Technical Evaluation Report MSG-107 meeting

Orlando (FL) December 2, 2011

Dr. Wouter Lotens
TNO Behavioural and Societal Sciences
Kampweg 5, 3769DE Soesterberg
THE NETHERLANDS

wouter.lotens@tno.nl

OVERVIEW

The NATO Modelling and Simulation Group MSG-107 "Human Behaviour Modelling for Military Training Application" organized an unclassified workshop in Orlando (FL) in the framework of a series of meetings that aim to further the proper use of Human Behaviour Modelling (or Representation) for military application. The audience was diversified and was composed of 29 attendees from the military (12), government R&D laboratories (13) and a significant representation from industry (4).The attendees originated from 10 nations (ITA, USA, CAN, NLD, FRA, GBR, GRE, AUS, TUR and GER) and from ACT and NATO. Six presentations were provided as introduction during a half day workshop, with a good share of the time reserved for discussion. This report presents a technical evaluation of the various presentations and the technical discussion that occurred during the workshop and provides overarching comments and recommendations based thereon.

1.0 INTRODUCTION

In 2010 the HFM panel organized a meeting on Human Behaviour Representation (HFM-202) with the explicit purpose of bringing ACT, NMSG, SAS and HFM together. These bodies each have their own visions on and interpretations of how human models could be exploited in simulations. The purpose of the simulation is a key aspect in these views. Human Factors (HF) specialists may want to understand and describe specific aspects of human behaviour (such as cultural values, decision making, tiredness, overheating, perception errors and many others), Operational specialist may be interested in how HF propagate in the military organisation and operations, NMSG may be focussing on how human behaviour could be captured in synthetic training partners and ACT may want to know how the forces can take advantage of the Human Behaviour Representation (HBR), for instance, increasing efficiency, in foresights on new ways of operation or in the assessment of the potential of new technologies. During HFM-202 there was a strong feeling that HBR indeed is a common interest and that the barriers for use should be taken down. Such barriers were identified as knowledge stove piping, lack of reusable tools, lack of a generic architecture and sometimes lack of scientific data. The latter pertains foremost to the field of social and cultural awareness that has become so relevant in the recent expeditionary operations.

RTO decided that a more common approach to HBR should be undertaken, integrating the interests of the various panels. This MSG-107 meeting is the first action to follow this up, focussing on military training and investigating the challenges and solutions of this specific application of HBR. This meeting will, hopefully, provide insight how to continue with the integration process and this point will obviously be addressed again in the recommendations.

2.0 WORKSHOP OBJECTIVES

There are two objectives regarding the type of audience and four regarding specific activities:

Organization of a Workshop that brings together domain experts from different disciplines: human modelling, social sciences (sociology, social psychology), cognitive psychology, physiology, anthropology, operational analysis, computer science, modelling standards and artificial intelligence and innovative man-machine interaction technologies (e.g. speech, gestures, mixed reality, natural language understanding).

The intention is inviting not only top experts with deep knowledge of these disciplines, but also military and industrial domain experts and scientists with a broad view, able to understand the integration demands from different domains in perspective. The activity will be organized by the NMSG but will require domain experts from R&T communities associated with HFM and SAS, academia and industry.

The activity is to deliver:

- Inventory of challenges and known pathways to solutions (papers 1, 3, 4)
- Identification of knowledge and technical barriers to integrated model development (papers 2, 3, 4, 5)
- High level guidelines for human modelling and its interaction with Live players in Military Training Simulations (papers 2, 4, 5)
- NATO perspectives of exploitation of human behaviour modelling advances (papers 2, 6)

Overlooking the list of participants not all the intended domains were represented, but the attendees were well spread over the NATO bodies, background (military, industry and research) and nations. The papers often addressed two or more activities and together gave a fair coverage of the intended activities, as shown between brackets at each activity.

3.0 WORKSHOP PROGRAMME

Following the opening of the workshop, by Chair Wim Huiskamp, six speakers presentated cases in a single session, followed by a general discussion. The presentation proceedings can be found at: http://myrto.rto.nato.int/msg/MSG-107/default.aspx.

Table 1: MSG-107 case presentations.

December 2, 2011: Presenters and titles, no co-authors	
1	Thomas Alexander: Applications and Limitations of today's Digital Human Models
2	Emilie A Reitz: Effect of Live Training on Virtual Environment Performance Title: Effect of Live Training on Virtual
3	Jan Joris Roessingh, Roel Rijken: Modelling CGF Behaviour in tactical air-to-air combat training Smart Bandits'
4	Philip Kerbusch: Socio-Cultural Behavior in Simulated Military Training
5	Jonathan A. Cohen: Toward Predicting Individual Soldier Cognitive Performance
6	Brad Cain: Validating virtual reality training simulations

4.0 EVALUATION

This section summarizes the presentations. The workshop organization was well chosen in conjunction with the IITSEC, and took place in the same premises. The chair gave everyone the opportunity to present them self, although many were already familiar. The quality of the presentations was quite good.

4.1 Presentations

Thomas Alexander presented an overview of anthropometrically correct models, including motion strategy and reach. He used those to attempt to mimic house clearing operations, using an engineering network approach for the teamwork, evaluating the impact of equipment on operational performance. He concluded that digital manikins are not moving like humans and that observation of real training would teach more. The engineered behaviour (complex motion build in simulations from small motion elements) is not realistic and there is a need for interfaces that allow the input of realistic movement, possibly exploiting motion trackers. Commercial games are no good drivers for this purpose.

Emilie Reitz did an experiment to compare the training effects of real and virtual environment (VE), concluding that training in a real environment does not help to 'read' the virtual environment better. Moreover, both trainings have the same positive effect on declarative knowledge when they are presented as 'classroom images'. This is surprising, as the software used (VBS2) is crippled in showing details, such as attitudes and facial expressions. Food for thought is also that the variation in Situational Judgement is smaller in VE than in reality, possibly pointing at more complete retrieval of information elements in VE. Interactions between agents in the VE are difficult to generate, requiring live exercise controllers and thus diminishing the advantage of computer generated team mates.

Jan Joris Roessingh and Roel Rijken made an attempt to include tactical intelligence in Computer Generated Forces (CGF) in an air combat scenario. The perception in this application is completely instrumental. They tried to master the potential complexity of adaptation of opponent behaviour to new conditions by using machine learning. This is effectuated by using neural networks parameters as input to a mixed cognitive model (including Belief, Desire, Intention model (BDI), Endsley type Situation Awareness (SA) and human constraints), Natural Decision Making (complex believes, future projection) and Theory of Mind (reasoning on opponent thinking). The resulting model has some similarity to Belief networks. The authors recognize the danger that student behaviour is adapting to wrong opponent behaviour, as there is no prior knowledge on which the opponent behaviour is based. Therefore dynamic scripting is planned as a follow on exercise. Validity has not been addressed so far. The hybrid model makes a good attempt to separate the human from the application, which may help greatly to reuse the model in other scenarios. Interested parties are invited to participate in an upcoming IST task group on the issues addressed.

Philip Kerbusch attempted to make smart use of existing theories (BDI, Culturally Affected Behaviour (CAB) and personality (Big Five)) to generate culturally determined reactions to events. The lack of a unique definition of culture and of its interaction with personality is not making this easier. The intended application is to train military to interact with persons of another culture by means of virtual players (agents), but this paper is focusing on the model architecture that may be needed. The agents are constantly evaluating the behaviour of the human player against their cultural norms and acts according to BDI and CAB, and according to their personalities and emotions. The Jadex software is used. The architecture works at face value, but no evaluation is made of the realism of the generated behaviour.

John Cohen showed how data regression can be used to model the effect of physical activity in Chemical Biological, Radiological Nuclear (CBRN) protective clothing on cognition. The effects run probably through a chain involving body heating, exhaustion and decreasing mental resources, but the regression goes straight from Moisture Vapour Transmission Rate (a clothing property) to reaction time (one measure of mental performance). By means of enough experimental conditions he was able to mathematically fit cognitive performance to several clothing and task parameters. The validity of the results for the domain in which data have been collected is undisputed, but extrapolation to other conditions is not based on knowledge of underlying relationships. The same method is planned for the perception of threats to crowds.

Brad Cain addressed the important methodological issue of validity, distinguishing between face validity, content validity and construct validity. These concepts are ordered nor exclusive. Face validity is often strived for, but for training purposes the appearance may be irrelevant or even counteractive. Given the limitations of simulations, it is better to handle the concept of Fitness for Purpose: does the simulation have a positive training transfer to reality for the intended learning goal. This must be assessed experimentally, which might be laborious. In an example assessment expert subjects were improving on simulator performance, which may mean that they adapted to the peculiarities of the simulation, rather than improved on the learning goal.

4.2 Observations based on the presentations

Realism of behaviour

There is a common strive for realistic behaviour of agents. The sort of behaviour was varying, involving moves, decisions, emotions, performance on a task, understanding opponent intent, etc. It is probably fair to say that behaviours, which are externally observable, are the outcomes of decisions on what behaviours, of all, to execute. The hidden drivers behind the decisions are complex in nature and here lies the kernel of the problem: how to generate the processing of drivers in the cognitive, emotional, cultural and kinesiological domains. Moreover, there is little information on which behaviours could be called realistic. Variation in behaviour is observed in the real world, but this is not random. Each variation has a reason. Many models are judged as TLAR (That Looks About Right, citing Brad), meaning that the produced behaviour could not easily be distinguished from the collection of observed real world behaviours. This does not guarantee that the model also follows the purposeful variations that real behaviour shows.

Training effect

Emilie Reitz pointed in her presentation implicitly to the importance of showing that the simulation indeed has a positive training effect. Even if training in reality is more effective, preparation by simulation may be important to save on real live training time. The design of the training process is to optimize throughput or cost of bringing students to a desired skill level. Too often design is interpreted as engineering, with no experimental evaluation of the claimed effectiveness. The cost and effort of the evaluation may be prohibitive, but the exact training transfer figures are needed to justify any claims. We are not yet in a stage where we can reliably predict transfer of training.

Model simplicity

Complex models have in the past been developed for cognition (ACT-R, Cognet), but there is a reluctance to incorporate those in full in the modelling architectures. Among the reasons may be that these models are not fully understood by designers with a different background, that these models are too heavy or that architectural differences discourage their use. Attempts have been made here and elsewhere to cut corners in order to simplify the model. Such attempts should be encouraged in a responsible way. The art of modelling lies in the selection and constructive use of scientific evidence that matters most, and the disposal of details which are not significant, how well confirmed they may be. Going back to the original theories (Theory of Mind, BDI, reinforced learning, personality, CAB, perception theories and other) and making attempts to use them in a more brief model is worthwhile, as long as the purpose and the desired fitness for purpose are set.

What was training specific?

Little educational engineering was presented. In this respect the focus on military training application was evident in artificial agents, but not in the careful decomposition of training in learning goals and dedication of simulations to specific goals. Such a focus would have highlighted the fitness for use concept. As presented, the search for how to generate behaviour or predict performance was more prominent.

Avoiding robotic behaviour

In several presentations engineering networks were used to implement complex tasks, such as team tasks and scenarios. The performance fell out of the TLAR range and this is a more common observation than just here. Engineering networks distribute task elements over team members and time, optimizing the task execution. These networks are based on a limited set of rules. Deviating behaviours are not included, whereas we know that errors, emotions, wrong judgements etc cause behaviours which are not on the task list. It is doubtful if any architecture solely relying on engineering networks will produce realistic behaviours under stressful conditions. To a lesser extent the same is true for scripted scenarios, in which a set of production rules is used to select behaviours. As these rules are not time dependent, they also do not respond to the acute condition of the human agent. Also scripted scenarios may thus lead to robotic behaviour. Other architectures may be needed, based on the behavioural choices individual humans make. This also means that modelling approaches that do not take these choices (decisions) into account are deemed to keep collecting data for all possible conditions.

Clear separation of human, task and environment

In presentations 3 and 4 attempts were made to indeed make their agents select behaviours based on internal cognitive drivers. Not only is this a versatile method, which can be extended to other drivers and behaviours, it also supports the reuse of developed software. The stricter the separation is between the 'human' agent and the task and environment, the easier it gets to replace human traits or tasks or environments and keep the rest of the software unchanged. If on the other hand any two of these three elements are entangled, transition to another case may involve extensive changes.

5.0 DISCUSSION

5.3 Summary of discussion

The discussion was seeded by questions posed by Chairperson Wim Huiskamp:

Which models can be brought together?

In presentation 1 Thomas Alexander listed some digital geometric models and showed a hypothetical scheme to connect these to the cognitive domain, including graphics, cognitive models, Artificial Intelligence, Social Cultural Models, Process models, Anthropometric models and Biomedical models. This does not answer the how-to-do question. In two presentations methods were proposed to combine scientific theories into a cognitive model (presentations 3 by Jan Joris Roessingh and 4 by Philip Kerbusch). They showed that even custom integration may demand simplification. No physiological constraints were included. Brad Cain asked what different communities need to get out of their models. That is a good question to start with. It is not sure that developments go along these lines. John Cohen confirmed that it is for an engineer difficult to imagine a behaviour model. He would rather use isolated models and combine the results.

Pascal Cantot thinks that the changes in doctrines and scenarios demand a dedicated language to make easier use of existing scenarios. Doctrine differences now prove a barrier to use loaned models, which would be obtained for free from partner nations. Brad Cain believes that component composed models would be helpful. Tijmen Muller believes that the content, focusing on specific learning goals, would make another driver for components. Philip Kerbusch thinks that decision support or training may require different cognitive models and thus also demand a component modelling architecture. Thomas Alexander states that software producers have their internal standards for reusing software. Commercial (game) models are sometimes surprisingly performing, but there are no open standards.

How adaptive can a model be?

Jan Joris Roessingh applied machine learning to adapt to new environments. However, this is a data regression method and machine learning can also obscure the vision on the behaviour drivers. Tijmen Muller agrees that machine learning is not transparent. For learning progress some variation is needed but the presented cases must also be consistent. Machine learning does not deliver consistency. Glenn Gunzelmann claims that despite these reservations machine learning may be one of the few methods that follow the changes in the real world fast enough. Niels Krarup-Hansen suggests that state-space elements might replace rules in machine learning.

What is the impact of new technology?

Wim Huiskamp gives massive computing power as an example: is it a game changer because many agents can run simultaneously? Tijmen Muller replies that this may be useful for serious games. Thomas Alexander observes that commercial games consume computing power to improve the looks. Insofar more agents are running they may be entertaining, but suffer from unknown errors. Also, commercial games consist of many pre-recorded scenes and computing power is used to bring these quick to the screen. Wim Huiskamp has another suggestion: what would cameras observing the players (motion tracking) bring? Emilie Reitz sees the advantage. Her simulations would look a lot smoother with gesture interaction, but this requires instrumented human players, who are not reproducible as learning demands.

Further

Philip Kerbusch asks if a validation methodology is applicable to training. Brad Cain responds that criteria for free play validity would be vague.

5.2 Observations regarding the discussion

Architecture

The architectural issues addressed in the presentations were limited to some cognitive issues (BDI, CAB, Theory of Mind, Decision making, personality). These were addressed by ad hoc architectures, based on existing software (Jadex) or custom developments. Further, philosophy on what more to integrate was superficially discussed (graphics, kinesiology models, anthropometry). This is a far from complete list of human functions, even of those that are frequently needed. In land operations various perceptual modes, physical work, environment, clothing, sleep and food are omnipresent. In marine and air operations vestibular and perceptual challenges are met. Effects of mental work rate, stress and discomfort are other important aspects of all operations. A generic architecture would let the developer choose from the human aspects those which are important for his case and allow him to combine these. A good inventory, supported with measures for the human state in each of these aspects, is still lacking. Neither is there a vision on combination of the effects of human states in the process of task execution. If showing behaviour is the result of individual decisions on the selection of behaviours, at least a mechanism is needed that allows for input from the various human aspects involved, maybe each with their own state variables.

Explicit doctrines

In the discussion also the lack of transferability of operations between doctrines was mentioned. This disadvantage results from the lack of standardization of software. Each developer is free to include doctrine at the level in the simulation that suits him. For opportunistic reasons this may be anywhere, but a more systematic method would feed doctrine in basic functions like Rules of Engagement, fire control, weapon and equipment libraries, unit composition, Tactical and Technical Procedures (TTP) and other. NAAG/LCG1 has devised a method to share operational information between NATO countries, exploiting a NATO neutral language. Each operational plan is converted to the neutral language and then to the national language of the receiving nation. The required instruction set has been developed and might serve as an example for our purpose.

Component modelling

The audience was well aware of the need for model components that could be combined. This is the basis for constructive modelling. The question is which level of granularity is desired and how the components would interact. Preferably, each component would be so fine, that no mutual impact between components results. Otherwise they could not be used independently. That was the reason to mention human states in the sections above. Human states are supposed to be unique for a specific human function/component. Interaction occurs at the level where states come together in individual task performance (Performance Shaping Functions, PSF). This may exemplify that component modelling for the human could be feasible. It is not so easy, however, to treat social, cognitive and physiological functions in a common way (see the HFM-202 proceedings). The definition of PSF is critical for the component architecture. Another critical definition is how individual task performance combines in operational performance. The components involved in tasks have been better developed than human components, as engineering networks are frequently applied, for instance in Integrated Performance Modelling Environment (IPME). The tri-partition human functions – individual task performance – operational performance may be a basis for a component modelling architecture. The arguments in the discussion: re-use of models, transfer across doctrines, implementation of specific learning goals (tasks), current architectural randomness could merit from such a vision. It also recognizes the concerns regarding integration bottlenecks.

Adaptation

Some discussion was spent, with indecisive outcome, on how agents could follow changes in the environment by adapting their behaviour. This is a complex method, which is not often applied in the military. Usually training initially involves the ‘school solution’. The school solution is the preferred behaviour as judged by the majority of Subject Matter Experts (SME). The next step is to develop variations of the case to enrich the behavioural options of the student. Finally, the student is using reasoning to analyse a case and select the appropriate option. Agent adaptation could possibly involve the reasoning, but this is at the edge of the desirable. The agent is not the student, he is the lesson. And it is doubtful if the students merit from lessons with inconsistent learning goals.

Validity of process

Brad Cain brought up the validation issue and it is an important concern. His statement is that a model must be fit for purpose, which means that the model can not be judged in isolation, but must be judged in the framework of the simulation purpose. A careful and explicit analysis of the case is needed, including steps as the in- and exclusions, domain definition, human and operational functions involved, model capabilities and result evaluation methods. Apparently it is better to talk about verification and validation of the simulation process, in which the transfer of training is the jewel in the crown.

5.3 Observations regarding the workshop's goal

Use of HFM-202

No single reference was made to the preceding HFM-202 symposium, under the title: HBR for Military Application, neither in the papers, nor in the discussion. The symposium had nearly the same purpose as the current workshop, but now the focus was on military training rather than all military application. One can only guess for the reasons of the lack of reference. Some attendees have been at both events, authors as well as from the audience. Whatever the reason is, the symposium has not notably helped this workshop and that may be an indication that RTO has to focus its targets.

Convergence process

Related to the remark direct above, the intended modelling integration process, as decided upon by HFM, SAS, NMSG and ACT in October 2010, has not gained momentum. From the point of view of the audience that is understandable. Each investigator (m/f) is reporting what and how he did investigate, presenting his ad hoc solutions for integration. Valuable examples of ad-hoc solutions have been presented during the workshop. However, investigators were never tasked to come with generic solutions. In the discussion the lack of oversight and of broad of angles of view is perceivable and it may not be expected that in a single compact session such views are developed. To gain more momentum another, more directive process is needed, with precise, progressive tasking, a time frame, involvement of all stakeholders (panels, ACT, industry) and at least one key specialist per knowledge field (various human factors, model developers, military domains and application fields, operations, modelling architectures). Such a composition would guarantee that the purposes are understood in all fields and that amateurish exercises are professionalized.

Workshop result

The workshop activities were stated in four points and are briefly evaluated as follows:

- Inventory of challenges and known pathways to solutions: A real inventory was not made, but at least three papers gave good input and showed how their specific cases were processed, notably involving levels of cognition.
- Identification of knowledge and technical barriers to integrated model development: The papers showed which knowledge was included and the discussion identified the lack of standardisation on doctrine implementation and of standardisation of component modelling as barriers to take down.
- High level guidelines for human modelling and its interaction with live players in Military Training Simulations: No real guidelines were distilled, but experience with players showed that generation of realistic behaviour is quite hard to achieve, due to the required behavioural detail, and that the variation in realistic behaviour may be confusing the training process.
- NATO perspectives of exploitation of human behaviour modelling advances: Validation of simulated behaviour is still a hot issue. Simulators must be fit for purpose, carefully comparing the simulation demands (based on the learning goals) with the capability of the simulator.

6.0 RECOMMENDATIONS

- There is a clear demand for cognition models that could provide virtual players in training with human-like intelligence. The full blown cognition models that have been developed need simple descendents for implementation in modelling environments. It is recommended that NATO guides a way to attain such compact, re-usable models.

- The need for a generic modelling architecture that allows for component selection of human and doctrinal functions is indicated. It is recommended to reinforce the initiative taken in Amsterdam, October 2010.
- The way forward in the modelling architecture initiative is too slow. It is recommended to set up a more directive process with involvement of stakeholders and specialists per scientific or engineering field.
- The information generated and disseminated through RTO was little used. It is recommended that RTO devices a way to improve on its use, at least for RTO’s own purposes.
- The risk that human behaviour as generated by HBR models is not valid for the intended purposes is considerable. It is recommended that a validation process for human behaviour in simulations is developed, involving steps in the analysis and simulation process and associated with validity criteria.