Earth, Life & Social Sciences
Kampweg 5
3769 DE Soesterberg
P.O. Box 23
3769 ZG Soesterberg
The Netherlands

www.tno.nl

T +31 88 866 15 00
F +31 34 635 39 77

TNO report

TNO 2016 R11848

Validating models of human behaviour (V1427)

Date	January 2017
Author(s)	Dr. K. van den Bosch Dr. J.E. Korteling
Classification report	Ongerubriceerd
Classified by	A.C. van Lier
Classification date	November 2016
Title	Ongerubriceerd
Managementuittreksel	Ongerubriceerd
Abstract	Ongerubriceerd
Report text	Ongerubriceerd
Appendices	-
Copy no	-
No. of copies	5
Number of pages	38 (excl. RDP & distribution list)
Number of appendices	-

The classification designation Ongerubriceerd is equivalent to Unclassified, Stg. Confidentieel is equivalent to Confidential and Stg. Geheim is equivalent to Secret.

Samenvatting

Probleemstelling
Moderne simulatoren en games bieden steeds meer mogelijkheden om het gedrag van mensen te simuleren: van individuen, groepen, en gemeenschappen. Dit biedt kansen voor militaire toepassingen, zoals bijvoorbeeld voor training, analyse van tactische situaties, missievoorbereiding en beslisondersteuning. Om die kansen te kunnen benutten is het belangrijk dat gedrag zo wordt gemodelleerd dat het geschikt is voor de specifieke toepassing (“fit for use”). Dit rapport behandelt de ontwikkeling en het gebruik van gedragsmodellen in simulaties, en de methoden en valkuilen bij het valideren.

Werkzaamheden en bevindingen
Er is literatuuronderzoek uitgevoerd naar de validatie van gedragsmodellen.
Het ontwikkelen van modellen van menselijk gedrag wordt als bijzonder moeilijk beschouwd, vanwege het gegeven dat niet altijd goed bekend is welke (soms onbewuste) factoren van invloed zijn op iemands gedrag. Daarnaast is er veel individuele variatie in de manier waarop mensen zich gedragen.
De aard van een gedragsmodel en de daaraan te stellen eisen is afhankelijk van de toepassing. Bijvoorbeeld, een gedragsmodel bedoeld voor training moet weliswaar plausibele en representatief gedrag van een individu of groep genereren, maar het gedrag hoeft niet altijd per se gelijk te zijn aan wat er in de werkelijkheid zou gebeuren. Om de trainingswaarde te vergroten kan het soms zelfs gunstig zijn om een gedragsmodel doelbewust te laten afwijken van de realiteit. Gedragsmodellen die daarentegen gebruikt worden voor ondersteuning bij missieanalyse en planning moeten wel realistische voorspellingen geven op basis van de situationele en sociale omstandigheden.
Bij het onderzoek naar de validiteit van gedragsmodellen moet het voorgenomen gebruik en de toepassing van het model centraal staan. Gedragsmodellen die bedoeld zijn voor analyse, verkenning en beslisondersteuning moeten de gebruiker helpen een beter inzicht te verkrijgen in situaties. Een belangrijke functie van zulke modellen is het verklaren van gedrag en gedragsvoorspellingen in termen van onderliggende causale verbanden (‘explainable AI’). Gedragsmodellen die bedoeld zijn voor simulatietrainingen moeten de militair voorbereiden op wat er kan gebeuren in een missiegebied, en gelegenheid bieden om te onderzoeken hoe de loop van gebeurtenissen door eigen handelen kan worden beïnvloed.
Bij het onderzoeken van de validiteit van een gedragsmodel is de overeenkomst met de werkelijkheid op verschillende dimensies van belang, te weten gelijkenis in domein, en fysieke, fysiologische, psychologische en sociologische gelijkenis.
Om te kunnen bepalen of de overeenkomsten voldoen aan de eisen van de toepassing waarvoor het model is bedoeld, moet validatie op verschillende aspecten plaatsvinden. Ten eerste is er de constructvaliditeit: zijn de componenten en de processen van het model representatief en geschikt voor de toepassing? Dit wordt meestal gedaan door ontwikkelaars en domeindeskundigen. Ten tweede moeten de uitkomsten van het model getoetst worden aan de hand van een serie testscenario’s en tegen vooraf gedefinieerde criteria (criteriumvaliditeit). Ten derde is er de vraag naar de externe- of toepassingsvaliditeit. Helpt het model de gebruiker om in de praktijk zijn doelen te bereiken?

Toepasbaarheid
Modellen van menselijk gedrag kunnen een essentiële bijdrage leveren aan de bekwaamheid en geoefendheid van personeel, en kunnen de Defensieorganisatie helpen om huidige en toekomstige militaire missies succesvol uit te voeren. Geschikte methoden en werkwijzen voor het evalueren van gedragsmodellen is een voorwaarde om investeringen in gedragsmodellering in de juiste richting te kunnen leiden.

Summary

Problem Statement
Moderns simulators and games more and more include the behaviour of humans: of individuals, groups and even societies. This development opens opportunities for the military to use this technology for purposes of e.g., training, tactics analysis, and mission preparation. Realizing this potential demands that the behaviours of the human(s) involved are adequately modelled for their purpose in the simulation, i.e. that the models are “fit for use”. This report discusses the use of Human Behaviour Models (HBMs) in simulations, and the opportunities and pitfalls of determining their validity.

Activities & Findings
A literature review has been carried out on the issue of validating human behaviour models. Developing models of human behaviour is considered especially difficult because it is often not known what (‘subconscious’) factors influence someone’s behaviour, and there tends to be much individual variation among humans.
The nature and requirements of a human behaviour model depend upon the application. For example, a HBM for a training application requires plausible behaviour of the modelled individual or group, but does not always necessarily have to be exactly as in real life. For optimization of training, deliberate deviations from reality may some-times even be desired. Instead, HBMs for mission analysis and forecasting for planning should be able to produce credible predictions for the given situational and social conditions.
When investigating the validity of a HBM, the intended use or purpose of the model has to be taken into account. Models used for analysis, exploration, and decision support, must foster a better situation under-standing by the user. An important function of such HBMs is to provide explanations about predicted events in terms of the principles causing the behaviour (‘explainable AI’). Behaviour models used for training must prepare a soldier what can happen in a mission area, and give a better understanding how own behaviour affects the course of events.
Several measures of correspondence can be used in the validation of a HBM, in particular: domain, physical, physiological, psychological, and sociological correspondence. To evaluate whether a HBM meets the level of correspondence required for the intended application, validation needs to be performed on different levels. First, validation of the model’s constructs (are the model’s internals suited for its purpose?). This is most commonly done by task and training analyses and by consulting domain experts. Secondly, the model’s performance should be evaluated in a series of referent scenarios against acceptability criteria (criterion validity). Thirdly, it should be evaluated whether a model actually helps to achieve the goals in practice (external or application validity).

Application of results
The technology of modelling human behaviour is considered essential for achieving and maintaining an adequately trained force, and for being able to execute current and future military missions. Appropriate methods and procedures for evaluating the validity of such models is a prerequisite for guiding the current investments in Human Behaviour Modelling into the right directions.

Contents

1 Introduction

Simulation is the imitation of a system over time. It requires models that represent the behaviours of a system's components. A simulation may involve one single model. It can also consist of many models that jointly represent the behaviour of the system as a whole. Simulations used by the military are generally of the latter type. For example, a flight simulator uses many models to imitate the world for a pilot, including a model of the aircraft's engine; aerodynamic models; models of airports; a weather model; and many more. One system component that has become increasingly important for military simulations, is the human. There are many reasons for this. One reason is that modern military operations tend to be staged in urban areas, amidst the local population. This feature stresses the importance for soldiers to learn assessing the nature and intentions of individuals and groups and learn to predict likely reactions to decisions and actions. These competencies are needed to decide upon appropriate action, which may include a wide range of options from hostile engagement to social communication. Simulation can be a very valuable tool for the training of soldiers in these competencies, as well as for providing decision support during operational missions. A prerequisite is, however, that the behaviours of the human(s) involved are adequately modelled for their purpose in the simulation, i.e. that the models are fit for use or "valid".

This report is about Human Behaviour Models in simulations, on the opportunities and pitfalls of determining the quality of Human Behaviour Models, and how the validity of Human behaviour Models can be assessed.

2 Human Behaviour in Military Simulations

Simulation has become a critical technology for current military analysis, planning, and training. It is an indispensable tool, which has benefited from vast improvements in computational power over the last decades. Given the complexity of modern operations, scarcity of resources (personnel and equipment, exercise ranges) and need for cost reduction necessitate continuous scientific efforts to investigate how technological developments can be applied to further enhance the effective use of Simulation. The project "*Effectiveness of Simulation*", part of the research program "*Simulation*" (V1427), investigates which factors determine the effectiveness of a simulation for different applications and target groups. An important goal is to review and develop methods and metrics to quantify and qualify training effectiveness.

The military has been using simulation and games for training, tactics analysis, and mission preparation for centuries (Smith, 2009). In the ages of the Roman Empire, military commanders used sand tables to create simulations of the environment and icons representing soldiers and units in battle. Modern technology allowed the development of machinery simulating the technical properties of devices and their interfaces. World Wars I and II boosted this development. Figure 1 shows two pictures of World War I simulators, enabling operators to practice the skills needed to operate the military platforms.

Figure 1 Military simulators during World War I.

Figure 2 shows simulators from during World War II, enabling pilots to practice operating the aircraft and to practice simple tactical manoeuvring patterns.

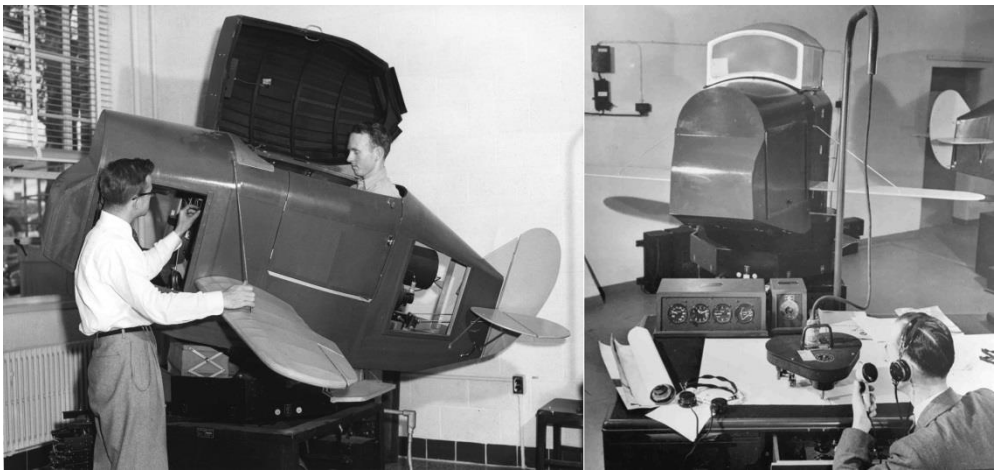

Figure 2 Military simulators during World War II.

The introduction of technology in the world of simulation meant significant improvements and new opportunities for the military. The simulators of that time used models that were able to represent the environment and the devices in a dynamic and interactive fashion to the human-in-the-loop, albeit very simplified representations. Simulations ran on models of the environment, on models of vehicles and weapon systems. However, they did not yet contain models of the behaviour of humans and groups, simply because there were at that time no models of human behaviour available. As a result, scenes in the simulators lacked presence of humans altogether. In some occasions, human role players (e.g. staff or instructors) were used to play higher and lower control, or they controlled the manoeuvring of other systems in the simulation.

The much fancier and advanced games and simulations of today (see e.g. Figure 3) do involve simulated other people.

Figure 3 VBS3, a contemporary military game.

The behaviour of the virtual characters in these games and simulators is driven by behaviour models that allow them, in principle, to act in an autonomous fashion, responding interactively and realistically to the events in the environment and to the actions of the human player(s).

The capability of modern simulators and games to simulate human behaviour makes it possible to expand the use of these devices. Whereas the use of simulators was traditionally limited to training procedural and technical skills only, modern systems can be used to prepare commanders for what they can expect of the behaviour of enemies and other people in their mission areas, and how to respond to that. In addition, the new generation of simulators can be used for other purposes than training as well, like exploring and analysing different courses of actions for a military problem. The simulated outcomes can then be used for e.g. decision support or for developing effective operational doctrines. Recently, the military investigate the use of behavioural models for training people how to act in social situations (e.g. foreign cultures) and the need for analysing the effects of policies in foreign country missions. The significance of being able to simulate human behaviour for the military is expressed by Van Hemel, MacMillan, & Zacharias (2008) in their report for the National Research Council:
" [...] the modelling of cognition and action by individuals and groups is quite possibly the most difficult task humans have yet undertaken. Developments in this area are still in their infancy. [...] It has become even more clear that human behavioural modelling at all levels is critical to DoD specifically and to the nation more generally." (p. 20).

2.1 The challenge of simulating and predicting human behaviour

These developments emphasize the importance of the human factor in simulations of military operations. It is important that in order to realize the potential of the required applications, the behaviour of the virtual people in the simulation must be simulated adequately. However, although we have become very proficient in developing accurate models of physical systems, developing models that simulate human behaviour accurately and realistically has proven to be quite another matter. It is often not known what ('subconscious') factors influence someone's behaviour. Furthermore, there tends to be much more individual variation among humans than in physical systems.

The difficulty to predict human behaviour can be explained by Chaos theory.
A central claim of this theory is that the behaviour of any dynamical system is highly sensitive to its initial conditions (Gregersen & Sailer, 1993). Small differences in initial conditions yield widely diverging outcomes, rendering reliable long-term prediction impossible. Even if human behaviour would be considered as deterministic rather than nondeterministic in nature (which is a continuing debate among psychologists and philosophers, e.g. Immergluck, 1964), chaos theory states that prediction of behaviour is still an illusion. This is nicely illustrated by a quote in Hosseini's book (see Figure 4).

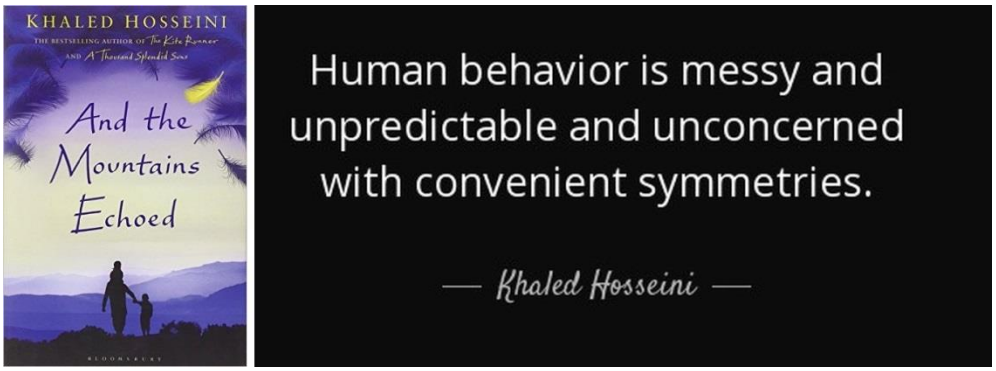

Figure 4 Quote in book by Hosseini (2013).

There is general consensus that it is indeed an illusion to claim that the available theories and models would be capable of accurately predicting the behaviour of individual humans or of groups. However, the opposite proposition, namely that human behaviour is only and fully the plaything of coincidental initial conditions, is also too simple. Knowledge about how regularities and rules in situations shape human behaviour (e.g. Suchman, 1986) can be used to develop human behaviour models. Thus, complete and perfect modelling of human behaviour is an illusion, but simulating and predicting human behaviour within ranges of predictability is possible (with its bandwidth determined by the complexity of the situation and the associated variability in possible initial conditions).

Goerger, McGinnis, & Darken (2005) summarize the principal reasons for why modelling human behaviour it is so difficult:

- The nonlinear nature of human cognitive processes;
- The large set of interdependent variables making it impossible to account for all possible interactions;
- Inadequate metrics for validating HBMs;
- The lack of a robust set of environmental data to run behavioural models for model validation;
- No uniform, standard method of validating cognitive models.

Harmon, Hoffman, Gonzalez, Knauf, and Barr (2002) argue that there is a critical feature that distinguishes models of human behaviour from models of physical systems: models of human behaviour are actually made up of two sets of computer programs, a behaviour engine and a knowledge base. The behaviour engine defines human information processing mechanisms in general; the knowledge base specifies the properties and capacities of a specific individual (virtual or human). According to Harmon et al. (2002, p.4), the combination of the inherent complexity of human behaviour (Appleget, Blaise, & Jaye, 2013) with the requirement to address it in two separate software components easily makes behaviour the most complex component of a simulation system. Because of these properties, human behaviour models are not as mature as physical models.

3 Human Behaviour Models

A Human Behaviour Model refers to describing, explaining and predicting the behaviour of individuals and groups as a function of their personal properties, their environmental conditions, and the interactions between the latter. All HBMs model the behaviour of people at some level.
Figure 5 depicts the classical architecture of a Human Behaviour Model, consisting of a behaviour engine interacting with a knowledge base to update an internal representation of a (simulated) world. Most HBMs follow this architectural structure, separating the stored mental and physical contents (including knowledge, beliefs, skills, physical abilities, emotions, goals, motivations, et cetera), and the mechanisms that operate upon them.

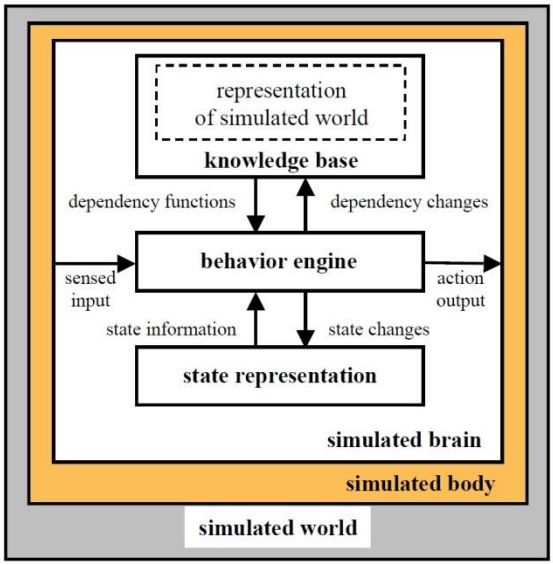

Figure 5 Canonical Model of Human Behaviour (Harmon et al., 2002).

The term HBM[1] may refer to representations of parts of individuals (e.g., hands operating controls), individuals (e.g., a specific terrorist or equipment operator), aggregates of individuals (e.g., a crowd, a command staff), and aggregates of organizations (e.g., several organizations responding in concert to an emergency situation). A HBM may include one or several classical cognitive functions (e.g., perception, inference, planning, control), human performance limitations (e.g., sensing bandwidth, decision latencies) and the effects of behaviour moderators (e.g., stress, injury, fatigue, discomfort, motivation and emotion). HBM may vary from simple scripts, to finite state machines to complex knowledge-based systems integrating multiple reasoning paradigms and augmented by simulations of the effects of various behaviour moderators (NATO RPG-ST12, 2001).

Van Hemel et al. (2008, p.302) emphasize that it is important to acknowledge that a model of human behaviour must always be viewed in the context of its purpose. Sometimes people criticize models for their lack of realism; that it not reflects the behaviour as perceived by the viewer. However, it depends upon the purpose of the

[1] Some use the term “human behaviour representation” (HBR). This is considered as synonymous with HBM in this report.

model whether this matters or not. Adding features to increase a model's realism makes the implications of a model more difficult to understand, and requires increasingly sophisticated techniques. Aiming for realism might also require an impractical amount of data to build the model or to specify parameter values and run the model. Consequently, if a simple model serves the intended purpose (i.e. if it is "fit for use"), then it should be preferred.

All HBMs model the behaviour of people at some level of abstraction (Harmon et al., 2002) potentially involving any combination of different facets of human behaviour including:

- Ability to reason (e.g., knowledge based systems);
- Ability to change the environment (e.g., operating equipment);
- Responds to comfort and discomfort (e.g., environmental safety);
- Susceptibility to injury and illness (e.g., injury models);
- Emotional responses (e.g., affective models);
- Ability to communicate with other humans (e.g. Virtual Pilot, Aliz-e[2]);
- Abilities to sense the environment (e.g., vision models);
- Physical capabilities and limitations (e.g., MANPRINT).

3.1 Contribution of Human Behaviour Models to military missions

HBMs may contribute to today's military missions in various application areas (Van Hemel et al., 2008, p.33). In this report we will discuss the following two:

1 analysis and forecasting for planning;
2 training and mission rehearsal.

3.1.1 Analysis and forecasting for planning

The nature and requirements of a human behaviour model depend upon the application. Take, for example, a model that should support a military commander in developing a policy for achieving social stability in a mission area. Then a model could help by making predictions on the behaviour of individuals and groups in response to interventions that the commander has in consideration (see Figure 6).

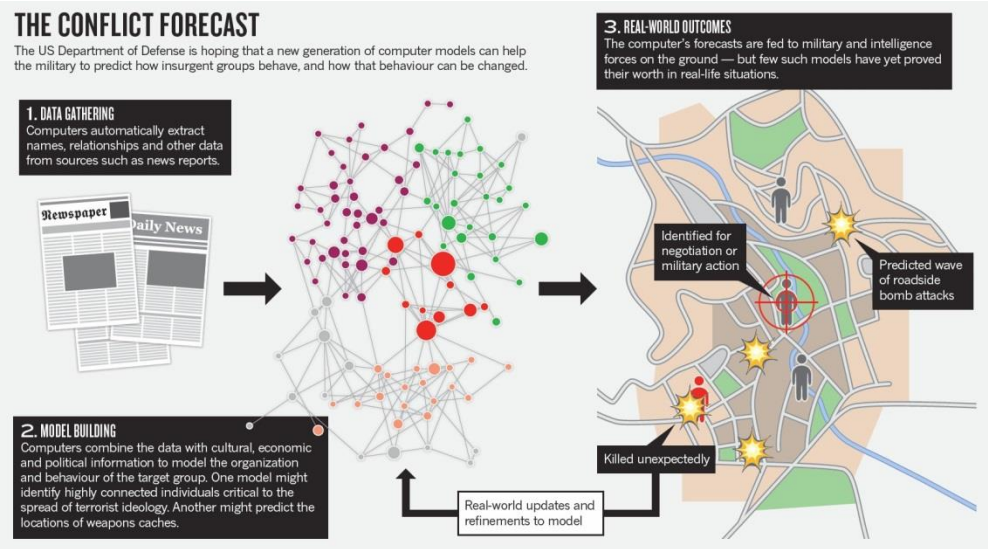

Figure 6 Illustration of conflicting forecast, taken from Weinberger, 2011.

[2] http://www.aliz-e.org/

In order to be truly helpful, such models should be able to produce predictions that are credible by taking into account the relevant situational and social conditions. Of course, in real life it is impossible to know all factors that play a role, let alone assess their true value (see also the discussion in §2.1) and their impact on the model's outcomes in interaction with multiple other variables. It is therefore necessary to let a model not present a single outcome, but instead provide multiple predictions, based upon different circumstances and scenarios. Furthermore, a likelihood estimate of predictions would also be helpful.

The application of Decision Support has similar demands as for Analysis and Planning. If a human behaviour model is utilized for supporting a military commander to decide between different Courses-Of-Actions to fight the enemy, then the commander too expects that the model presents viable and reliable predictions for the situational conditions.

3.1.2 Training and mission rehearsal

Using behaviour models for training and mission preparation, has again its own requirements.

Figure 7 Impression of military training with a virtual team mate.

Of course, training too requires that the simulated behaviour of modeled individuals and groups is sufficiently plausible for the trainee to accept the training as useful. But the generated behaviour does not always necessarily have to be exactly as it would be in real life. What counts in training is whether and to what degree the behaviour model generates behaviour of the simulated individual(s) that helps to achieve the learning objectives. For instance, in a training scenario, a virtual team mate may deliberately act inaccurately because this enables the trainee to achieve the learning goal: "detecting and correcting errors made by team mates". Just because this behaviour occurs seldom in real life, embedding it in a simulation will bring about exactly the situation that enables the trainee to learn the behaviour associated with the learning objective.

3.2 Approaches to Human Behaviour Modelling

In most applications, virtual character behaviour is controlled by defining a list of rules and contingencies. This is a successful approach for tasks that are straightforward (e.g. procedural tasks) and in simulated worlds that can be strictly controlled, so that no situations emerge for which the model of the virtual character cannot produce adequate behaviour. The entertainment gaming industry has been using this approach to develop elaborated and complex scripts of input-output rules to control the behaviour of virtual characters in their games. They accomplish that with great success. Using input-output contingencies makes it possible to let a virtual character behave in a quasi-intelligent fashion, provided that the developer anticipated the situation and developed a behavioural rule for it. The advantage of this approach is extended control over the game, and has proven to be a robust, error-resistant technology. A disadvantage is, however, that the behaviour of virtual players tend to become fairly predictable (especially in the eyes of experienced and skilled human players), hence lose their believability and credibility as a (virtual) person.

Some applications, however, require more freedom on the part of the human player or the user. This can, for example, be for training purposes (application area 2, see §3.1), and for decision support, e.g., commander who uses a model as a support tool for planning or decision-making in military operations (application area 1). In such situations it is hard or even impossible to create a 'spanning set' of input-output contingencies specifying appropriate behaviour for all possible states that may occur during a scenario (Silverman, 2001). Even in relatively simple tactical scenarios the number of states tend to be very high (Klein, 1998). An approach that is more suitable for this type of applications is to model behaviour as a function of fundamental underlying processes (Zachary, Ryder & Hicinbothom, 1998). Such a model represents the knowledge and processes of an individual or entity in a certain domain, task or scenario.

The psychological validity of the knowledge representation and information processing may vary across models. We discuss here: (a) Belief, Desire, Intentions (BDI) models, and (b) cognitive models.

BDI-models: A popular approach is to model behaviour as a function of beliefs, desires and intentions (Bratman, 1987; Rao & Georgeff, 1991). BDI (see Figure 8) is fundamentally different from modelling behaviour as input-output contingencies. In BDI models, a virtual character is not instructed to act upon a certain state in the scenario, but rather upon the *interpretation* of that state. An event in the world brings about a belief in "the mind" of a character (e.g. hearing a fire alarm creates the belief that the house is on fire). The belief triggers a goal. What goal is triggered by the event depends upon the context and the role of the character: a mother, for instance, may adopt the goal to search for her child; another person may adopt to goal to leave the building quickly; a fire fighter may adopt the goal to locate and fight the fire. The advantage of BDI over input-output contingencies is that BDI-models are more flexible and reusable. For example, instead of specifying separate actions for each of the following events: "person threatening with rifle"; "person threatening with knife"; "person threatening with hand grenade", and so on, in BDI all these events activate the belief "I am in danger" that subsequently invokes a goal e.g. "escape".

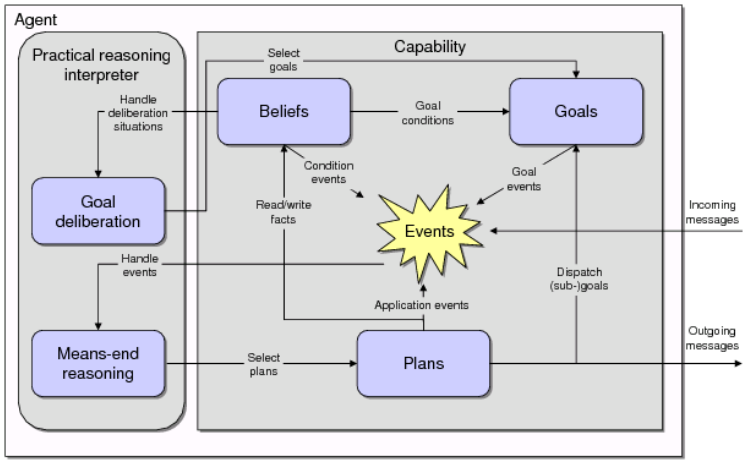

Figure 8 JADEX, an example of a BDI-architecture (Figure taken from Pokahr & Braubach, 2007).

Although people may themselves believe that their behaviour is a function of their beliefs and goals, psychology learns that behaviour is caused and influenced by many other (often subconscious) factors (e.g. emotion, bias, fatigue, stress, et cetera). And these causal and moderating factors are typically not included in BDI-models (although there have been attempts to include emotion in BDI-models, e.g., Jiang, Vidal, & Huhns, 2007). A more fundamental approach at modelling human behaviour is ‘cognitive modelling’.

Cognitive models: A cognitive model is a representation of human cognitive processes for the purposes of comprehension and prediction. Cognitive models may focus on a single cognitive phenomenon or process (e.g., pattern recognition), how two or more processes interact (e.g., pattern recognition and decision making), or to make behavioural predictions for a specific task (e.g., performance in a tactical picture compilation task). Cognitive models can function independently, or they may be embedded in a cognitive architecture.

A cognitive architecture represents the conceptual and structural properties of the human mind. There exist many different theories about how the human mind functions; and these theories are associated with different cognitive architectures. Cognitive architectures can be symbolic (e.g., SOAR, Laird, Newell, & Rosenbloom, 1987), connectionist (e.g., PDP, Rumelhart, McClelland, & the PDP Research Group, 1986), or hybrid (e.g., CLARION, Sun 1996).

A cognitive architecture can also refer to a blueprint for intelligent agents by means of implementing the processes and the identified relations in computer software. Such a computational cognitive architecture represents the knowledge and cognitive processes of an individual in such a fashion that, when provided with input, the system produces realistic behaviour as output. Some argue that cognitive architectures are most suitable for developing models that demonstrate intelligent and autonomous behaviour (e.g., Jones, 2004).

Some of the architectures have proven to be capable of detailed modelling of human cognitive processes, e.g. ACT-R (Anderson & Lebiere, 1998), EPIC (Kieras & Meyer, 1997), CLARION (Sun, 2006),and SOAR (Laird, Newell, & Rosenbloom, 1987). The latter has gained acceptance as a practically successful tool for modelling behaviour in military simulations (Taatgen, & Anderson, 2010).

4 Purposes for Validation of Human Behaviour Models

The value of a simulation application depends upon the quality of the underlying models. Assessing a model's quality is called validation. 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" (ITT Research Institute, 2001, p. 10). Stated more intuitively, validation asks "Did I build the right thing?" In building the right thing, there are two elements: the degree of real-world representation and the intended uses of the model. They are related but are not the same thing. A realistic representation may not meet the intended use. It is a frequent error to put primary emphasis on a realistic representation, assuming it will meet the purpose. This mishap may result in an unending quest for realism without considering the intended use or purpose of the model. When one, however, begins with the intended use or purpose, then the required degree of realism of the representation follows (Van Hemel et al., 2008).

Validation of human behavioural models is gaining importance, as the military depends on such models more and more (Gonzalez & Murillo, 1999). It is of crucial importance for users, because validation can prevent the costly consequences of using incorrect models and simulations.

Campbell & Bolton (2005) proposed the concept of "application validity" to suggest that validation of a Human behaviour Model must be specific to an application rather than extensively generalizable. In line with this argument, Van Hemel et al (2008) argue that "without a prior specification of intended purpose, there are no clear-cut a priori criteria for deciding which features of a phenomenon to stress in its modeled representation" (p.302). Marks (2006) states that in order to be able to validate a human behaviour model, its "purpose" must be defined. For example, the purpose of a model could be to *explain* observed behaviour, to *predict* possible behaviours of individuals and/or groups that might occur with or without an intervention, or to *generate* behaviour of a virtual character in a simulation or game. A different model would typically be required to meet each of these different purposes.

Military applications of HBMs typically include models for (Van Hemel et al., 2008):

- understanding, exploration and decision support (task support & doctrine development/ testing);
- intelligent virtual characters in simulation and intelligent training.

The first purpose of human behaviour modelling requires what might be termed an understanding approach to validation. The second type requires what might be termed an action approach to validation.

4.1 Validation of models for Understanding, Exploration & Decision Support

Ideally, an explanation of human behaviour would entail a complete understanding of both the necessary and sufficient conditions for its occurrence. In practice, a complete understanding is unfeasible. Some even argue that it is principally impossible (see §2.1). One way to nevertheless achieve a model that is useful for

purposes of explanation is proposed by Epstein (2006). It entails to first construct a model that can generate the behaviour of interest in the same or similar context. This model is then used to derive a potential explanation for the phenomenon. For example, a model may be used to forecast a range of possible locations where the enemy may cross a river (see Figure 9).

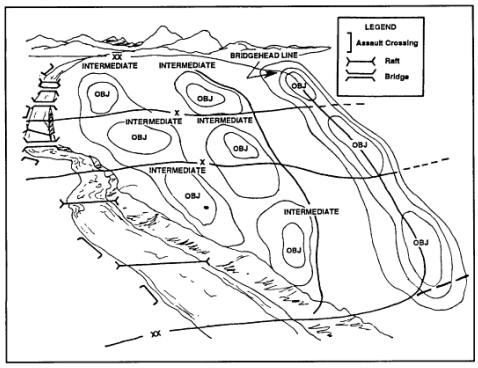

Figure 9 Where will the enemy cross the river?

The model may predict a densely forested river-crossing location if the assumption is adopted that the enemy demands a covered area for the operation. Another possible outcome, a shallow section in the river, may be predicted if the assumption is that the enemy demands to bring along open and low-platform vehicles. Depending on the assumptions and their weighing, the model may predict even more alternative locations. The robustness of the derived explanations can be tested by deliberately and systematically applying variations to the model and asking subject matter experts to consider and evaluate the likelihood of the generated outcomes.

According to Marks (2006), predicting is a simpler purpose of a model than explanation, in that only sufficient conditions for the occurrence of a phenomenon are sought. An important value of predictive models for exploration purposes can be the uncovering of nonobvious insights into complex phenomena that could not have been obtained without the model. For example, a commander in a foreign mission area may be tasked with maintaining a stable security situation in the region (see Figure 10).

Figure 10 How to maintain stability in a foreign society?

This requires the critical skill of being able to pick up the cues that indicate when a fragile state is likely to erupt in violence. A societal behaviour model would be able to support the commander in this task by predicting the level of stability in a society. Research from DARPA (Kettler & Hoffman, 2012) has shown that societal (in)stability is predicted by e.g., the number of gatherings in the area, and the circulation of protest posters. Other factors prove not to predict stability in a region, e.g. tribal customs. Models that capture these observed relationships may be able to assist the commander by making predictions about the expected stability in the region. Such models may predict stability more or less successfully, they do, however, not explain *why*. Nevertheless, the predictions of such a model may be actively used by the commander to gain a better understanding of the situation by exploring various possible interventions. The output of the model helps the user decide which action to take (or to refrain from taking action at all).

A model of behaviour (e.g. pertaining to individual or societal behaviour) needs to relate actions of interest to outcomes of interest. The model does not necessarily need to reveal deep understanding. However, such a model must be timely and accurate relative to its purpose. Thus the validation of such a model must include a careful consideration of the possible action choices to be modeled (including no intervention). Appropriate modelling of action choices will not eliminate the uncertainty inherent in a situation, but it should help to clarify the possible action alternatives and hence provide useful guidance regarding the best action to take (Van Hemel et al., 2008). Thus, a central concern of using predictive models in this way is the model's accuracy: does the model predict outcomes with appropriate likelihoods? Does the model exclude outcomes that could never actually be observed in reality?

In addition to developing predictive models that are able to forecast possible and likely behaviour of individuals and groups, it is also possible to design and develop behaviour models that provide explanations to the user about the events and principles causing the behaviour (Core, Lane, Van Lent, Gomboc, Solomon, & Rosenberg, 2006). This is called Explainable Artificial Intelligence, and can be of significant value for military analysis, and for military training (Gomboc, Solomon, Core, Lane, & Van Lent, 2005). In a training context it can help the trainee to

understand the relationships between his own behaviour and the responses of other (virtual) people in the scenario. This can be important as trainees do not always understand why other players behave in the way they do. For instance, virtual team members (whose behaviour are generated by their human behaviour models) that do not follow the instructions of their leader (a human trainee) may have misunderstood the instructions, or they may disobey them on purpose for some reason. Due to this obscurity, the trainee does not know whether he should communicate clearer, be more persuasive, or give better or safer instructions. This problem could be solved if virtual agents would be able to explain the reasons behind their behaviour.

It may sound intuitive to think that if a model uses knowledge and principles to generate behaviour, it should be possible to follow the trace back and use the information to produce an explanation for the generated behaviour. Unfortunately, this is a too simple impression of things (Harbers, van den Bosch, & Meyer, 2010).

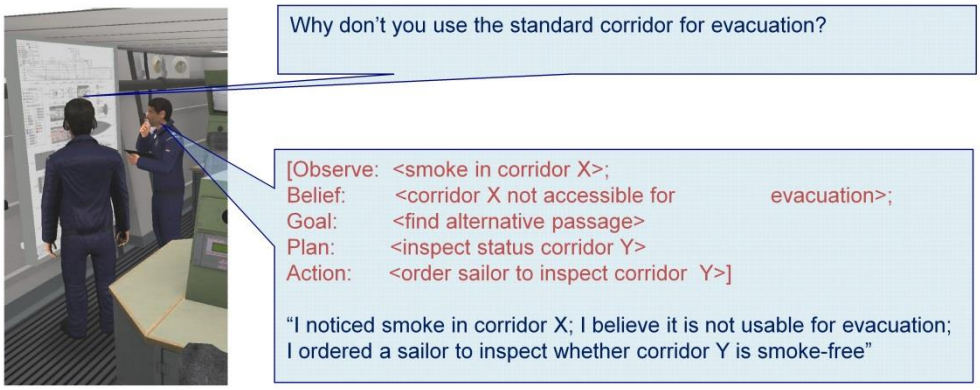

Figure 11 Example of Explainable AI by Van den Bosch, Harbers, Heuvelink, & van Doesburg, 2009. The human trainee controls the left avatar, that of the Officer of the Watch in this game for training on-board fire-fighting command. A human behaviour model controls the right avatar, the leader confinement team. The human player asks the virtual player to clarify his choice. The Explainable AI embedded in the behaviour model provides logical and useful explanations, based upon its beliefs, goals and plan repertoire. The formal notation of the representation in the model needs to be transformed into natural language in order for a trainee to be able to understand it. This can be achieved, for example, by using predefined templates (Muller, Heuvelink, van den Bosch, & Swartjes, 2012).

In order to be of use, the self-explanations of human behaviour models needs to be valid for the given context, needs to have the right level of aggregation (not too abstract, not too trivial), and should be tuned to the needs of the learner (Harbers, Van den Bosch, & Meyer, 2011). Validation is needed to determine whether explanations generated by human behaviour models comply with these demands.

4.2 Validation of models for simulation and intelligent training

Current simulators and games offer contextually rich and flexible environment for training purposes. These provide the environment for the soldier to prepare for what can happen in a mission area, and how to respond to that. To make trainees understand how their own behaviour affects the course of events, the behaviour of the virtual people present in the simulation must be simulated adequately.

What “adequately” means depends upon the learning objectives. In general it can be argued that the behaviour of a virtual character should be sufficiently plausible to create a suspension of disbelief, implying that the trainee is willing to accept the obviously designed and artificial nature of a training environment as representative for a real situation in the future. In other words, the trainee should be willing, for the sake of learning, to set aside his awareness that the training setting is not real.
In order to achieve a suspension of disbelief, the behaviour of virtual characters should meet the following general requirements (Livingstone, 2006):

- behaviour should be consistent with human information processing characteristics (e.g. bias in decision making);
- behaviour should be human-like and understandable to people;
- behaviour should be appropriate for its purpose (i.e. learning by the trainee(s)).

Human behaviour models are frequently criticized for lack of realism. That is, criticasters emphasize that they experience the behaviour of a model not exactly the same as they themselves observe the world, or note that the behaviour leaves out some aspect of reality. However, any model is a simplification of the real world, hence inherently departing from the real phenomenon. The challenge is to include the behaviour components that matter to the purpose, and omit or simplify the components that do not. Including features in a model just for the sake of increasing its realism runs the risk that the model becomes more and more difficult to understand. Thus, if a simple model meets its purpose, then it should be preferred over a complex one (see Table 1 for an example).

Table 1 Example illustrating the value of a simple HBM.

CARIM is stand-alone low-cost desktop simulation trainer for the training of the Officer of the Watch (OW) in on-board fire fighting command (Van den Bosch et al., 2009). The simulation is equipped with virtual team mates of the OW. The behaviour models of these virtual characters allow them to act independently and intelligently to events in the simulation (e.g. alarms going off, presence of smoke, blocked passages, et cetera) and to actions of the human player (e.g. given commands; requests for information, et cetera). The virtual team mates behave as a function of their virtual expertise (i.e. their knowledge base) and the information that the simulation environments makes available to them. Maritime experts judged that the virtual team mates behaved adequately and realistically for under “normal” circumstances. However, some of them indicated that in reality, tasks sometimes need to be carried out under demanding circumstances (e.g. a wildly moving ship, low oxygen atmosphere, while being severely tired, et cetera). They pointed out that the virtual team mates do not show the behaviour that can be expected under such circumstances. They were right, because the behaviour models did not include the components that allows this. However, as the *purpose* of the application was *initial training* rather than advanced training, a simple rather than a complex model was chosen.

4.3 What makes validating HBMs difficult?

In real life, human behaviour emerges from numerous interacting factors, often responding in a nonlinear fashion. Small situation changes therefore often create wildly different responses. Validation of human behaviour models is difficult because of the large number of behavioural paths that must be explored for any given purpose (van den Bosch & Doesburg, 2005). Whenever a simulation application involves the modelling of humans, or groups of humans, the quality of the HBMs are of great importance to the over-all value of the simulation. Despite this impact, organizations validate HBMs only sporadically (Harmon et al., 2002). The authors suggest that this may partly be caused by a number of myths concerning HBM-validation (see Table 2).

Table 2 Prevailing myths concerning HBM validation (adapted from Harmon et al., 2002).

Myth	Explanation
Users are good sources of HBM requirements.	When users are asked what is needed to develop a human behaviour model for an application, they tend to focus on the breadth and depth of their domain, rather than what is minimally required for a model. As a result, users tend to overstate the requirements. Thus, even when users appreciate the need for a HBM in their simulation (not a common occurrence), they tend to state that need in unrealistic terms.
A good referent for a HBM is a human doing the same job.	Identifying the corresponding human as a referent for its simulation may seem like a good idea but everything about that human is probably not well known. Human referents also tempt one to expect that the models performs exactly like that specific human, rather than its abstraction.
A valid HBM is as realistic as possible.	Again, HBMs are necessarily abstractions of real humans. We do not understand human nature well enough to accurately model all of human behaviour. Like other simulations, HBMs should only represent what the purpose requires.
A good HBM is stochastic just like humans.	Human behaviour often appears stochastic due to its high complexity and chaotic nature. Some aspects of human behaviour lend themselves to stochastic representation but treating all of human behaviour as random ignores the adaptive and goal-directed nature inherent to all humans, a key property.
A good HBM is logical just like humans.	Humans seldom behave logically. They can, however, explain their behaviour as if it were logic, but that explanation rarely agrees with the real phenomena underlying their behaviour.
Different users produce univocal and testable criteria for HBM validity.	Specifications of validation criteria often depend substantially upon the observer and are therefore not objective. This will lead to irreconcilable differences between observers about whether the HBM actually met the criterion. Good validation criteria are both observable and observer-independent.

Myth	Explanation
The experts will recognize (in)valid HBM behaviour when they see it.	Experts may recognize some invalid behaviour when they see it, but the complex nature of human behaviour will lead to many false positives. Experts have often declared quirky HBM behaviour as distinctly “human” when it actually resulted from implementation errors.
Validating HBMs is too hard so why do it or even try to understand it.	This defeatist perspective only leads to accepting poorly performing HBMs. Like any validation task, a reasonably simple discipline can produce acceptable and cost effective results. Good understanding of HBM validity can even simplify the difficulty of abstracting the parts of human behaviour necessary to achieve a purpose thereby reducing the developmental costs and risks.

5 Validating Human Behaviour Models

Validation of a Human Behaviour Model involves testing the model within the simulated environment in which it will be used and determining whether the output of the model meets the requirements of its intended use. Establishing the level of *face validity* is by far the technique most often applied to validate simulation models. In this technique, a subject matter expert (SME) runs a model in various scenarios, observes the resulting behaviour, and determines, often qualitatively, whether that behaviour meets the requirements. Although face validity is certainly important as a measure of user's confidence and trust in the model, as a validation of a model it is insufficient (Holden, 2010; Korteling & Van den Bosch, 2015; Korteling & Sluimer, 1999). A practical drawback is that requirements should be explicit and testable, but in face validity tests they often consist of implicit and indefinite expectations on realism in the reviewer's mind. Another drawback is that the evaluating SMEs tend to focus on the experienced realism of the model's output, rather than on the question whether the models supports the intended purpose (e.g. learning complex relationships in a particular domain, or acquiring a particular skill) (Caro, 1977). Thus, face validity testing is an important aspect in validation, but is by itself not a sufficient coverage of the validation process.

At the most general level, to validate a HBM one should (NATO RPG-ST12):

1 develop an adequate statement of requirements (see §5.1);
2 identify the referents that define the standards for determining accuracy or error (see §5.1.1);
3 assess the capabilities of the HBM (i.e. knowledge base and information processing mechanisms) (see §5.2);
4 compare those capabilities against the requirements to determine the fitness of the HBM for the intended purpose (see §5.3).

In addition to the levels identified by the NATO-RPG, another level of validity can be introduced:

5 check for evidence on whether the HBM actually supports the intended application (see §5.4).

These elements of the validation process are discussed below.

5.1 Model requirements and validation

As with any other simulation model, a clear requirement-specification of a human behaviour model is needed to be able to address its validity. Typically, getting complete requirements statements from its users can be challenging. It is necessary that users provide information on both the purpose (intended use) of the model, as well as on the behaviour of the to be modeled individuals or groups. Users should be able to define what contextual and human effects they feel should be incorporated into the simulation. Contextual effects refer to how the environment shapes the behaviour. Human effects refer to moderators that affect behaviour such as various kinds of physical and psychological stress, fatigue, emotion, motivation and various other personality differences (e.g. Silverman, 2001). Much of the required capabilities of a HBM can be derived from that information.

It may often not be possible to develop a HBM that meets all defined user requirements completely. Therefore, it is recommended to formulate acceptability criteria. Acceptability criteria define the testable standards of functionality and performance that the HBM must meet in order to be considered adequate to achieve the intended objectives. For example, for a particular training program that uses a simulator it may be required to develop HBMs that generate the behaviour of simulated (virtual) players. The acceptability criteria for these HBMs should then specify:

- the human roles that must be represented;
- the level of performance of those roles;
- the human aspects that must be represented (e.g., which behaviour moderators should be included?, what are the necessary performance limitations?).

The training needs analysis should provide the information to be able to determine the acceptability criteria for the HBM(s), at least in a qualitative form. In order to implement HBMs in the simulator system, the HBMs must have quantitative measures of the acceptability criteria.

5.1.1 Referents for model validation

Referents, like requirements, contribute knowledge essential for validating any simulation model. A referent is an information source to be used as a standard against which to measure the properties and output of a model. This comparison identifies where the models coincide with the referents and where they deviate from them. This information can then help to determine how well the models serve specific purposes. Referents can come from expert opinions, experimental observations and theoretical approximations (NATO RPG-ST12).

Human Behaviour Models can be validated against many different referents. Each referent provides standards against which to compare an HBM behaviour to test its correspondence with the referent (see e.g. Groen, Valk, Bijl, Korteling, & Ledegang, 2016; NATO RPG-ST12). The following type of referents can be distinguished:

- Domain Correspondence: Domain experts, also known as SMEs, know what kind of human behaviours are typically required to represent their particular domain. Their knowledge permits SMEs to examine the model's knowledge base(s); to observe the model's output in simulations; and to assess -often qualitatively- how realistically the model's output represent the necessary human behaviour for a purpose. In addition to SMEs as an information source, referents for domain correspondence can also come from experimental data, in the form of systematically recorded performance data.
- Physical Correspondence: Basic physical laws limit the performance of humans. Consequently, the performance of a HBM can be compared against these limits. A believable representation of human behaviour responds to events with human-like reaction times and abilities (Livingstone, 2006). Representations that exceed these limits inaccurately predict the behaviour manifested by the human.
- Physiological Correspondence: Models that have physiological correspondence are more likely to behave like real people under conditions where physiology contributes to the performance (e.g., response speed, fatigue, and injury) (see e.g. Tolk, 2012).

- Psychological Correspondence: An important aspect of validation can be the testing whether a model's properties and its resulting behaviour correspond to psychological theories and to empirical data appropriate for the problem domain. Testing the psychological correspondence creates stronger validation than domain correspondence testing alone because of its linkage to the underlying psychological phenomena. Psychological correspondence testing enables validation of all of the model components (e.g. perception, memory, reasoning, et cetera), both as separate functions and as an integrated whole.
- Sociological Correspondence: If a HBM is to be applied in a setting that includes interacting people, then sociological correspondence is likely to be an important issue in the validation of the model. This interaction may take place in disordered groups, such as crowds, as well as in groups operating within some organizational structure (see Figure 12).
  To test the sociological correspondence of a model, sociological theories and empirical and historical observations may be useful. If these are not available, a model can also be tested against the opinions of psychology and sociology professionals that act as experts on this issue.

Figure 12 The incident at the remembrance ceremony, Amsterdam, 2010[3] (source: nu.nl).

Clearly, the validity of a HBM can be tested against a number of referents. It should be noted that in real life the different types of referents are not as distinct as the categorization above suggest. There is not always a clear boundary. Some properties may be considered as psychological correspondence, but also as sociological correspondence. Similarly, it is not always crystal clear whether a model property represents a physical or a psych-physiological property. A HBM that corresponds to all referents is likely to approximate human behaviour to the fullest. Fortunately, most purposes only require correspondence in only a few of these areas. This can reduce the complexity, cost and risk of the validation process.

[3] A well-known example of social influence upon human behaviour occurred during the war remembrance ceremony in Amsterdam, 2010. An activist suddenly broke the solemn silence with loud and alarming cries, causing panic in the crowd. People tried to run away, running over each other. Bosse, Hoogendoorn, Klein, Treur, van der Wal, & van Wissen (2013) showed how the dynamics in a crowd's behaviour can be modeled using the concept of "contagious emotions", implemented in a BDI-E model.

All referents for validation of HBMs have their limitations. Testing the domain correspondence of a model may require unrealistic searches of very large and nonlinear behaviour spaces, especially if it concerns behaviour in complex situations. Validating a HBM for the full breadth of a domain may be necessary for doctrine-testing applications or for decision support. For training applications this is generally not required, as the kind of situations emerging in a training program are under the control of the training developer and the instructor. Testing the domain correspondence of a HBM for training purposes needs to cover only the scope of the training program.

Testing the physical, physiological, psychological, and social correspondences of a specific behaviour model requires validated theories and models of the respective behavioural phenomena. While models have been developed within each of the disciplines, these models are often not comprehensive, but apply to very restricted behaviour spaces only. Data on the validation of such separate models (if at all available) cannot be used to make claims about the validity of a HBM that integrates models from different disciplines. Thus, the outcomes of separate models cannot be automatically aggregated to other separate models, or to the overarching model. For example, modelling groups of people cannot be done by simply aggregating individual behaviours as if each individual contributes equally to the behaviour of the overall group. While it is clear that the group behaviour is influenced by its members, accounting for the individual contributions is not straightforward (Appleget et al., 2013). In other words: the validity of the whole can be different than the validity of its parts.

Referent measures for testing the validity of a model can be obtained from expert opinions (SMEs), experimental observations and theoretical approximations. In the validation of HBMs, experts (SMEs) are practically almost always used. SMEs generally offer a broader but more qualitative source of referent measures than experimental data, and they perform particularly well when the situation closely resembles their experience and education. Nevertheless, care should be taken because the use of SMEs also brings its own problems. The NATO RPG-ST12 addresses the following issues:

- Expertise: only SMEs with the right expertise may understand the form of the information used and produced by the HBM. Using SMEs where they are not trained to interpret HBM data can only give a false sense of security and may lead to inappropriate conclusions about a model's validity.
- Availability: SMEs with considerable experience and, therefore, utility to their operational unit, tend to have limited availability to the developer and V&V agent. Often the best SMEs are also the least available resources.
- Funding: Many projects fail to reserve funding for SME involvement.
- Time: SMEs require time to become familiar with the project, the user requirements and acceptability criteria, and the HBM itself. Failure to allocate sufficient time in the schedule for all aspects of SME involvement will limit their utility and effectiveness.

5.2 Validating a model's constructs

The architecture of a Human Behaviour Model typically consists of a behaviour engine and a knowledge base to update an internal representation of a (simulated) world. The behaviour engine defines the various cognitive functions (e.g., perception, memory , reasoning) and the effects of behaviour moderators (e.g., stress, injury, fatigue, discomfort, motivation and emotion). The knowledge base contains the stored mental and physical contents (including knowledge, beliefs, skills, physical abilities, emotions, goals, motivations, et cetera). Determining the qualities of the individual and integrated model components are issues related to construct validation (assessing whether the model captures the appropriate constructs) and content validation (assessing whether the model concerns the intended types of behaviour) (Cronbach & Meehl, 1995).

5.2.1 Validating a model's behaviour engine

The first stage of developing a human behaviour model generally has the form of a schematic, conceptual model of the relevant processes involved in the behaviour. This overview of the model may be on paper or digital, often documented using symbolic language (e.g. Unified Modelling Language (UML)).
This offers the user a first opportunity to evaluate a HBM while its development is still in progress. A conceptual model of the behaviour engine describes the developer's interpretation of what is needed, and defines, among others, tasks; goals associated with each task; objects and their properties that the HBM can sense; effects of internal factors that can moderate the HBM's responses (e.g., fatigue, injury, fear); et cetera. The conceptual model of the behaviour engine also specifies (in global qualitative terms) how these concepts relate to each other. SMEs, familiar with the user's objectives with the model, should examine the conceptual HBM to assure its completeness and consistency with user doctrine.

Figure 13 depicts a classical architecture of human behaviour, consisting of a behaviour engine (upper level) interacting with a state representation (lower level). Many architectures follow this structure, separating the stored mental and physical contents (including knowledge, beliefs, skills, physical abilities, emotions, goals, motivations, et cetera), and the mechanisms that operate upon them (Harmon et al., 2002).

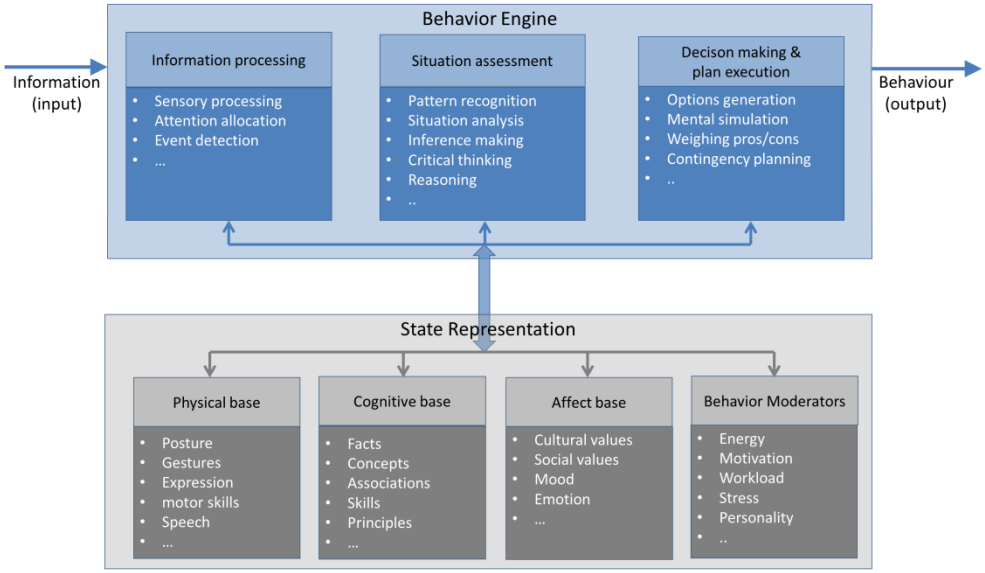

Figure 13 Example of a classical general architecture of human behaviour.

As SMEs are seldom familiar with formal and conceptual notations that modelers often use, they may require assistance from the development team in reading and understanding the conceptual model representation.

5.2.2 *Validating a model's knowledge base*

An HBM's knowledge base (see Figure 13) contains the information and skills that determines the model's response to events in the (simulated) world, and thus largely determines its behaviour. An incorrect or incomplete knowledge base will likely generate invalid behaviour. A valid database, however, does by itself not automatically guarantee valid behaviour, since other components too contribute to the validity of the generated behaviour (e.g. processing of behaviour moderating variables, like fatigue, et cetera).

The nature of the knowledge base determines to a large extent the opportunities for its validation, as well as the selection of tools for doing so. HBMs that define their knowledge in terms of scripts or states (e.g. Finite State Machines, scripts, IF-THEN-ELSE rules) can often be understood by domain experts, and thus be accessible to evaluation by them. Other knowledge representations, such as Bayesian Networks, Markov models, neural networks are much less accessible for evaluation.

Validating a knowledge base gives insight into the knowledge that "drives" the model's behaviour. This insight can not only help to assess whether the knowledge base warrants the intended use of the model, but it can also help to identify possible inconsistencies, gaps and misconceptions that subsequently can be used to correct and/or complete the model. Knowledge base validation can be achieved by employing a VV&A protocol (e.g. Harmon, 1998; Voogd & Smits, 2013).

5.3 Criterion Validation

In the previous sections we argued that the specification of requirements should be validated (§5.1) and that appropriate referents should be selected (§5.1.1).
The specification of the HBM's constructs should be checked (§5.2), including the knowledge base and behaviour engine. The next phase in validation concerns the performance of the model. It entails the questions whether the models' output meets the requirements. These are issues related to Criterion Validity
the relationship between the model and criterion performance).

The validation of a HBM generally proceeds as follows (NATO RPG-ST12):

- developing a test plan for the HBM;
- designing the testing scenarios;
- conducting the testing and collection of data;
- assessing the test results against the acceptability criteria to determine the HBM's validity.

The protocol of testing a HBM's validity should strictly reflect the user objectives in the context of the defined (simulation) scenarios. It is important to realize that a complete test of a HBM, for all possible courses that a scenario may take, is usually not possible. Each test only supplies information on the system behaviour for that particular (set of) scenarios. Extrapolating the results of that test to other, untested scenarios may be unjustified, especially for complex applications (NATO RPG-ST12).

What if the outcomes of a results validation demonstrate that the model produces implausible or incorrect output on the testing scenarios? Then an organization or user has the following options (NATO RPG-ST12):

- do not use scenarios where the anomalous behaviour occurs. In other words: prevent that the scenario ever comes forward in the intended application (the easiest of options);
- adjust the contents of the model's knowledge base to correct the anomalous behaviour (the next easiest option);
- modify the model's behaviour engine to correct the problem. This may involve the mechanics of single information processing component of the model, but it may also require a modification of the way various components of the model interact. In other words: modify the model's architecture. Modifications in the behaviour engine to repair validity problems is the hardest option.

5.4 External or Application validity

In the validation of a model to its criterion (see §5.3), a series of testing scenarios is used to verify whether the developed human behaviour model meets the specified requirements. Results may make users confident that the model is appropriate and accurate for the goals they use the model for. However, whether the goals are actually achieved in practice, requires another level of validation. This is the issue of external or application validity. It is about the extent to which the model can be used or generalized to the situations for which the model is intended (Aronson, Wilson, Akert, & Fehr, 2007).

As noted earlier, HBM may be developed for a variety of purposes. For example, a HBM may be developed to be integrated in a training simulator, to enable new and/or better training exercises. They may also be used to present commanders with possible outcomes of decisions under consideration and thus enable better decision making. They may also be used to test how a new system or interface is likely to be used by humans, and what possible errors will be made. Whatever the specific purpose of the HBM, the question whether the intended application is in fact realized in practice, can be empirically evaluated by conducting an application validity study.

The best way to do this is to conduct an “experimental-versus-control-group method”-study. This method uses an experimental and a control group with randomly allocated subjects. The experimental group uses the newly developed application; the control group uses the old system or method. Afterwards, task performance is measured using a predetermined criterion task resembling operational task performance.

However, such studies are seldom carried out (Korteling & Sluimer, 1999; Korteling, 2016). One important reason is that such studies tend to cost lots of time and money. Due to this, organizations sometimes use other methods that may perhaps provide less conclusive answers, but do nevertheless provide valuable insights into the qualities and profits of the developed application. On the basis of an analysis by Korteling and Sluimer (1999), Korteling (2016) describes the following alternative methods for application validity (see Table 3):

Table 3 Alternative designs to application validity (Korteling, 2016).

Self-control-transfer method	In this method, the experimental group is also the control group (a “within-subjects” design) . Data are collected on subject performance on the real task or system, and also on performance with the developed system.
Pre-existing-control-transfer method	In this method, the newly developed application is introduced in the organization. Performance is compared to recorded performance prior to introduction. Conclusions based on this method are tentative because of time-related changes (e.g., changes in the trainee group, training methods or circumstances, or the training staff).
Uncontrolled-transfer method	If forming a control group is not possible, the effectiveness of a newly developed application can be established by determining how subjects perform with it the first time. Data from such an uncontrolled-transfer method study are tentative, since one cannot conclusively determine the contribution of the application to subjects’ performance (Caro, 1977).
Quasi-transfer-of-training method	For training applications, the quasi-transfer-of-training method (QToT) is a relatively popular method thanks to its efficiency. In this method, the experimental group is trained until criterion using the newly developed training application. A control group is trained until criterion using a high-fidelity replica of the system. Both are tested afterwards on a criterion task in the high-fidelity replica of the system. Performance differences between groups are used to draw conclusions about the added value of the training application. The major limitation of this design is the absence of performance measurement under real-task (i.e., operational) conditions.
Backward-transfer method	In a backward transfer study, a proficient operator performs the task using the newly developed application. If this expert can instantaneously perform the task successfully, backward transfer has occurred. The assumption here is that transfer in the other direction (forward transfer) for novices and inexperienced performers will also occur.

Simulator-performance-improvement method	In this method, the performance of subjects being trained to work with the newly developed application is measured across a series of sessions in a simulation . The assumption is that an application is effective if it increases subjects' performance over sessions substantially. The assumption is, however, only warranted if the simulation has a high degree of physical, functional, and psychological fidelity to the real-task environment (Korteling, Helsdingen, & Theunissen, 2012).

In order to prevent falling prey to the shortcomings of the above methods, Korteling (2016) advocates to combine multiple methods in validation research. For example, surveys, questionnaires, ratings, and checklists may be used in combination with quantitative measurements (e.g., time, speed, error). Such a combined approach may provide the best mix of reliable and relevant information on model performance in a relatively pragmatic way. In general, sound conclusions on a model's validity require multiple sources of converging data. Examples of such converging measurement methodologies are described by Bell and Waag (1998), Schreiber and Bennett (2006), Schreiber, Bennett, and Stock (2006), and Schreiber, Schroeder, and Bennett (2011).

6 Discussion

The military uses HBMs to attain a potentially wide range of objectives, like training; decision support; doctrine testing; system acquisition; human interface design; et cetera. Validation of human behaviour models is very important, not only for the quality of the models itself, but also as evidence that the models are fit for their intended purposes ('fit for use'). However, the issue of validation is not always given proper importance. One reason may be that research program managers frequently see validation as a drain on resources (Van Hemel et al., 2008). Practitioners or model users, however, typically do view validation as a worthy investment of time and effort, since it can prevent the costly consequences of using incorrect models and simulations. If the intended use is not fully considered, then the model is not as useful as it might be.

This report addressed the issue of validating HBMs in the context of simulation applications. It has been emphasized by many researchers that the validation of a HBM must be specific to an application rather than extensively generalizable (e.g. Campbell & Bolton, 2005; Marks, 2006; Van Hemel et al., 2008). Furthermore, it has been argued that this can best be achieved by a model architecture consisting of a separate knowledge base and a behaviour engine (Harmon et al., 2002; Van Hemel et al., 2008) (see Figure 13). This partitioning creates the flexibility needed to represent the behaviour of different individuals performing in different roles without requiring building a new execution infrastructure each time (Harmon et al., 2002).

The recommended procedure to validate a human behaviour model, as discussed in Chapter 5, is based upon the assumed use of this typical architecture, and corresponds to recommended practices as published by the Defense Materiels Systems Office (DMSO). The general approach is that validation of HBM should begin with the defining of purpose and scope, and then consider the action set, scenarios, and if-then relations in the specific situation. The US National Research Council (NRC) study on behaviour modelling (Van Hemel et al., 2008) provide the following additional practical recommendations to facilitate the validation process for a specific model:

- Check with multiple experts: modelers should not examine a model by themselves; they tend to focus on the verification with less emphasis on the purpose of the model than other experts. The NRC-committee therefore advises to involve at least three other experts in the examination of a model: (a) the users of the model (e.g. instructors, trainees, decision maker); (b) the scenario experts (e.g. training developer); and (c) the domain experts.
- Keep the model as simple as possible for its purpose: A HBM does not necessarily always have to conform to reality in order to be of value. If the model is fit for its purpose, then a parsimonious model is to be preferred over a complex one.
- Examine not only what is known for inclusion in a model, but also to "what might be": Of course, a model has to have correspondence with the real world to be of relevance. However, many relevant action-scenario combinations may not have been observed yet in the past, but may nevertheless be likely to happen in the future. The committee therefore advises developers (within limits) to examine data beyond to what is already known, to "what might be".

- Combine multiple methods and multiple data sources: A combined approach is a pragmatic solution to obtain the best mix of reliable and relevant information on model performance against acceptable efforts.

A final word on the importance of data for validation is in order here. Data issues are an essential component for assessing the ultimate success for model development, validation, and applications (Van Hemel et al., 2008). For some models it is possible to make use of statistical data, a form of data that is considered to be objective or “hard”. Many HBMs, however, rely on data that are tacit and often non-observable. Such data typically tend to become available through SMEs, a form of data that is considered to be soft or subjective. In general, the acquisition of reliable, appropriate, and accurate data is important, because trust in the data ultimately determines ones trust in a model’s outcomes or predictions (Van Hemel et al., 2008).

The military considers the capability to model human behaviour as an essential technology for achieving and maintaining an adequately trained force, and for being able to execute current and future missions to the required standards of operation. This need of the military is an important force behind the research and development of HBM technology and its applications. As the value of an application stands or falls with the quality of its underlying models, the growing interests in HBM validation is a good thing. Moreover, it is a prerequisite for guiding the current investments in Human Behaviour Modelling into the right directions.

7 Referenties

Anderson, J. R., & Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Appleget, J., Blais, C., & Jaye, M. (2013). Best practices for US Department of Defense model validation: lessons learned from irregular warfare models. *Journal of Defense Modelling and Simulation: Applications, Methodology, Technology*, pp 1–16.

Aronson, E., Wilson, T. D., Akert, R. M., & Fehr, B. (2007). *Social psychology*. (4 ed.). Toronto, ON: Pearson Education.

Bell, H. H., & Waag, W. L. (1998). Evaluating the effectiveness of flight simulators for training combat skills: A review. *The International Journal of Aviation Psychology*, *8*(3), 223–242.

Bosse, T., Hoogendoorn, M., Klein, M.C.A., Treur, J., Wal, C.N. van der, & Wissen, A. van (2013). Modelling Collective Decision Making in Groups and Crowds: Integrating Social Contagion and Interacting Emotions, Beliefs and Intentions. *Autonomous Agents and Multi-Agent Systems Journal*, vol. 27, pp. 52-84.

Bratman, M. (1987). *Intention, plans, and practical reason*. Harvard University Press, Cambridge, MA.

Campbell, G. E., & Bolton, A. E. (2005). HBR validation: Integrating lessons learned from multiple academic disciplines, applied communities, and the AMBR project. Modelling human behaviour with integrated cognitive architectures: Comparison, evaluation, and validation, 365-395.

Caro, F. G. (Ed.). (1977). *Readings in evaluation research*. Russell Sage Foundation.

Core, M. G., Lane, H. C., Van Lent, M., Gomboc, D., Solomon, S., & Rosenberg, M. (2006, July). Building explainable artificial intelligence systems. In *Proceedings of the national conference on artificial intelligence* (Vol. 21, No. 2, p. 1766). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281.

Epstein, J.M. (2006). Remarks on the foundations of agent-based generative social science. In: L. Tesfatsion and K.L. Judd (Eds.), *Handbook of computational economics, volume 2: Agent-based computational economics*. Amsterdam, The Netherlands: Holland/Elsevier.

Goerger, S. R., McGinnis, M. L., & Darken, R. P. (2005). A validation methodology for human behaviour representation models. *The Journal of Defense Modelling and Simulation: Applications, Methodology, Technology, 2(1),* 39-51.

Gomboc, D., Solomon, S., Core, M. G., Lane, H. C., & Van Lent, M. (2005). Design recommendations to support automated explanation and tutoring. *Proc. of BRIMS*.

Gonzalez, A. J., & Murillo, M. (1999, March). Validation of Human Behavioural Models. In *FLAIRS Conference* (pp. 489-493).

Gregersen, H., & Sailer, L. (1993). Chaos theory and its implications for social science research. *Human relations*, *46*(7), 777-802.

Groen, E.L., Valk, P., Bijl, P., Korteling, J.E., Ledegang, W.D. (in preparation). Metrics to characterize the effectiveness of simulator (training) under extreme physiological conditions. *TNO-report*. TNO, Soesterberg.

Harbers, M., Bosch, K. van den, & Meyer, J. J. Ch. (2010). Design and Evaluation of Explainable BDI Agents. In: *Proceedings of the International Conference on Intelligent Agent Technology* (pp. 125-132). Held at: Toronto, Canada. WI-IAT.

Harbers, M., van den Bosch, K., & Meyer, J. J. C. (2011, January). A Theoretical Framework for Explaining Agent Behaviour. In *SIMULTECH* (pp. 228-231).

Harmon, S. Y. (1998). Bibliography of Verification, Validation, Evaluation and Testing of Knowledge-Based Systems, Defense Modelling and Simulation Office, Alexandria, VA, 1998.

Harmon, S. Y., Hoffman, C. W. D., Gonzalez, A. J., Knauf, R., & Barr, V. B. (2002, October). Validation of human behaviour representations. *In Foundations for V&V in the 21st Century Workshop (Foundations' 02)* (pp. 22-24).

Holden, Ronald B. (2010). "Face validity". In: Weiner, Irving B.; Craighead, W. Edward. *The Corsini Encyclopedia of Psychology* (4th ed.). Hoboken, New Jersey: Wiley. pp. 637–638.

Hosseini, K. (2013). *And the mountains echoed*. Penguin Canada.

Immergluck, L. (1964). Determinism-freedom in contemporary psychology: An ancient problem revisited. *American Psychologist*, *19*(4), 270.

ITT Research Institute. (2001). Modelling and Simulation Information Analysis Center (MSIAC). *Verification, validation, and accreditation (VV&A) automated support tools: A state of the art report, Part 1-Overview*. Chicago, IL: Author.

Jiang, H., Vidal, J.M., and Huhns, M.N. (2007). EBDI: An Architecture for Emotional Agents. In: *Proc. of the 6th Int. Joint Conf. on Autonomous Agents and Multi-Agent Systems*, AAMAS'07. ACM Press, pp. 38-40.

Jones, R. M. (2004, December). An introduction to cognitive architectures for modelling and simulation. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference. I/ITSEC, Orlando, FL*.

Kettler, B., & Hoffman, M. (2012). Lessons learned in instability modelling, forecasting, and mitigation from the DARPA integrated crisis early warning system (ICEWS) program. In *2nd International Conference on Cross-Cultural Decision Making: Focus*.

Kieras, D. & Meyer, D.E. (1997). *An overview of the EPIC architecture for cognition and performance with application to human-computer interaction*. Human-Computer Interaction, 12, 391-438.

Klein, G. (1998). *The source of power: how people make decisions*. Cambridge, MA: MIT Press.

Korteling, J. E., & R.R. Sluimer, (1999). A critical review of validation methods for man-in-the-loop simulators. (*Report No. TM-99-A023*). Soesterberg, the Netherlands: TNO-TM

Korteling, J.E. (2016). Determining training effectiveness. In: Z. Wang (Ed.) *Cost-benefit Analysis of Military Training.* Report No. SRTO-TR-SAS-095. Paris: North Atlantic Treaty Organization [NATO] Research & Technology Organisation [RTO].

Korteling, J.E., & Bosch, K. van den (2015). Validatie van Educatieve Games (in Dutch). Homo Ludens.

Laird, J. E., Newell, A., and Rosenbloom, P. S. (1987). *SOAR: an architecture for general intelligence*. Artificial Intelligence, 33(1), 1-64.

Livingstone, D. (2006). Turing's test and believable AI in games. Computers in Entertainment (CIE), 4(1), 6.

Marks, R.E. (2006). *Validation and complexity.* Working paper, Austrailian Graduate School of Management, University of New South Wales, Sydney.

Muller, T. J., Heuvelink, A., Bosch, K. van den, & Swartjes, I. (2012). Glengarry Glen Ross: Using BDI for Sales Game Dialogues. In: *The Eighth Annual AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. Held at: Palo Alto, CA.

NATO RPG-ST12, 2001. Validation of Human Behaviour Representation. [http://www.msco.mil/documents/RPG/ST12_Human_Behav_Validation.pdf]

Pokahr, A., & Braubach, L. (2007). Jadex user guide. https://download.actoron.com/docs/releases/jadex-0.96x/userguide/index.html.

Rao, A. S., & Georgeff, M. P. (1991). Modelling rational agents within a BDI-architecture. *KR*, 91, 473-484.

Rumelhart, D., McClelland, J. L., & the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Volume 1: Foundations. Cambridge: The MIT Press.

Schreiber, B. T., & Bennett, W., Jr. (2006). *Distributed mission operations within-simulator training effectiveness baseline study: Summary report* (Report No. AFRL-HE-AZ-TR-2006-0015, Vol. 1). Mesa, AZ: Warfighter Readiness Research Division.

Schreiber, B. T., Bennett, W., Jr., & Stock, W.A. (2006). *Distributed mission operations within-simulator training effectiveness baseline study: Metric development and objectively quantifying the degree of learning* (Report No. AFRL-HE-AZ-TR-2006-0015, Vol. 2). Mesa, AZ: Warfighter Readiness Research Division.

Schreiber, B. T., Schroeder, M., & Bennett, W., Jr. (2011). Distributed mission operations within-simulator training effectiveness. *The International Journal of Aviation Psychology, 21*(3), 254–268. doi: 10.1080/10508414.2011.582448

Silverman, B.A. (2001). More realistic human behaviour models for agents in virtual worlds: emotion, stress and value ontologies. (Report No. Technical Report). Philadelphia, PA: Univ. of Penn/ACASA.

Smith, R. (2009). The long history of gaming in military training. Simulation & Gaming.

Suchman, L. (1986). Plans and situated actions. *New York, Cambridge University*.

Sun, R. (2006). *The CLARION cognitive architecture: Extending cognitive modelling to social simulation* In: Ron Sun (ed.), Cognition and Multi-Agent Interaction. Cambridge University Press, New York.

Taatgen, N., & Anderson, J. R. (2010). The past, present, and future of cognitive architectures. *Topics in Cognitive Science*, *2*(4), 693-704.

Tolk, A. (Ed.). (2012). Engineering principles of combat modelling and distributed simulation (pp. 263-294). New Jersey: Wiley.

Van den Bosch, K., & Doesburg, W. A. van (2005). Training Tactical Decision Making Using Cognitive Models. In: *Proceedings of the Seventh International NDM Conference*. Held at: Amsterdam, the Netherlands.

Van den Bosch, K., Harbers, M., Heuvelink, A., & Doesburg, W.A. van (2009). Intelligent Agents for Training On-board Fire Fighting. In: *Digital Human Modelling, HCII 2009* (pp. 463-472). Held at: San Diego, CA. Berlin Heidelberg: Springer-Verlag.

Van Hemel, S. B., MacMillan, J., & Zacharias, G. L. (Eds.). (2008). Behavioural Modelling and Simulation: From Individuals to Societies. National Academies Press.

Voogd, J.M., & Smits, C.S. (2013). Verification and Validation of Live, Virtual and Constructive Simulation. TNO-report.

Weinberger, S. (2011). Web of war. *Nature*, *471*(7340), 566-568.

Zachary, W.W., Ryder, J.M., & Hicinbothom, J.H. (1998). Cognitive task analysis and modelling of decision making in complex environments. In: J. A. Cannon-Bowers, & E. Salas (Eds.), *Making decisions under stress: implications for individual and team training* (pp. 315-344). Washington, DC: APA.

REPORT DOCUMENTATION PAGE
(MOD-NL)

1. DEFENCE REPORT NO (MOD-NL)	2. RECIPIENT'S ACCESSION NO	3. PERFORMING ORGANIZATION REPORT NO
-	-	TNO 2016 R11848
4. PROJECT/TASK/WORK UNIT NO	**5. CONTRACT NO**	**6. REPORT DATE**
060.07968	-	January 2016
7. NUMBER OF PAGES	**8. NUMBER OF REFERENCES**	**9. TYPE OF REPORT AND DATES COVERED**
38 (excl RDP & distribution list)	55	Final

10. TITLE AND SUBTITLE

Validating models of human behaviour

11. AUTHOR(S)

Bosch, K. van den, Korteling, J.E.

12. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

TNO, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands
Kampweg 5, Soesterberg, The Netherlands

13. SPONSORING AGENCY NAME(S) AND ADDRESS(ES)

Ministry of Defense, DSMO/JIVC/KIXS

14. SUPPLEMENTARY NOTES

The classification designation Ongerubriceerd is equivalent to Unclassified, Stg. Confidentieel is equivalent to Confidential and Stg. Geheim is equivalent to Secret.

15. ABSTRACT (MAXIMUM 200 WORDS (1044 BYTE))

Moderns simulators and games more and more include the behaviour of humans: of individuals, groups and even societies. This development opens opportunities for the military to use the technology of behaviour modelling for purposes of training, tactics analysis, and mission preparation. Realizing this potential demands that the behaviours of the human(s) involved are adequately modelled for their purpose in the simulation, i.e. that the models are “fit for use”. This report discusses the use of Human Behaviour Models (HBMs) in simulations, and the opportunities and pitfalls of determining their validity.

16. DESCRIPTORS **IDENTIFIERS**

Human behaviour modelling; training; validation

17a. SECURITY CLASSIFICATION (OF REPORT)	17b. SECURITY CLASSIFICATION (OF PAGE)	17c. SECURITY CLASSIFICATION (OF ABSTRACT)
Ongerubriceerd	Ongerubriceerd	Ongerubriceerd

18. DISTRIBUTION AVAILABILITY STATEMENT	17d. SECURITY CLASSIFICATION (OF TITLES)
Subject to approval MOD-NL	Ongerubriceerd

Distribution list TNO 2016 R11848 (V1427)

The following agencies/people will receive a complete copy of the report.

DEFENSIE

hardcopy NLDA/Projectbureau K&I, Defensie Programma procesbegeleider
Dr. M.J.P. van Veen

hardcopy NLDA/Bibliotheek KMA

hardcopy Defensie Programmabegeleider DMO/JIVC/KIXS
Lkol P. van Onzenoort

pdf PLANNEN/K&I, Kol. J.C. Dicke

pdf DMO/Joint IV Commando/C4I&I/InformatieBeheer/PDB

pdf Projectbegeleider DMO/JIVC//KIXS
A.C van Lier

TNO

pdf TNO Referent
Drs. M.P. van Esch-Bussemakers

pdf TNO Programmaleider (PGL) Dr W. Huiskamp

pdf Afdelingshoofd TNO PGL
Drs. W.S.M. Piek
Ir. J.P. Dezaire

pdf TNO medewerkers op aangeven van de TNO PGL
Dr. J.E. Korteling
Dr. J. Voogd
Dr. K. van den Bosch

hardcopy TNO Archief (locatie Soesterberg)