

An Approach to Evaluating Human Characteristics in Agents

Emma Norling¹ and Liz Sonenberg²

¹ Computer Science and Software Engineering
The University of Melbourne ejn@cs.mu.oz.au

² Department of Information Systems
The University of Melbourne

Abstract. When complex cognitive or social characteristics, such as emotions or organisational behaviour, are incorporated into agent frameworks a rich dynamic environment is needed to test and evaluate these modifications. Simple scenarios simply cannot provide the complexity required to test these behaviours. Not only does increased interest in dynamic behaviour of agents require the investigation of new modelling concepts, but it also demands a new perspective on strategies for evaluation, for often there are no objective measures of success for these models. What, for example, is to say that adding emotion to an agent makes it more realistic, or that one model of emotion is superior to another? To answer these and similar questions, statistical techniques borrowed from the social sciences must be employed to obtain meaningful evaluations from a collection of subjective judges.

This paper discusses the use and benefits of a commercial multi-player game as a testing environment for models of human behaviour. It describes a project in which the BDI (*belief-desire-intention*) agent framework has been modified to have more human-like decision-making strategies, the process of selecting a testbed, and the evaluation procedure being applied. The issues that have arisen in this project are then discussed in a more general context.

1 Introduction

Human-like qualities are being added to software agents in many different guises for many different reasons: emotions are being added to agents in training simulations to make their behaviour more realistic [5]; models of perception and action are being added to agents for interface testing [16]; social learning has been added to agents as an alternative to classical learning approaches [4]; and there are many other examples. In the area of human modelling alone, the growth in popularity of agent-based models has driven extensive work in this area because there has been a recognition that existing agent frameworks are lacking many characteristics that are generally applicable to human modelling [14]. For example, a model of personality, or of teamwork, will be broadly applicable to human modelling in many domains. Researchers are now commonly addressing

these sorts of issues. With the recent foray of the games industry into this area (see *Black & White* by Lionhead Studios [10]), there is likely to be even more of a push for a standard framework, as there has been previously for game engines.

One of the issues that researchers in this area face is that they are developing models to incorporate human characteristics that are *beyond* the needs of current applications. The problem here is that this means the existing application is unlikely to test the modification. How then can they be evaluated? One possibility is to enhance the agent frameworks and then just sit back and wait until the applications catch up to test the enhancements, but obviously this is not a satisfactory way of going about research. A better approach is to test them in another environment that does exercise these new characteristics, but where is such an environment found? One possibility is to build it from scratch, but this could be a mammoth task – one only has to look at the size and complexity of existing environments to realise this. A second option is to extend an existing environment, which may be viable if one has access to the code base. A third option, and that which is advocated here, is to find an alternative existing environment that will exercise the new features of the model without needing any modifications.

Once a testbed is built, modified or found, the challenges are not over. Many of the characteristics that researchers want to incorporate into their models cannot be easily quantified and so the question of qualitative evaluation also deserves considerable attention. Researchers in AI are increasingly familiar with empirical methods, thanks to the work of Cohen [3] and others, but as well as the uncertainty and variability involved in many AI studies there is also the question here of how to measure improvements. Making a better model of a human will not necessarily lead to better performance in an environment – indeed a plausible enhancement would be to make the agent make the same sort of mistakes that a human would – so this cannot be used to evaluate the modification. For many modifications, the only way to judge their efficacy is to solicit people’s opinions, and people will naturally be in some way subjective judges.

In this paper the issues involved in the evaluation of this type of agent are examined in the context of a project which is currently in progress. This project was inspired by work undertaken at Air Operations Division (AOD) of Australia’s Defence Science and Technology Organisation (DSTO), where agents have been used to simulate people in operations analysis for a number of years. The project aims to enhance the agent framework that they use, so that agents built in the framework display more realistic behaviours than those currently in use. We discuss the testing and evaluation of a modification to the decision-making strategy of the agents to illustrate the issues raised.

2 Selecting a Testbed

To evaluate human characteristics in agents, the testing environment must be sufficiently rich to test the characteristics of interest – that is, the characteristics that need to be evaluated should be exercised in the course of normal tasks in

that environment. A second requirement is that the humans who would normally operate in that environment – the Subject Matter Experts (SMEs) – be accessible, because the modified framework needs to be populated with domain-specific knowledge so that the agents can operate in the environment. If the test environment simulates a nuclear power plant, for example, getting access to the plant operators to get domain knowledge for the agents could be challenging. Access to the SMEs is also important for evaluation purposes – see Sec. 3.

When the work is being undertaken for a specific target application, this application might be suitable as a testbed, but there are also a number of reasons that it may be unsuitable. As well as the potential problem with SME access, the target application might only be envisioned and not yet implemented, and the researchers may have limited access to the application because of commercial and/or security concerns. In some cases minor changes can be made to an existing application to make it a suitable testbed, but the researchers may not have access to the code base, or the changes needed may be too significant to be feasible.

If the target application is unsuitable, or indeed there is no specific target application, it may seem appropriate to build a new environment for testing the agents. The difficulty with this approach is that building an environment that has the complexity to test the characteristics of interest can take more time and effort than the development of the agents themselves. Examining the development history of existing application environments should convince one of the effort required.

The final option is to find some alternative existing environment that will be a suitable testbed. If such an environment can be found, it can greatly reduce the time and effort in establishing a testbed. For our project, we found such an environment in a commercial computer game, as have other researchers – see the section on “Computer Games in Scientific Research” in a recent edition of *Communications of the ACM* [7], or Capps et al’s summary of use of computer games by the U.S. Department of Defence [2] for examples.

Computer games provide some of the richest simulated environments that are generally accessible. For a number of years, one of the selling points of computer games has been the realism or immersiveness of their environments. This is not to say that they are necessarily perfect copies of real-world environments, but they are believable “possible” worlds. They usually contain sophisticated physical models (so that crashes, projectiles, explosions, etc are realistic), detailed terrain, and often other environmental factors such as weather. Even though these environments are not copies of the real world, they allow exploration of many of the same issues that are encountered in the real world. The use of these environments is facilitated by games companies who often distribute source code with the game, which allows users to “plug in” their agents to the game, or alternatively agents can interface with networked games through the same links that a human player would use.

A second benefit of using a computer game as a testbed is that SMEs are usually plentiful. While it may be extremely difficult to access a fighter pilot, or intensive care nurse, or nuclear power plant operator, there are many computer

game players around, who will usually be quite open to the idea of talking about their skills or playing a few games in the interests of research. It is necessary to explain *why* the tasks in the game will exercise the same characteristics as the task in the domain that will ultimately be simulated, but if this case can be made, it will be a far less expensive way of testing the characteristics.

In the case of the project described here, Quake II was selected as the testbed. This game can be played in both single- and multi-player forms, with the multi-player version being far more popular. Using the multi-player version for evaluation allows people to interact with the agents, which is important for our procedure, see Sec. 4.4. Although the focus of our research at this stage is on decision-making strategies, the game also provides an opportunity to study complex social behaviours, in the interaction between players (both agent and human) in the game.

Many other games would also have provided a suitable testbed for our work. The deciding factors for Quake II was that researchers at the University of Michigan were also using this game for similar research [17], and that there is a wealth of information available about the game, strategies for playing it, and how it works at the program level.

3 Evaluation Procedures

The major difficulty in evaluating human characteristics in agents is that there is often no simple quantitative measure of their benefit. It is essential to consider *why* the characteristic has been added – was it to make the agents more accurate, more realistic, easier to interact with, or something else? If it was to make them more accurate, evaluation is relatively straightforward: give them a set of tasks and measure their accuracy. However, if it was for realism, or ease of interaction, the only way to judge these sorts of things is to ask the opinions of people. These people will usually be subject matter experts in the test domain, because they will be the best judge of what is realistic, or easier to interact with, or whatever the measure might be. However any person, expert or not, is naturally a subjective judge, and so qualitative evaluations like this must be handled carefully.

This inherent subjectivity of the judges means that care needs to be taken in analysing their responses to any surveys. There should be sufficient participants that their responses can be generalised, so that at the end of the evaluation one can say with confidence whether or not the characteristic that was added was effective in improving the realism of the agents. McBreen et al give an example of this type of analysis [8]. In their case, they were evaluating the acceptability of human-like agents, rather than the realism of them, but this changes the questions to be asked of the judges, rather than the procedure to be followed. While this type of rigorous evaluation is common and well-understood in the social sciences, unfortunately it is far less common in the agent community.

4 Case Study: A Realistic Decision-Making Strategy

These issues involved in evaluation of human characteristics are illustrated here in the context of a particular project, which aims to enhance the realism of agent-based models of Air Force personnel. Discussions with analysts working on different projects at AOD have highlighted several human characteristics which are not part of the framework being used but are likely to be important in their studies. This project addresses one of these issues: that the framework does not use human-like decision-making strategies.

4.1 Background

Analysts at Air Operations Division use agents to model a wide range of Air Force personnel. The existing framework is an implementation of the BDI model of agency [15], in which agents are characterised by their beliefs about the world, their desires, and the plans that they can use to attempt to fulfil those desires. Each BDI agent is situated in some environment, which it senses (usually incompletely and imperfectly), and in which it acts. Based upon the sensor data that it receives, the agent updates beliefs, manages goals (desires), and selects plans (from a fixed plan library) to form intentions, which in turn lead to actions. A person building an agent in this framework needs to specify the plans for each agent, along with the types of goals and beliefs that it can have.

A major advantage of the BDI framework is that subject matter experts (SMEs), who are generally *not* programmers, can easily express their expertise in a form closely matching what is needed to construct the agents, and, with appropriate visualisation tools, can also understand what the agent is doing (and *why* it is doing it) while it is running. This greatly speeds up development and testing time for the agents [9].

Although the BDI framework provides this high-level abstraction of human reasoning, this also leads to its major disadvantage: that there are many human characteristics which are not part of the framework, but may be important to model in a range of applications. These shortcomings include: unrealistic decision-making strategies; no concept of time taken to perform actions, process information, etc; workload has no effect on performance; agents have no notion of team or social structures; and so on. While this is not an exhaustive list, it serves to illustrate the types of things that analysts at DSTO consider to be important. Some of these issues are already being addressed, and others will be in the near future. The project described in this paper addresses the first issue, incorporating a more human-like decision-making strategy into the BDI framework.

Another advantage of the BDI framework is that many human characteristics are commonly described in similar folk psychological terms – for example, Ortony et al’s model of emotions [13] talks of goals, beliefs and intentions, and Klein’s model of decision-making [6] refers to goals, cues (mapping to beliefs) and “courses of action” (mapping to plans). These types of correspondence should

simplify incorporating these models with the BDI agent framework. The example of adaptive decision-making described in Sec. 4.2 is one such implementation.

4.2 The Enhancement

In this project, a model of naturalistic decision-making called the *Recognition-Primed Decision model* (RPD) was incorporated into the BDI agent framework. RPD is a descriptive model of decision-making developed by Gary Klein [6] that has been shown to be highly applicable in the type of domain used by AOD. It is characterised by the fact that people learn from their mistakes: when a “bad” decision is made (one that leads to some lack of success), a human operator will attempt to determine the reason for failure, and avoid making the same choice in those circumstances.

For example, when an agent has several plans available that will allow it to achieve a particular goal, how should it choose which one to use? Formal descriptions of the BDI architecture usually describe a mechanism using utility theory – where the relative merits of each option are considered and the one with the highest utility is selected. Practical implementations of the BDI architecture are usually much more simplistic, often just selecting the first applicable plan. These strategies are rarely used in real life, particularly in the type of domain described above.

Agents using this decision-making strategy have been implemented in the JACK [1] programming language. Meta-level reasoning (i.e. the plan selection strategy) uses Q-learning, with the environment characterised by the cues that the SME considers pertinent [11]. Further details of the implementation will be available in a forthcoming publication. If the evaluation of these agents indicates benefits using this alternate decision-making strategy, it may be incorporated as an extension to the implementation language.

4.3 Testbed Selection

It would have been difficult to use any existing application as a testbed for several reasons in this project. Firstly, current applications are designed so that they avoid scenarios that are likely to need this sort of decision-making. It may have been possible to extend them so that they did, but this would have meant developing new scenarios, which would have required the input of SMEs. As the SMEs in this case are Air Force personnel, this would have been difficult to arrange. Also, these applications often contain classified information. Removing or aliasing this information so that results could be presented to the research community at large would have taken considerable effort. These factors combined to encourage us to find an alternative testbed.

At an early stage of this project, a simple testbed was built in which an agent had to make its way across town, choosing between a number of different modes of transport, and influenced by a number of different conditions [12]. While this testbed was sufficient to show the limitations of that implementation of the decision-making strategy, it was also clear that it was not a sufficiently

rich environment to test a more sophisticated model. More importantly, the effort taken to construct that simple environment indicated that building our own testbed would be extremely time-consuming, reinforcing our argument that it is inadvisable to build a test environment from scratch.

As mentioned previously, the testbed that has been finally selected for evaluation of these agents is the commercial game of Quake II. The game meets the two key requirements for an evaluation testbed. Firstly, the domain characteristics closely match those of the target domain. In both cases, the environment is dynamic, information about the environment is incomplete and uncertain, there is time pressure, it is high risk, there are action/feedback loops, and there are shifting, ill-defined and sometimes competing goals. Secondly, the tasks in the testbed do include opportunities to exercise the decision-making strategy. This can be seen by the way that human players adapt their strategies based on their opponents. In the game the ultimate goal is to kill as many players as possible, but this can be achieved with many different strategies.

4.4 Evaluation of the RPD Agents

The evaluation stage of this project is only just beginning, but we describe the procedure that we are using as an example. To combat the subjectiveness of human judges, a range of SMEs – in this case, experienced Quake II players – will be asked to evaluate the agents, allowing us to statistically analyse the results to get confidence measures. In addition, although all participants will be informed that the second agent that they interact with will be the modified agent (which should be more human-like) half of the participants will actually interact with the same agents for both sessions. This control group should show if any expectations arise from the instructions given to the participants. For this experiment we have chosen not to divulge the type of modification that was made, because we are interested more in whether the performance of the agents as a whole is more human-like, not just that particular characteristic. If the decision-making is more human-like but the overall performance of the agents is effectively the same, it would not be worth incorporating the modified decision-making strategy permanently into the agent framework (at this stage at least).

The participants in the evaluation will be asked to complete a three stage questionnaire. Before interacting with any agents, they will be asked questions about their experience playing Quake II, their style of playing the game, and the sorts of improvements that they might expect to find in a more human-like computer generated character. The first two questions (experience and playing style) will be used to determine if there is any correlation with their answers to the subsequent sets of questions. The final question comes about because all game players will have played against computer generated characters at some stage, if not in Quake II then in some other game. We expect that they will already have opinions about the issue of human-like computer generated characters.

The second set of questions will be asked after playing a game against the agents using the original framework. In these questions, they will be asked to

classify the agents that they played against (i.e. the agents' playing styles), to rate them against a human player (using a five point scale) and give reasons for these ratings, and finally, given their experience playing against these agents, to list the sorts of improvements that they will be looking for in the modified agents. It will be interesting to see if this corresponds to their answer to the pre-game question, or if the original framework has advantages or disadvantages over standard computer generated characters.

The final set of questions, which will follow the game against the modified agents (or unmodified for the control group), will ask them to rate these agents against the ones in the first game using a five point scale – are they 1) far less human-like; 2) slightly less human-like; 3) no different; 4) slightly more human-like; or 5) far more human-like? – and to give reasons for their ratings.

Hopefully the results of these questionnaires will show conclusively that the modification leads to more realistic agents. If the results are less conclusive, there should also be sufficient information in the survey responses to determine the reason(s). Possibly some other characteristic will be more important, in which case it will be necessary to determine whether this is also true for the Air Operations domain. The questions about expectations are expected to provide this sort of information, as well as provide focus for future research.

5 Conclusions

Testing and evaluating agents with human characteristics is a difficult and time consuming task. The unwary can spend more time developing the testbed for the agents than they devote to the agents themselves, and evaluation can sometimes be little more than “feel-good” appraisals. We believe that the approach we have taken in our project, both in selecting a testbed and in evaluating the modification, is generally applicable to attempts to add human-like characteristics to agents.

As discussed, existing applications can be unsuitable testbeds for the agents for a number of reasons. It is sometimes possible to modify these applications so that they are suitable, but more often researchers will need to use an alternative application as a testbed. In this case we strongly discourage the thought of building a testbed from scratch. Instead we advocate finding an existing alternative application that will not require any modification. Many commercial computer games fit these requirements, and addition, because of the popularity of computer games, there is usually a large pool of subject matter experts available, which can be important for evaluation purposes.

This brings us to the subject of evaluation. When evaluating human-like characteristics in an agent, it is essential to consider what the goal of including those characteristics was. If it was to improve accuracy, or speed, or a similar quantifiable measure, standard quantitative techniques can be used for evaluation. However it is increasingly the case that the aim of adding such characteristics is something that cannot easily be quantified. In these cases, it is necessary to

employ qualitative evaluation procedures, as are often used in the social sciences. The procedure we have described is an example of such an evaluation.

With the growing complexity of agent-based models of people, we expect that more and more researchers will be facing these issues in testing and evaluating their systems. We stress that finding a suitable testbed and evaluating the system are steps that take considerable time and thought, and should be taken into consideration in the early stages of a project. Once a modification or enhancement has been added to a system, it needs to be shown to be worthwhile – there is no point advocating a modification that makes little or no change to the overall performance of a system, or even degrades its performance. We hope that our experiences in the project described here will help others working on similar projects.

References

1. Agent Oriented Software Pty. Ltd. JACK Intelligent Agents. <http://agent-software.com.au/jack.html>.
2. Michael Capps, Perry McDowell, and Michael Zyda. A future for entertainment-defense research collaboration. *IEEE Computer Graphics and Applications*, January/February 2001.
3. Paul R. Cohen. *Empirical Methods for Artificial Intelligence*. The MIT Press, Cambridge, Massachusetts, 1995.
4. Rosaria Conte and Mario Paolucci. Intelligent social learning. *Journal of Artificial Societies and Social Simulation*, 4(1), 2000.
5. Jonathon Gratch and Stacy Marsella. Tears and fears: Modeling emotions and emotional behaviors in synthetic agents. In *Proceedings of the Fifth International Conference on Autonomous Agents*, pages 278–285, Montreal, Canada, May 2001. ACM Press.
6. Gary A. Klein. A recognition-primed decision (RPD) model of rapid decision making. In Gary A. Klein, Judith Orasanu, Roberta Calderwood, and Caroline E. Zsombok, editors, *Decision Making in Action: Models and Methods*, pages 138–147. Ablex Publishing Corporation, 1993.
7. Michael Lewis and Jeffrey Jacobson. Game engines and scientific research (introduction to the special section). *Communications of the ACM*, 45(1), 2002.
8. Helen McBreen, Paul Shade, Mervyn Jack, and Peter Wyard. Experimental assessment of the effectiveness of synthetic personae for multi-modal E-retail applications. In *Proceedings of the Fourth International Conference on Autonomous Agents*, pages 39–45, Barcelona, Catalonia, Spain, 2000. ACM Press.
9. David McIlroy and Clinton Heinze. Air combat tactics implementation in the Smart Whole AiR Mission Model (SWARMM). In *Proceedings of the First International SimTecT Conference*, Melbourne, Australia, 1996.
10. Peter Molyneux. Postmortem: Lionhead Studios' Black & White. *Game Developer*, June 2001.
11. Emma Norling. Learning to notice: Adaptive models of human operators. In *Second International Workshop on Learning Agents*, Montreal, Canada, May 2001. ACM.
12. Emma Norling, Liz Sonenberg, and Ralph Rönquist. Enhancing multi-agent based simulation with human-like decision-making strategies. In *Multi-Agent-Based Simulation, Second International Workshop, MABS 2000 Boston, MA, USA. Revised*

- Papers*, volume 1979 of *Springer Lecture Notes in Artificial Intelligence*. Springer, 2000.
13. Andrew Ortony, Gerald L. Clore, and Allan Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, 1988.
 14. Richard W. Pew and Anne S. Mavor, editors. *Modelling Human and Organizational Behavior: Application to Military Simulations*. National Academy Press, 1998.
 15. Anand S. Rao and Michael P. Georgeff. BDI agents: From theory to practice. In Victor Lesser, editor, *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95)*. MIT Press, 1995.
 16. Robert St. Amant and Mark O. Riedl. A perception/action substrate for cognitive modeling in HCI. *International Journal of Human-Computer Studies*, 55:15–39, 2001.
 17. University of Michigan Artificial Intelligence Laboratory. SoarBot Project Overview.
<http://ai.eecs.umich.edu/~soarbot/>.