Philosophical Foundations of AI

David Vernon[1] and Dermot Furlong[2]

[1] Etisalat University College, UAE
[2] Trinity College Dublin, Ireland

Abstract. Artificial Intelligence was born in 1956 as the off-spring of the newly-created cognitivist paradigm of cognition. As such, it inherited a strong philosophical legacy of functionalism, dualism, and positivism. This legacy found its strongest statement some 20 years later in the physical symbol systems hypothesis, a conjecture that deeply influenced the evolution of AI in subsequent years. Recent history has seen a swing away from the functionalism of classical AI toward an alternative position that re-asserts the primacy of embodiment, development, interaction, and, more recently, emotion in cognitive systems, focussing now more than ever on enactive models of cognition. Arguably, this swing represents a true paradigm shift in our thinking. However, the philosophical foundations of these approaches — phenomenology — entail some far-reaching ontological and epistemological commitments regarding the nature of a cognitive system, its reality, and the role of its interaction with its environment. The goal of this paper is to draw out the full philosophical implications of the phenomenological position that underpins the current paradigm shift towards enactive cognition.

1   Philosophical Preliminaries

Realism is a doctrine which holds that the objects of our perceptions are what is real and that reality is what is directly perceived; it is through our perceptions that we apprehend the actual real external world. The tradition of modern realism has an long pedigree, beginning with Ockham and continuing through Gallileo, Hobbes, Locke, Hume, Moore, and Russell. Gallileo, along with, *e.g.*, Copernicus, Descartes, and Kepler, heralded the beginning of the scientific age which placed all empirical measurement and quantification along with rigourous mathematical (or logical) reasoning as the cornerstones for the construction of knowledge. This empiricist ethos was strengthed by John Locke, a quintessential realist, who viewed perception as a causal process whereby physical stimuli act on the sensory apparatus to produce ideas (concepts or representations, in the modern terminology). Much of today's common understanding of reality is a legacy of this Lockean frame of mind. In realistic positions, there is the underpinning assumption that reality exists absolutely and, whether rationally by reason or empirically by sense, we apprehend it and thus come to understand its form and structure.

Idealism, on the other hand, is a doctrine which posits that reality is ultimately dependent on the mind and has no existence outside of it. If Locke was

the quintessential realist, then Berkeley was the quintessential idealist. Berkeley developed the philosophy that nothing exists save that which is perceived by a mind. This is neatly summarized by his famous aphorism '*esse est percipi*' — to be is to be perceived. Berkeley's position is that our idea about the world are based on our perceptions of it. In this sense, Berekeley is also taking an empirical position — that our knowledge of the world is gained exclusively from our senses. On the other hand, Berkeley denied the existence of matter: what exists is that which is perceived, and it exists because it is perceived. Reality pervades all perception but corporeal matter has no place in this scheme. This denial of the reality of matter distinguishes Berkeley's empirical idealist notions of perception from the realist, empirical, notion that perception is an abstraction or apprehension of the (material) world *via* a causal process of sensing.

Kant (1724-1804) was also an idealist, but his views differed significantly from those of Berkeley. Kant differentiated between *noumena*, the domain of 'things in themselves' and *phenomena*, or the 'appearances' of things as they are presented to us by our senses. Kant argued that noumena are not accessible to us, and cannot be known directly, whereas the phenomena — the contact we have with these things via our senses and perceptions — are the basis for knowledge. Kant refers to noumena as 'trancendental objects' and his philosophy is sometimes referred to as 'trancendental idealism'. Thus, Kant admits the 'reality' of a domain of objects, the unknowable noumenological domain. On the other hand, he maintains that the objects of our experience are the only knowable objects and it is the mind that shapes and forms these sense data and, hence, for us, these objects are the only objects that really exist and they exist *because* of us and our minds. Reality, then, exists as an unknowable, non-sensible, noumenal domain which gives rise to the phenomenal domain of our senses.[1] The idealist tradition did not stop with Kant and has been added to by, *e.g.*, Schopenhauer, Nietzsche, and Hegel.

There are many variations on these two themes of idealism and realism, perhaps the most well-known of which is *dualism* which holds that reality comprises two distinct 'substances': one physical and one mental. Dualism was first propounded as a philosophical system by Descartes who argued for the existence of *two* domains of reality: one corporeal and one non-corporeal. Both mutually-exclusive domains exist concurrently. It is this mutual exclusivity which has caused dualism most of its problems for, if they are truly mutually exclusive, it is not clear how they can interact. This difficulty has been transposed into modern philosophical debate as the 'mind-body' problem: the paradox that if the body and mind are mutually exclusive entities, then how do they 'communicate'?

In the above, we have attempted the impossible: to summarize five hundred years of philosophical thought in a few paragraphs. Nonetheless, from this cursory look at the history of western philosophy, we can see that the philosophical

---

[1] Although Kant is best known as an idealist, his particular brand of philosophical idealism is sometimes referred to as constructive realism due to the central role played by the observer in shaping phenomenal reality.

positions on reality have been dominated by realism (including dualism). Additionally, the philosophies that have been most closely aligned with the scientific method have also been those of realism. In a sense, this isn't surprising since realism is the more immediately common-sense view: things exist — we perceive them. This world-view has been copper-fastened in the last century by the logical positivists, *e.g.* Schlick and Carnap, who held that reality is exactly that which yields to empirical investigation and anything that is not verifiable by empirical investigation is meaningless.

There were, of course, other developments in philosophical thinking, which begin with Kant's distinction between noumena and phenomena, and which evolved into a type of reconciliation of the idealist and the realist positions. The one that interests us here was developed by Husserl, who held that reality is personally and fundamentally *phenomenological* but is set against an objective spatio-temporal world. However, it was best espoused by Heidegger who denied the dichotomy between the world and 'us' and saw existence or 'being in the world' as our activity in a constitutive domain. Reality does not exist 'outside us'; we are beings in a world, not disjoint from it. From a phenomenological perspective, what we perceive depends on what it is we are. The position taken by phenomenology is subtly, but significantly, different to that taken by either realism or idealism. The position is as follows. We play a role in defining reality, but only insofar as it affects us as individuals (the idealist aspect), that is, insofar as it affects our experience of reality; the reality that we perceive does exist (the realist aspect) but our perception and conception of it is conditioned by our experience. Thus, reality is for us a personal experience, though it derives from a common source and this reality is our experience and is contingent upon the current ontological status of us as entities in that universe. As perceivers, our perceptions of the world are a function of what we are: reality is conditioned by experience and experience is conditioned by the nature of the system and its history of interaction with reality.

The dependence of reality on the ontogenetic state of an individual is the essential characteristic of phenomenology and is often referred to as radical constructivism: we construct our reality as a consequence of our perceptions and experiences. Unfortunately, the term constructivism is also sometimes used to denote an entirely different realist position taken by advocates of the cognitivist approach to artificial intelligence whereby representations of the external world are constructed through perception. Consequently, one must be careful when interpreting the term constructivism to be clear exactly what is meant: the radical constructivism of phenomenology or the representational constructivism of realism.

2   The Birth of AI

The development of cybernetics in 1943 heralded the birth of cognitive science and an attempt to create a formal logical model of cognition and a science of mind [1]. The year 1956 saw the emergence of an approach referred to as

cognitivism which asserts that cognition involves computations defined over internal representations, in a process whereby information about the world is abstracted by perception, and represented using some appropriate symbolic data-structure, reasoned about, and then used to plan and act in the world.

For cognitivist systems, perception is concerned with the abstraction of faithful spatio-temporal representations of the external world from sensory data. Reasoning itself is symbolic: a procedural process whereby explicit representations of an external world are manipulated to infer likely changes in the configuration of the world that arise from causal actions.

In most cognitivist approaches concerned with the creation of artificial cognitive systems, the symbolic representations are typically the descriptive product of a human designer. This means that they can be directly accessed and understood or interpreted by humans and that semantic knowledge can be embedded directly into and extracted directly from the system. These programmer-dependent representations 'blind' the system [2] and constrain it to an idealized description that is dependent on and a consequence of the programmer's own cognition. Arguably, it is this *a priori* designer- or programmer-dependent knowledge that is embedded in the system that limits the adapability of the cognitive system since this knowledge intrinsically encapsulates the designer's assumptions about the system's environment, it operation, and its space of interaction.

Cognitivism makes the realist assumption that 'the world we perceive is isomorphic with our perceptions of it as a geometric environment' [3]. Today, cognitivist systems will deploy an arsenal of techniques including machine learning, probabilistic modelling, and other techniques in an attempt to deal with the inherently uncertain, time-varying, and incomplete nature of the sensory data that is being used to drive this representational framework. However, ultimately the representational structure is still predicated on the descriptions of the designer.

AI is the direct descendent of cognitivism [4] and represents the empirical side of cognitivist cognitive science. A major milestone in its development occured in 1976 with Newell's and Simon's 'Physical Symbol System' approach [5]. In their paper, two hypotheses are presented:

1. *The Physical Symbol System Hypothesis:* A physical symbol system has the necessary and sufficient means for general intelligent action.
2. *Heuristic Search Hypothesis:* The solutions to problems are represented as symbol structures. A physical-symbol system exercises its intelligence in problem-solving by search, that is, by generating and progressively modifying symbol structures until it produces a solution structure.

The first hypothesis implies that any system that exhibits general intelligence is a physical symbol system *and* any physical symbol system of sufficient size can be configured somehow ('organized further') to exhibit general intelligence.

The second hypothesis amounts to a assertion that symbol systems solve problems by heuristic search, *i.e.* 'successive generation of potential solution structures' in an effective and efficient manner. 'The task of intelligence, then, is to avert the ever-present threat of the exponential explosion of search'.

A physical symbol system is equivalent to an automatic formal system[6]. It is 'a machine that produces through time an evolving collection of symbol structures.' A symbol is a physical pattern that can occur as a component of another type of entity called an expression (or symbol structure): expressions/symbol structures are arrangements of symbols/tokens. As well as the symbol structures, the system also comprises processes that operate on expressions to produce other expressions: 'processes of creation, modification, reproduction, and destruction'. An expression can *designate* an object and thereby the system can either 'affect the object itself or behave in ways depending on the object', or, if the expression designates a process, then the system *interprets* the expression by carrying out the process. In the words of Newell and Simon,

> 'Symbol systems are collections of patterns and processes, the latter being capable of producing, destroying, and modifying the former. The most important properties of patterns is that they can designate objects, processes, or other patterns, and that when they designate processes, they can be interpreted. Interpretation means carrying out the designated process. The two most significant classes of symbol systems with which we are acquainted are human beings and computers.'

What is important about this explanation of a symbol system is that it is more general than the usual portrayal of symbol-manipulation systems where symbols designate only objects, in which case we have a system of processes that produces, destroys, and modifies symbols, and no more. Newell's and Simon's original view is more sophisticated. There are two recursive aspects to it: processes can produce processes, and patterns can designate patterns (which, of course, can be processes). These two recursive loops are closely linked. Not only can the system build ever more abstract representations and reason about those representation, but it can modify itself as a function both of its processing, *qua* current state/structure, and of its representations.

Symbol systems can be instantiated and the behaviour of these instantiated systems depend on the the details of the symbol system, its symbols, operations, and interpretations, and *not* on the particular form of the instantiation.

The *physical symbol system hypothesis* asserts that a physical symbol system has the necessary and sufficient means for general intelligence. From what we have just said about symbol systems, it follows that intelligent systems, either natural or artificial ones, are effectively equivalent because the instantiation is actually inconsequential, at least in principle.

To a very great extent, cognitivist systems are identical to physical symbol systems.

The strong interpretation of the physical symbol system hypothesis is that not only is a physical symbol system sufficient for general intelligence, it is also necessary for intelligence.

It should be clear that cognitivism, and the classical AI of physical symbol systems, are dualist, functionalist, and positivist. They are dualist in the sense that there is a fundamental distinction between the mind (the computational processes) and the body (the computational infrastructure and, where required,

the plant that instantiates any physical interaction). The are functionalist in the sense that the actual instantiation and computational infrastructure is inconsequential: any instantiation that supports the symbolic processing is sufficient. They are positivist in the sense that they assert a unique and absolute empirically-accessible external reality that is apprended by the senses and reasoned about by the cognitive processes.

3   Enaction

Cognitivism is not however the only position one can take on cognition. There is a second class of approaches, all based to a lesser or greater extent on principles of emergent self-organization [1,7] and best epitomized by enactive approaches.

   The enactive systems research agenda stretches back to the early 1970s in the work of computational biologists Maturana and Varela [8,9,10,11,1,2,12]. In contradistinction to cognitivism, which involves a view of cognition that requires the representation of a given objective pre-determined world [13,1], enaction asserts that cognition is a process whereby the issues that are important for the continued existence of a cognitive entity are brought out or enacted: co-determined by the entity as it interacts with the environment in which it is embedded. Thus, nothing is 'pre-given'. Instead there is an enactive interpretation: a real-time context-based choosing of relevance. Cognition is the process whereby an autonomous system becomes viable and effective in its environment. It does so through a process of self-organization through which the system is continually re-constituting itself in real-time to maintain its operational identity through moderation of mutual system-environment interaction and co-determination [12]. Co-determination implies that the cognitive agent is specified by its environment and at the same time that the cognitive process determines what is real or meaningful for the agent. In a sense, co-determination means that the agent constructs its reality (its world) as a result of its operation in that world. Thus, for emergent approaches, perception is concerned with the acquisition of sensory data in order to enable effective action [12] and is crucially dependent on the richness of the action interface [14]. It is not a process whereby the structure of an absolute external environment is abstracted and represented in a more or less isomorphic manner.

   In contrast to the cognitivist approach, many enactive approaches assert that the primary model for cognitive learning is anticipative skill construction rather than knowledge acquisition and that processes that both guide action and improve the capacity to guide action while doing so are taken to be the root capacity for all intelligent systems [15]. While cognitivism entails a self-contained abstract model that is disembodied in principle, the physical instantiation of the systems plays no part in the model of cognition [16,17]. In contrast, enactive approaches are intrinsically embodied and the physical instantiation plays a pivotal constitutive role in cognition.

   With enactive systems, one of the key issues is that cognitive processes are temporal processes that 'unfold' in real-time and synchronously with events in

their environment. This strong requirement for synchronous development in the context of its environment is significant for two reasons. First, it places a strong limitation on the rate at which the ontogenetic learning of the cognitive system can proceed: it is constrained by the speed of coupling (*i.e.* the interaction) and not by the speed at which internal changes can occur [2]. Second, taken together with the requirement for embodiment, we see that the consequent historical and situated nature of the systems means that one cannot short-circuit the ontogenetic development. Specifically, you can't bootstrap an emergent dynamical system into an advanced state of learned behaviour.

For cognitivism, the role of cognition is to abstract objective structure and meaning through perception and reasoning. For enactive systems, the purpose of cognition is to uncover unspecified regularity and order that can then be construed as meaningful because they facilitate the continuing operation and development of the cognitive system. In adopting this stance, the enactive position challenges the conventional assumption that the world *as the system experiences it* is independent of the cognitive system ('the knower'). The only condition that is required of an enactive system is *effective action*: that it permit the continued integrity of the system involved. It is essentially a very neutral position, assuming only that there is the basis of order in the environment in which the cognitive system is embedded. From this point of view, cognition is exactly the process by which that order or some aspect of it is uncovered (or constructed) by the system. This immediately allows that there are different forms of reality (or relevance) that are dependent directly on the nature of the dynamics making up the cognitive system. This is not a solipsist position of ungrounded subjectivism, but neither is it the commonly-held position of unique — representable — realism. It is fundamentally a phenomenological position.

The goal of enactive systems research is the complete treatment of the nature and emergence of autonomous, cognitive, social systems. It is founded on the concept of autopoiesis – literally *self-production* – whereby a system emerges as a coherent systemic entity, distinct from its environment, as a consequence of processes of self-organization.

In the enactive paradigm, linguistic behaviours are at the intersection of ontogenetic and communication behaviours and they facilitate the creation of a common understanding of the shared world that is the environment of the coupled systems. That is, language is the emergent consequence of the structural coupling of a socially-cohesive group of cognitive entities. Equally, knowledge is particular to the system's history of interaction. If that knowledge is shared among a society of cognitive agents, it is not because of any intrinsic abstract universality, but because of the consensual history of experiences shared between cognitive agents with similar phylogeny and compatible ontogeny. A key postulate of enactive systems is that reasoning, as we commonly conceive it, is the consequence of reflexive[2] use of the linguistic descriptive abilities to the cognitive agent itself [12]. Linguistic capability is in turn developed as a consequence of the consensual co-development of an epistemology in a society of phylogenetically-identical cognitive agents. This

---

[2] Reflexive in the sense of self-referential, not in the sense of a reflex action.

is significant: reasoning in this sense is a descriptive phenomenon and is quite distinct from the self-organizing mechanism (*i.e.* structural coupling and operational closure [12]) by which the system/agent develops its cognitive and linguistic behaviours. Since language (and all inter-agent communication) is a manifestation of high-order cognition, specifically co-determination of consensual understanding amongst phylogenetically-identical and ontogenetically-compatible agents, symbolic or linguistic reasoning is actually the product of higher-order social cognitive systems rather than a generative process of the cognition of an individual agent.

4   Conclusion

The chief point we wish to make in this paper is that the differences between the cognitivist and emergent positions are deep and fundamental, and go far beyond a simple distinction based on symbol manipulation. It isn't principally the symbolic nature of the processing that is at issue in the divide between the cognitivist and the emergent approaches — it is arguable that linguistically-capable enactive systems explicitly use symbols when reasoning. Neither is it the presence or use of a physical body or situated perceptual agents. Cognitivists now readily admit the need for embodiment; in Anderson's words: 'There is reason to suppose that the nature of cognition is strongly determined by the perceptual-motor systems, as the proponents of embodied and situated cognition have argued' [18]. Elsewhere they are compared on the basis of several related characteristics [19] but in this paper, we have contrasted the two paradigms on the basis of their philosophical commitments: the functionalist, dualist, and positivist ground of cognitivist cognition versus the phenomenological agent-specific mutual-specification of enactive cognition.

In the enactive paradigm, the perceptual capacities are a consequence of an historic embodied development and, consequently, are dependent on the richness of the motoric interface of the cognitive agent with its world. That is, the action space defines the perceptual space and thus is fundamentally based in the frame-of-reference of the agent. Consequently, the enactive position is that cognition can only be created in a developmental agent-centred manner, through interaction, learning, and co-development with the environment. It follows that through this ontogenic development, the cognitive system develops its own epistemology, *i.e.* its own system-specific knowledge of its world, knowledge that has meaning exactly because it captures the consistency and invariance that emerges from the dynamic self-organization in the face of environmental coupling. Thus, we can see that, from this perspective, cognition is inseparable from 'bodily action' [20]: without physical embodied exploration, a cognitive system has no basis for development. Despite the current emphasis on embodiment, Ziemke notes that many current approaches in cognitive/adaptive/epigenetic robotics still adhere to the functionalist dualist hardware/software distinction in the sense that the computational model does not in principle require an instantiation [21,22]. Ziemke suggests that this is a real problem because the idea of embodiment in the enactive sense is that the morphology of the system is actually a key component of the systems

dynamics. In other words, morphology not only matters, it is a constitutive part of the system's self-organization and structural coupling with the environment and defines its cognition and developmental capacity.

There are many challenges to be overcome in pushing back the boundaries of AI research, particularly in the practice of enactive AI. Foremost among these is the difficult task of identifying the phylogeny and ontogeny of an artificial cognitive system: the requisite cognitive architecture that facilitates both the system's autonomy (*i.e.* its self-organization and structural coupling with the environment) and its capacity for development and self-modification. To allow true ontogenetic development, this cognitive architecture must be embodied in a way that allows the system the freedom to explore and interact and to do so in an adaptive physical form that enables the system to expand its space of possible autonomy-preserving interactions. This in turn creates a need for new physical platforms that offer a rich repertoire of perception-action couplings and a morphology that can be altered as a consequence of the system's own dynamics. In meeting these challenges, we move well beyond attempts to build cognitivist systems that exploit embedded knowledge and which try to see the world the way we designers see it. We even move beyond learning and self-organizing systems that uncover for themselves statistical regularity in their perceptions. Instead, we set our sights on building enactive phenomenologically-grounded systems that construct their own understanding of their world through adaptive embodied exploration and social interaction.

Acknowledgments

References

1. Varela, F.J.: Whence perceptual meaning? A cartography of current ideas. In: Varela, F.J., Dupuy, J.-P. (eds.) Understanding Origins – Contemporary Views on the Origin of Life, Mind and Society. Boston Studies in the Philosophy of Science, pp. 235–263. Kluwer Academic Publishers, Dordrecht (1992)
2. Winograd, T., Flores, F.: Understanding Computers and Cognition – A New Foundation for Design. Addison-Wesley Publishing Company, Inc., Reading, Massachusetts (1986)
3. Shepard, R.N., Hurwitz, S.: Upward direction, mental rotation, and discrimination of left and right turns in maps. Cognition 18, 161–193 (1984)
4. Freeman, W.J., Núñez, R.: Restoring to cognition the forgotten primacy of action, intention and emotion. Journal of Consciousness Studies 6(11-12) ix–xix (1999)
5. Newell, A., Simon, H.A.: Computer science as empirical inquiry: Symbols and search. Communications of the Association for Computing Machinery 19, 113–126 (1976)

6. Haugland, J.: Semantic engines: An introduction to mind design. In: Haugland, J. (ed.) Mind Design: Philosophy, Psychology, Artificial Intelligence, pp. 1–34. Bradford Books, MIT Press, Cambridge, Massachusetts (1982)
7. Clark, A.: Mindware – An Introduction to the Philosophy of Cognitive Science. Oxford University Press, New York (2001)
8. Maturana, H.: Biology of cognition. Research Report BCL 9.0, University of Illinois, Urbana, Illinois (1970)
9. Maturana, H.: The organization of the living: a theory of the living organization. Int. Journal of Man-Machine Studies 7(3), 313–332 (1975)
10. Maturana, H.R., Varela, F.J.: Autopoiesis and Cognition — The Realization of the Living. Boston Studies on the Philosophy of Science. D. Reidel Publishing Company, Dordrecht, Holland (1980)
11. Varela, F.: Principles of Biological Autonomy. Elsevier North Holland, New York (1979)
12. Maturana, H., Varela, F.: The Tree of Knowledge – The Biological Roots of Human Understanding. New Science Library, Boston & London (1987)
13. van Gelder, T., Port, R.F.: It's about time: An overview of the dynamical approach to cognition. In: Port, R.F., van Gelder, T. (eds.) Mind as Motion – Explorations in the Dynamics of Cognition, pp. 1–43. Bradford Books, MIT Press, Cambridge, Massachusetts (1995)
14. Granlund, G.H.: The complexity of vision. Signal Processing 74, 101–126 (1999)
15. Christensen, W.D., Hooker, C.A.: An interactivist-constructivist approach to intelligence: self-directed anticipative learning. Philosophical Psychology 13(1), 5–45 (2000)
16. Vernon, D.: The space of cognitive vision. In: Christensen, H.I., Nagel, H.-H. (eds.) Cognitive Vision Systems: Sampling the Spectrum of Approaches. LNCS, pp. 7–26. Springer, Heidelberg (2006)
17. Vernon, D.: Cognitive vision: The case for embodied perception. Image and Vision Computing, 1–14 (in press, 2007)
18. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y.: An integrated theory of the mind. Psychological Review 111(4), 1036–1060 (2004)
19. Vernon, D., Metta, G., Sandini, G.: A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. IEEE Transaction on Evolutionary Computation 11(1), 1–31 (2006)
20. Thelen, E.: Time-scale dynamics and the development of embodied cognition. In: Port, R.F., van Gelder, T. (eds.) Mind as Motion – Explorations in the Dynamics of Cognition, pp. 69–100. Bradford Books, MIT Press, Cambridge, Massachusetts (1995)
21. Ziemke, T.: Are robots embodied? In: Balkenius, Z., Dautenhahn, K., Breazeal (eds.) Proceedings of the First International Workshop on Epigenetic Robotics — Modeling Cognitive Development in Robotic Systems, Lund University Cognitive Studies, Lund, Sweden. vol. 85, pp. 75–83 (2001)
22. Ziemke, T.: What's that thing called embodiment? In: Alterman, Kirsh (eds.) Proceedings of the 25th Annual Conference of the Cognitive Science Society. Lund University Cognitive Studies, pp. 1134–1139. Lawrence Erlbaum, Mahwah, NJ (2003)