# Machine Education - the way forward for achieving trust-enabled machine agents

**George Leu** Erandi Lakshika Jiangjun Tang Kathryn Merrick
Michael Barlow
Trusted Autonomy Group
School of Engineering and IT, UNSW Canberra, Australia
g.leu@adfa.edu.au

## Abstract

This position paper proposes the machine education concept, a novel point of view for enabling trust capabilities in machine agents in the context of human-machine teaming, via emergence of inherent systems of values solely from interaction with environment and peers. Thus, the paper aims to generate the foundation and a potential road-map of future research on how machine education concept can contribute to building trust-capable machine agents within heterogeneous human-machine teams, in order to ensure team success. (*An extended version of this work is close to submission to AI Journal. At the time of this workshop, the work will be under review. There is some amount of overlapping content, however, a substantial shift in position and scope exists; therefore, the two papers can be considered as different pieces of work.*)

## 1 Introduction

With the advent of more and more capable machine learning concepts and technologies, machine agents are today closer to autonomous operation than ever before, and the levels of autonomy they reach are increasing rapidly in all aspects implied by the definition of autonomy [1]. However, with increased autonomy also comes the problem of trust capabilities of these agents, which is closely related to and conditions their social acceptance and inclusion, as well as the overall performance of the team/society they operate in. On the one hand, since they are inherently adaptive and operate based on learning from environmental and peer inputs, it becomes problematic whether they will achieve their intended purposes, and more importantly, whether they will do so without harming or impeding their peers (and the society at large) to operate efficiently. It is important thus that their behaviour is trustworthy from the perspective of others (peers, designers, etc.). On the other hand, while operating in a designated team for a particular task, or in society for general purposes, these agents should also be able to trust/distrust their peers and take from them or give them control over tasks when needed in order to achieve (either individual or collective) goals. Trust is therefore at the core of social intelligent behaviour [2], and from a machine intelligence perspective this is a fundamental issue in relation to intelligent agents, which are supposed to operate in real-life conditions, in our society. Socially, heterogeneous teams of human and machine agents require first of all that trust between team members is enabled at each moment in time [1]. Without trust, delegation of tasks, acceptance of tasks, sharing of tasks and any collaborative intelligent behaviour [9] in general may be hardly achievable, and thus, the functionality of the team may be further affected, with significant impact on its success.

In the case of humans, trust and distrust are inherently embedded in the complex fabric of human social behaviour, and come from long term learning either through formal education or through various life experiences and interactions. Displaying trustworthy behaviour or trusting the behaviour of others can

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

be, arguably though, considered trivial from a human perspective. In the case of machines, it is less clear how or if the two facets of trust can be achieved in the same manner as in humans' case; thus most of the existing approaches employ various constraints or prescriptive behaviours when critical thresholds are reached [37]. However, ensuring trust at the interaction between heterogeneous agents (machines and humans) in complex and dynamic environments may become extremely difficult when the preferred approaches rely on imposing hard and explicit constraints, or prescriptive behaviour. The issues become even more serious when these behaviours need to be redesigned for new tasks and new environmental niches, where agents have to be reused and still exhibit trust capabilities within a designated team, or in society in general.

In this paper we contribute to the concept of trust-enabled autonomous systems, studied under the umbrella of trusted autonomy [1, 2], by promoting a research position and a potential research field in which machines are educated towards exhibiting trust capabilities, rather than constrained or programmed. In this way, trust can be injected in a machine agent directly in its learning capabilities, and can be developed through systematic training programs that lead to the emergence of implicit systems of values. These systems of values can further guide agent's behaviour so that the agent (1) is trustworthy and (2) can trust/distrust other agents.

The obvious source of inspiration for the Machine Education (ME) concept is human education, under the assumption that machine agents possess enough learning capabilities so they can be subject to education in similar (or as close as possible) ways the humans are. Humans perform tasks requiring various types of skills or combinations of skills, at all levels from physical to mental. There are many types of learning as well, with some learning types more appropriate than others for certain tasks. Also, the same activity may require different learning types in different contexts. Therefore, when educating people for a certain type of task, a systematic curriculum is used, which trains the relevant skills in isolation, then trains their integration at various levels, until the necessary 'heterarchy' [28] of skills is acquired. This heterarchy of skills allows the performance of the required task, and also embeds the system of values necessary for exhibiting trust capabilities. Therefore, enabling trust through education means that an agent (be it natural or artificial) is guided towards inherently adhering to certain systems of values. A naive but intuitive example from human domain can be: children should not smoke. This can be achieved by enforcing a crisp constraint/rule, such as "Smoking is not allowed. Do not smoke!", which equates in the machine agents domain with controlled or constrained systems operating based on prescribed behaviours. Another way, which is the one we promote through this position paper, is to expose these children to an active and natural lifestyle where they enjoy the benefits of fresh air. As a consequence, smoking will simply not attract them. Inhaling smoke will be against their nature, without any other need for us to describe what smoking is and why should it be forbidden. In the machine agents domain, this is what we would call Machine Education.

However, in the case of humans, education is possible due to the fact that all its ingredients or enablers are in place. Numerous learning types are available, as well as the bio-psycho-physiological substrate that facilitates this learning, e.g. memory and computation capabilities along with the relevant sensing and embodiment. In the case of machines, these enablers are not yet sufficiently investigated in the context of machine education; therefore several aspects need to be clarified for allowing the very existence of the concept. The main purpose of this position paper is to pin down the broad issues coming along with the machine education concept, and discuss the necessary ingredients, which are, we believe, similar to those required for humans. Thus, the paper will focus (though very briefly) on three major directions which we consider as the main enablers of future machine education research. These directions are:

- **the substrate issue:** what is the substrate that allows learning to happen in machine agents;
- **the learning issue:** what are the learning types available to machine agents;
- **the task issue:** what is the task set and/or structure that agents need to be exposed to and how their performance is to be assessed, so that as a result they acquire the needed systems of values that guide their behaviour, given the existing learning capabilities and the subsequent substrate.

We argue that, if these three are sufficiently investigated, systematic curricula can be envisaged to educate agents towards gaining trust capabilities via implicit environmentally induced systems of values. We also assume that, designing a good education program for a machine, with the purpose of obtaining trust capabilities, works in a context where we consider machine agents that can gradually

accumulate and integrate skills, including the skill of continuous skill maintenance and refinement, i.e. a developmental approach.

Thus, this paper aims to generate the foundation of the machine education concept, and establish the main directions of investigation that enable this concept to contribute towards building trusted interaction within heterogeneous teams (and, ultimately, a society) of machines and humans. We believe that machine education and the associated skill-based developmental view are relevant to multiple applications, such as, but definitely not limited to, surveillance, exploration or disaster recovery, where human-machine teaming is the pertinent mean of achieving present and future goals in a society.

## 2 The substrate issue

Just like in the case of humans, there are certain conditions that facilitate the very existence of potential skill-based trust capable machine agents, and thus, certain substrates that enable machine agents to be subject to machine education. In relation to these substrates we identified two essential aspects that need to be investigated; these are (1) *memory* and (2) *sensory computing*.

The former aspect refers to the intrinsic ability of an agent to accumulate experience, from which it can later learn. When speaking about educating machines towards accumulating and combining skills in relation to a specific task, there is the immediate question of how these skills are represented, stored, and eventually retrieved and used. Therefore, it becomes pertinent to ask what is the best way of implementing the memory effect [11], and clarify what memory is and how it operates. This first aspect is situated at conceptual level, and investigating it may lead to methods or algorithms usable in general, with any kind of agents.

The second aspect narrows the domain to the relevant type of agent (e.g. embodied or disembodied, and their respective sub-categories), where the focus is on particular entities operating in mixed human-machine teams with clearly established purposes. Consequently, the general methods and algorithms previously envisaged need the appropriate technology support to ensure optimal operation. Since the targeted entities are autonomous, they operate based on continuous interaction with their environment; thus, they operate by conveniently storing and processing the perceived sensory information. Thus, it becomes pertinent to investigate the type of sensory inputs agents should handle, and the computational effort needed to do so, in order to find the most appropriate implementation paradigm [23].

We believe that matching the conceptual operation of memory with the computational handling of sensory inputs is the essential enabler of machine agents in the context of machine education, and we reckon that research in this direction should investigate whether the two have a common ground on which the skill-based agents can be built.

## 3 The learning issue

With a substrate in place, the next question is what are the types of learning that are available to machines and can be hosted on this substrate. Machine learning technologies employ a variety of techniques in order to endow artificial entities (machines) with the ability to autonomously learn facts about various phenomena, in order to uncover and understand the inherent structures and causal relations coming along with the sensory perception.

We consider several classes of computational intelligence methods which are of high importance in enabling learning in machine agents: statistical machine learning, non-symbolic learning through neural networks, and evolutionary techniques. The methods in these categories can be used either individually or combined in order to implement agents capable to act autonomously in various environments.

### 3.1 Statistical machine learning

Numerous statistical methods have been used over the years for extracting both structure and causal relations from sensory perception [15]. Statistical analysis can be seen as the mathematical way to capture and disseminate data, with the purpose of defining models for prediction. Rule sets (crisp,

rough or fuzzy), k-means techniques, regression analysis, or decision tree analysis are some of the most popular approaches, which generated a significant support for various methods and algorithms used in general machine learning. Comprehensive reviews of these techniques can be found in [3] in relation to their mathematical foundation, and in [16] in relation to sensory data acquisition.

In machine agents statistical analysis methods are mainly used as part of the inference modules, especially for the data preprocessing stages, but also for tasks such as clustering or rule mining. Inductive Logic Programming, Support Vector Machines, and Bayesian Learning are popular integrated approaches combining various statistical techniques. Thorough reviews of these approaches can be found in [16], [29] and [24].

Methods in the statistical machine learning category benefit from a strong mathematical foundation, through which they can provide well defined and reliable insights into the mechanisms underpinning the sensory data and the problems at hand. However, [4] most of them can infer information mainly from static and well structured data, while they produce less clear results when dealing with highly non-linear data, dynamic environments or multidimensional sources.

### 3.2 Non-symbolic learning - neural networks

Typical artificial neural networks adopt a punctual (mathematical) representation of the neuron, where the focus is primarily on the relation between inputs and outputs and lesson the biological plausibility. They assume that a neuron state, and hence its output, is based on a threshold function which takes as input the weighted sum of the connections incoming to it from other neurons. The pioneering work of McCulloch and Pitts [25] proposed the simplest mathematical model, a binary punctual neuron acting as a basic threshold gate and producing binary output based on the simple weighted sum of the inputs. The simple approach of McCulloch and Pitts further evolved towards the more versatile Perceptron [34], which in turn generated advanced models that included various linear or non-linear threshold functions [33, 18] and non-binary, discrete- or continuous-valued outputs [33].

Numerous reviews from different historical periods, discuss both the single neuron models and the resultant ANNs in relation to the mathematical underpinnings of their operation [41, 26, 18]. The classic connectionist approach on artificial neural networks (ANNs) has its roots in the mathematical neuron models. The resultant ANNs are used in a vast majority for computing in data mining, pattern recognition and in other related fields that are part of the modern computational intelligence research under the umbrella of neural information processing [8]. Consequently, they employ mainly the processing side of the neural network operation. Computation based on punctual neuron models has been employed by all major connectionist approaches, such as Feed-Forward multi-layer networks with back-propagation, Radial Basis Function (RBS) networks [14], Adaptive Resonance Theory (ART) [6, 7], Self Organising Maps (SOM) [17], Hopfield associative networks [13], or the more general concept of deep neural networks (deep learning) [36].

The networks based on neurons with simple threshold gates with binary outputs are considered ANNs of 1st generation (e.g. McCulloch-Pitts, standard Perceptron, Hopfield, Botlzmann machine), while those based on neurons implementing activation functions with continuous outputs (e.g. MLP, RBF etc.) are considered ANNs of the 2nd generation [33].

Learning with neural networks can be either through a supervised learning process, when certain amount prior domain knowledge about the environment and data exists, or through unsupervised learning, when no prior domain knowledge exists. In either cases the networks can be used in conjunction with symbolic production systems, e.g. in the form of sets of rules, statistical learning/analysis or evolutionary computation techniques.

### 3.3 Evolutionary techniques

The motivation for applying evolutionary computation techniques to learning is that they are robust and adaptive search techniques that perform global search in solution spaces. Evolutionary computation techniques can be used either as standalone learning methods, or in conjunction with other machine-learning tools to evolve parameters of these methods in order to improve the quality of overall learning [31, 30]. Evolutionary computation techniques have been found particularly useful in processing of large quantities of raw noisy data, where large numbers of parameters used by various other learning techniques needed to be optimally set in order for those methods to be able to generate meaningful

learning behaviour. In [31] and [30] the authors identify the most used evolutionary techniques for feature selection, classification, clustering and association rule mining, and provide detailed guidance for adapting the design of each component of the evolutionary algorithms (e.g. encoding, genetic operators, selection strategies, objective functions) to the desired learning task. Also in a different study, Ngai and colleagues [32] investigate how various methods based on evolutionary algorithms can be employed to evolve features of neural networks used as primary means in learning contexts.

## 4 The task issue

We have seen in the previous two sections how the concept of machine education can be enabled by the existence of sufficient learning capabilities, together with the substrate (conceptual and technological) on which this learning and its outcomes occur. However, another aspect subsequent to educating machine agents is related to the tasks they need to be exposed to, in order to gain trust capabilities.

The task issue has been traditionally addressed by task and cognitive task analysis in relation to human activity, as discussed in-depth in [20], and by system dynamics in relation to artificial (technical) systems, as described in [12]. However, both directions treat tasks from a top-down perspective, where a complex task of interest is decomposed in simple primitive tasks.

In the case of machine education the interest is in the opposite direction, that is, a developmental endeavour towards the emergence of general intelligence. In this view, through a bottom-up approach, a machine agent or system learns to perform simple tasks by acquiring the relevant skills, and then gradually acquire through systematic training new and more complex (and transferable) skills that, in turn, allow the performance of new and more complex tasks. This aspect has been recently addressed in several studies [38, 39], where the authors noted the need of a task theory in relation to artificial systems, which would reflect the developmental perspective. Thorrisson et al. [39] position the task concept at the core of artificial systems, due to the fact tasks are used for both training and evaluation of these systems, and therefore, we note, tasks are at the foundation of curriculum design. Thus, the need of an unifying framework that would integrate the relevant aspects of intelligent behaviour in relation to assessment and development of skills necessary for exhibiting proficiency in certain contexts of interest. In [38] a framework with 11 design principles is proposed in relation to both the task and the environment in which an artificial agent operates, with the purpose of facilitating a consistent path for development of artificial general intelligence capabilities. Later, in [39] a framework with 6 design principles is proposed, taking into account only the tasks and not the environment.

The task issue have been only recently addressed from a skill-based perspective, under the umbrella of curriculum learning. In [19], the authors describe a designer-based curriculum applied to a skill-based computational Sudoku solver. The solver learns to play Sudoku based on a set of non-symbolic primitive skills (neural networks) trained in isolation to overfit the corresponding primitive tasks. The primitive skills are then aggregated through a symbolic production system to solve Sudoku boards in a cognitively plausible manner. An entirely non-symbolic approach to curriculum learning is presented in [5], where the authors use deep neural networks that learn initially from simple examples and then progress towards learning from more complex ones in order to gradually increase proficiency in a task of interest. The purpose in this study however, is not to advance towards multiple task learning and skill transfer for generalisation, but rather to speed up the learning, based on the assumption that by choosing the order in which examples are presented to the learner, one can guide training towards better and faster learning. In a more recent study [42], the authors refine the approach presented in [5], and use a similar curriculum learning method in conjunction with a Recurrent NN (RNN) with long short-term memory (LTSM) units which provides the memory substrate needed for potential application to skill transfer and generalisation to new tasks. In another recent study [35], the authors propose the so-called "progressive neural networks", claiming advanced ability to learn skills for solving multiple tasks. They note that the proposed progressive network is able to perform skill acquisition and transfer, while avoiding catastrophic forgetting. The proposed networks are tested in multiple reinforcement learning tasks on various 3D maze games, confirming that transfer occurs at both low-level sensory and high-level control layers of the learned policy.

The use of curricula for with a focus on trust is still in its infancy, with very few studies grouped around the concept of adversarial sampling. Curriculum learning for adversarial contexts has been studied mainly with a focus on adversarial alterations of the training data, known as Causative Availability

[40], where alterations can be done based on various sample selection methods. Examples are targeted alterations focusing on particular samples that are of interest for an adversary, or indiscriminate alterations produced with randomly selected samples across the training data set. Regardless of the way the samples to be altered are chosen, the alteration itself plays an important role in learning accuracy and subsequent trustworthiness of the learner agent [22].

## 5 Discussion

Evolution and learning are two of the essential ingredients that facilitated the existence of biological life as we know it today, especially when referring to life forms exhibiting highly intelligent behaviour, such as humans. It is nowadays common understanding that humans are born with particular predispositions due to genetic inheritance, and then learn continuously during their lifetime in order to expand their skills and capabilities. With respect to a particular individual, these are typically known at a philosophical level as nature and nurture, respectively. The former is concerned with the substrate that allows learning, which consists of genetic information, neuro-biological aspects and other innate psycho-physiological features. The latter is concerned with how learning occurs on this substrate, since this learning comes in a variety of forms. Further, the development of the individual receives substantial contribution from both formal education and incidental self-education through continuous socio-environmental exposure.

The broad field of AI envisaged artificial agents (either embodied or not) capable of exhibiting intelligent behaviour similar or above to that seen in humans. However, since Dartmouth manifesto established the field of AI six decades ago [9], various subsequent research fields expended large amounts of resources on attempts to 'build' the intelligence per se, as a designer-based deliverable product, and endow various artificial agents with this 'product' so they perform the required tasks in the desired way. Classic intelligent agents operating under the assumption of rationality relied on utility-related enablers and have been often instantiated via rule, constraint, and other knowledge-based constructs for achieving the relevant behaviours. The very essence of intelligent behaviour, i.e. learning and its subsequent substrate, has been out of the mainstream machine agents research for a long period of time.

In the last few decades though, a significant shift took place towards learning and adaptation as the underpinnings of artificial intelligent behaviour. Recent work [10, 9] thoroughly explains this shift in paradigm, and notes how it becomes clear that today's machine agents rely heavily on adaptation, learning and skills, as opposed to rules, control and knowledge, and are the result of a continuous (self)development via interaction with their environment and peers. In this case, the resultant behaviour becomes questionable from a trust perspective, given that these agents continuously learn from and adapt to their environment rather than being controlled in a programmatic manner.

The first important question arising is: can these agents be constrained to become trustworthy (by rules or other prescriptive means), since they are meant to be adaptive? Or any attempt to control an adaptive entity via constraints is generating a paradox in relation to the very idea of adaptive behaviour, and is therefore ineffective. While not dismissing the former, we do support the latter statement, due a number of issues applicable to constrained trustworthiness, as follows:

1. difficulty to formulate such a constraint when the situation requiring the constraint is complex (designer's point of view);
2. even if a good formulation is found, a complex operational environment may prevent the agent to recognise either or both the situation and the constraint that applies (agent's point of view);
3. even when (1) and (2) are satisfied, the agent may still have difficulties in deciding *how* to apply the constraint in the current context (agent's point of view).

However, if trust capabilities are obtained through education instead of constraints, this leads to a rather soft approach to trust, and to another important question: if hard constraints are not in place, is trustworthiness by education robust against environmental and/or peer pressure? Continuing the naive example of smoking children, this would equate with the following situation: "if you do not smoke you do not belong to our fancy group; therefore, you will be alone". Thus, children may indeed enjoy fresh air, but under peer pressure they may fail to refrain from smoking. This question does not have an immediate answer, and arguably, is the question that defines the proposed research on

machine education as a significant emerging topic in AI and related fields. Indeed, we can argue that, in general, the way to ensure a high likelihood that one delivers the expected outcome, is consistent training following relevant systematic approaches. The better trained one is, the higher the pressure he/she can sustain. How well the training program is designed and applied is the key ingredient to success (along with trainee's quality, of course, which touches the substrate/learning issue). Thus, we can speculate that, strictly related to machine agents, this could be seen as employing a certain mix of machine learning methods with the purpose of over-fitting machine's behaviour (i.e. the relevant skills) to training samples. Yet, substantial investigation needs to be performed in order to make such a claim based on clear scientific grounds.

## 6 The "Machine Education" road-map

In the light of the above discussions, it becomes pertinent to say that the concept of machine education assumes a skill-based view [19] on intelligent machine agents, in which these agents have the ability to acquire new skills, maintain existing skills, and rewire the 'heterarchical' organisation of the skill-set [28, 21], through continuous training by exposure to or interaction with relevant training material. Therefore, machine education is different from machine learning through that it wraps conveniently the variety of machine learning types into meaningful aggregated skill acquisition programs for machines. Since numerous learning types exist in relation to machine agents, the acquisition and/or maintenance of different skills may require different learning types. Also, different learning types may be required by the same skill in different contexts. Thus, machine education is a way of handling in a systematic and consistent manner, either designer-based or through self-education capabilities, the various training programs applicable to each skill, to integration of these skills, and to refinement of this integration, in order to enable agents' success in multiple environments and tasks.

In essence, educational curricula specialised for particular tasks are to be obtained by conveniently aggregating interdependent individual skill training programs, which may be similar in a certain extent to the way education principles operate in case of humans. The skill-based view on agents has been recently receiving increased attention, with numerous studies concentrating on designing architectures for agents with cognitive skill development capabilities. Several discussions relevant to the above can be found in: [21] - on cognitive agent architectures, [19] - on skill acquisition and integration using neural networks, [2] - on foundations of trust and sensor-based autonomy/adaptation.

From a trust perspective, through a curriculum, machine education focuses on enabling machine agents to achieve training-induced value systems [27] that guide them to behave in the intended way and not otherwise. While not entirely excluding hard constraints and prescriptive behaviours, in machine education it should be desired that agents gain trust capabilities purely by extracting the value systems through exposure to external world and learning. In human-machine teaming, trust is concerned not only with gaining an agent's own trustworthy behaviour, but also with the agent gaining trust in behaviours of others in the team [1]. Machine education can contribute to this when the system of values for an agent emerges from observing both the operational environment and peer agents, as summarised in Figure 1. Therefore, an agent can be educated or can self-educate for both sides of trust: become trustworthy, and trust others (humans or machines). In addition, the agent can be re-educated or can self-educate continuously towards expansion of its skills and value system, in order to operate in new teams for accomplishing new purposes. In this manner, the machine education paradigm seeks to create machine learning systems with trust capabilities.

## 7 Conclusions

In this paper we contribute to the concept of trusted autonomy adopting a position in which trust capabilities in human-machine teaming contexts are nurtured in machine agents via education. We postulated that designing a good education program for a machine with the purpose of obtaining trust enabled behaviour, works in a context where machine agents can gradually accumulate and refine skills; that is a skill-based developmental view.

Based on this view we proposed Machine Education, an emerging research topic that envisages machine agents capable to acquire via skills certain systems of values needed for them to exhibit trust capabilities. We then performed a brief analysis of the essential ingredients allowing the very

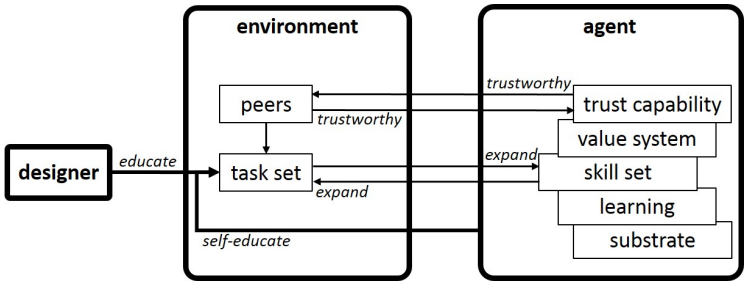


Figure 1: Machine education: the perspective

existence of machine education concept, proposing that future research on this topic should build on a foundation with three pillars: the substrate, the learning and the task.

## Acknowledgements

This work has been funded by the Defence Science and Technology Group (DSTG), Australia, through Competitive Evaluation Research Agreement grant number MyIP7319: *Trust in human-machine teaming through machine education: a skill-based agent design*. The authors acknowledge the valuable feedback of DSTG collaborators: Dr. Peyam Pourbeik, Dr. Rowland Dickinson, Dr. Robert Hunjet and Mr. Thomas Stevens.

## References

[1] H. A. Abbass, G. Leu, and K. Merrick. A review of theoretical and practical challenges of trusted autonomy in big data. *IEEE Access*, 4:2808–2830, 2016.

[2] H. A. Abbass, E. Petraki, K. Merrick, J. Harvey, and M. Barlow. Trusted autonomy and cognitive cyber symbiosis: Open challenges. *Cognitive computation*, 8(3):385–408, 2016.

[3] A. A. Afifi and S. P. Azen. *Statistical analysis: a computer oriented approach*. Academic Press, NY, 1972.

[4] E. Begoli and J. Horey. Design principles for effective knowledge discovery from big data. In *Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), 2012 Joint Working IEEE/IFIP Conference on*, pages 215–218, Aug 2012.

[5] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 41–48, New York, NY, USA, 2009. ACM.

[6] G. A. Carpenter and S. Grossberg. *Adaptive Resonance Theory*, pages 87–90. MIT Press, Cambridge, MA, 2 edition, 2003.

[7] G. A. Carpenter and S. Grossberg. *Adaptive Resonance Theory*. 2009.

[8] A. C. Coolen, R. Kuhn, and P. Sollich. *Theory of neural information processing systems*. Oxford University Press, 2005.

[9] S. L. Epstein. Wanted: Collaborative intelligence. *Artificial Intelligence*, 221:36 – 45, 2015.

[10] T. Froese and T. Ziemke. Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence*, 173(3):466 – 500, 2009.

[11] C. R. Gallistel and P. D. Balsam. Time to rethink the neural mechanisms of learning and memory. *Neurobiology of Learning and Memory*, 108(0):136 – 144, 2014.

[12] C. Gonzalez and V. Dutt. Learning to control a dynamic task: A system dynamics cognitive model of the slope effect. In *Proceedings of The 8th International Conference on Cognitive Modeling*, pages 61 – 66, Oxford, UK, 2007. Taylor & Francis - Psychology Press.

[13] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

[14] R. J. Howlett and L. C. Jain. *Radial basis function networks 2: new advances in design*, volume 67 of *Studies in fuzziness and soft computing*. Springer-Verlag Berlin Heidelberg, 2001.

[15] D. H. Jonassen, K. Beissner, and M. Yacci. *Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge*. Routledge, 2013.

[16] M. Kantardzic. *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, NJ, 2 edition, 2011.

[17] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

[18] E. Kuriscak, P. Marsalek, J. Stroffek, and P. G. Toth. Biological context of hebb learning in artificial neural networks, a review. *Neurocomputing*, 152(0):27 – 35, 2015.

[19] G. Leu and H. Abbass. Computational red teaming in a sudoku solving context: Neural network based skill representation and acquisition. In *Intelligent and Evolutionary Systems*, pages 319–332. Springer, 2016.

[20] G. Leu and H. Abbass. A multi-disciplinary review of knowledge acquisition methods: From human to autonomous eliciting agents. *Knowledge-Based Systems*, 105:1 – 22, 2016.

[21] G. Leu, N. J. Curtis, and H. A. Abbass. Society of mind cognitive agent architecture applied to drivers adapting in a traffic context. *Adaptive Behavior*, 22(2):123–145, 2014.

[22] G. Leu, J. Tang, E. Lakshika, K. Merrick, and M. Barlow. Curriculum optimisation via evolutionary computation, for a neural learner robust to categorical adversarial samples. In *The 4th Asian Conference on Defence Technology*, number accepted, in press. IEEE Explore, 2017.

[23] B. J. MacLennan. Natural computation and non-turing models of computation. *Theoretical Computer Science*, 317(1 - 3):115 – 145, 2004. Super-Recursive Algorithms and Hypercomputation.

[24] S. Marsland. *Machine learning: an algorithmic perspective*. Macine Learning & Pattern Recognition. CRC press, FL, 2 edition, 2014.

[25] W. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[26] T. M. McKenna, J. L. Davis, and S. F. Zornetzer. *Single neuron computation*. Academic Press, 1992.

[27] K. E. Merrick. A comparative study of value systems for self-motivated exploration and learning by robots. *IEEE Transactions on Autonomous Mental Development*, 2(2):119–131, 2010.

[28] M. Minsky. *The society of mind*. Simon and Schuster, New York, 1985.

[29] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[30] A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and C. Coello. Survey of multiobjective evolutionary algorithms for data mining: Part ii. *Evolutionary Computation, IEEE Transactions on*, 18(1):20–35, Feb 2014.

[31] A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and C. Coello Coello. A survey of multiobjective evolutionary algorithms for data mining: Part i. *Evolutionary Computation, IEEE Transactions on*, 18(1):4–19, Feb 2014.

[32] E. W. T. Ngai, L. Xiu, and D. C. K. Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36(2, Part 2):2592 – 2602, 2009.

[33] H. Paugam-Moisy and S. Bohte. Computing with spiking neuron networks. In G. Rozenberg, T. Back, and J. N. Kok, editors, *Handbook of Natural Computing*, pages 335–376. Springer Berlin Heidelberg, 2012.

[34] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[35] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell. Progressive neural networks. *CoRR*, abs/1606.04671, 2016.

[36] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61(0):85 – 117, 2015.

[37] G. Taylor, R. Mittu, C. Sibley, and J. Coyne. *Towards Modeling the Behavior of Autonomous Systems and Humans for Trusted Operations*, pages 11–31. Springer US, Boston, MA, 2016.

[38] K. R. Thórisson, J. Bieger, S. Schiffel, and D. Garrett. *Towards Flexible Task Environments for Comprehensive Evaluation of Artificial Intelligent Systems and Automatic Learners*, pages 187–196. Springer International Publishing, Cham, 2015.

[39] K. R. Thórisson, J. Bieger, T. Thorarensen, J. S. Sigurðardóttir, and B. R. Steunebrink. *Why Artificial Intelligence Needs a Task Theory*, pages 118–128. Springer International Publishing, Cham, 2016.

[40] S. L. Wang, K. Shafi, C. Lokan, and H. A. Abbass. Adversarial learning: the impact of statistical sample selection techniques on neural ensembles. *Evolving Systems*, 1(3):181–197, Oct 2010.

[41] B. Widrow and M. Lehr. 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. *Proceedings of the IEEE*, 78(9):1415–1442, Sep 1990.

[42] W. Zaremba and I. Sutskever. Learning to execute. *arXiv:1410.4615v3*, 2015.