Machine Learning for Adversarial Agent Microworlds

J. Scholz[1], B. Hengst[2], G. Calbert[1], A. Antoniades[2], P. Smet[1], L. Marsh[1], H-W. Kwok[1], D. Gossink[1]

[1]DSTO Command and Control Division, [2]NICTA, E-Mail: Jason.Scholz@defence.gov.au

Keywords: *Wargames, Reinforcement learning, Hierarchy; Representation.*

EXTENDED ABSTRACT

We describe models and algorithms to support interactive decision making for operational national and military strategic courses of action.

Abstract representations or 'microworlds' have been used throughout military history to aid in conceptualization and reasoning of terrain, force disposition and movements. With the introduction of digitized systems into military headquarters the capacity to degrade decision-making has become a concern with these representations. Maps with overlays are a centerpiece of most military headquarters and may attain an authority which exceeds their competency, as humans forget their limitations both as a model of the physical environment, and as a means to convey command intent (Miller 2005). The design of microworlds may help to alleviate human-system integration problems, by providing detail where and when it is appropriate and recognizable abstractions where it is not. Ultimately, Lambert and Scholz (2005 p.29) view "the division of labor between people and machines should be developed to leave human decision making unfettered by machine interference when it is likely to prove heroic, and enhance human decision making with automated decision aids, or possibly override it with automated decision making, when it is likely to be hazardous." Cognisant of these issues, we state a formalism for microworlds initially to manage representation complexity for machine learning. This formalism is based on multi-agent stochastic games, abstractions and homomorphism.

According to Kaelbling *et al* (1996), "Reinforcement learning (RL) is the problem faced by an agent that must learn behaviour through trial and error interactions with a dynamic environment." As RL is frustrated by the curse of dimensionality, we have pursued methods to exploit temporal abstraction and hierarchical organisation. As noted by Barto and Mahadevan 2003 (p. 23), "Although the proposed ideas for hierarchical RL appear promising, to date there has been insufficient experience in experimentally testing the effectiveness of these ideas on large applications."

We describe the results of RL approaches to microworld applications of progressively increasing complexity, as determined by the state-action space dimension. These include discrete space-time (grid-world) models of modified Chess and Checkers, and a military campaign-level game with concurrent moves, called "TD-Island". We also detail complexity reduction through hierarchical decomposition in a continuous space model of an operational level air-combat model as is shown in figure 1.

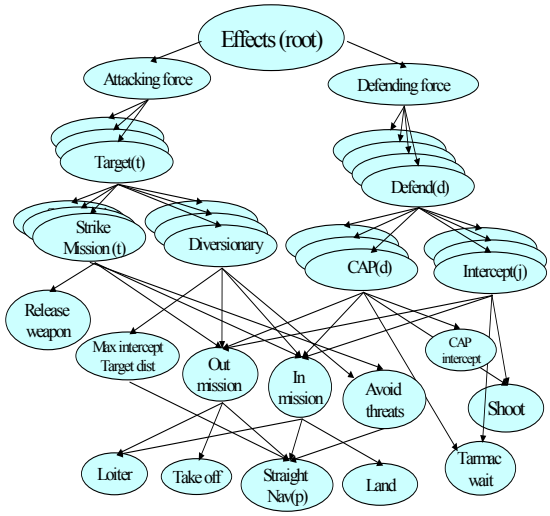

Figure 1: Hierarchical decomposition of a complex operational air-game.

Given this decomposition, current hierarchical RL approaches are not sufficient in an adversarial context as game-theoretic notions such as mixed or randomised strategies, to keep the opponent "second-guessing" are absent. We briefly describe a new approach to hierarchical learning. This combines both the "win-or-learn-fast" algorithm which learns mixed strategies and the MAXQ algorithm developed to learn through the structure of the hierarchy.

Finally, we compare the performance of human guessed parameter importance with that of machine learned importance in a modified game of Checkers with hidden-pieces. Our results show that in new or novel domains, such as the hidden Checkers game, machine-learned parameters fare better in play, as the algorithms can gain experience faster than humans.

1. INTRODUCTION

During Carnegie Mellon University's Robotics Institute 25th anniversary, Raj Reddy said, "the biggest barrier is (developing) computers that learn with experience and exhibit goal-directed behaviour. If you can't build a system that can learn with experience, you might as well forget everything else". This involves developing agents characterized by the ability to:

- learn from their own experience and the experience of human commanders,
- accumulate learning over long time periods on various microworld models and use what is learned to cope with new situations,
- decompose problems and formulate their own representations, recognising relevant events among the huge amounts of data in their "experience",
- exhibit adaptive, goal directed behaviour and prioritise multiple, conflicting and time varying goals,
- interact with humans and other agents using language and context to decipher and respond to complex actions, events and language.

Extant agent-based modelling and simulation environments for military appreciation appear to occupy extreme ends of the modelling spectrum. At one end, Agent-Based Distillations (ABD) provide weakly expressive though highly computable (many games per second execution) models using attraction-repulsion rules to represent intent in simple automata (Horne and Johnson 2003). At the other end of the spectrum, large scale systems such as the Joint Synthetic Armed Forces (JSAF) and Joint Tactical Level Simulation (JTLS) require days to months and many staff to set up for a single simulation run (Matthews and Davies 2001). We propose a microworld-based representation, which falls between these extremes in that it may be used to achieve statistically-significant results yet also allows for decision interaction.

2. INTERACTIVE MICROWORLDS

Microworlds have been used throughout military history to represent terrain, force disposition and movements. Typically such models were built as miniatures on the ground as depicted in figure 2 or on a board grid, but these days are more usually computer-based.

Figure 2. Example of a Microworld.

Creating a microworld for strategic decisions is primarily a cognitive engineering task. A microworld (e.g. like a map with overlaid information) is a form of symbolic language that should represent the necessary objects and dynamics, but not fall into the trap of becoming as complex as the environment it seeks to express (Friman and Brehmer 1999). It should also avoid the potential pitfalls of interaction, including the human desire for “more”, when more is not necessarily better. Omedei *et al* (2004) have shown that humans under certain conditions display weakness in self-regulation of their own cognitive load. Believing for example, that “more detailed information is better”, individuals can become information overloaded, fail to realise that they are overloaded and mission effectiveness suffers. Similarly, a belief that “more reliable information is better”, individuals can be informed or observe that some sources are only partially reliable, fail to pay sufficient attention to the reliable information from those sources and mission effectiveness suffers. Cognisant of these issues, we seek a formalism which will allow the embodiment of an austere, adaptable and agreed representation. Though there is a formal definition of agreement in a mathematical sense based on homomorphism as we describe below, in practice, humans are the purveyors of authority for agreed models.

2.1. Microworld Formalism

We use the formalisation of multi-agent stochastic games (Fundenberg and Tirole 1995) as our generalised scientific model for microworlds. A multi-agent stochastic game may be formalised as a tuple $\langle N, S, A, T, R \rangle$ where N is the number of agents, S the states of the microworld, $A = A_1 \times A_2 \times ... \times A_N$ concurrent actions, where A_i is the set of actions available to agent i, T is the stochastic state transition function for each action, $T : S \times A \times S \to [0,1]$, and

$R = (R_1, R_2, ..., R_N)$, where R_i is the immediate reward for agent i, $R_i : S \times A \to \Re$ .The objective is to find an action policy $\pi_i(S)$ for each agent i to maximise the sum of future discounted rewards R_i. This is effectively, a multi-agent Markov Decision Process (MDP).

In multi-agent stochastic games each agent must choose actions in accordance with the actions of other agents. This makes the environment essentially non-stationary. In attempting to apply learning methods to complex games, researchers have developed a number of algorithms that combine game-theoretic concepts with reinforcement learning algorithms, such as the Nash-Q algorithm (Hu and Wellman 2003).

Abstractions, Homomorphism and Models

To make learning feasible we need to reduce the complexity of problems to manageable proportions. A challenge is to abstract the huge state-space generated by multi-agent MDPs. To do this, we plan to exploit any structure and constraints in the problem to decompose the MDP.

Models can be thought of as abstractions that generally reduce the complexity and scope of an environment to allow us to focus on specific problems. A good or homomorphic model may be thought of as a many-to-one mapping that preserves operations of interest as shown in figure 3. If the environment transitions from environmental states w_1 to w_2 under environmental dynamics, then we have an accurate model if the model state m_1 transitions to m_2 under a model dynamic and states w_1 map onto m_1, w_2 map onto m_2 and environmental dynamics map onto the model dynamics.

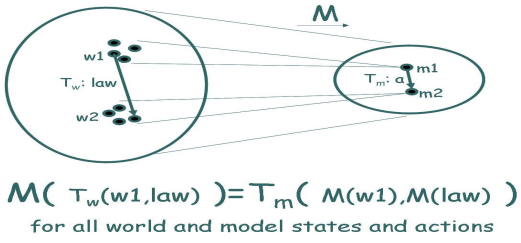

Figure 3: Homomorphic models

The purpose of a microworld is to represent an agreed model of a situation. The following models are constructed on the basis of our observations of military planning and general strategic thought.

2.2. Grid World Examples

Chess and Checkers Variant Microworlds

We initially focused on variants of abstract games in discrete time and space such as Chess and Checkers. We deliberately introduced asymmetries into the games and conducted sensitivity analysis through Monte-Carlo simulation at various ply depths up to seven. The six asymmetries were either materiel, where one side commences with a materiel deficit, planning, with differing ply search depths, tempo, in which one side was allowed to make a double move at some frequency, stochastic dynamics where pieces are taken probabilistically and finally hidden pieces, where one side possessed pieces that could be moved, though the opponent could only indirectly infer the piece location (Smet *et al* 2005).

The TDLeaf algorithm is suitable for learning value functions in these games (Baxter et al, 1998). TDLeaf takes the principal value of a minimax tree (at some depth) as the sample used to update a parameterised value function through on-line temporal differences. Independently discovered by Beal and Smith in 1997, TDLeaf has been applied to classical games such as Backgammon, Chess, Checkers and Othello with impressive results (Schaeffer, 2000).

For each asymmetric variant, we learned a simple evaluation function, based on materiel and mobility balance. The combined cycle of evaluation function learning and Monte Carlo simulation allowed us to draw general conclusions about the relative importance of materiel, planning, tempo, stochastic dynamics and partial observability of pieces (Smet *et al* 2005) and will be described further in section 4.

Our approach to playing alternate-move partially-observable games combined some aspects of information theory and value-function based reinforcement learning. During the game, we stored not only a belief of the distribution of our opponent's pieces, but conditioned on one of these states, a distribution of the possible positions of our pieces (a belief of our opponent's belief). This enabled the use of 'entropy-balance,' the balance between relative uncertainty of opposition states as one of the basis functions in the evaluation function (Calbert and Kwok 2004). Simulation results from this are described in section 4.

TD-Island Microworld

TD-Island stands for 'Temporal Difference Island', indicating the use of learning to control elements

in an island-based game. Space and time are discrete in our model. Each of 12 discrete spatial states is either of type land or sea (figure 4).

Figure 4: TD-Island microworld.

Elements in the game include jetfighters, army brigades and logistics supplies totaling twenty per side. Unlike Chess, two or more pieces may occupy the same spatial state. Further from Chess, in TD-Island, there is concurrency of action in that all elements may move at each time step. This induces an action-space explosion with around one million move options per turn (compared with Chess with thirty six moves per turn). This makes centralized control difficult, particularly as the number of elements may be large. The TD-Island agent uses domain symmetries and heuristics to prune the action-space, inducing the homomorphic model.

Some symmetries in the TD-Island game are simple to recognise. For example, if considering the set of all possible concurrent actions of jetfighters, striking targets is the set of all possible permutations of jetfighters assigned to these targets. One can eliminate this complexity by considering only a single permutation. By considering this single permutation, one "breaks" the game symmetry, reducing the overall concurrent action space dimension which induces a homomorphic model (Calbert 2004).

The approach to state-space reduction in the TD-Island game is similar to that found in Chess based reinforcement learning studies, such as the KnightCap game (Baxter *et al* 1998). A number of features or basis functions are chosen and the importance of these basis functions forms the set of parameters to be learned. The history of Chess has developed insights into the appropriate choice of basis functions (Furnkranz and Kubat 2001). Fortunately there are considerable insights into abstractions important in warfare, generating what is termed 'operational art'. Such abstractions may include the level of simultaneity, protection of forces, force balance and maintenance of supply lines amongst others (US Joint Chiefs of Staff 2001). Each of these factors may be encoded into a number of different basis functions, which form the heart of the value function approximation.

2.3. Continuous Time-Space Example

Tempest Seer

This is a microworld for evaluating strategic decisions for air operations. Capabilities modelled include various fighters, air to air refuelling, surface to air missiles, ground based radars, as well as logistical aspects. Airports and storage depots are also included. The model looks at a maximum 1000 km square of continuous space, with time advancing in discrete units. It is illustrated in figure 5. The Tempest Seer microworld will enable users to look at the potential outcomes of strategic decisions. Such decisions involve the assignment of aircraft and weapons to missions, selection of targets, defence of assets and patrolling of the air space. While these are high-level decisions, modelling the repercussions of such decisions accurately requires attention to detail. This microworld is therefore at a much lower level of abstraction than TD-Island.

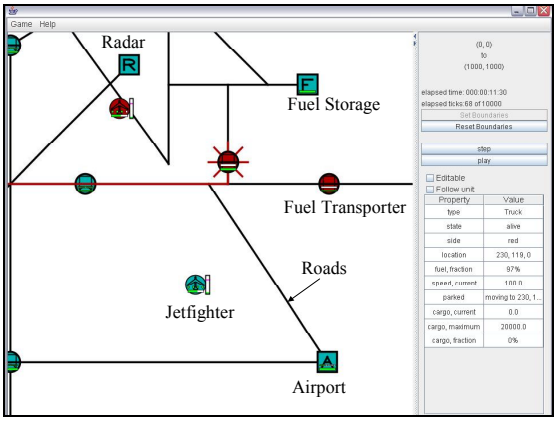

Figure 5: Tempest Seer microworld.

Given the complexity of this microworld, it is natural to attempt to systematically decompose its structure into a hierarchically organized set of sub-tasks. The hierarchy would be decomposed into higher-level strategic tasks such as attack and defend at the top, with lower level tactical level tasks near the bottom of the hierarchy, with primitive or atomic moves at the base. Such a hierarchy has been composed for the Tempest Seer game as shown in figure 1.

Given this decomposition, one can apply the MAXQ-Q algorithm to learn game strategies (Dieterrich, 2002). One can think of the game being played through the execution of differing subroutines, these being nodes of the hierarchy as differing states are encountered. Each node can be considered as a stochastic game, focused on a

particular aspect of play such as targeting, or at an elemental level taking-off or landing of aircraft.

Though the idea of learning with hierarchical decomposition is appealing, in order to make the application practical, several conditions must apply. First, our human-designed hierarchy must be valid, in the sense that at the nodes of the hierarchy capture the richness of strategic development used by human players. Second, crucially, one must be able to abstract out irrelevant parts of the state representation, at differing parts of the hierarchy, in a process called structural abstraction. As an example, consider the "take-off" node. Provided all aircraft, including the enemies, are sufficiently far from the departing aircraft, we can ignore their states. The states of other aircraft can effectively be safely abstracted out, without altering the quality of strategic learning. Without such abstraction, hierarchical learning is in fact far more expensive than general "flat" learning (Dietterich, 2002). Finally, the current theory of hierarchical learning must be modified to include explicit adversarial reasoning, for example the randomisation of strategies to keep the opponent "second-guessing" so to speak.

We have completed research into incorporating adversarial reasoning into hierarchical learning through a vignette in which taxi's compete for passengers, and are able to "steal" passengers. The hierarchical decomposition for this game is shown in figure 6.

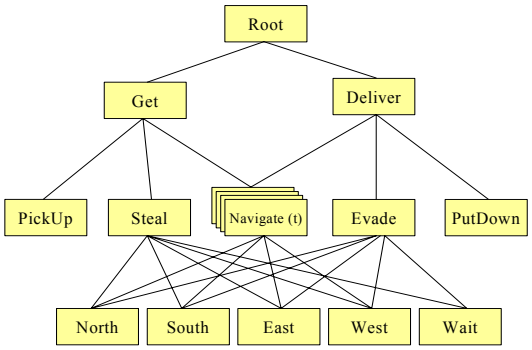

Figure 6: Hierarchical decomposition for the competitive taxi problem.

In an approach to combining hierarchy with game-theory, one of the authors has combined MAXQ learning with the "win-or-learn fast" or WOLF algorithm (Bowling and Veloso, 2003) which adjusts the probabilities of executing a particular strategy according to whether you are doing well in the strategic play or losing.

3. SUMMARY ISSUES

We have found the use of action and state symmetry, extended actions, time-sharing, deictic representations, and hierarchical task decompositions useful for improving the efficiency of learning in our agents. Action symmetries arise, for example, when actions have identical effects in the value or Q-function (Ravindran 2004, Calbert 2004). Another form of action abstraction is to define actions extended in time. They have been variously referred to as options (Sutton *et al* 1999), sub-tasks, macro-operators, or extended actions. In multi-agent games the action space $A = A_1 \times A_2 \times ... \times A_N$ describes concurrent actions, where A_i is the set of actions available to agent *i*. The number of concurrent actions is $|A| = \prod_i |A_i|$.

We can reduce the problem by constraining the agents so that only one agent can act at each time-step. State symmetry arises when behaviour over state sub-spaces is repetitive throughout the state space (Hengst 2002), or when there is some geometric symmetry that solves the problem in one domain and translates to others. Deictic representations only consider state regions important to the agent, for example, the agent's direct field of view. Multi-level task decompositions use a task hierarchy to decompose the problem. Whether the task hierarchy is user specified as in MAXQ (Dietterich 2000) or machine discovered as in HEXQ (Hengst 2002), they have the potential to simultaneously abstract both actions and states.

Part of our approach to address achieving human agreement and trust in the representation has been a move from simple variants of Chess and Checkers to more general microworlds like TD Island and Tempest Seer. We advocate an open-interface component approach to software development, allowing the representations to be dynamically adaptable as conditions and hence models need to change.

4. RESULTS OF A MICROWORLD GAME

We illustrate learning performance for a Checkers-variant microworld. In this variant, each side is given one hidden piece. If this piece reaches the opposition baseline, it becomes a hidden king. After some experience in playing this game, we estimated the relative importance of standard pieces, hidden unkinged and hidden kinged pieces. We then machine learned our parameter values using the belief based TDLeaf algorithm described earlier.

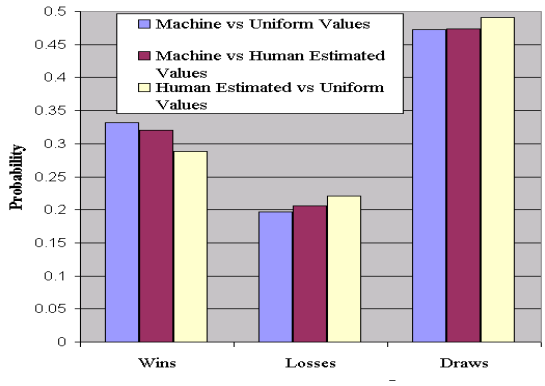

Figure 7. Results for Checkers with hidden pieces.

Figure 7 details the probability of wins, losses and draws for three competitions each of 5000 games. In the first competition, an agent with machine-learned evaluation function weights was played against an agent with an evaluation function where all weights were set as uniformly important. In the second competition, an agent with machine-learned evaluation function weights was played against an agent with a human-estimated evaluation function. Finally, an agent using human estimated weights played against an agent using uniform values. The agent with machine-estimated evaluation function weights wins most games. This result points towards our aim to construct adequate machine learned representations for new unique strategic situations.

5. FUTURE WORK AND CONCLUSIONS

We will continue to explore the theme of microworld and agent design and fuse our experience in games, partially-observable planning and extended actions. Though we have identified good models as homomorphic, there is little work describing realistic approaches to constructing these homomorphisms for microworlds or adversarial domains.

We are also interested in capturing human intent and rewards as perceived by military players in microworld games. This may be viewed as an inverse MDP, where one determines rewards and value function given a policy, as opposed to determining a policy through predetermined rewards (Ng and Russell 2000). We believe that augmentation of human cognitive ability will provide an improvement in the ability for military commanders to achieve their intent.

6. ACKNOWLEDGEMENTS

National ICT Australia is funded through the Australian Government's Backing Australia's Ability initiative, in part through the Australian Research Council. We wish to acknowledge the inspiration of Dr Jan Kuylenstierna and his team from the Swedish National Defence College.

7. REFERENCES

Alberts, D., Gartska, J., Hayes, R., (2001) Understanding information age warfare, CCRP Publ., Washington, DC.

Barto, A.G., Mahadevan, S., (2003) Recent advances in hierarchical reinforcement learning. *Discrete-Event Systems*, 13:41-77. Special Issue on Reinforcement Learning.

Baxter, J.; Tridgell, A.; and Weaver, L, (1998) TDLeaf(λ): combining temporal difference learning with game tree search. Proceedings of the Ninth Australian Conference on Neural Networks*:* 168-172.

Beal, D.F. and Smith, M.C., (1997) Learning piece values using temporal differences, *ICCA Journal* 20(3): 147-151.

Bowling, M and Veloso, M., (2002) Multiagent learning using a variable learning rate. *Artificial Intelligence,* Vol. 136, pp. 215-250.

Calbert, G., (2004), Exploiting action-space symmetry for reinforcement learning in a concurrent dynamic game, Proceedings of the International Conference on Optimisation Techniques and Applications.

Calbert, G and Kwok, H-W, (2004), Combining entropy based heuristics, minimax search and temporal differences to play hidden state games. Proceedings of the Florida Artificial Intelligence Symposium, AAAI Press.

Dietterich, T.G, (2000), Hierarchical reinforcement learning with the MAXQ function decomposition, *Journal of Artificial Intelligence Research*, Vol. 13, pp. 227-303.

Friman, H. and Brehmer, B., (1999), Using microworlds to study intuitive battle dynamics: a concept for the future, ICCRTS, Rhode Island, June 29 - July 1.

Fundenberg, D. and Tirole, J., (1995), Game Theory. MIT Press.

Furnkranz, J., Kubat, M. (eds), (2001), Machines that learning to play games. *Advances in Computation Theory and Practice,* Vol.8. Nova Science Publishers, Inc.

Hengst, B., (2002), Discovering hierarchy in reinforcement learning with HEXQ, Proceedings of the Nineteenth International Conference on Machine Learning, pp. 243-250, Morgan-Kaufman Publ.

Horne, G., Johnson, S., Eds., (2003), Maneuver Warfare Science, USMC Project Albert Publ.

Hu, J., Wellman, M.P., (2003), Nash-Q learning for general sum stochastic games*, Journal of Machine Learning Research,* Vol. 4, pp. 1039-1069.

Kaelbling L.P., Littman M.L., Moore A.W., (1996), Reinforcement learning: a survey, *Journal of Artificial Intelligence Research*, Vol. 4, pp. 237-285.

Kuylenstierna, J., Rydmark, J. and Fahrus, T. (2000), The value of information in war: some experimental findings", Proc. of the 5th International Command and Control Research and Technology Symposium (ICCRTS), Canberra, Australia.

Lambert, D., Scholz, J.B., (2005), A dialectic for network centric warfare, Proceedings of the 10th International Command and Control Research and Technology Symposium (ICCRTS), MacLean, VA, June 13–16. http://www.dodccrp.org/events/2005/10th/CD/papers/016.pdf

Matthews, K, Davies, M., (2001), Simulations for joint military operations planning, Proceedings of the SIMTEC Modelling and Simulation Conference.

Miller, C.S., (2005), A new perspective for the military: looking at maps within centralized command and control systems, Air and Space Power Chronicles, 30 March. http://www.airpower.maxwell.af.mil/airchronicles/cc/miller.html

Ng, A.Y. and Russell, S., (2000), Algorithms for inverse reinforcement learning, Proceedings of the Seventeenth International Conference on Machine Learning.

Omodei, M. M., Wearing, A. J., McLennan, J., Elliott, G. C., & Clancy, J. M., (2004), More is better? Problems of self regulation in naturalistic decision making settings, in Montgomery, H., Lipshitz, R., Brehmer, B. (Eds.), Proceedings of the 5th Naturalistic Decision Making Conference.

Ravindran, B., (2004), An algebraic approach to abstraction in reinforcement learning". Doctoral Dissertation, Department of Computer Science, University of Massachusetts, Amherst MA.

Schaeffer, J., (2000), The games computers (and people) play. *Advances in Computers* 50 , Marvin Zelkowitz editor, Academic Press, pp. 189-266.

Smet, P., Calbert, G., Kwok, H-W, Scholz, J., Gossink, D., (2005), The effects of materiel, tempo and search depth in win loss ratios in chess, Proceedings of Australian Artificial Intelligence Conference.

Sutton, R., Precup, D. and Singh, S., (1999), Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, Vol. 112. pp. 181-211.

US Joint Chiefs of Staff, (2001), Joint Doctrine Keystone and Capstone Primer, Appendix A, 10 Sept. http://www.dtic.mil/doctrine/jel/new_pubs/primer.pdf