must pay a close watch to the universe they are coming from. Whatever form we are using to inject data into our model (and surely statistical measures are one), we must ensure that data are representative of the universe for which we are designing the model. Representativeness is again hard to define formally. However, by using data in a mediated manner, the representativeness problem arises twice. In the following sections we provide some procedures for how to handle data for the purposes of social simulation.

### 4.2 Handling Data: the Procedures

Once it has been decided that data will be used to drive the simulation, the next questions are, what type of data, and where could the data be obtained?

It is desirable to have data from some representative sample of the target population. In practice, this usually means survey data from a large random sample of individuals, although it needs to be recognised that large representative samples, while statistically advantageous, also have some disadvantages:

1. if the sample is large, it is likely that the researcher will not be the person who designs or carries out the survey. More likely, the data will come from a government or market research source. This means that the survey will probably not include exactly the right questions phrased in the right way for the researcher's interests, and compromises will have to be made.
2. if the sample is random, it is unlikely that it will include much or any data about interconnections and interactions between sample members, so studying networks of any kind is likely to be impossible. This can be a serious problem when the topic for investigation concerns matters such as the diffusion of innovation or information, or friendship relations.

3. some data are inherently qualitative and not easily gathered by means of social surveys. For example if one is interested in workplace socialisation (e.g. [13]), a survey of employees is a very crude and ineffective method as compared with focused interviews, focus groups or participant observation (for more details on these standard methods of social research, see [5]).

Despite these disadvantages, survey data can be valuable. It is particularly valuable when it is collected from panels, i.e. if the same individuals are interviewed at several times at intervals, such as every year. Panel studies are more or less the only way of collecting reliable data about change at the individual level. Such data are valuable because they can be used to calculate transition matrices, that is the probability that an individual in one state changes to another state (e.g. The probability of unemployment). With a sufficient amount of data, one can calculate such transition matrices for many different types of individual (i.e. for many different combinations of attributes). So for example, it becomes possible to calculate the rates of unemployment for young men, old men, young women and old women. However, if one tries to take this too far - differentiating according to too many attributes - the reliability of the computed probabilities will drop too far, because there will be too few cases for each combination of attributes. These probabilities provide the raw material for constructing probability distributions that may be used to simulate the effect of the passage of time on individuals.

We have stressed the importance of obtaining data repeatedly over periods of time. This is because generally agent-based models are concerned with dynamical processes, and snapshots of the situation at one moment in time are of limited value and can sometimes even be misleading as the data basis for such models. While panel survey data is relatively rare compared with cross-sectional data, other forms of data collection about social phenomena are often more attuned to measuring processes. This is particularly the case with ethnography where the researcher observes a social setting or group continuously over periods of days, weeks or months. A third form of data collection is to use official documents, internet records and other forms of unobtrusive data that are generated by participants as a byproduct of their normal activities, but that can later be gathered by researchers. Examples are newspaper reports, web pages, and government reports. In these cases, it is often possible to collect a time series of data (e.g. using the Internet Archive `http://www.archive.org/` to recover the history of changes to a web site) and thus to examine processes of change.

Regardless of whether the data is quantitative or qualitative, it is often the case that they do not have to be collected afresh, but rather that data previously collected by another organisation, possibly for another purpose, can be used. Enormous quantities of survey and administrative data are stored in national Data Archives (European archives are listed at `http://www.nsd.uib.no/cessda/archives.html`) and increasing Archives are extending their scope to include qualititative data (e.g. in-depth interviews) as well (see for example, `http://www.esds.ac.uk/qualidata/` ).

# 5 The Data-driven Flow

## 5.1 Changing the Diagram

In this section we propose an alternative logic of simulation that could be used in data-driven modelling. The main change is the focus on collected data. In the classical logic of simulation presented in section2.1 the data gathering could be done after building the model and the simulation, because it was used just for validation. However, in the diagram presented in Fig. 2 the new arrows represent a twist in the sequence. The new flow forces the data gathering to be before the simulation. This is due to the two processes represented by the new arrows: the influence of collected data in the design of the model and the initialisation of the model based on some of this data. Building the model is not finished until the abstraction, data-driven design and initialisation are all completed. Only then can the simulation can be executed and the output obtained. The last stage, the validation process, must be done with data not used previously in initialisation. The there may be a need for feedback and modification of the ABM again.

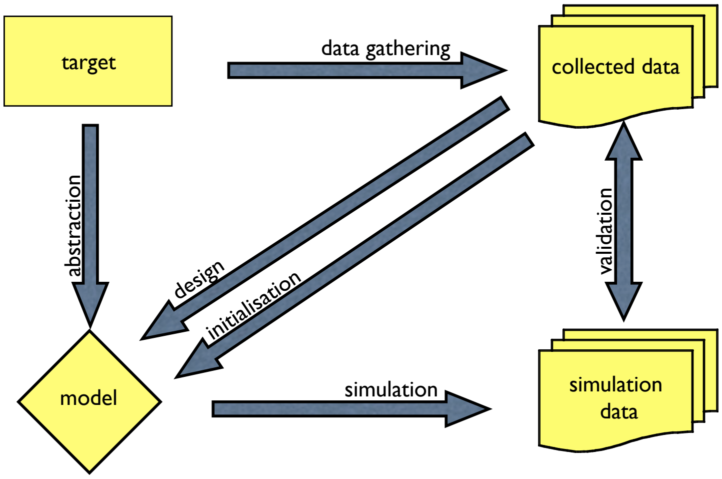

**Fig. 2.** Diagram showing a modified data-driven logic of simulation.

## 5.2 Complete Diagram

The proposed agent-based modelling process is therefore driven by data. Data plays a role in most of the activities. Some activities are devised specifically for collecting and preparing data, and others make use of these data, more specifically for building and initialising the model, and for its validation. Fig. 3 shows which kinds of data participate in these activities. The figure does not show the iterative nature of the process. The role of data in the main stages is as follows:

- Data collection from different sources. Note that there are different kind of data, such as qualitative data, or quantitative data that could be equation-like, from surveys, etc. As it has been mentioned in section 4.2, all necessary data is not always available, or do not have the structure that would facilitate conceptualisation, abstraction, and correspondence with the model. It is also possible to find contradictory data or data that is difficult to correlate. In this sense, this activity implies the selection of representative data sources and analysis of its structure, which could also be valuable for the other activities. All this is naturally guided by the supporting theory and hypotheses.
- Data pre-processing. In order to have usable data, these have to be processed and adapted to structures in the agent-based model. Some techniques to assist in structuring and inferring are mentioned in section 5.3.
- Build model. The resulting data is an important input to characterise agents in the model. They can be used to systematically identify agent attributes, agent relationships, and variables of the environment. Some results from data pre-processing can provide also information for the design. For instance, clustering can be useful for identifying some kind of relationships to implement. Theories and hypothesis are also fundamental for building the model. In fact, they are the subject of validation by using the simulation. Theories drive the abstraction process, i.e., the consideration of which attributes are relevant for the model, and imply relationships, which should be validated by simulation, for instance. Moreover, agent behaviour is derived from them.
- Model initialisation. Empirical data are also used for this activity, as discussed above in section 4. Data structures were used for building the model, and data values should be used to define the initial population of agents, with their initial values, and the initial parametrization of the environment.
- Simulation. This is the process of exploration through the executions of the model in a simulation environment, from certain initial conditions, with the possibility of changing its configuration. The modeller is guided by the objectives, theory and intentions in this process.
- Post-processing of simulation results is required for the alignment of data generated from several executions. It may also be used to facilitate interpretation by analysis tools and the social scientist. Therefore, it can be advisable to use some data mining tools at this point to discover emergent patterns; to extract graphs, statistics, natural language reports, logs, network diagrams, etc; to compare the results with the clustering or ontologies of pre-processing. The technologies mentioned are addressed in section 5.3.
- Analysis and Validation of results. At this point, collected data are used to contrast with results. Also, analysis is driven by the theories and hypothesis that are to be validated.

### 5.3 Technologies Can Help

Collected data usually needs some kind of treatment before it can be useful in the design and initialisation processes. Moreover, there are multiple issues concerning ABM that should be addressed carefully, specially in the case of data-driven

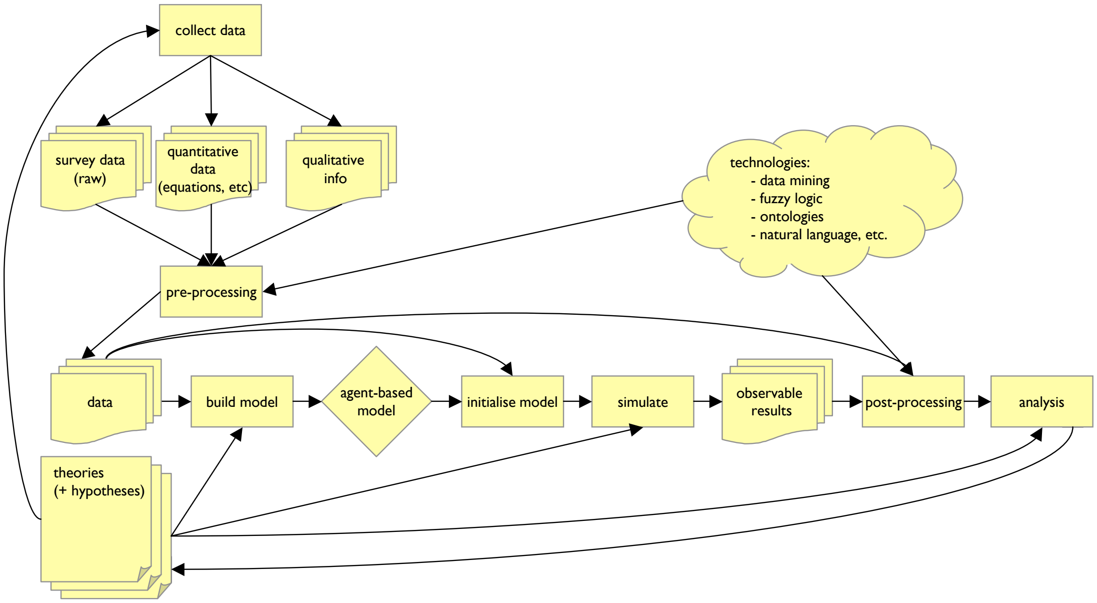

**Fig. 3.** Proposed flow diagram for the data-driven agent-based modelling.

simulation. The social phenomena usually follow a smooth behaviour, which is hard to represent with standard programming algorithms. Soft computing techniques [11] including neural or bayesian networks, fuzzy systems, and evolutionary computation can be helpful. Depending on the case, one or another can be used with good results. For example, neural networks are useful for adaptive learning behaviours; fuzzy logic is helpful in modelling social processes; evolutionary algorithms usually substitute agents as another way of doing simulation, but they can also be used to optimise agent behaviour.

Other problematic issues that can be solved are the search for patterns and characterized groups (clusters) in the input data or in the simulation output. The larger the amount of data, or the more complex is the phenomena represented, the more difficult it is to find patterns and clusters. However, there are several Artificial Intelligence (AI) tools, such as classifiers and data mining, which can make it rather easy. Representation of the concepts is another complication. Ontologies represent a easy-to-handle interface with experts, and a formal view that can be inserted in the ABM. Natural Language Processing can be proposed for a better representation of the simulation output, prepared for non-experts.

## 6 Discussion and Difficulties

The introduction of empirical data in ABM implies some costs. Here we discuss the main issues that arise:

- In some cases complicating the model with empirical data will not benefit the results. Then, a KISS model would be the ideal approach. It is advisable to use empirical data but the way it can be applied in the different activities of the ABM process depends on several factors such as the type of available data, its structure, its reliability, its quantity, its structure, its ability to be processed, etc.
- The ease of understanding and communication associated with KISS is partially lost with this kind of modelling. However, a modular well-defined specification should be helpful in this sense. Moreover, in the deepening approach, the structured gradual process of increasing complexity facilitates understanding together with extensibility.
- Data-driven modelling demands a special effort in gathering data. Although this process is frequently required for validation, it may not have the intensity required here. The additional costs may not be worthwhile in certain cases. Moreover, validation of very abstract ABMs can be theoretical or through a process of sensitivity analysis, not requiring a deep comparison with collected data. In those cases to turn to a data-driven approach means a high cost that may be difficult to justify, in spite of the expected improvement of results.
- In subsection 4.2 several specific difficulties related to the procedures have been addressed: surveys not providing exactly the required data; lack of information; qualitative or subjective data not easily gathered. Besides, if the data is extracted from several sources, it can be quite difficult to match it: different indicators, data not complementary or even contradictory. And handling huge amounts of data makes still more complicated the process of deciding what is relevant. In all those cases representativeness and hypothesis should play an important role.
- About the technologies mentioned in 5.3, each one can be useful only in a limited range of cases. For instance, data mining needs large amounts of data to be effective. Fuzzy logic requires blur properties or concepts to deal with. Ontologies may be useless in cases where the classification is too simple. The output in natural language can be considered non-crucial for the implementation effort that it requires, although there are already several tools for NLP that could be useful depending on the context.

## 7 Concluding Remarks

The debate between abstraction and descriptiveness has been going on in the philosophy of science for quite some time. In Social Simulation, two different perspectives have been proposed to place empirical-based models correctly in this discussion: KISS and KIDS, one starting from simplicity, the other from full descriptiveness. Models informed by data used in social sciences go through verification, validation, sensitivity analysis, calibration, and so on. However, many unfounded assumptions and design options still have an important influence on how the exploration of the models will yield and support conclusions that can be used to explain, predict, and even prescribe solutions for the problems addressed. The Mentat model [8] was used as a stereotypical project in which a combination of data use and iterative deepening on details of the design could inspire a middle way approach between KISS and KIDS. For instance, when considering data, radical simplifications such as the use of theoretical distributions, supported by statistical tests on empirical data, need to be considered. In this paper, we have shown how to go through the injection of data through the design, construction, exploration, and analysis of a model designed to generate deeper insights into complex social problems. Building on results from [9], we put forward a new simulation logic that uses data to help build and initialise the model, and not only for validation purposes. We have provided guidelines and suggested the integration of technologies to insert multiply-sourced, unrelated, and/or unstructured data into the iterative exploration of series of increasingly complex computational social models.

**Acknowledgments.** We acknowledge support from the projects: *NEMO: Network Models, Governance and R&D collaboration networks* funded by the European Commission Sixth Framework Programme - Information Society and Technologies - Citizens and Governance in the Knowledge Based Society, and *Methods and tools for modelling multi-agent systems*, supported by Spanish Council for Science and Technology, with grant TIN2005-08501-C03-01.

## References

1. L. Antunes, H. Coelho, J. Balsa, and A. Respicio. e*plore v.0: Principia for strategic exploration of social simulation experiments design space. In S. Takahashi, D. Sallach, and J. Rouchier, editors, *Advancing Social Simulation: the First World Congress*, pages 295–306. Springer, Kyoto, Japan, 2006.
2. R. Axelrod. Advancing the art of simulation in the social sciences. *Complexity*, 3(2):16–22, 1997.
3. B. Edmonds and S. Moss. From KISS to KIDS - An 'Anti-simplistic' Modelling Approach. In P. Davidsson, B. Logan, and K. Takadama, editors, *MABS*, volume 3415 of *Lecture Notes in Computer Science*, pages 130–144. Springer, 2004.
4. A. Geller. Power, resources and violence in contemporary conflict: Artificial evidence. In *2nd World Congress of Social Simulation*, Washington, D.C., 2008. (To appear).
5. N. Gilbert. *Researching Social Life.* SAGE Ltd., third edition edition, Mar. 2008.
6. N. Gilbert and K. G. Troitzsch. *Simulation for the Social Scientist.* Open University Press, Apr. 1999.
7. P. Guyot and A. Drogoul. Designing multi-agent based participatory simulations. In H. Coelho and B. Epinasse, editors, *Proceedings of 5th Workshop on Agent-Based Simulation*, pages 32–37, Lisbon, 2004. Publishing House.
8. S. Hassan, L. Antunes, and M. Arroyo. Deepening the demographic mechanisms in a data-driven social simulation of moral values evolution. In N. David and J. S. Sichman, editors, *MABS 2008: Multi-Agent-Based Simulation*, pages 189–203, Estoril, Portugal, 2008. Springer. To appear in Springer LNAI.
9. S. Hassan, J. Pavon, and N. Gilbert. Injecting data into simulation: Can agent-based modelling learn from microsimulation? In *2nd World Congress of Social Simulation*, Washington, D.C., 2008. (To appear).

10. C. Kennedy and G. Theodoropoulos. Intelligent management of data driven simulations to support model building. In *20th Workshop on Principles of Advanced and Distributed Simulation, 2006. PADS 2006*, page 132, 2006.
11. S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach.* Prentice Hall, 2 edition, Dec. 2002.
12. A. Sloman. Explorations in design space. In *Proc. of the 11th European Conference on Artificial Intelligence*, 1994.
13. L. Yang and N. Gilbert. Getting away from numbers: Using qualitative observation for agent-based modelling. In F. Amblard, editor, *ESSA'07: Fourth Conference of the European Social Simulation Association*, pages 205–214, Toulouse, France, 2007.