Certification of On-Line Learning Neural Networks

James T Smith
Institute for Scientific Research, Inc.
jsmith@isr.us

ABSTRACT

This paper presents the results of research related to the verification and validation (V&V) of on-line learning neural networks (OLNN) that continually adapt and evolve. Such is the situation with OLNN-based air-flight controllers developed at the Institute for Scientific Research, Inc. (ISR).

For such systems to be employed in commercial aircraft, a certification process is required that addresses the capabilities and risks associated with a system that is continually adapting and evolving. This research is addressed by the NASA funded project *"Development of Methodologies for Independent Verification and Validation of Neural Networks,"* Research Grant NAG5-12069.

ISR has devoted much effort to understanding the complexity of this problem and to the development of approaches to address this V&V requirement.[1] Because of adaptation with use, an OLNN cannot be pre-certified at release time, simply based on the analysis of initial training sets.[2]

The findings and approaches presented here are based upon an analysis of the foundation principles and techniques that underpin the pilot certification process by which pilots are deemed sufficiently prepared to operate an aircraft. Such areas as human factors analysis and accident analysis have been extrapolated to identify corresponding problems and opportunities in addressing the certification of an OLNN-based system.

1.0 INTRODUCTION

Humans and neural networks have much in common, both functionally and architecturally. The historical development and success of artificial neural network systems has depended heavily upon efforts to model and mimic at the biological level the neural systems of living things. This study attempts to incorporate higher-level psychological considerations. The certification of humans provides a framework for understanding additional considerations as to how neural network systems could and should be verified and validated.

The defining characteristics of complex systems such being considered here include the attributes: (1) adaptive, (2) autonomous, and (3) non-deterministic. Formal definitions of each are:

- *Adaptive* – the capacity or suitability for, or the tendency toward change, modification, etc.
- *Autonomous* – being free from external control and constraint in action and judgment, independent in mind or judgment, self-directed.
- *Non-deterministic* – the property that a computation or execution may yield multiple plausible results.

Complex systems are developed with these attributes because of the potential benefits and functionalities enabled. However, such attributes also represent a system engineer's conundrum. They inherently introduce a risk of uncertainty and failure leading to considerations such as:

- *Fault-tolerance* – ability to detect and recover from failure.
- *Graceful degradation* – ability to perform less than optimally, rather than cessation of that function.

2.0 THE HUMAN-IN-THE-LOOP

The introduction of the human-in-the-loop not only imbues a system with the aforementioned attributes, but the human also contributes to the mechanisms of fault-tolerance and graceful degradation. However, the human-in-the-loop introduces yet another class of problem, specifically human error—humans make mistakes.

Various studies have implicated human error in a variety of occupational accidents.[3] With safety-critical system, human lives may be at risk. The prevention of such undesirable outcomes, becomes a top priority that may necessitate less desirable

solutions and performance levels. To address this problem, special tools, techniques, and procedures are available: (1) the pilot certification process; (2) human factors analysis (HFA); (3) accident analysis; and (4) human error theory.

2.1 Pilot Certification

The purpose of pilot certification stated by the Federal Aviation Administration (FAA) and the Department of Transportation[4] is "… to enhance the ability of pilots to meet the evolving demands of the National Airspace System and operate safely and effectively in this environment." This process is to verify and validate that the pilot (the human-in-the-loop) meets his system design requirements.

The pilot certification process was developed and has been analyzed from various perspectives. In this study, the pilot certification process is viewed in terms of how it addresses *performance*—normal or standard operation of the aircraft, and *safety*—continued fault-tolerant, possibly degraded, operation under abnormal conditions.

2.2 Human Factors Analysis

The human-in-the-loop introduces a significant limitation to how other system components can be designed to address systems and operational requirements. The systems engineer must work within the physical, physiological, mental, and psychological limitations of the human being. The minimax challenge—to maximize system performance yet minimize the possibility of failure—motivates such disciplines as human factors analysis and accident analysis—the study of the circumstance and causality of accidents.

Simply stated, *human factors analysis* (HFA) is the study of people in their working and living environments.[5] HFA is concerned with relationships between people and machines, people and environments, and people and other people. HFA seeks to optimize the effectiveness of the system with respect to safety and efficiency, and to optimize the well being of the individual.

2.3 Accident Analysis

Comprehensive theories have been developed to characterize system accidents and failures, especially those that are human-related. Safety-critical system domains, such as aeronautics, have provided fertile areas to study. The aerospace industry was the originator and has been the largest contributor and beneficiary.[6]

Accidents generally are complicated events that result from a myriad of interrelated causes and circumstances—called *failure events*, the last of which culminates in the failure.[7] These events may be *active failures*—those actions or inactions that ultimately cause an accident, or *latent failures*—other circumstantial errors that affect the sequence of events that characterize an accident

While a mesh of active and latent failures may be adequate to describe the chronology of an accident, it does not adequately capture deeper relationships among those failure events. Previously mentioned system attributes (adaptive, autonomous, and non-deterministic), together with the innate capabilities and error proneness of the human must be juxtaposed with an accident's chronology mesh of active and latent failures to explain its cause(s) and to develop reasonable preemption or remediation approaches for that accident type.

2.4 Human Error Theory

In support of this juxtaposition, accident analysis practitioners and human factors analysts have developed comprehensive frameworks of human error that organize and explain an accident's mesh of failure events from a human perspective.[8]

One early approach is Frank Bird's *Domino Theory,* which promoted the idea that, like dominos stacked in sequence, mishaps are the end result of a series of errors made throughout the chain of command[9]. James Reason extended this theory to his *Swiss Cheese* model that identified a taxonomy of multiple levels at which active and latent failures could occur and interact.[10]

The United States military has developed a comprehensive framework, the *Human Factors Analysis and Classification System* (HFACS), from which to identify and analyze aeronautical accidents.[11] The HFACS framework has since been applied to commercial aviation, as well as in other safety critical problem domains, such as the nuclear and medical industries.[12]

3.0 HFACS—HUMAN FACTORS ANALYSIS AND CLASSIFICATION SYSTEM

The HFACS treats individual operators as an elements in a larger safety critical system. It analyzes error events by considering relationships between elements in the system. The HFACS describes a taxonomy consisting of four first-level tiers of failure, namely: (1) Unsafe Acts; (2) Preconditions for Unsafe Acts; (3) Unsafe Supervision; and (4) Organizational Influences. This paper considers only the first tier.

The project has analyzed all four tiers. Two major types of relationships were considered: (1) those related to the taxonomy of actual unsafe acts, and (2) the causal-effect relationships among the various tiers of the HFACS framework. This paper focuses on the first type.

While the HFACS focuses on the roles of humans in the causation of accidents and failures, many insights gleaned from the HFACS are applicable to other system components. Other researchers have contributed extensions to the HFACS as they adapted the HFACS to their particular analysis requirements.[13]

Tier 1 Unsafe acts are operator actions or inactions that occur immediate to, and often trigger, an adverse event, previously termed an active failure. Unsafe acts are further classified into two categories, violations and errors. The differentiation between violations and errors is based on whether the action is currently considered acceptable, or legal, behavior.

- *Violations*, in contrast to errors, are willful deviations of accepted regulations, whether or not they actually result in failures. Violations are further divided into the following sub-types:
  - *Routine* violations are part of a behavior pattern. They are known to be unsafe acts that often do not result in immediate failure.
  - *Exceptional* violations are not typical of an individual nor condoned by management. These isolated offenses may or may not involve malice, the intention to cause harm or failure.
- *Errors* are legal mental and physical activities that, nevertheless, fail to achieve their intended outcome. Errors are further decomposed into the following sub-types:
  - *Skill-based* errors occur during execution of a familiar procedure that require little or no conscious thought.
  - *Perceptual* errors are misinterpretations of what is seen, heard, or received through the senses.
  - *Decision* errors represent conscious, goal-intended behavior that proceeds as designed, yet proves inadequate or inappropriate for that situation.

4.0 HFACS ANALYZED & APPLIED TO OLNN SYSTEMS

An analysis of the HFACS framework provides a sound scientific basis from which to approach the enhancement of other non-human components of complex systems that exhibit the same human-like behaviors (adaptive, autonomous, and non-deterministic) that have been previously discussed. In this section, Tier 1 of the HFACS framework is examined, applied to understanding the pilot certification process, and ultimately is extrapolated to provide new insight and guidance into how the V&V of neural networks could be improved.

4.1 Neural Network Violations

Violations are determined by rules and regulations, which are externally imposed constraints, independent of what a neural network system is capable of learning. For some problems, OLNN systems may be unencumbered by external rules and regulations. An example of this situation would be the application of OLNN technology to general data mining tasks where supposedly *a priori* illegal patterns do not exist.

In the case of the OLNN developed for the Intelligent Flight Control System (IFCS) project at ISR, external rules and regulations do exist. Two major issues with respect to Neural networks are: (1) the representation of such regulations so that the NN may operate within them, and (2) under what circumstances could, or should, an NN nevertheless violate them.

4.1.1 Neural Network Routine Violations

Technically, an OLNN, unconstrained by rules and regulations, will learn to fly the aircraft in otherwise unsafe ways. An explanation of how this can occur is quite simple. The components of most systems typically are over-specified, over-designed, and over-engineered rather than merely being adequate to meet specifications. Otherwise, systems may become unacceptably brittle at their specification boundaries. Under normal circumstances, the actual system should be able to perform better than the specified system.

An OLNN that is unconstrained in its learning to improve the actual system's performance may push the boundaries of that system, exceeding

capabilities of the specified system. At some point, the OLNN could be flying the aircraft in a manner that would be in violation of the aircraft's specified capabilities. Thus, an OLNN can learn to routinely commit a violation.

This introduces an interesting dichotomy regarding the OLNN embedded in the IFCS. This system is expected to control the aircraft properly under normal conditions. It also is expected to adapt and learn to control the aircraft's flight under abnormal circumstances, even to the point of possible failure of various aircraft components and subsystems. Thus, the OLNN is learning to fly the aircraft routinely in an otherwise unsafe mode. A policy issue is thus raised: should the OLNN be allowed to perform in what are otherwise considered unsafe circumstances?

4.1.2 Neural Network Exceptional Violations

Consider the situation of a human who breaks rules for a higher purpose. In such a circumstance, the human may be operating in an abnormal situation that is not adequately addressed by the current rules. Perhaps a *rule taxonomy* is required that differentiates what is conditionally or arbitrarily illegal—reflecting current technical or management limitations—versus what is absolutely illegal, reflecting violations of the laws of science.

Sometimes, two wrongs do make a right, at least in that the latter somehow compensates for the former in a fault-tolerant, error-correcting sense. From this perspective, the latter action is illegal unless it is the only means of correcting a prior error that could lead to worse consequences.

To the extent that some abnormal situations are more likely to occur and are of higher risk than others, *a priori* preemptive analysis, preparation, training, etc. can result in that situation being normal. This is an example of *risk mitigation*. Similarly, the OLNN of the intelligent flight control system is expected to learn to perform under such known-to-be unsafe conditions.

The handling of rules and regulations by an OLNN might be addressed in several ways, each with its own issues. They may be represented within the OLNN, so that the OLNN is *self-regulating*. They may be captured internally; so that the V&V process must access correctness of this embedded rule-regulation set. They may be applied by monitoring the learning and functioning of the OLNN, thereby enabling the anticipation of a violation and the determination possible recovery.

4.2 Neural Network Errors

In addition to violations, the HFACS framework identified three general types of errors, those legal activities that fail to achieve their intended outcome: (1) skill-based, (2) perceptual, and (3) decision. The assessment of an NN must consider all three, because they are interrelated.

4.2.1 Neural Network Skill-Based Errors

Generally, a neural network is trained to perform skill-based functions that involve execution of a familiar procedures that normally require little or no conscious thought. Some skill-based functions may be quite complex, involving multiple skills.

The human becomes proficient at skill-based functions through repetitive practice of the skill-based task until conscious thought is not required. Neural networks also require appropriate training and evaluation for the skill-based task to be performed. Training sets must be sufficiently encompassing, including abnormal situations of the operation space where the skill is to be used.

From the human's perspective, skill-based errors generally result from a lapse in memory, such as forgetting, or otherwise omitting a step due to loss of focus or attention, or a distraction. Similar conditions may exist when the neural network is performing skill-based tasks, but the manifestations and consequences are different.

Distracting situations generally occur due to unexpected or unanticipated events. *Distractions* are events that could interfere, if noticed, with performing the task at hand, but for which ignoring them poses no undesirable consequences. Both the human and the neural network may be trained to recognize an oft-occurring distraction and so to dismiss it, as an unconscious skill. The solution to the distraction problem for a neural network may seem apparent. However, complications may arise that carry the risk of introducing another problem. There is the risk that a future occurrence is ignored by habit that in fact should not have been ignored.

In the early 1980's, a flight crew was practicing landing a C-5 super cargo military aircraft. The crew went through all steps leading up to landing the aircraft, except for actually landing on the runway. Before contact with the runway, the crew would pull up to practice another pass.[14] Since they did not plan actually to land, they did not activate the landing gear; a violation that generated a warning alert, which became a distraction. So, they disabled the alert, another violation.

The magnitude of their cumulative failure events became apparent only when they finally did land the aircraft. They had not re-armed the alert, as there was no procedure for re-arming a supposed always-on alert. They also failed to activate the landing gear, doing just as they had practiced. The consequence of this series of errors and violations was the crew landed the aircraft on its belly!

With humans, one solution to unintentional conditioning involves bringing the otherwise automated event to the conscious level for confirmation that it indeed can be ignored. One tool commonly used for this purpose is the *checklist*. Critical steps and milestones are explicitly called out for conscious note.

The general approach of explicit subtask decomposition, recognition, and conscious-level checklists presupposes that the skill-based task implemented by the neural network does lend itself to such decomposition. The checklist manager must be able to recognize when the neural network has achieved its subtasks. This can support the black-box V&V of neural networks that lack such internal monitoring and reporting capabilities.

The checklist serves several purposes. During training, self-feedback that the total task is being learned correctly is provided. During normal execution of a skill-based task, the checklist confirms specific subtasks are addressed correctly. In particular, the concept of the checklist can support real-time V&V procedures.

This explicit elevation of the skill-based task to the conscious level provides an opportunity for real-time re-evaluation of how well a task is proceeding and if it should be modified, aborted, etc. A recorded checklist also provides an audit trail to support *ex post facto* analysis.

4.2.2 Neural Network Perceptual Errors

Perceptual errors are misinterpretations of what is received via the senses. They generally occur when sensory input is degraded, or when actual input is correct, but is misinterpreted by the perceiver.

As previously noted, distractions are events that may interfere with performing a task, but for which otherwise ignoring them poses no undesirable consequences. For practical purposes, distractions may be treated as system noise. On the other hand, not all confusing situations are so benign.

In the case of the neural network, a strategy exists that is better than training to ignore distractions. The neural network may be trained not only to respond correctly to events involving distractions, but also to report, as a status output, detection of ignored distractions. From a V&V perspective, knowledge of what is ignored can be as important as what it recognized. Status recognition may be based on information already present in the neural network, or it could require additional inputs.

Status recognition of distractions also represents the beginning develop of *self-awareness*, since the neural network is aware of where it is procedurally. This information could be useful in comparing what the neural network perceives itself to be doing with what the outside world perceives.

Events incorrectly treated as distractions by a neural network could have serious consequences. The problem is to determine whether the event is indeed a distraction or a significant event. This leads to the discussion of confusing situations.

A *confusing situation* generally is due to mixed signals including inconsistent inputs, conflicting requirements, and complex events. Two general cases may be considered: (1) misinterpretation due to incorrect processing of correct information, and (2) misinterpretation due to incorrect inputs.

Confusion can be due to a lack of experience, where a system has not been exposed to the given confusing combination. Neural network training and sets may include such confusing inputs. In fact, *over-training* might be a consideration with exposure to situations beyond what is considered likely or even realizable in a real-life setting. This represents a form of stress training and testing.

4.2.3 Neural Network Decision Errors

Decision errors represent conscious, goal-intended behavior that proceeds as designed, yet proves inadequate or inappropriate for that situation. They tend to occur when a familiar situation is not recognized, is misdiagnosed, or when an unfamiliar situation occurs and generally result in the application of an unsuccessful procedure.

The more complicated scenario involves confusing situations in which the total set of inputs, while correctly received, form an inconsistent or conflicting view. The tasks may be performable if taken individually, and the information sources may be plausible if taken individually; however, taken together they are neither performable nor plausible. The set of possible themes may be extendible to multiple plausible interpretations.

Several general methods or approaches have been developed to attack this problem. This scenario is an example of the *data fusion*—the seamless integration of data from disparate sources. On the

other hand, identification of those sub-themes is the domain of methods such as *data mining*, which analyzes data for trends or anomalies without knowledge of the meaning of the data. The data fusion process could be viewed as *looking at trees and seeing a forest (an ecosystem);* while the data mining process could be viewed as *looking for a needle, the correct needle, in a haystack.*

Neural networks have been employed extensively for both data fusion and data mining applications. At issue is how to certify that neural networks have appropriate background knowledge. In a neural network, knowledge exists in a compiled form embedded in its weights, links, and structure. The neural network's processing must be taken as a whole. It does not yield partial results or lend itself to any obvious explanation mechanism.

Various efforts have been explored to capture the underlying knowledge in some human-readable, recognizable, and understandable format. These efforts include methods, such as rule extraction and decision-tree extraction, which lend themselves to visualization methods that assist the human in identifying potential relationships between the nuggets of knowledge from rules and decision tree branches.[15] This knowledge must correctly generalize to other situations. Furthermore, new knowledge, not previously stated in human understandable terms, might be gleaned from such an effort.

5.0 CONCLUSION

An analysis of the pilot certification process has yielded new mechanisms and paradigms for the verification and validation of an OLNN-based system. The OLNN is found to share many characteristics with human-in-the-loop systems.

The historical development of artificial neural network systems that has depended heavily upon efforts to model and mimic at the biological level the neural systems of living things can be extended to incorporate higher-level psychological considerations of human intelligence.

This paper has considered how an OLNN may manifest many human-like behaviors such as susceptibility to distraction, and confusion when there exist multiple plausible solutions. Fortunately, many of the mechanisms humans use to address such short-comings suggest analogous mechanisms for the design, training, operation, and certification of an OLNN-based system.

---

[1] Institute for Scientific Research, Inc. (ISR). 2000. *Software Verification and Validation Plan for the Airborne Research Test System II Intelligent Flight Control Program.* IFC-SVVP-F001-UNCLASS-120100.

[2] Mili, A., 2003. B. Cukic, Y. Liu, and R. Ben Ayed. *Towards the Verification and Validation of On-Line Learning Adaptive Systems*, In Computational Methods in Software Engineering. T. Khoshghoftaar, editor. Kluwer Scientific Publishing, 2003.

[3] O'Hare, D., M. Wiggins, R. Batt, and D. Morrison. 1994. Cognitive failure analysis for aircraft accident investigation. *Ergonomics*, 37, 1855-69.

[4] Department of Transportation (DOT). 1997. Pilot, Flight Instructor, Ground Instructor, and Pilot School Certification Rules. Federal Aviation Administration. RIN 2120-AE71.

[5] Southeastern Oklahoma State University (SOSU). Aviation Sciences Institute. 2002. Course notes for AVIA 4643 - Physiology: Human Factors & The SHEL Model. Fundamentals of Safety Engineering and Human Factors.

[6] Ford, C., T. Jack, V. Crisp, and R. Sandusky. 1999. Aviation accident causal analysis. Advances in Aviation Safety Conference Proceedings, (P-343). Warrendale, PA: Society of Automotive Engineers Inc.

[7] Wiegmann, D., and S. Shappell. 1999. Human error and crew resource management failures in Naval aviation mishaps: A review of U.S. Naval Safety Center data, 1990-96. *Aviation, Space, and Environmental Medicine,* 70: 1147-51.

[8] Geller, E. 2000. Behavioral safety analysis: A necessary precursor to corrective action. *Professional Safety*, 29-32.

[9] Bird, Jr. Frank E. 1974. *Management Guide to Loss Control.* Atlanta: Institute Press.

[10] Reason, James. 1990. *Human Error*. New York: Cambridge University Press.

[11] Naval Safety Center. 1996. Quality Management Board Charter - Reducing Human Error in Naval Air Operations.

[12] Gutierrez, Maria. 2002. HFACS Reports, Quarterly Newsletter of Marine Facility Incidents using The Human Factors Analysis & Classification System.

[13] ibid.

[14] Smith, James. Personal account.

[15] Boz, Olcay. 2002. Extracting Decision Trees From Trained Neural Networks. Paper presented at the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 2002.