Chapter 1

How do agents represent?

Alex Ryan
DSTO, Australia
alex.ryan@dsto.defence.gov.au

Representation is inherent to the concept of an agent, but its importance in complex systems has not yet been widely recognised. In this paper I introduce Peirce's theory of signs, which facilitates a definition of representation in general. In summary, representation means that for some agent, a model is used to stand in for another entity in a way that shapes the behaviour of the agent with respect to that entity. Representation in general is then related to the theories of representation that have developed within different disciplines. I compare theories of representation from metaphysics, military theory and systems theory. Additional complications arise in explaining the special case of mental representations, which is the focus of cognitive science. I consider the dominant theory of cognition – that the brain is a representational device – as well as the sceptical anti-representational response. Finally, I argue that representation distinguishes agents from non-representational objects: agents are objects capable of representation.

1.1 Introduction

Representation is an essential concept for understanding the behaviour of agents in a complex system. Consider traders in a stock exchange market as agents. If every agent has unmediated access to the value of a company (including its exact future profits discounted to present value), then the market cannot exist, since shareholders would only be willing to sell above this value, a price no rational buyer would pay[1]. Only when partial information on value is allowed

[1] One might expect trades to be made exactly at the value of the stock. However, once a financial or time cost is included no rational buyer can exist. Why would an agent buy shares that never increase in real value and incur an exit fee?

and different agents have access to different information is it possible to predict the formation of a market. In this case, each agent must construct a model representing the perceived value of a company. By communicating, agents can modify their models to take into account the representations other agents in their social network have constructed. Because there is a benefit in being connected to agents who are better at predicting future value, some agents may specialise in developing predictive models and charging other agents for access to their expectations (such as financial advisors). Markets would not exist if there were not differences between agents in their representations. Variety in representation allows the simultaneous existence of buyers and sellers, as well as the potential for a secondary market based on constructing representations and selling advice. Even though this is quite obvious, imperfect information, bounds on rationality, and consequently the need for constructing representations did not feature in the theories and models of classical economics.

It turns out that an account of representation is just as important in understanding the role of the discipline of complex systems, as for understanding the behaviour of agents within a complex system. This is because the systems approach is a way of representing the world. When this is overlooked, systems applications may be blind to the limitations of the representations they employ. This discussion of representation is intended to be interpreted on two levels. On one level, when an analyst uses a complex systems approach, they invariably construct systems representations. On another level, when the system contains agents that also represent their environment, this must be accounted for in any model of the system.

Section 1.2 makes the metaphysical assumptions of this paper explicit. Then in Section 1.3, Peirce's theory of signs is used as a basis for a theory of representation in general. When agents represent their environment, they may use either external or internal models. Section 1.4 surveys accounts of external representation across several disciplines, while Section 1.5 surveys internal representation, which has been discussed mostly in philosophy of mind and cognitive science. This paper concludes by defining 'agent' in Section 1.6, which demonstrates the strong link between agency and representation.

1.2 Metaphysical Assumptions

Before representation is discussed in detail, it is prudent to make the metaphysical assumptions of this paper explicit. The metaphysical position I will adhere to is known as physicalism, the view that there are no kinds of things other than physical things. In particular, I assume that the relationship between macroscopic and microscopic phenomena is one of supervenience. The Stanford Encyclopedia of Philosophy offers the following definition:

Definition 1 (Supervenience) *A set of properties A supervenes upon another set B just in case no two things can differ with respect to A-properties without also differing with respect to their B-properties [40].*

Supervenience, along with physicalism, entails that in principle, all of the bookkeeping regarding forces can be accounted for in purely physical terms between arbitrarily small entities, when the set B is taken to be the properties of fundamental physics. This is because every time there is a change in a macro level property, there must be a corresponding change in the micro level properties. That the physcial forces fully account for the dynamics at the micro level tells us little about what physical predictions *mean*. Semantics is always relative to an agent's subjective experience of the world, a concept which does not feature in, and cannot be fully explained by, the elementary particles and fundamental forces of physics. First-person experience is just one example of an emergent property, the general reason why descriptions at other levels cannot be eliminated. I will assume that forces in chemistry, biology, psychology and sociology do not add anything to the physical: that the laws of physics are *conservative*. This is consistent with Anderson's [3] twin assertions that all ordinary matter obeys simple electrodynamics and quantum theory, but that "the ability to reduce everything to simple fundamental laws does not imply that ability to start from those laws and reconstruct the universe". This assumption can be argued with, but it cannot be proved either way. I assume supervenience regarding the relationship between macro and micro phenomena because to do otherwise is to place some entities outside the domain of scientific explanation, and it is difficult to see what is achieved by doing so. Descartes' [21] non-physical mind that provided the basis for substance dualism in Meditations VI, and Bergson's [8] *elan vital* that animated the evolution and development of organisms, are examples of non-physical entities that have been postulated in science, and history suggests both acted as barriers to progress. Consequently, I only consider representations that supervene on the physical as meaningful.

1.3 Representation in General

> *Things don't mean: we construct meaning using representational systems – concepts and signs.*
>
> Stuart Hall

There are a number of reasons why unmediated interaction with the world can be undesirable. Some entities are distinctly unfriendly, others are inaccessible, and sometimes the process of interaction is too costly or time consuming. In order to understand anything about a solar flare on the surface of the Sun, mediated access, via the construction of models, is necessary to avoid the undesirable consequences of unmediated contact. A model acts as a representation because it *stands in* for unmediated interaction with the system of interest. Other situations where representations stand in for unmediated interaction include predicting properties of previously unrealised configurations; designing artifacts that do not exist; facilitating comparison of structural similarities between apparently dissimilar phenomena; and generalising knowledge to apply beyond a single entity at a single moment in time[2]. As will be discussed in Section 1.5.1, the dominant theory of human cognition assumes that the mind is a representational device, and that the brain has representational content.

In counterpoint to the important and varied roles of representation, there exists little formal work on representation in general. What are the necessary and sufficient conditions for representation to occur? What kinds of representations exist? One such general theory was proposed by Peirce, which he named the theory of signs or "semiotics". However, because much of his work was misplaced and posthumously edited non-chronologically into highly fragmented volumes, and since Peirce's unique and subtle philosophy requires explication before the finer points of his theory of signs can be appreciated, it remains under-utilised as a general theory of representation. Fortunately, Von Eckardt [63, p. 143-159] has performed considerable work to situate Peirce's theory of signs as a foundation for the more specialised debate on mental representation in cognitive science. I will draw heavily on Von Eckardt's interpretation of Peirce, since it is better oriented towards contemporary concerns in the theory of representation. Unlike Peirce or Von Eckardt, my interests apply to the field of complex systems, and so I will abuse the semiotic and cognitive science terminology by translating it into more general language.

For Peirce, representation was an irreducible triadic relation between objects, signs and interpretants. This implies that something is a sign only if it is a sign *of* an object *with respect to* an interpretant [63, p. 145]. It also implies that the representation relation cannot be decomposed into diadic relations between entities, objects and signs. According to Peirce, representation can only be fully understood by considering the three components of the triadic relation simultaneously.

In the triadic relation, the sign is a token that signifies the object. The more general term I will substitute for sign is model. Peirce's object is already quite general: it may be abstract or concrete, a singular object or a set of objects (a complex object). However, since objects do not in fact need to be objective (or concrete), I will use the more general term entity. The interpretant exists in the mind of the interpreter for whom the sign is a sign [63, p. 148]. Whereas Peirce and Von Eckardt limit their attention to human interpreters, I will generalise this to consider agents. The cost of this generalisation is that the precise nature of interpretants cannot be specified in a way that applies to all agents. Consequently, I can only induce the existence of an interpretant by an observable effect on the agent's behaviour. The triadic relation between entity E, model M and interpretant I is illustrated in Figure 1.1.

Von Eckardt [63, p. 158] summarises the value of Peirce's theory of signs for understanding representation in general as follows:

1. Peirce's distinction between a representation and a representation bearer;
2. His insistence that something can be a full-blown representation only if it is both grounded and interpreted;

[2] This list paraphrases Kline [35, p. 19].

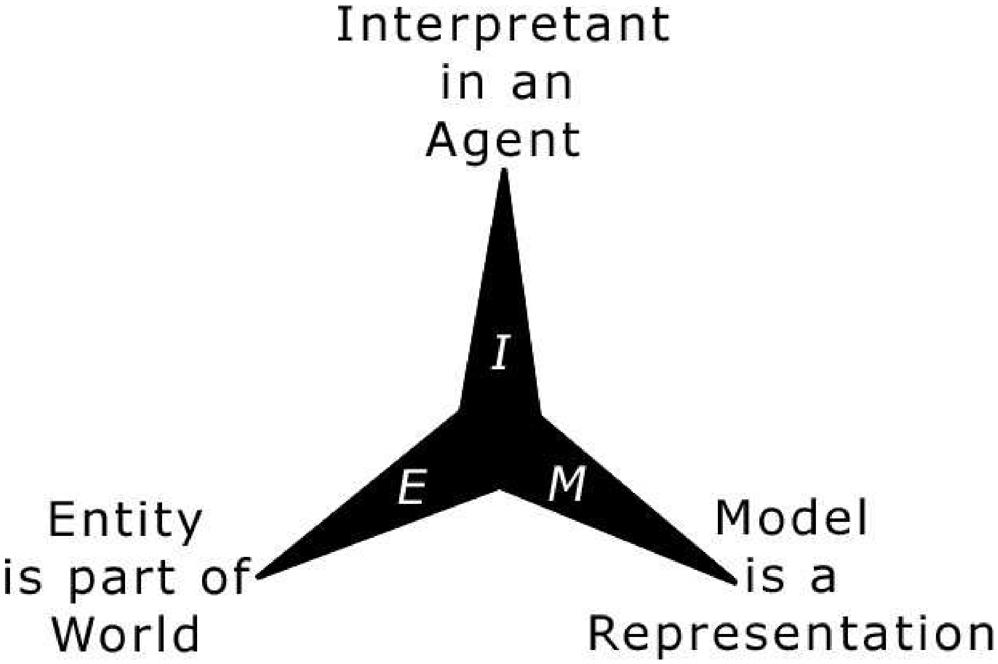

Figure 1.1: Representation as a triadic relation between entity, model and interpretant.

3. His attempt to understand what makes a mental effect an interpretant of some particular representation;

4. His struggle with the problem of interpretation for mental representation;

5. The idea that model and entity are related by two very different sets of relations—semantic relations (such as representing, signifying, referring to, and expressing) and the ground relations in virtue of which those semantic relations hold;

6. His taxonomy of kinds of ground; and

7. His apparent interest in ultimately understanding representation in a completely naturalistic way.

Examining each of these points in turn, the first distinction leads Peirce to consider the character of a model itself. Von Eckardt uses the term "representation bearer" to refer to the properties of the model that belong to the model itself, and not to the entity it represents. In a similar vein, Kline [35] brings attention to the essential difference between a model and an entity by invoking what he calls Korzybski's Dictum, after Alfred Korzybski's [36] warning that "the map is not the territory". It is important to remember that Korzybski's Dictum applies to all representations. It implies that a representation must behave differently to the entity in some contexts. Representations are not perfect substitutes, which means there always exist limits to their ability to stand in for the entity they represent. The other implication of this distinction is that it is both possible and useful to understand the properties of representation bearers as distinct from the properties of the entity they represent.

The second item addresses the interpretation and grounds of a model. The reason that the interpretant is a necessary component of the triadic relation is because a representation is more than just a logical similarity between two entities. If a tree casts a shadow, it is not telling the time until an agent *uses it* to tell the time [1]. Or in the words of Dennett, "Nothing is intrinsically a representation of anything; something is a representation only *for* or *to* someone" [19, p. 101]. By shaping the agent's behaviour, I brings M into the appropriate relation as a representation of E. That is, when an agent interprets (and therefore understands) the model, this grounds the model as a representation of E.

The third item links interpretation with a change in behaviour. If I is in agent A, then the representation must be capable of shaping the behaviour of the agent through the presence of I. Peirce classified interpretants as emotional (feelings), energetic (efforts) and logical (habit-changing) effects. According to Von Eckardt, logical effects, which modify the interpreter's disposition to behave, were considered the primary effect.

The fourth item refers to the problem of infinite regress. While non-mental representation is relatively straightforward, issues arise when M is internal to the agent. The difficult question one faces is "what interprets a mental representation?" If mental representations are interpreted in the same way as non-mental representations, this gives rise to an infinite regress of thoughts interpreting thoughts [63, p. 282].

Separating semantic and ground relations, as noted in point five, allows one to account for how a semantic relation can come to exist. In order for a model M to produce an interpretant I in an agent A, it is necessary for A to understand the representation, which requires A to have knowledge of what the ground is. The following example clarifies this account [63, p. 156]:

> For example, suppose I see a photograph. To understand that photograph I must know (in some sense) that there are both a causal relation and a similarity relation between the photograph and its subject, and I must know (in some sense) the respects in which the photograph is a causal effect of and is similar to its subject. If I know all that, then I will be able to form a belief or a thought about the subject of the photograph (that is, who or what the photograph represents)—specifically, that there was such a subject and that this subject looked a certain way at the time the photograph was taken. In other words, by considering the photograph in conjunction with its ground I come to be in a relation to the object it represents.

With respect to item six, according to Peirce, there exist three kinds of pure ground: iconic, indexical, and symbolic [63, p. 150]. Icons, such as diagrams and images, are models grounded by their intrinsic (first order) similarity to the entity they represent. An index, such as a weathervane, signifies an entity because of a causal or spatiotemporal connection between the index and the entity. Symbolic representations, such as words, are grounded by convention. Symbols act as models only because of the way they are consistently interpreted, which can then generate regular effects on the behaviour of the agent.

In the final item, Von Eckardt interprets Peirce's theory of signs as naturalistic, meaning closely connected to natural science. The naturalistic approach fits neatly with the metaphysical assumption of supervenience outlined in Section 1.2.

I will now propose a definition of representation that reflects Peirce's triadic relation.

Definition 2 (Representation) *A triadic relation between a model, entity and agent. The model substitutes for at least one entity, shaping the behaviour of at least one agent.*

The model *stands in* for an entity, and it always does so *for* an agent, thereby modifying the agent's predisposition to behave. In this definition, the interpretant is implicit in the ability of the model to shape the behaviour of an agent. The model may refer to a class of entities, and may also be shared by multiple agents. However, at least one entity and one agent are necessary for a representation relation.

1.4 External Representation

According to Peirce's triadic relation, the entity E is part of the world, the interpretant I is in the agent, but the location of the model M is unspecified. For the case that M is external to the agent, the triadic relation is relatively straightforward, since the problem of infinite regress does not need to be addressed.

Peirce's typology describes three 'pure' types of grounding relations, which is important for a theory of representation in general. However, in practice, models may incorporate some combination of iconic, indexical and symbolic grounds. The aim of this section is to provide concrete examples of external models, and then show how entities, external models and interpretants have been classified within disparate academic disciplines.

1.4.1 Three kinds of external models

Common models used in representation can be distinguished by implementation rather than pure type. Here, I will assume that a model is somehow simpler than the entity it represents. Although not necessarily true, in practice this is reasonable, since a $1:1$ mapping in complete detail is in general completely useless (consider a life-sized map of the world, then consider trying to maintain the accuracy of every detail). Even if the representation bearer is not itself simple, practical models must confer some benefit, such as ease of manipulation[3]. Models have deliberate differences and may accentuate salient features, in order to retain only those aspects that are necessary to stand in for the entity. A caricature of a politician and a scale model of an aeroplane are examples of representations that can be understood and manipulated efficiently, which makes them useful substitutes for direct experience under certain conditions.

One special kind of representation is a mathematical model. For example, two contained gas particles can be modelled mathematically by two hard uniform spheres with no internal energy except velocity, in an enclosed continuous four dimensional space (including time). The dynamics of the model constrain its behaviour by conserving momentum and energy, which is transferred along the axis joining the spheres' centres of mass when they collide elastically. The spheres are reflected by collisions with the containing walls. In principle, the model, in conjunction with initial measurements of position and velocity, can be used to predict the outcome of measuring the position and velocity of the particles at any future time. The model is a representation when someone (or more generally an agent) uses the model to stand in for a system of interest. For example, the agent could deduce the value of variables associated with the particles in place of direct observations of the gas particles at future times. In Peirce's typology, mathematical models have symbolic grounds. A mathematical model can always be interpreted as manipulating symbols in a formal system according to syntactic rules[4].

The gas particle dynamics can be represented in at least two other ways. Predictions could also be derived using a physical model, such as two balls on a billiard table. The billiard balls are analog representations, which are not arbitrary and abstract like symbols, but are in some way analogous to their subject. Formally, an analog model must exhibit systematic variation with its task domain [41]. This means the analog model does not have to represent every aspect of the entity in the same units – consider a sun dial, which represents the passage of time as the movement in space of a shadow. Animal testing of pharmaceuticals, architects' scale models and pictures are examples of analog representations, although note that the last two examples can also contain symbolic content, which is usually of a secondary nature. Note that in some analog models, it is possible to view the model from multiple perspectives, while other analog models may fix the perspective in the process of representing an entity. Either way, an agent must use the analog model in place of real world measurements in order to fulfil the representation relation. In Peirce's typology, analog models have iconic grounds.

Another way the gas particles could be represented is using the English lan-

[3] An example of a useful $1 : 1$ mapping is the conversion between Polar and Cartesian coordinates, which is practical because performing this conversion can often improve the ease of manipulation.

[4] For some areas of mathematics, the corresponding formal system may have an infinite number of axioms, rules or symbols, however these areas of pure mathematics are not practical for forming representations, and for my purposes can safely be ignored.

guage. For the task of predicting particle dynamics, language is quite limited. However, if the particles were at sufficiently low temperature that their movement was frozen, an English description of their configuration could provide a useful representation for an agent. Language can be used to arrange words, which function as labels, to represent objects. Nouns are labels for entities or classes of entities, while the verb phrase of a predicate with two arguments (two nouns) refers to the relationship between the corresponding entities. Labels in isolation can act as signs, which constitute the most primitive form of representation, capable of standing in for only a single idea. Signs are formalised in semiology, whose contemporary form follows much more closely from the work of Saussure [54] than Peirce [29]. When a set of signs is organised into a language with syntactic rules for manipulation and intricate networks of relationships between components, its representational power is qualitatively increased, and is rich enough to be studied in the distinct but related fields of linguistics and structuralist philosophy. An important observation is that the structure of sign systems (languages) does not need to represent the structure of the world. The structure in language is based on the difference between terms, rather than a reflection of structure in the world. This decoupling both provides flexibility of expression within language, while at the same time necessarily limiting its representational nature. Following Saussure, this constructionist view of language is the dominant view in structuralist and post-structuralist philosophy. Note that while mathematics is also a language[5], and both mathematics and language have symbolic grounds, I consider formal systems separately from linguistic representations, because they can play significantly different roles in representation.

Many disciplines have developed explanations of the way external models – mathematical, analog and linguistic – are used by agents to represent their world. The three disciplines I now consider are metaphysics, military theory and systems theory. The terminology and the scope of representations under consideration varies significantly. In spite of this, it is found that Peirce's triadic relation provides a common structure for explaining representation in each case, and also that the external models conform to the three kinds identified in this subsection. Further, the theory of representation in general reveals shortcomings in each of the disciplinary accounts.

1.4.2 Representation in metaphysics

Popper advocated an ontological pluralist doctrine from 1967, which is detailed in *Objective Knowledge* [49] and concisely summarised in [50]. According to Popper, there exists three worlds:

- World 1 is the physical universe, including both living organisms and non-organic matter.

[5] This interpretation is made precise in formal language theory.

- World 2 is the world of individual psychology, of mental events, raw feels and thoughts.
- World 3 is the world of abstract products of the human mind, including language, scientific theories, mathematics, paintings and symphonies.

The use of 'world' is indicative of the ontological nature of Popper's distinction. He clearly views each world as consisting of different kinds of stuff, proposing "a view of the universe that recognizes at least three different but interacting sub-universes." [50, p. 143]. The nature of the interactions are causal, and the abstract world is always linked to the physical world via the human mind [50, p. 165]:

> If I am right that the physical world has been changed by the world 3 products *of the human mind*, acting through the intervention *of the human mind* then this means that the worlds 1, 2, and 3, can interact and, therefore, that none of them is causally closed.

Popper contrasts his three world hypothesis with ontological monism (materialism or physicalism) and ontological dualism (mind-body dualism) by saying that the monist only admits world 1, while the dualist only admits worlds 1 and 2. When Popper refers to say a symphony or a sculpture in world 3, this is separate from the world 1 instantiation of the entity. It is only the abstract ideal of the entity that exists in world 3. Thus, world 3 entities are types, which may have many corresponding real world tokens that are imperfect embodiments of their type. The key to Popper's defence of world 3 are the claims that a) abstract entities exist that are not embodied in world 1 or 2, such as the infinite members of the set of natural numbers $\mathbb{N}$; and b) abstract entities have a causal influence on world 1, such as Einstein's equation $e = mc^2$ resulting in the development and use of an atomic bomb.

The three world hypothesis is of interest to us, because it neatly separates the real world entities E that are being represented (world 1), mental interpretations I of those entities (world 2), and external models M that are products of the human mind (world 3), in a way that is compatible[6] with Peirce's triadic relation. However, Popper's cosmology directly contradicts our understanding of physics. In particular, conservation laws and symmetry imply that world 1 is closed, and current theory requires only four fundamental forces (the strong and weak nuclear forces, electromagnetism and gravity) to explain every causal physical interaction. Popper claims that world 3 entities that cannot be embodied in world 1 can nevertheless exert a causal influence on world 1, because they are apprehended by human minds in world 2, which then control causal events back in world 1. But then any physical explanation of a system that includes humans is causally incomplete. Even if one accounted for all of the interactions of the four fundamental forces, there would be a 'residual causality'

[6] For both Peirce and Popper, M could be private or shared. However, there is a difference in emphasis: Peirce is mostly concerned with private use, whereas Popper concentrates on shared uses of M.

that remained unaccounted for. This is because abstract entities are not subject to the four fundamental forces, and yet if they have their claimed causal powers, their absence or presence will change the aggregate force acting on bits of world 1 matter. Consequently, one can ask whether it is conceivable that an experiment exists that could test for a residual causality leak from world 1. This would require us to ascertain the presence or absence of an abstract entity, which would require a human mind, without affecting the physical state in the experiment. But in order to say whether the abstract entity was present, the memory would have to be stored in the brain, thus changing the physical state in the experiment (assuming supervenience). In fact, Popper's claim is metaphysical, and unfalsifiable, in contrast to the ideal of scientific conjecture that he advocated. For our purposes, Popper's ontological distinction is stronger than is justified. The same argument applies to the similar, but less sophisticated distinction that Penrose [46] proposes between the physical, Platonic forms, and the human mind.

1.4.3 Representation in military theory

At the other extreme of the academic spectrum, one finds a position that as far as I can ascertain, is advocated only within the relatively isolated discipline of military theory. A central idea in Network Centric Warfare (NCW) [2] and the closely associated, but broader Effects Based Operations (EBO) [58] concepts, is that military actions occur in three domains: the physical, information and cognitive domains. They are based largely on "common sense" and are not rigorously defined. For example, Garstka [25] provides the circular definition: "The information domain is the domain where information lives." This definition is perpetuated in [2]. More sense can be made of Smith's [58, pp. 160-173] interpretation:

> The three domains provide a general framework for tracing what actually goes on in the stimulus and response process inside human minds and human organizations, and how physical actions in one domain get translated into psychological effects and then into a set of decisions in another domain. Understanding this process is important because with it, we can begin to comprehend how people and organizations perceive a stimulus or action and why they respond or react in the way they do and thus, how we might shape behavior.

Smith then defines each domain.

> . . . the physical domain encompasses all the physical actions or stimuli that become the agents for the physical and psychological effects we seek to create. . . . the actions in the physical domain may be political, economic, and/or military in nature, and all must be equally considered to be objects or events. . .
>
> The information domain includes all sensors that monitor physical actions and collect data. It also includes all the means of collating or

> contextualizing that data to create an information stream, and all the means of conveying, displaying, and disseminating that information. In essence, the information domain is the means by which a stimulus is recognized and conveyed to a human or to a human organization…
>
> The cognitive domain is the locus of the functions of perceiving, making sense of a situation, assessing alternatives, and deciding on a course of action. This process relies partially on conscious reasoning, the domain of reason, and partially upon sub-conscious mental models, the domain of belief. Both reason and belief are pre-conditioned by culture, education, and experience.

It is clear from these definitions that the physical domain contains the entities E that one needs to represent; the information domain is where external models M are displayed and disseminated; and the cognitive domain is where the models are made sense of – where the interpretants I exist. In both Smith [58] and Alberts *et al.* [2], the relationship between the domains is seen to be a flow from the physical domain to the cognitive domain via the information domain. Alberts *et al.* [2, pp. 12-13] establish this flow, and then use it to motivate the central importance of information:

> With the exception of direct sensory observation, all of our information about the world comes through and is affected by our interaction with the information domain. And it is through the information domain that we communicate with others (telepathy would be an exception). Consequently, it is increasingly the information domain that must be protected and defended to enable a force to generate combat power in the face of offensive actions taken by an adversary. And, in the all important battle for Information Superiority, the information domain is ground zero.

Disregarding the reference to telepathy and the sales speak, what is Alberts' claim? Direct sensation is claimed to be an exception to the usual flow of understanding from the physical world to the cognitive domain, via the information domain. This description, along with the accompanying diagram – reproduced in Figure 1.2, conjures up visions of the information domain as populated by automated sensors collecting, fusing and disseminating data unaided by human cognition and judgement. Smith is again more cautious, describing the cognitive domain as the locus where data is interpreted and decisions are made, not the information network. However, he maintains the same connections from the physical domain to the information domain, and the information domain to the cognitive domain. This is most explicit in the layered diagrams Smith uses to depict the domains, with the physical domain layer at the bottom, the information domain layer in the middle, and the cognitive domain layer on top.

Interestingly, the three domain model is derived[7] from Fuller's [24] book *The Foundations of the Science of War*, in which Fuller described a trinity between

[7] In [25], Garstka notes that "A key element of the model is a focus on three domains: the physical domain, the cognitive domain, and the information domain. This conceptual model

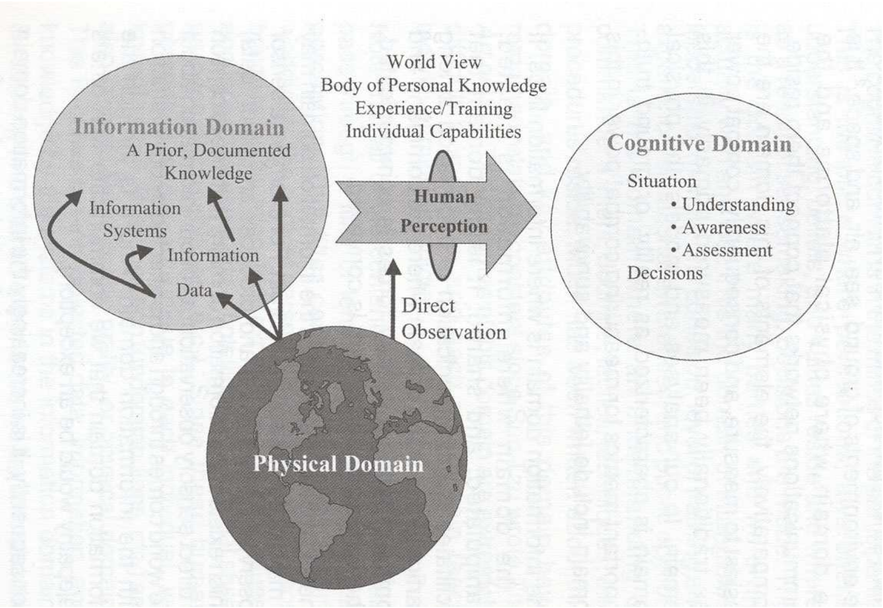

Figure 1.2: The three domains of warfare, after Alberts *et al.* [2].

the three spheres of man. Because they are both based on a triadic relation, the structure of Fuller's three sphere theory of warfare is structurally analogous to Peirce's theory of signs. However, in the modern day interpretation of Fuller, the triadic relation between the domains of warfare has been reduced to two dyadic relations: the physical–information and information–cognitive relations. This intended "refinement" of Fuller's conceptual model has only concealed the essential nature of representation as an irreducible triadic relation.

1.4.4 Representation in systems theory

In his book on multidisciplinary thinking, Kline [35, p. 16] notes three uses of the word 'system' within science, which are relevant to the role of external representations. So that they can be compared, he gives them three separate labels. The first conception, the most common use by scientists outside the systems community, is:

Definition 3 (System (1)) *The object of study, what we want to discuss, define, analyse, think about, write about, and so forth.*

Kline refers to this understanding with the label '**system**', which for example could refer to the solar system, a communicating vessel, an ecosystem, or an

builds upon a construct proposed initially by J.F.C. Fuller in 1917, and refined in *Measuring the Effects of Network-Centric Warfare*."

operating system. In fact, according to Kline, a **system** can be anything, as long as there is a well defined boundary associated with the **system**. In this thesis, I use 'system of interest' to denote this meaning of system. In Peirce's triadic relation, the system of interest corresponds to the entity E.

The second usage is defined as:

Definition 4 (System (2)) *A picture, equation, mental image, conceptual model, word description, etc., which represents the entity we want to discuss, analyse, think about, write about, etc.*

Kline coins the term 'sysrep' to mean representations of systems. Sysreps are "one of three basic types of representation: words, pictures and mathematics": that is, sysreps are models M. The types of representations Kline identifies correspond to the categories of language, analog and mathematical models I proposed in Section 1.4.1, except that pictures are only one of several possible analogs. According to Kline, the ideal aim of a sysrep is to perfectly mirror a **system**, where "[b]y 'perfect mirror' we mean not only that the sysrep will fully represent each and every characteristic of the **system** with total accuracy, but also that it will represent nothing more" [35, p. 18]. The common – but misguided – conception of representation as a perfect mirror is critically examined in Section 1.5.2.

The third usage is the most general conception, which is consistent with attempts to define the meaning of system within the systems community:

Definition 5 (System (3)) *An integrated entity of heterogeneous parts which acts in a coordinated way.*

Kline uses the label 'system' or 'systemic' for this conception, where a systemic property is an emergent property, which is a property of the whole but not a property of the components of the system.

The final concept Kline invokes is a schemata, which denotes "all the ideas in a person's head which are used to represent and interact with the world" [35, p. 31]. Example schema include words, relational ideas, behavioural routines and medical diagnosis. Kline then answers the question: "What is the relation of a sysrep to schemata in the mind? A sysrep is a particular kind of schemata, a very special class of the totality of the schemata we construct in our minds." Kline defines non-mental representation as a special class and an extension of mental representation. This approach is problematic, because mental representation is actually more difficult to understand than external representation. It makes more sense to explain mental representation in terms of the more straightforward case of external representation, even if mental representation precedes external representation from a chronological view.

In Kline's view, the relation between entities, models and their interpretants is as follows. Scientists view the world as being comprised of **systems**, which are interpreted using mental schemata. Schemata enable complex interactions with the world, but are formed using largely non-conscious mechanisms, and may be fuzzy and unstructured. When we go to the trouble of making a schemata explicit and shared in a structured social environment, it becomes a sysrep (which must still be interpreted by people). The goal of forming sysreps is to mirror the **system**, so that ideally the behaviour of the sysrep and the **system** are identical. While Kline is right to distinguish between systems, system representations, and system interpretations, the details of how they interact are not consistent with Peirce's triadic relation.

Burke [14] has formalised and refined the systems approach to understanding representation, in a clearly articulated conceptual model. He offers the following definitions for entity, system, system description and model [14, pp. 9-12]:

Definition 6 (Entity) *An entity is any object that has existence in the physical, conceptual or socio-cultural domains.*

Definition 7 (System (4)) *A system is an idealisation of an entity as a complex whole.*

Definition 8 (System description) *A system description is a representation of a system.*

Definition 9 (Model) *A model is an idealisation and/or representation of an entity.*

Four implications follow from these definitions. Firstly, because systems are idealisations of entities, they are abstractions that have no physical existence [15]. Systems are not part of the furniture of the world, they only exist inside minds. Stated another way, a system is a way of looking at the world [64]. Secondly, an entity can be idealised as a system in multiple ways: there is no unique systems view for any entity. Thirdly, and most importantly for this discussion on representation, both systems and system descriptions are considered to be models by Burke. The difference is that external models (a system description) presuppose the existence of a corresponding idealisation (a system). This is equivalent to requiring that external models M require an interpretant I in order to represent an entity E. Therefore, Burke's system theoretic interpretation is consistent with Peirce's triadic relation for external representation. Fourthly, Burke defines system descriptions to be derived from systems (idealisations of entities), rather than directly from entities. This implies that the system description can only capture aspects of the entity that have already been captured in the system. Consequently, a system description can be interpreted as a system that has been further abstracted from the entity it represents.

1.4.5 Summary of external representation

I will conclude the discussion of external representation by comparing the distinctions that have been identified above in different disciplines. The most notable commonality is that in each case, exactly three categories have been necessary to explain external representation, and furthermore these categories can be aligned with the entities, models and interpretants of Peirce's triadic relation. Of course, this has more to do with the selective nature of my literature survey than uniformity of approach. Descartes' [21] dualism was unconcerned with external representations, while Rosen's [53] Modeling Relation between the formal systems of science and the natural systems they represent attempted to explain external representations without explicit reference to the human mind or interpretants. Nevertheless, each of the approaches I have covered supposes that things are naturally considered as belonging either to the physical, the mental, or the social products of the mental. The physical world contains the entities that one would like to represent, external models are social products that can be shared, but they must be interpreted by someone or some agent to count as a representation.

There is an important way in which the domains in military theory differ from the accounts of representation by Popper, Kline and Burke. Interactions between the physical domain and the cognitive domain are mediated by the information domain. In contrast, the other accounts explain external models as products of the human mind. Physical entities must be conceptualised before they can be externally represented. Because militaries functionally separate the collection of information from decision-making, the role of human conceptualisation in information collection that mediates between the physical and information collection is easily ignored. But without human intervention and judgement there is only data, not representation or information, and automation can reduce but not eliminate human participation in constructing representations[8]. In view of Peirce's triadic relation, each of the alternative accounts considered aspects of this relation, but none are as comprehensive as Peirce's theory of signs.

External models have been variously held to be: abstract products of the human mind; information bearing artifacts; the socio-cultural environment; a specially precise subset of mental representation; a mirror that reflects part of the world; and a mental representation reduced by additional simplifying assumptions, which is explicit and shared. However, most of these assertions are not entirely accurate. Definitions, such as Kline's, that attempt to explain external representations with respect to mental representation are not enlightening, because the cause is more complicated than the effect. Peirce's typology of iconic, indexical and symbolic pure forms, and my categories of formal, analog and linguistic models, provide a framework for understanding external representation, which is sufficiently general to account for representation across disparate disciplines. Within this framework, an external model is most accurately conceived of as a grounded representation bearer external to the agent who interprets the model. Less formally, an external model is an equation, analog or description that represents something for an agent and thereby modifies its behaviour.

[8] This is a point that Polanyi [48, p. 20] makes well, and an example is the automation of the photo-finish for horse races, which still required human judgement in a case where one horse was fractionally in front, but the other extended further past the finish line due to a thick long thread of saliva coming from the horse's mouth. It would seem that such semantic ambiguities cannot be satisfactorily resolved by syntactic processors.

1.5 Internal Representation

Representation plays an important explanatory role in biology. From the perspective of a living agent, the world contains limited essential resources of energy and matter for survival and reproduction, as well as threats to survival such as predators and other harmful energy sources. It is easy to see that the ability to sense qualities of the immediate environment and to control locomotion with context sensitive behaviour confers a significant relative selective advantage. Bacteria that follow a chemical or light gradient can be viewed as performing very basic representation: chemical reactions triggered by the local environment stand for greater expected concentrations of non-local useful energy which cannot be directly detected. An agent that can sense distal features of its environment, using passive or active sensors to detect patterns of incoming energy such as light photons or sound waves, can secure an even greater selective advantage. Whereas proximal sensory information requires an agent to 'bump' into a threat before it can react to it, an agent that can sense a threat at a distance can avoid the threat entirely.

However, distal information is noisy, incomplete and intermittent. Just because a predator becomes occluded by vegetation does not secure the safety of its prey. Current sensory input alone is inadequate for determining the best action in any context. By constructing an internal representation of its environment, an agent can continue to act appropriately in the absence of direct sensory stimuli.

This story of representation in biology is inspired by the accounts of Dennett [20, pp. 177-182] and O'Brien and Opie [42], which suggest that representation is the problem that the brain is intended to solve. There is some empirical support for this conjecture in the form of the sea squirt *Ciona intestinalis*. The tadpole larva has a central nervous system of about 330 cells that controls locomotion. Once it attaches to a permanent object, it undergoes a metamorphosis that has been loosely described as eating its own brain (the cerebral ganglion is broken down and reused), since it no longer needs sensorimotor control, and therefore has no need to represent its environment.

Given this story, one may ask how the brain represents. This question has generated the most sophisticated conversation about internal representations, and has been especially preoccupied with the human brain. The Representational Theory of Mind, or representationalism, dates back at least to Aristotle [47]. The proposed answers of contemporary cognitive science divide into three main camps. They are Good Old Fashioned Artificial Intelligence (GOFAI), also known as symbolic, classical, or conventional cognitive science; connectionism; and the dynamical systems hypothesis.

When the representation is internal to the agent, one is faced with the question of what interprets the model. If internal models are interpreted in the same way as external models, then this leads to infinite regress, because the interpretant is also an internal model that requires its own interpretant, and so on. Von Eckardt [63, p. 283] describes two alternative resolutions to the infinite regress problem Peirce considered, and relates these to analogous moves in contemporary cognitive science.

The first solution is to weaken the definition of interpretant, to be a *potential* rather than an *actual* interpretant. The regress still consists of an infinite series of representations, but it is now easier to reconcile the associated interpretants, because they do not need to actually exist. This solution is reiterated by Cummins in cognitive science.

The second solution is "to find something that can function as an interpretant but which is not, itself, also representational and therefore in need of interpretation" [63, p. 283]. Peirce suggests that the only candidate for this is a habit-change. Specifically, Von Eckardt argues it must be a modification in the tendency to act in ways dependent on the content of the representation. The habit-change does not need to affect external behaviour; changes to mental habits (processes that generate other internal representations) also count. However, in order to eventually curtail the regress, internal models must ultimately be interpreted by shaping the agent's external behaviour. A very similar solution is suggested by Dennett, which Von Eckardt claims is the widely endorsed solution in cognitive science. Further, Von Eckardt [63, p. 290-302] shows in detail how this solution can handle the regress problem. Briefly, this involves demonstrating that:

- Interpretant I of model M is producible by M; and

- I is related to both the agent A and M, such that by means of I the content of M can make a difference to the internal states or the external behaviour of A towards the entity E.

Von Eckardt establishes this is the case for both conventional (symbolic) and connectionist machines. I will now provide a short introduction to GOFAI and connectionism, the two strongest advocates of representationalism.

1.5.1 Representationalism

> *How can a particular state or event in the brain represent one feature of the world rather than another? And whatever it is that makes some feature of the brain represent what it represents,* how does it come *to represent what it represents?*
>
> Daniel Dennett

These are the questions of representationalism, a position that assumes that the mind is a representational device, and that the brain has representational content. They are exceptionally difficult questions, because the mechanisms behind brain functions such as learning, memory and computation in the brain are currently poorly understood. Consequently, the mechanisms underlying representation are equally opaque. Also, under almost any metric, the human brain rates as the one of the most complex entities studied in science[9]. For a deeper discussion of representationalism than I can afford here, see Cummins [17].

As is the case for most enduring themes of Western philosophy, the first records of representational theories of mind are found in the writings of Aristotle [4]. In Book III, part 4, Aristotle describes the part of the human soul that thinks and judges: $\nu o\hat{\upsilon}\varsigma$ or the mind. According to Aristotle, the mind is "capable of receiving the form of an object; that is, must be potentially identical in character with its object without being the object." This statement clearly demonstrates Aristotle's use of the distinction between a model and the entity it represents. By form, Aristotle refers to the properties of the object, as opposed to its material substance. In Aristotle's metaphysics, the immaterial mind knows something when it takes on the form of that object, such that it represents the object in virtue of their similarity, in exactly the same way that a picture can represent a scene (Peirce's iconic grounds). "To the thinking soul images serve as if they were contents of perception . . . That is why the soul never thinks without an image." Berkeley [9] and Hume [33, 34] both extended this Aristotelian conception to argue that all mental contents are images in the mind, and that they are representations in virtue of their resemblance to perception. The inherent weakness of basing mental content on similarity can be seen by probing the mechanisms that could imbue mental images with the same properties as the objects they represent. Images presented to an immaterial mind are not so much an explanation as a metaphor, where thinking is like putting on a theatre for the Eye of the Mind.

In contrast, Hobbes [27, Chapter V] and Leibniz [37] advanced the idea that everything done by the mind is a computation. In this view, thought proceeded by symbolic manipulations analogous to the additions and subtractions of the new calculating devices – in modern parlance the mind was seen as an "automatic formal system" [26]. Notably, this reframed the question of representationalism to propose a mechanical and material explanation of mental processes. This provided a crucial step towards a science of cognition, because it opens up the possibility that certain features of cognition could be reproduced artificially.

The link between computation and representation is important but subtle. Because of the universality of Turing's conceptual model of digital (symbolic) computation, it is a common assumption that all computation is equivalent to a Universal Turing Machine. However, as O'Brien and Opie [43] correctly point out, this does not account for analog computation. They propose a definition of computation in general, which is broad enough to capture both analog and digital computation, but still sufficiently constrained to differentiate computation from the vast majority of physical systems – intestines, microwave ovens, cups of tea, etc. – that are not involved in computation.

> [T]here are two distinctive features of computational processes (as opposed to causal processes in general). First, they are associated with representing vehicles of some kind. Second, and more import-

[9] See Bar-Yam [7] for estimates of the complexity of the brain compared with other systems.

antly, computational processes are shaped by the contents of the very representations they implicate. We thus arrive at the following characterisation:

> *Computations are causal processes that implicate one or more representing vehicles, such that their trajectory is shaped by the representational contents of those vehicles.*

This characterisation of computation makes explicit the link between computation and representation. Computations are those processes involving representing vehicles (models), such that the outcome of the process depends on the content of the model. Representation is inherent in computational processes, and computation is the mechanism that causally links the contents of models to changes in the behaviour of the agent that interprets the model. Representation and computation are a package deal: a commitment to a computational theory of mind entails a commitment to representationalism.

Although conceived in the 17$^{\text{th}}$ century, it was not until the mid 20$^{\text{th}}$ century that the computational idea rose to prominence. The initial hype associated with the AI movement had a profound impact on 20$^{\text{th}}$ century cognitive science, such that computational theories of mind were predominantly based on algorithmic symbol manipulation. The Universal Turing Machine [61] provided a theoretical basis for universal symbol-based simulators of human intelligence, while exponential increases in computing power dramatically expanded the application of computer algorithms towards focussed engineering tasks that had previously required the application of the human mind.

Yet simulations that could be confused with intelligent humans have not materialised. AI researchers began to hit some fundamental walls: general intelligence appeared to require fast, situated, unencapsulated reasoning, where automated formal systems were slow, abstract, and only capable of manipulating the initial axioms they were given according to fixed rules. Coinciding with a growing dissatisfaction with the ability of the products of AI to live up to expectations, several alternatives have been advanced within cognitive science, and symbolic computational models of cognition began to be referred to as GOFAI.

Connectionism, or Parallel Distributed Processing (see for example [39]), which is based on highly abstract networks of artificial neurons, presents an alternative paradigm for modelling cognition, which can be interpreted as performing analog computation. Connectionist models are inspired by current understanding of the architecture of the brain, and are described by Dennett [20, p. 269] as blazing the first remotely plausible trails of unification between the mind sciences and the brain sciences. Different kinds of connectionist networks have been shown to have content addressable memory [30]; provide universal function approximation [31]; degrade gradually when damaged; and distribute representations across the set of connection weights, which decouples representations from individual symbols. Due to their parallel processing, connectionist networks are also very fast. The theoretical model of connectionism, an artificial neural network with real connection weights, has been proven to be capable of hypercomputation [57] – that is, able to compute functions that Turing machines cannot. Of course, such machines are not practical, since real numbers in general require infinite information, and there are also a number of issues artificial neural network implementations suffer from. They are almost always simulated on a digital computer, which implies these instantiations are equivalent to Turing machines; they learn reliably only under supervision; and they are usually treated as black boxes, because their behaviour is not currently well understood. Of course, there are philosophical concerns as well. For example, Fodor and Pylyshyn [23] criticise connectionism because it cannot explain systematicity: the feature of human cognition whereby the ability to think one thought entails the ability to think of numerous logically related thoughts, such as its converse.

The most recent alternative to both GOFAI and connectionism is the dynamical systems hypothesis [62]. However, advocates of the dynamical systems account are often explicitly critical of explanations involving representation, so a discussion of dynamical systems is deferred to Section 1.5.2 on anti-representationalism.

In summary, representational theories of mind have been proposed that are based on symbolic manipulation and analog covariance. GOFAI and connectionism agree that mental contents can stand in for, and stand in relation to real world objects. They also assume that psychological processes are computations that represent aspects of the external world.

1.5.2 Anti-representationalism

> *There is no harm in saying of good tools and good moves that they are also good representations, but nothing interesting is conveyed by this choice of idiom, and its employment should not tempt us to construct theories about how representation works.*
>
> Richard Rorty

Accounts of representation in cognitive science and artificial intelligence have been criticised as a basis for biological behaviour on a number of fronts. Brooks [12] summarises one key idea against representation in the physical grounding hypothesis: the world is its own best model, the trick is to sense it appropriately and often enough. The first part is true but uninteresting, because it is the trivial case where the representation relation degenerates into a diadic relation between A and $E \equiv M$. The second part is important, because it emphasises the need for agent decisions to be grounded. However, the physical grounding hypothesis does not and can not dispose of representation entirely. Even if the world is used as its own model, the agent needs to interpret the meaning of its observations. Constructionist accounts of vision (see for example [45]) argue that the process of perception involves significant construction by the observer. In their critique of a simplistic but common conception of "pure vision" – essentially the idea that the visual system is a bottom-up hierarchy designed to fully mirror the visual scene – Churchland *et al.* [16] provide an alternative account that they label interactive vision. Some of the constructive characteristics of interactive vision are: visual fields are highly non-uniform; vision is exploratory and predictive; the motor system and the visual system are entangled; sensory processing is more like a recurrent network than a hierarchy; and vision cannot be neatly separated from other brain functions. Consequently, the process of sensing the world appropriately is in fact one of the major sources of representational activity in the brain [44]. Also, in Section 1.3 I gave a number of reasons why internal representations can be convenient, even if they are not perfect substitutes for unmediated access to the real world.

Van Gelder [62] denies that cognition involves computation or representation, advancing an alternative dynamical systems hypothesis. In this account, rather than interpreting cognitive states as symbols, they are treated as quantifiable states of a nonlinear dynamical system. Van Gelder uses the Watt governor, depicted in Figure 1.3, to illustrate his thesis. The Watt governor is a mechanical device that maintains a constant speed for a flywheel despite fluctuations in both steam pressure from the boilers and the engine workload. The Watt governor was a pivotal invention during the industrial revolution that allowed the generation of reliable, smooth and uniform power. The Watt governor works because a spindle is geared into the flywheel such that the spindle rotates proportionally to the speed of the flywheel. The faster the spindle rotates, the more centrifugal force it generates, raising the spindle of the flywheel. Because the spindle is directly linked to the throttle valve, the faster the spindle rotates, the higher its arms rise, the more the valve is closed, restricting the flow of steam. As the speed of the flywheel decreases, so too does the spindle, the arms fall, opening the valve and increasing the flow of steam. Thus, a steady state for the speed of the flywheel exists and the Watt governor maintains the steady state by exerting negative feedback on any deviation from the steady state.

Van Gelder compares this mechanical device, which he classifies as a dynamical system, with a hypothetical computational device capable of performing the same function. The computational device would follow an algorithm that depends on representation.

> The very first thing it does is measure its environment (the engine) to obtain a symbolic representation of current engine speed. It then performs a series of operations on this and other representations, resulting in an output representation, a symbolic specification of the alteration to be made in the throttle valve; this representation then causes the valve adjusting mechanism to make the corresponding change [62, p. 350].

In contrast, the mechanical device is non-representational. Van Gelder gives four reasons: representation is not needed to fully account for the operation of the Watt governor; the obvious correlation between arm angle and engine speed is not representational because representation is more than mere correlation; the simple correlation only obtains in the steady state; and the arm angle cannot

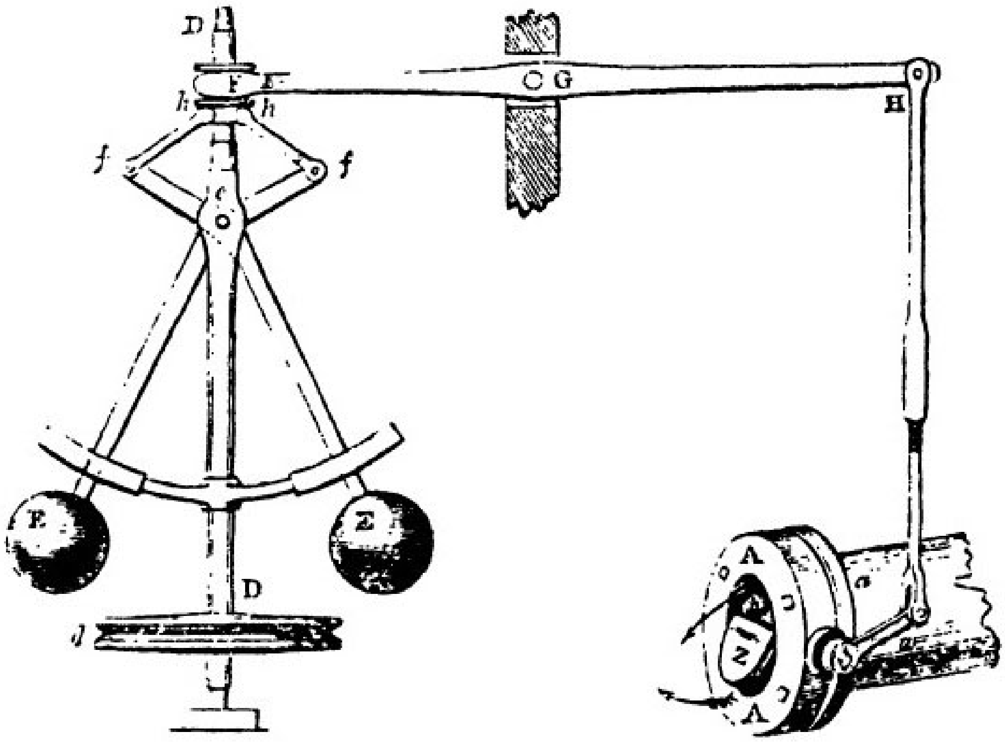

Figure 1.3: The Watt centrifugal governor for controlling the speed of a steam engine, after [22] as reproduced in [62].

represent engine speed because the two quantities are coupled.

Of these, the first three reasons are not persuasive. Just because an explanation of the Watt governor within some frameworks do not need the concept of representation does not imply that representation *cannot* be used to explain the same process. After all, none of the compound objects – such as spindles and throttles – are necessary concepts in the quantum mechanical wavefunction of a Watt governor. The second point does nothing to disprove representation occurs, it merely demands a higher standard of proof than demonstrating correlation, while the third point only notes that any correlation is not simple.

The forth reason is the most interesting. Van Gelder observes that "the angle of the arms is at all times determining the amount of steam entering the piston, and hence at all times both determined by, and determining, each other's behaviour." Because of this circular causality, Van Gelder claims that representation is "the wrong sort of conceptual tool to apply". When representation is thought of as a mirror, it does indeed seem wrong for the mirror to determine any part of the mirrored entity, because there is an asymmetry in their relationship. However, under the conception of representation as a triadic relation, it is *necessary* for the model to change the behaviour of an agent, and *possible* for the agent to be acting upon the represented entity. Peirce's triadic relation does not preclude the formation of feedback loops, although it does provide an incomplete explanation for such tightly coupled variables as the arm angle and engine speed.

The important criticisms of both Brooks and van Gelder are directed at the cognitive science community's early preoccupation with explicit symbolic representation. However, Section 1.3 demonstrates that representation in general can have iconic and indexical – not just symbolic – grounds. Brooks' situated robots do not do away with representation altogether – they actually encode significant amounts of their behaviour symbolically on finite state machines. Dynamical systems, as advocated by Van Gelder, can still function as representations with iconic grounds. The analog model in Section 1.4.1 is one such example. Rather than undermining representation, these critiques serve to highlight differences between formal systems and other possible bases for biological representation.

Maturana and Varela's [38] ground-breaking second order cybernetics approach to the biological basis of cognition is also critical of representationalism, which they claim is inadequate for a scientific explanation. They use an analogy reminiscent of Searle's [55] Chinese room argument to claim that living systems do not represent [38, p. 136]:

> Imagine a person who has always lived in a submarine. He has never left it and has been trained how to handle it. Now, we are standing on the shore and see the submarine gracefully surfacing. We then get on the radio and tell the navigator inside: 'Congratulations! You avoided the reefs and surfaced beautifully. You really know how to handle a submarine.' The navigator in the submarine, however, is perplexed: 'What's this about reefs and surfacing? All I did was push some levers and turn knobs and make certain relationships between indicators as I operated the levers and knobs. It was all done in a prescribed sequence which I'm used to. I didn't do any special maneuver, and on top of that, you talk to me about a submarine. You must be kidding!'

This analogy works by specifying an overly narrow system boundary. The adequacy of the navigator in avoiding the reefs cannot be explained unless the boundary is expanded to include the process that generated the prescribed sequence of knob turns and lever pushes. Specifically, the person in this example only becomes a navigator once they have been *trained*. But then it is easy to see that the precise purpose of training the navigator in the sequence of actions is *to stand in for* observations of the reefs and the depth below sea level, and thereby modify the submarine's behaviour.

A more serious threat to representationalism is anti-representationalism, which has been advocated by Davidson, and even more forcefully by Rorty [51].

Anti-representationalism holds that any statement about the world is an inseparable cohabitation of subject and object, rather than correspondence between an object and a model. Rorty rejects the 'mirror' metaphor of knowledge, where knowledge is a reflection of the mind-external world. According to Rorty this metaphor, which we have already seen used explicitly by Kline above, is the central metaphor for representationalism. Rorty criticises what he calls the Aristotle-Locke analogy of knowledge to perception,

> . . . the original dominating metaphor as being that of having our beliefs determined by being brought face-to-face with the object of the belief (the geometrical figure which proves the theorem, for example). The next stage is to think that to understand how to know better is to understand how to improve the activity of a quasi-visual faculty, the Mirror of Nature, and thus to think of knowledge as an assemblage of accurate representations. Then comes the idea that the way to have accurate representations is to find, within the Mirror, a special privileged class of representations so compelling that their accuracy cannot be doubted. These privileged foundations will be the foundations of knowledge, and the discipline which directs us toward them—the theory of knowledge—will be the foundation of culture. The theory of knowledge will be the search for that which compels the mind to belief as soon as it is unveiled. Philosophy-as-epistemology will be the search for the immutable structures within which knowledge, life, and culture must be contained—structures set by the privileged representations which it studies. The neo-Kantian consensus thus appears as the end-product of an original wish to substitute *confrontation* for *conversation* as the determinant of our belief [51, p. 163].

It should be noted that Rorty's attacks are not directly focussed on the cognitive science debate on mental contents, which the previous critiques have participated in. Rorty, who is trained in analytic philosophy, is more concerned with structuralist and linguistic attempts to ground knowledge as representation. For example, Rorty [52] cites Brandom's [11] characterisation of the representationalist school as saying that "the essential feature of language is its capacity to represent the way things are." Proponents of this school are taken to be Frege, Russell, Tarski and Carnap, who are contrasted with Dewey, Wittgenstein and Sellars, who view language as a set of social practises. However, it was noted in Section 1.4.1 that there is no necessity for language to be representational. Further, Rorty's target is much larger than just the philosophy of mind – he seeks to question the legitimacy of transcendental (Kantian) epistemology as distinct from psychology, and advocates that the only constraints on knowledge are essentially conversational in nature. Rorty rejects the very idea of a theory of knowledge, truth or rationality.

Is it possible to salvage representationalism from these attacks? Rorty seeks an *a priori* defeat of representation, but there is an empirical component to the question of whether minds represent. Languages are not reflections of reality, in the sense that they are not mirrors that can be polished to provide a True representation of the world. However, it does not immediately follow that formal systems, analogs or minds can not represent in any meaningful way.

A stronger defence of representationalism can be made by carefully articulating what work the representation relation needs to perform. From the perspective of an agent faced with a decision, consider two different processes for choosing between the alternatives. For concreteness, suppose the agent is a frog near the edge of a cliff choosing whether to jump forward to the left or right. One choice is safe but the other choice will result in certain death. Using the first process, suppose that the frog, like one of Brooks' situated robots, is able to make its decision by using the world "as its own best model". Brooks' idea of sensing the world often enough is similar to Van Gelder's claim that dynamical coupling and feedback can do the same job as representation, without requiring extensive planning or computation. In the case of the frog, it is a pretty simple and efficient process: both alternatives are sensed and the apparently less perilous alternative is immediately acted upon. The process is memoryless[10], so it can be repeated every time the frog lands. It can work when several conditions are met: if the frog can sense at least as far as it leaps; if the sensory comparison is reliable; and if the environment is sufficiently stationary while the frog is airborne.

If the latter condition is violated, nothing can guarantee the safety of the frog if it continues to leap. However, by using an alternative representational process, the frog may stand a chance even when the first two conditions do not hold. Suppose that it is a dark and foggy night, so that the frog cannot reliably sense the relative merits of jumping left and right. Fortunately, the frog has taken this path many times before, and remembers the sequence of left and right jumps that have got it home safely in the past. Then recognition of the starting location and recollection of this sequence can substitute in the decision-making process for sensing the alternatives at every step. The sequence can act as a model which when interpreted by the frog can stand in for the currently unreliable sensory information. In this process, representation must still be grounded by having previously sensed the alternatives using the first process. However, it can no longer be memoryless. By maintaining an internal model, the frog amplifies the applicability of sensory information beyond immediate local sense-response reflexes, to affect behaviour non-locally in space or time. Representation allows the agent to do more with less sensory information, to fill in gaps and to generalise new information. In other words, the real work that representation is doing is inference.

Rorty is right to reject the metaphor of representation as a mirror, reflecting the nature of reality for Descartes' Eye of the Mind. But representation is not a mirror, its purpose is not to *reflect* but to *infer*. In this section, I have argued that

[10] I should clarify that I mean memoryless in the mathematical (Markovian) sense: the probability of future states only depends on the current state. This is significantly more abstract and general than the meaning of memory in cognitive science.

perception is to some extent constructed, which involves representational activity. I have shown that representation can have non-symbolic grounds, meaning dynamical systems can be a basis for representation. I have examined Maturana and Varela's argument against representation, which relies on an overly narrow definition of system boundary. Finally, I have argued that Rorty's attack on representation is largely directed towards language as representational, and to representation as mirroring. These critiques have merit, but do not challenge the general theory of representation provided by Peirce's triadic relation as an explanation of internal representation.

1.6 Agents

So far I have not specified what I mean by an agent. However, the preceding discussion on representation offers a precise way of characterising agency. Because of the ubiquity of agents in complex systems, this section contains the most important implications of this chapter.

As Peirce has argued, representations require an interpretant, and therefore an agent to perform the interpretation. Thus, there is a sense in which representations (and computations) are relative to a subject – that is, they are subjective [56, p. 92]. But there is an equally objective way that a subject plus a model either does, or does not, represent. By redrawing the system boundary to include the model's user, representation is an intrinsic feature of this system of interest. This is the importance of the triadic relation: because it is irreducible, the system boundary must always extend to include the agent in order to understand representation.

This is why, unless one can identify who is using the model and how it shapes their behaviour, the model cannot be considered to be a representation. Peirce, Von Eckardt, Popper, Smith, Kline and Burke all identify the 'who' with a human mind, which is more generally the case in the literature. I have generalised this to say that a model is always a model employed by an agent. An agent can be a person, but it can also be a group of people, an animal, a cell, a certain kind of robot or a certain kind of physical process (after all, each member of this list is a physical process). The ability to form and use representations appears to be the principle difference[11] between an object and an agent – it is the difference between kicking a rock and kicking a cat (not that either experiment is condoned, except as a thought experiment).

More formally, if an entity's response to a stimulus is directly determined by its current state, and the current state does not include any models, then the entity is not an agent. If stimulus and response are indirectly related because they are mediated by representation, then the entity is an agent.

[11] For example, Aristotle [4] says "The soul of animals is characterized by two faculties, (a) the faculty of discrimination which is the work of thought and sense, and (b) the faculty of originating local movement." In my account, representations both encode distinctions and shape movement or behaviour.

Definition 10 (Agent) *An entity that constructs and uses representations to shape its own goal-directed behaviour.*

More will be said about goal-directed behaviour below, but for now note that goal-directed behaviour does not imply that agents only have a single goal: it is merely intended to distinguish between directed and undirected behaviour. It seems there is a continuum, such that entities may have a degree of agency, depending on how indirect the relationship between stimulus and response is, and how sophisticated the representations can become, which is often called the plasticity of the representing medium. I am not overly concerned about the precise demarcation between agent and non-agent. The definition is more useful for comparative purposes, in order to investigate if the degree of agency has increased, and to say that a human has more agency than a cockroach, which has more agency than a virus, which has more agency than the robot Cog [13], which has more agency than a cyclone[12].

The degree of autonomy of an agent refers to freedom of choice or variety, which is made precise by the notion of source coding in information theory. The degree of autonomy is evident in the sensitivity of changes in the behaviour of the agent to changes in its representations. For example, if an agent's model is replaced by any other model (such as its inverse) and yet this has no causal influence on the behaviour of the agent, then the model does not contribute to the autonomy of the agent. If this holds for all models, then the agent is not autonomous. A model must shape behaviour to be a representation and provide the agent with autonomy. In contrast, if any arbitrary desired feasible state within the agent's environment can be achieved by changing only the agent's representations, which then realise the desired state by modifying the agent's behaviour, the agent has maximum autonomy. When autonomy is shared between two or more agents, this is the subject of game theory, and the degree of autonomy of a player is the number of available strategies[13], and any mixed strategy on this set constitutes a model.

The autonomy of agents can lead to philosophical debate about free will and teleology. In view of Hume's compatibilism [33, 34], the autonomy of an agent does not imply the agent is necessarily nondeterministic – that with *exactly* the same internal and external states, two distinct responses to the same stimulus are possible. Instead, a weaker condition holds, namely given different representations, an agent is capable of choosing different actions. To confuse the matter, often it is useful to explain the behaviour of a system as an autonomous agent, even when it is clearly not purposive. For example, Dawkins [18] describes genes as selfish molecules, as if they have minds, which is a form of teleonomic

[12] Most people do not consider a cyclone to be an agent, even though it is a self-maintaining, non-equilibrium entity with unpredictable behaviour. The anti-cyclonic Great Red Spot on Jupiter is a structure with a diameter significantly greater than Earth, which has persisted since it was observed by Cassini in 1665. I believe the main reason cyclones are not considered agents is because it is not possible to sustain an interpretation of either goal-directed behaviour or representation for a cyclone and although unpredictable, they are not autonomous.

[13] This game theoretic interpretation assumes the set of strategies is countable, where only strategies that affect the value of the game are counted.

explanation. For my purposes, I will always assume that agency entails a degree of autonomy in Hume's sense, and also implies that the agent is capable of exhibiting goal-directed behaviour.

The notion of goal-directed behaviour has been formalised by Sommerhoff [59], who observed that the essence of goal-directed activity is:

> that the occurrence of the goal-event G is *invariant* in respect of certain initial state variables ($\mathbf{u}_0$) *despite* the fact that G depends on action factors and environment factors that are *not* invariant in respect of $\mathbf{u}_0$. The invariance of G being due to the fact that the transitional effects of changes in $\mathbf{u}_0$ mutually compensate, so to speak.

Sommerhoff realised it was possible to treat goal-directedness as an objective property, independently of the subjective notion of purposiveness of interest to the psychologist. He established three necessary and sufficient criteria for goal-directed behaviour. Firstly, for at least one variable **a** associated with the action, and one variable **e** associated with the environment, for at least one time t_k,

$$F(\mathbf{a}_k, \mathbf{e}_k) = 0. \tag{1.1}$$

This ensures that the action is capable of compensating for environmental variability. Secondly, **a** and **e** must be mutually orthogonal, meaning that the value of one of the variables does not determine the value of the other for the same instant. This allows the mechanism for goal-directed behaviour to realise Equation (1.1) for a range of initial conditions. And thirdly, there must be a set S_0 of initial environmental conditions (where $|S_0| \geq 2$), such that each initial condition requires a unique action $\mathbf{a}_k$ which satisfies Equation (1.1). This criterion ensures that the goal could have been achieved from an ensemble of initial conditions, rather than only from the actual initial conditions. $|S_0|$ provides a measure of the degree of goal-directed behaviour: the greater $|S_0|$ is, the more environmental variety the agent can destroy and still achieve its goal. Thus, goal-directed behaviour is underpinned by Ashby's [5, 6] law of requisite variety.

From a stimulus-response perspective, an agent can be thought of as sensing stimuli and acting to produce an appropriate response. The function that maps from sensory inputs and models to output actions is its decision map. The sense, decide and act functions of agents are roughly analogous to detectors, rules and effectors in Holland's [28] complex adaptive systems terminology; perceptual, cognitive and motor components in cognitive science; input (actuating signal), control unit (dynamic element), and output (controlled variable) in control theory; state, policy and action in reinforcement learning [60]; stimulus, organism and response in Hull's [32] version of behavioral psychology; state, mixed strategy and move in game theory; and observe, orient/decide and act (OODA) in Boyd's [10] decision cycle.

The sense, decide, act trinity is a pervasive characterisation of agency that can be related back to Peirce's triadic relation. Without sensing, there is no way to ground representations. Without acting, representation cannot shape external behaviour. Without deciding, representations cannot be interpreted, the agent cannot be autonomous, and its behaviour is not goal-directed. A necessary and sufficient condition for agency is the possession of sense, decide, and act functions. But this is exactly equivalent to requiring that an agent be able to construct and use representations to shape goal-directed behaviour.

In summary, representation and agency have been co-defined. Representations always involve an agent, and agents always represent their environment. The triadic nature of the representation relation is the reason that these definitions cannot be separated. Due to this intimate relationship, a theory of representation is essential to an understanding of the behaviour of agents in a complex system.

Bibliography

[1] Abbott, R., "If a tree casts a shadow is it telling the time?", *Unconventional Computation, UC2006* (York, UK,) (C. S. Calude, M. J. Dinneen, G. Paun, G. Rozenberg, and S. Stepney eds.), Springer (2006).

[2] Alberts, D. S., J. J. Garstka, R. E. Hayes, and D. A. Signori, *Understanding Information Age Warfare*, CCRP Publication Series (2001).

[3] Anderson, P. W., "More is Different", *Science* **177**, 4047 (1972), 393–396.

[4] Aristotle, *De Anima*, transl. J. A. Smith, eBooks@Adelaide, originally published 350BC Adelaide, Australia (2004).

[5] Ashby, W. R., *An Introduction to Cybernetics*, Chapman & Hall London, UK (1956).

[6] Ashby, W. R., "Requisite variety and its implications for the control of complex systems", *Cybernetica* **1** (1958), 83–99.

[7] Bar-Yam, Y., *Dynamics of Complex Systems*, Westview Press Boulder, Colorado (1997).

[8] Bergson, H., *Creative Evolution*, transl. A. Mitchell, Macmillan London (1911).

[9] Berkeley, G., *A Treatise concerning the Principles of Human Knowledge*, eBooks@Adelaide, originally published 1710 Adelaide, Australia (2006).

[10] Boyd, COL J. A., "A Discourse on Winning and Losing" (1987).

[11] Brandom, R., "Truth and Assertability", *Journal of Philosophy* **73** (1976), 137.

[12] Brooks, R. A., "Elephants Don't Play Chess", *Robotics and Autonomous Systems* **6** (1990), 3–15.

[13] Brooks, R. A., C. Breazeal, M. Marjanovic, B. Scassellati, and M. M. Williamson, "The Cog Project: Bulding a Humaniod Robot", *Computation for Metaphors, Analogy and Agents*, (C. L. Nehaniv ed.). Springer Berlin, Germany (1999).

[14] Burke, M., "Robustness, Resilience and Adaptability: Implications for National Security, Safety and Stability (Draft)", *Tech. Rep. no.*, DSTO, (2006).

[15] Checkland, P., *Systems thinking, systems practice*, John Wiley and Sons Chichester UK (1981).

[16] Churchland, P. S., Ramachandran V. S., and T. J. Sejnowski, "A Critique of Pure Vision", *Large-Scale Neuronal Theories of the Brain*, (C. Koch and J. L. Davis eds.). A Bradford Book, The MIT Press Cambridge, USA (1994).

[17] Cummins, R., *Meaning and Mental Representation*, The MIT Press Cambridge, USA (1989).

[18] Dawkins, R., *The Selfish Gene*, Oxford University Press Oxford, UK (1976).

[19] Dennett, D. C., *Brainstorms: Philosophical Essays on Mind and Psychology*, MIT Press Cambridge, USA (1978).

[20] Dennett, D. C., *Consciousness explained*, Little Brown and Co. Boston, USA (1991).

[21] Descartes, R., *A Discourse on Method: Meditations and Principles*, transl. J. Veitch, J. M. Dent and Sons London (1912).

[22] Farey, J., *A Treatise on the Steam Engine: Historical, Practical, and Descriptive*, Longman, Rees, Orme, Brown, and Green London, UK (1827).

[23] Fodor, J. A., and Z. Pylyshyn, "Connectionism and Cognitive Architecture: A Critical Analysis", *Cognition* **28** (1988), 3–71.

[24] Fuller, J. F. C., *The Foundations of the Science of War*, Hutchinson and Co. Ltd. London (1925).

[25] Garstka, J. J., "Network Centric Warfare: An Overview of Emerging Theory", *Phalanx* **33**, 4 (2000), 1–33.

[26] Haugeland, J., *Artificial Intelligence: The Very Idea*, MIT Press Cambridge, USA (1985).

[27] Hobbes, T., *Leviathan*, University of Oregon Oregon, USA (1651).

[28] Holland, J. H., *Hidden Order: How Adaptation Builds Complexity*, Helix Books, AddisonWesley Reading, USA (1995).

[29] Hoopes, J. ed., *Peirce on Signs: Writings on Semiotic by Charles Sanders Peirce*, University of North Carolina Press Chapel Hill, USA (1991).

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

[31] Hornik, K., M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators", *Neural Networks* **2**, 5 (1989), 359–366, 70408.

[32] Hull, C., *Principles of Behavior*, Appleton-Century-Crofts New York, USA (1943).

[33] Hume, D., *A Treatise of Human Nature*, Reprint Oxford University Press, 1978 Oxford, UK (1741).

[34] Hume, D., *An Enquiry Concerning Human Understanding*, Ed. P. H. Niditch, Reprint Clarendon Press, 1975 Oxford, UK (1777).

[35] Kline, S. J., *Conceptual Foundations for Multidisciplinary Thinking*, Stanford University Press California, USA (1995).

[36] Korzibski, A., *Science and Sanity*, The International Non-Aristotelian Library (1948).

[37] Leibniz, G. W., *Dissertatio de Arte Combinatoria*, Leipzig, Germany (1666).

[38] Maturana, H., and F. Varela, *The Tree of Knowledge: The Biological Roots of Human Understanding*, transl. J. Young, Shambhala Publications, originally published 1988 Boston, USA (1984).

[39] McClelland, J., and D. Rumelhart eds., *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, MIT Press Cambridge, USA (1986).

[40] McLaughlin, B., and K. Bennett, "Supervenience", *The Stanford Encyclopedia of Philosophy*, (E. Zalta ed.). The Metaphysics Research Lab (2006).

[41] O'Brien, G., "Connectionism, Analogicity and Mental Content", *Acta Analytica* **22** (1998), 111–131.

[42] O'Brien, G., and J. Opie, "Notes Toward a Structuralist Theory of Mental Representation", *Representation in Mind: New Approaches to Mental Representation*, (H. Clapin, P. Staines, and P. Slezak eds.). Elsevier Science (2004).

[43] O'Brien, G., and J. Opie, "How do connectionist networks compute?", *Cognitive Processing* **7**, 1 (2006), 30–41.

[44] Opie, J., "Personal communication" (2006).

[45] Palmer, S. E., *Vision Science: Photons to Phenomenology*, MIT Press Cambridge, USA (1999).

[46] Penrose, R., *Shadows of the Mind*, Oxford University Press Oxford, UK (1994).

[47] Pitt, D., "Mental Representation", *The Stanford Encyclopedia of Philosophy* (E. Zalta ed.), (2005).

[48] Polanyi, M., *Personal Knowledge: Towards a Post-Critical Philosophy*, Routledge London, UK (1962).

[49] Popper, K. R., *Objective Knowledge*, Clarendon Press Oxford, UK (1972).

[50] Popper, K. R., "Three Worlds", *The Tanner Lecture on Human Values* (The University of Michigan,), (1978).

[51] Rorty, R., *Philosophy and the mirror of nature*, Princeton University Press Princeton (1979).

[52] Rorty, R., "Representation, social practise, and truth", *Objectivity, Relativism and Truth: Philosophical Papers Volume 1.* Cambridge University Press Cambridge, UK (1991).

[53] Rosen, R., *Anticipatory Systems: Philosophical, Mathematical & Methodological Foundations*, Pergamon New York, USA (1985).

[54] Saussure, F. de, *Course in General Linguistics*, transl. W. Baskin, The Philosophical Library, originally published 1916 New York, USA (1959).

[55] Searle, J. R., "Minds, Brains and Programs", *Behavioral and Brain Sciences* **3** (1980), 417–424.

[56] Searle, J. R., *Mind*, Oxford University Press Oxford, UK (2004).

[57] Siegelmann, H. T., *Neural Networks and Analog Computation: Beyond the Turing Limit*, Springer Verlag (1999).

[58] Smith, E. A., *Effects Based Operations: Applying Network Centric Warfare In Peace, Crisis, And War*, CCRP Publication Series (2002).

[59] Sommerhoff, G., "The Abstract Characteristics of Living Systems", *Systems Thinking: Selected Readings*, (F. E. Emery ed.). Penguin Books Harmondsworth, UK (1969).

[60] Sutton, R. S., and A. G. Barto, *Reinforcement Learning: An Introduction*, A Bradford Book, The MIT Press Cambridge (1998).

[61] Turing, A. M., “On Computable Numbers, With An Application To The Entscheidungsproblem”, *Proceedings of the London Mathematical Society* **42**, 2 (1936-7), 230–265.

[62] Van Gelder, T., “What might cognition be, if not computation?”, *Journal of Philosophy* **92**, 7 (1995), 345–381.

[63] Von Eckardt, B., *What Is Cognitive Science?*, A Bradford book, The MIT Press Cambridge, USA (1993).

[64] Weinberg, G. M., *An Introduction to General Systems Thinking* Silver Anniversary ed., Dorset House Publishing New York, USA (2001).