An Overview of Strategies for Neurosymbolic Integration

Melanie Hilario
CUI - University of Geneva
24 rue Général-Dufour
CH-1211 Geneva 4
hilario@cui.unige.ch
Switzerland

Abstract

At the crossroads of symbolic and neural processing, researchers have been actively investigating the synergies that might be obtained from combining the strengths of these two paradigms. Neurosymbolic integration comes in two flavors: unifed and hybrid. Unified approaches strive to attain full symbol-processing functionalities using neural techniques alone while hybrid approaches blend symbolic reasoning and representational models with neural networks. This papers attempts to clarify and compare the objectives, mechanisms, variants and underlying assumptions of these major integration approaches.

1 Introduction

Throughout its brief history, the field of artificial intelligence (AI) has been the arena of jousts between two *frères ennemis*, symbolicism and connectionism. No sooner had connectionism recovered from [Minsky and Papert, 1969]'s devastating blows than Fodor and Pylyshyn charged to the fore in the name of symbolic AI. They argued that connectionism cannot be a valid theory of cognition, since it fails to account for the combinatorial syntactic and semantic structure of mental representations: at best, connectionism is just another implementation technology, an alternative means of implementing classical symbolic structures and processes [Fodor and Pylyshyn, 1988]. This implementationalist viewpoint has since been the traditional defense of symbolic AI against connectionism's cognitive claims. At the other extreme, according to [Pinker and Prince, 1988]'s classification, eliminativism rejects the symbol level as a valid level of description of cognitive phenomena: symbolic theories are no more than crude approximations of what really takes place in the brain and must give way to connectionist or neural theories.

Between these two radical stances, a number of more subtle philosophies have emerged at the interface of connectionist and symbolic AI. Their origins have been inextricably linked with the proliferation of attempts at integrating neural and symbolic processing. This paper will give an overview of the various approaches to neurosymbolic integration. Roughly, these can be divided into two strategies: *unified* strategies aim at combining neural and symbolic capabilities using neural networks alone, while *hybrid* strategies combine neural networks with symbolic models like expert systems and decision trees. These two approaches are depicted as the main subtrees of classification hierarchy in Figure 1; they are discussed in detail in the next two sections.

2 Unified strategies

Unified strategies are premised on the claim that there is no need for symbolic structures and processes as such: full symbol processing functionalities emerge from neural structures and processes. Two trends can be distinguished among unified strategies: neuronal symbol processing and neural (or connectionist) symbol processing. This distinction is based on a terminological convention adopted in [Reeke and Edelman, 1988], where the term *neuronal* implies a close identification with the properties of actual (biological) neurons and the term *neural* implies only a general similarity to actual neurons.

Neuronal symbol processing (NSP) is a special case of the neuronal approach, a broader research strategy which claims to ground all cognitive processing in biological reality. NSP's specific objective is to model the brain's high-level functions. The neuronal approach is a bottom-up approach: its mandatory starting point is the biological neuron. Perhaps the most brilliant example of the neuronal approach is the theory of neuronal group selection (TNGS), better known as neural darwinism [Edelman, 1992]. Built on three fundamental tenets—developmental selection, experiential selection and reentrant mapping—this theory attempts to provide a biological account of the full range of cognitive phenomena, from sensorimotor responses all the

This work was partially supported by the EU and the Swiss OFES under Esprit Basic Research Project MIX.

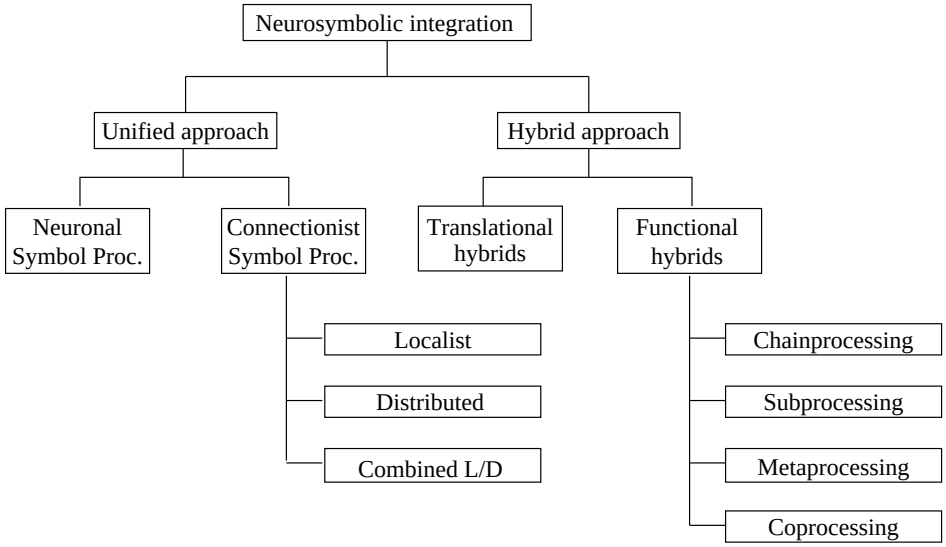

Figure 1: Classification of integrated neurosymbolic systems

way up to concept formation, language, and higher-order consciousness. The consistency of the TNGS has been demonstrated in a series of automata which avoid the preestablished categories and programming of standard AI. Constructed as networks of neuronlike units undergoing a process of natural selection, these automata carry out categorization and association tasks in a dynamic environment. In Darwin III, for example, recognition and categorization networks are combined with motor circuits and effectors that act upon the environment. Objects are categorized on the basis of internal values like "light is better than no light"; the result of the automaton's neuronal activity becomes apparent as motor responses to categorized objects. The processes demonstrated in these automata—perceptual categorization, memory and learning—are precisely the fundamental triad of higher-order brain functions, according to the TNGS. However, neuronal symbol processing remains to be demonstrated in the Darwin or its descendant series. Neuronal symbol processing may yet be the ultimate proof-of-concept of the neuronal approach; however, it may take some time before it can even envisage real-world applications.

Connectionist symbol processing (CSP) or neural symbol processing lays no claim to neurobiological plausibility: the neuron in question here is generally a formal neuron. Artificial neural networks are used as building blocks to create a cognitive architecture capable of complex symbol processing. Typically, model construction starts with an idea of some high-level symbolic function to be performed and proceeds with the design of the appropriate connectionist infrastructure. In this sense, the neural approach can be thought of as a top-down strategy, despite the opposite thrust of its claim that complex symbolic functions emerge from neural structures and processes. However, CSP is not inherently top-down: in principle, nothing precludes it from actually starting out with neural networks from which non-predetermined symbolic structures and processes can emerge in unforeseen ways.

Historically, Fodor and Pylyshyn's critique has been a significant if negative driving force behind CSP: one of its persistent motivations has been to show that neural networks exhibit a combinatorial constituent structure—precisely what Fodor and Pylyshyn declared wanting in connectionist architectures. For instance, BoltzCONS is a connectionist model that dynamically creates and manipulates linked lists; according to its author, its aim is not to show that neural networks can implement complex symbol structures, but rather to show how neural networks can exhibit compositionality and distal access, two distinguishing properties of high-level symbol processing [Touretzky, 1990].

From the point of view of the underlying representation scheme, CSP architectures can be localist, distributed, and combined localist/distributed. In localist architectures, each node in a neural network represents a concept [Shastri, 1988; Feldman and Ballard, 1982]. In distributed architectures like DCPS [Touretzky and Hinton, 1988], the most elementary concepts emerge from the interaction of several different nodes. Finally, combined local/distributed architectures couple these two representations [Sun, 1991]. From the point of view of system tasks, the CSP approach has been actively investigated in the areas of logic and inferencing [Hölldobler and Kurfess, 1991], natural language understanding [Bookman, 1987; Dyer, 1991] and connectionist expert systems [Gallant, 1988].

3 Hybrid strategies

The hybrid approach rests on the assumption that only the synergistic combination of neural and symbolic models can attain the full range of cognitive and computational powers. Hybrid neurosymbolic models can be either translational or functional hybrids.

Translational hybrids can be viewed as an intermediate class between unified models and functional hybrids. Like unified models, they rely only on neural networks as processors, but they can start from or end with symbolic structures. Most often, the symbolic structures used are rules [Kuncicky *et al.*, 1992; Fu and Fu, 1990], though attempts have also been made to extract schemas from neural networks [Crucianu and Memmi, 1992]. However, symbolic structures are not processed as such in translational models; for instance, rules are not applied by an inference engine but only serve as source or target representations of network input or output data. They can thus be considered semi-hybrid systems in the sense that they use symbolic structures without the corresponding symbolic processors. Typically, translational models compile symbolic structures into neural networks before processing, or extract symbolic structures from neural networks after processing. Often, compilation into neural networks is followed by refinement of preexisting symbolic knowledge via connectionist learning, then by extraction of symbolic structures in view of communicating the knowledge thus refined to other systems (either humans or symbol-processing systems) [Towell, 1992]. Translational hybrids have also been called transformational models [Medsker, 1994].

Functional hybrids incorporate *complete* symbolic and connectionist components: in addition to neural networks, they comprise both symbolic structures and their corresponding processors—e.g., rule interpreters, parsers, case-based reasoners and theorem provers. Functional hybrids are so-called because, contrary to translational hybrids, they achieve effective functional interaction and synergy among the combined components. However, since translational hybrids can be viewed as a degenerate case of functional hybrids, we shall be using the term 'hybrid' to designate a complete or functional hybrid, unless indicated otherwise.

Hybrid systems can be distinguished along different dimensions such as their target problem or task domain, the symbolic (e.g., rule-based reasoning, case-based reasoning) and neural (e.g., multilayer perceptrons, Kohonen networks) models used, or the role played by the neural (N) and symbolic (S) components in relation to each other and to the overall system. Though such dimensions allow for more or less clear distinctions between individual systems, they have little bearing on the central issues of neurosymbolic integration. We therefore propose a taxonomy of hybrid systems based on the mode and level of integration of the N and S components.

We distinguish two **integration levels**—loose and tight coupling. In *loosely coupled* systems, interaction between the two components is clearly localized in space and time: control and data can be transferred directly between N and S components (e.g., by function or procedure calls), or via some intermediate structure (e.g., domain or control blackboards accessible to both components) or agent (e.g. a supervisor), but interaction is always explicitly initiated by one of the components or by an external agent. In *tightly coupled* systems, knowledge and data are not only transferred, they can be shared by the N and S components via common *internal* structures. Thus a change in one of the components which affects these common internal structures has immediate repercussions on the other component without need for explicit interaction initiatives. Within this category, too, coupling is not uniformly tight from one system to another: whereas the shared structures are often simple links or pointers between the N and S components as in SYNHESYS [Giacometti, 1992], they can be significantly more important in number and function (e.g., nodes shared by a semantic marker-passing network and a distributed neural network, as in [Hendler, 1989]).

The **integration mode** or scheme refers to the way in which the neural and symbolic components are configured in relation to each other and to the overall system. Four integration schemes have been identified: chainprocessing, subprocessing, metaprocessing and coprocessing. To define them, we suppose a system comprising one neural and one symbolic module, with the understanding that for more complex systems, there can be as many integration schemes as pairs of neural and symbolic components.

In *chainprocessing*, one of the (N or S) components is the main processor whereas the other takes charge of pre and/or postprocessing tasks. In [Hayes *et al.*, 1992], for instance, a neural network preprocesses data from a respiratory monitor to determine qualitative states which are then fed as facts into a classical expert system. Conversely, a neural network can be assisted by a symbolic preprocessor; e.g., a decision tree generator selects significant features to be input into a backpropagation network, thus reducing learning and processing time [Piramuthu and Shaw, 1994]. Another example is a system which uses a Hopfield network to solve a wastewater treatment optimization problem: to accelerate convergence, a relevant solution is retrieved by a case-based reasoner and used to initialize the Hopfield net instead of a randomly generated state [Krovvidy and Wee, 1992].

In *subprocessing*, one of the two components is embedded in and subordinated to the other, which acts as the main problem solver. Typically, the S component is the main processor and the N component the subprocessor. It is an open question whether the reverse setup is at all possible. An example of neural subprocessing is INNATE/QUALMS: the main processor, a fault diagnosis expert system, calls on a set of multilayered perceptrons to generate a candidate fault, then either confirms their diagnosis or offers an alternative solution

	Loose coupling	Tight coupling
Chainprocessing	[Hayes *et al.*, 1992] [Piramuthu and Shaw, 1994] WATTS [Krovvidy and Wee, 1992]	
Subprocessing	INNATE/QUALMS [Becraft *et al.*, 1991]	[Hendler, 1989]
Metaprocessing	RSA^2 [Handelman *et al.*, 1989] [Gutknecht and Pfeifer, 1990]	
Coprocessing	DDT [Gutknecht *et al.*, 1991] [Jabri *et al.*, 1992]	SYNHESYS [Giacometti, 1992]

Table 1: Classification dimensions and instances of of hybrid NS systems

[Becraft *et al.*, 1991].

In *metaprocessing*, one component is the baselevel problem solver and the other plays a metalevel role (such as monitoring, control, or performance improvement). Symbolic metaprocessing is illustrated in the Robotic Skill Acquisition Architecture [Handelman *et al.*, 1989], where a rule-based system supervises learning and performance in a baselevel hybrid composed of neural networks and rules. A case of neural metaprocessing is a system in which a rulebase solves physics problems, guided by a backpropagation network which chooses the next unknown variable to solve for [Gutknecht and Pfeifer, 1990].

In *coprocessing*, the N and S components are equal partners in the problem-solving process: each can interact directly with environment, each can transmit information to and receive information from the other. They can compete under the supervision of a metaprocessor, or they can cooperate in various ways, e.g., by performing different subtasks, or by doing the same task in different ways and/or under different conditions. An example of cooperative neurosymbolic coprocessing by execution of specialized subtasks is a system where a decision tree (timing classifier) and a neural network (morphology classifier) work together to detect arrythmia in heart patients [Jabri *et al.*, 1992]. In SYNHESYS [Giacometti, 1992], on the contrary, the same diagnostic task is executed by a rule-based system and a prototype-based neural network that learns incrementally; if the neural component comes up with a diagnostic, this output is validated by the rulebase in backward chaining mode; otherwise, the rulebase is activated in forward chaining mode and its diagnostic is used to train the neural network.

Table 1 situates these representative hybrid systems along the two classification dimensions.

4 The big picture

CONNECTIONISM	NEUROSYMBOLIC INTEGRATION				SYMBOLICISM
	Unified approaches		Hybrid approaches		
	Neuronal Symbol Proc.	Connectionist Symbol Proc.	Functional hybrids	Translational hybrids	
Segregation	*Neuronal eliminativism*	*Connectionist eliminativism* *Limitivism* *Revisionism*	*Hybridization or cohabitation*		*Segregation* *Implementation–alism*

Table 2: Synoptic view of neural, symbolic and neurosymbolic approaches

To sum up this overview of approaches to neurosymbolic integration, we will relate the different computational strategies to cognitive stances in the symbolic/connectionist debate (see Table 2).

First of all, both unified approaches can be eliminativist, but in different ways. To clarify this, we adopt [Smolensky, 1988]'s distinction between neural (read *neuronal* for consistency with our terminology) and subsymbolic (or connectionist) models. Following this distinction, eliminativism—which denies symbolic models any scientific standing—comes into two flavours. Smolenksy's neural eliminativism, which we rename *neuronal eliminativism*, sees in neuronal models the only scientifically valid cognitive models, while *connectionist eliminativism* also recognizes the validity of connectionist models. It is clear that neuronal sym-

bol processing rests on neuronal eliminativism, while connectionist eliminativism is one of the possible positions that can be taken by exponents of the connectionist symbol processing approach.

However, the CSP approach can map onto other, more ecumenical positions. One is Smolenksy's *limitivism*, which recognizes the validity of neuronal, subsymbolic and symbolic theories, while observing that symbolic models can only provide restricted and approximate descriptions (they cannot, for instance, provide complete and precise accounts of intuitive processing). Another is *revisionist connectionism* which acknowledges the scientific validity of symbolic models *after* revision by connectionist theory. However, the precise nature of this revision varies. In [Pinker and Prince, 1988]'s definition, this revision does not consist in simply adding connectionist models alongside symbolic models; rather connectionist models will implement symbol-processing schemes in ways that have important emergent properties. Revisionist connectionism thus defined is the stand taken explicitly by a number CSP researchers like [Touretzky, 1990].

The revisionist position as interpreted by Smolensky seems to correspond more closely to the hybrid approach. In this view, the revision called for in symbolic models is a kind of division of labor: perception, memory, pattern matching and other "low-level" operations are relegated to connectionist networks while symbolic models retain control of hard, rational symbol processing. This brand of revisionism, which Smolensky calls by the French term "cohabitation", is no other than "hybridization" [Memmi, 1992]. Finally, resistance to all integration efforts can be founded on what Memmi calls segregationism—the claim that symbolicism and connectionism apply to different, non-overlapping domains and can pursue their respective tasks in peaceful coexistence, if not in mutual indifference.

References

[Becraft *et al.*, 1991] W.R. Becraft, P.L. Lee, and R.B. Newell. Integration of neural networks and expert systems. In *Proc. 12 th International Joint Conference on Artificial Intelligence*, pages 832–837. Morgan-Kaufmann, 1991.

[Bookman, 1987] L.A. Bookman. A microfeature based scheme for modelling semantics. In *Proc. 10 th International Joint Conference on Artificial Intelligence*. Morgan-Kaufmann, 1987.

[Crucianu and Memmi, 1992] M. Crucianu and D. Memmi. Extraction de la structure implicite dans un réseau connexionniste. In *Neuro-Nîmes 92. Neural Networks and their Applications*, pages 491–502, Nanterre, 1992. EC2.

[Dyer, 1991] M.G. Dyer. Symbolic neuroengineering for natural language processing: A multilevel research approach. In J. A. Barnden and J. B. Pollack, editors, *Advances in Connectionist and Neural Computation Theory. Vol.1: High-Level Connectionist Models*, pages 32–86. Ablex Publishing, 1991.

[Edelman, 1992] G. Edelman. *Bright Air, Brilliant Fire. On the Matter of the Mind.* Basic Books, 1992.

[Feldman and Ballard, 1982] J. Feldman and D. Ballard. Connectionist models and their properties. *Cognitive Science*, 6:205–254, 1982.

[Fodor and Pylyshyn, 1988] J. A. Fodor and Z. W. Pylyshyn. Connectionism and cognitive architecture: a critical analysis. *Cognition*, 28:2–71, 1988.

[Fu and Fu, 1990] L.M. Fu and L.C. Fu. Mapping rule-based systems into neural architecture. *Knowledge-Based Systems*, 3(1):48–56, March 1990.

[Gallant, 1988] S. I. Gallant. Connectionist expert systems. *Communications of the ACM*, 31(2):152–169, February 1988.

[Giacometti, 1992] A. Giacometti. *Modèles hybrides de l'expertise.* PhD thesis, Ecole Nationale Supérieure des Télécommunications. Paris, France, November 1992.

[Goodman *et al.*, 1989] R.K. Goodman, J.W. Miller, and P. Smith. An information theoretic approach to rule-based connectionist expert systems. In D.S. Touretzky, editor, *Advances in Neural Information Processing 1*, pages 256–263. Morgan-Kaufmann, San Mateo, CA, 1989.

[Gutknecht and Pfeifer, 1990]
M. Gutknecht and R. Pfeifer. An approach to integrating expert systems with connectionist networks. *AICOM*, 3(3):116–127, 1990.

[Gutknecht *et al.*, 1991] M. Gutknecht, R. Pfeifer, and M. Stolze. Cooperative hybrid systems. In *IJCAI Proceedings*, pages 824–829, 1991.

[Handelman *et al.*, 1989] D. A. Handelman, S. H. Lane, and J. J. Gelfand. Integrating knowledge-based system and neural network techniques for robotic skill acquisition. In *Proc. 11 th International Joint Conference on Artificial Intelligence*, pages 193–198, Detroit, MI, 1989. Morgan-Kaufmann.

[Hayes *et al.*, 1992] S. Hayes, V. B. Ciesielski, and W. Kelly. A comparison of an expert system and a neural network for respiratory system monitoring. Technical Report TR #92/1, Royal Melbourne Institute of Technology, March 1992.

[Hendler, 1989] J.A. Hendler. Problem solving and reasoning: A connectionist perspective. In R. Pfeifer, Z. Schreter, and F. Fogelman-Soulié, editors, *Connectionism in Perspective*, pages 229–243. Elsevier, 1989.

[Hölldobler and Kurfess, 1991] S. Hölldobler and F. Kurfess. CHCL—a connectionist inference system. In B. Fronhoefer and G. Wrightson, editors, *Parallelization in Inference Systems.* Springer-Verlag, 1991.

[Jabri *et al.*, 1992] M. Jabri, S. Pickard, P. Leong, Z. Chi, B. Flower, and Y. Xie. Ann based classification for heart difibrillators. In *Advances in Neural Information Processing 4*, pages 637–644. Morgan-Kaufmann, San Mateo, CA, 1992.

[Krovvidy and Wee, 1992] S. Krovvidy and W. G. Wee. An intelligent hybrid system for wastewater treatment. In A. Kandel and G. Langholz, editors, *Hybrid Architectures for Intelligent Systems*, chapter 17, pages 358–377. CRC Press, Boca Raton, 1992.

[Kuncicky *et al.*, 1992] D. C. Kuncicky, S. I. Hruska, and R.C. Lacher. Hybrid systems: The equivalence of rule-based expert system and artificial neural network inference. *International Journal of Expert Systems*, 4(3):281–297, 1992.

[Medsker, 1994] L. R. Medsker. *Hybrid Neural Network and Expert Systems.* Kluwer Academic Publishers, Boston, 1994.

[Memmi, 1992] D. Memmi. Connectionism and artificial intelligence as cognitive models. In A. Clark and R. Lutz, editors, *Connectionism in Context*, pages 145–165. Springer-Verlag, 1992.

[Minsky and Papert, 1969] M. Minsky and S. Papert. *Perceptrons.* MIT Press, 1969.

[Pinker and Prince, 1988] S. Pinker and A. Prince. On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28:73–193, 1988.

[Piramuthu and Shaw, 1994] S. Piramuthu and M.l J. Shaw. On using decision tree as feature selector for feed-forward neural networks. In *International Symposium on Integrating Knowledge and Neural Heuristics*, pages 67–74, Pensacola, Florida, May 1994.

[Reeke and Edelman, 1988] G. N. Reeke and G. M. Edelman. Real brains and artificial intelligence. In S. Graubard, editor, *The Artificial Intelligence Debate. False Starts, Real Foundations*, pages 144–173. MIT Press, 1988.

[Shastri, 1988] L. Shastri. A connectionist approach to knowledge representation and limited inference. *Cognitive Science*, 12:331–392, 1988.

[Smolensky, 1988] P. Smolensky. On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11:1–74, 1988.

[Sun, 1991] R. Sun. *Integrating Rules and Connectionism for Robust Reasoning. A Connectionist Architecture with Dual Representation.* PhD thesis, Brandeis University, Waltham, MA 02254, 1991. Technical Report CS-91-160.

[Touretzky and Hinton, 1988] D. S. Touretzky and G. E. Hinton. A distributed connectionist production system. *Cognitive Science*, 12:423–466, 1988.

[Touretzky, 1990] D. S. Touretzky. Boltzcons : Dynamic symbol structures in a connectionist network. *Artificial Intelligence*, 46(1-2), 1990.

[Towell, 1992] G. G. Towell. Symbolic knowledge and neural networks : insertion refinement and extraction. Technical Report 1072, Univ. of Wisconsin-Madison, Computer Science Dept., January 1992.