Continuous Computation and the Emergence of the Discrete[*]

Bruce J. MacLennan
Computer Science Department
University of Tennessee, Knoxville
MacLennan@CS.UTK.edu

March 24, 1994

> Over many years I have searched for the point where myth and science join. It was clear to me for a long time that the origins of science had their deep roots in a particular myth, that of *invariance*.
>
> — Giorgio de Santillana

1 INTRODUCTON

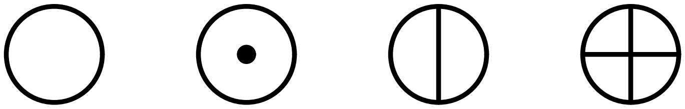
In this paper I'll address the emergence of the discrete from the continuous, first in mythology and psychology, then in cognitive science and artificial intelligence. This will provide a context for considering some continuous neural processes that can result in (approximately) discrete behavior.

Many traditional cosmologies begin with a separation of the primordial *massa confusa* into opposites. For example, Euripedes said,

> And the tale is not mine but from my mother, how *Ouranos* (Sky) and *Gaia* (Earth) were one form; and when they had been separated apart from each other they bring forth all things, and gave them up into the light: trees, birds, beasts, the creatures nourished by the salt sea, and the race of mortals. (fr. 484, *Melanippe the Wise*)

[*] Based on an invited presentation, *2nd Appalachian Conference on Behavioral Neurodynamics: Processing in Biological Neural Networks* (October 3–6, 1993), Radford University, Radford, VA. To appear in *Origins: Brain & Self-Organization* (tentative title), edited by Karl Pribram, Hillsdale, NJ: Lawrence Erlbaum.

This separation, which is often associated with the creation of recognizable *things*, corresponds to the emergence into the world of the faculty of *discrete categorization*. With definite properties come definite things, and conversely the definiteness of things seems to depend on definite oppositions. Early philosophers enumerated many oppositions (e.g., hot/cold, dry/wet, light/dark, straight/curved, odd/even), but in myth they are often equated to the opposition most salient to all people: male/female. Thus the primordial separation creates a god and a goddess, whose subsequent union creates the world of things (though they retain their separate identities). In an editor's forward [30] Alan Watts observes that this process is reflected in an ancient series of images:



The images represent, respectively, the undifferentiated matrix, the male seed in the cosmic womb, polarization into opposites, and the cosmos as a complex system of polarities.

Discrete categories are very comfortable, for they bring a sense of security: everything either *is* or *isn't*. This is captured by Aristotle's Law of the Excluded Middle, which applies to every space that is topologically discrete: there is *nothing* in the middle, between the points of the space, and there are no matters of degree; two points are either identical or as different as they can be. It is significant that the Greek word *chaos* originally referred to a *gap* and, specifically, to the primordial gap between earth and the heavens [10]; and indeed the categorizing mind finds only chaos in the grey areas between categories.[1] The Pythagoreans were quite explicit about the role of the gap in creating discrete things, including the integers. Stobaeus reported that they said:

> The void (*kenon*) distinguishes the natures of things, since it is the thing that separates and distinguishes the successive terms in a series. This happens in the first instance in the case of numbers; for the void distinguishes their nature. (DK 58B 30)

Another traditional theme in many common accounts of creation is the efficacy of the spoken word in beginning the creation of things; this perhaps reflects the close connection between language and the conceptual faculty.

There is also an old tradition that wisdom comes with the transcending of categories. This is a familiar theme in many Eastern traditions, such as Taoism, but we also find it in the West. For example, Heraclitus recognized that categories are often relative and contextual (DK 22B 9, 13, 37, 58, 59–61). He also recognized the importance of the continuum that unifies the opposition:

[1] In the Babylonian creation myth, the creation of the organized world does not begin until Enlil (the air god) creates a gap between An (Heaven) and Ki (Earth). In one scholar's reading of the text, the gap is created by an act of cognition, for it is effected by the god Mummu, whose name means "mind," "reason" or "consciousness" [29].

> The teacher of most is Hesiod; they are sure he knows most, who did not recognize Day and Night — for they are one. (DK 22B 57)

That is, Hesiod considered the goddesses Day and Night to be different — an archetypal opposition — and failed to acknowledge their underlying unity, for they are just the extremes of a cycle [32]. Heraclitus also stressed that the *harmonia* (structure) of the world results from a flow between opposites:[2] "through opposing currents existing things are made a *harmonia*" (Diogenes Laertius 9.7). Thus, from a deeper perspective, reality is a *structured flow* rather than a discrete structure. Furthermore, the wise realize that this law governing the macrocosm applies equally to people:

> Those speaking with sense need to rely on what is common to all, as a city on its laws — and with much more reliance. For all human laws are nourished by one, the divine.[3] (DK 22B 114)

The message seems to be that we do not have a simple dichotomy between a chaotic continuum and a structure of discrete parts. The third alternative is a *structured continuum* (*harmonia*), which result from a flow between opposites. This is the law of the macrocosm, but also the law of the microcosm, that is, the principle by which an organism's behavior can move in conformity with the universal process.

Similarly, in Jungian psychological theory, the process of conscious individuation results in a personality that is "firm in its flexibility" [4]. This process corresponds to the *coniunctio oppositorum* of spiritual alchemy, which is primarily a union of male and female, but symbolically a reunification of all the opposites [8]. Thus the evolution from an initial undifferentiated state to an inflexible discrete state is superseded by a flexible reintegration of the units that preserves their differentiation. Specifically, an individualized personality requires the unification of the subconscious mind, which is intuitive, synthetic and flexible, but opaque, with the conscious mind, which is transparent, rational and analytic, but rigid. The resulting "conscious spontaneity" combines the light of the conscious mind with the intuitive flow of the unconscious, and so in alchemy it is symbolized by the "fiery water" that results from a conjunction of elemental fire (illuminating but rigid and disintegrative conscious rational thought) and elemental water (veiled but flowing and integrative subconscious intuition).

In artificial intelligence, cognitive science and epistemology, we are presently in the progress of recapitulating this evolution. The half century that ended about 1980 was the heyday of discrete knowledge representation; it showed the capabilities of symbolic cognitive models, but also their limitations. Connectionist knowledge representation, as embodied in neurocomputation, promises to fill in the gaps of discrete knowledge structures,

[2] 'Harmony' is not the primary sense of ancient Greek *harmonia*, which refers to a continuous structure obtained by joining discrete parts.

[3] There is a three way pun in the Greek between "with sense" (*xun noōi*), "what is common" (*tōi xunōi*) and "laws" (*toi nomōi*) [9]. The nominalized adjectives in this translation reflect the (apparently) intentional ambiguities of Heraclitus' maxim.

providing flexibility while retaining differentiation and structure. That is, by allowing (approximately) discrete symbolic representations to emerge from continuous neurocomputational processes, connectionism will provide a basis for flexible symbolic processing.

In this paper I will discuss several simple, neurally plausible mechanisms by which approximately discrete structures can emerge from continuous computational processes. The goal is to better understand how brains manipulate approximately discrete symbols, as in language and logic, without losing the flexibility of the underlying neural processes.

2 CONTINUOUS COMPUTATION

2.1 What is Computation?

One might suppose that the answer to this question is obvious, but it becomes problematic when one looks for computation in places other than manufactured computers, such as in the brain, and one considers kinds of computation that are less familiar than digital computation, such as analog computation. Indeed, a forthcoming issue of the journal *Minds and Machines* is devoted to the question "What is Computing?" in the context of psychology and philosophy, and exhibits a variety of opinions on the matter. I'll briefly summarize my position, which is presented in more detail elsewhere [21, 26].

The meaning of "computation" is more apparent if we consider the use of the term before it became widely recognized that computers can manipulate nonnumerical data. Then, computation meant doing arithmetic by means of physical procedures. These procedures included the arrangement of written figures in tableaux (as in long division), the position of beads in an abacus or similar "digital" device, and the position of scales in a slide rule, nomograph or similar "analog" device. Here already we can see the essential characteristics of computation, which still hold.

First, the purpose of computing is to operate on *abstract objects*. Originally numbers were the only objects manipulated, but once it became apparent that computers could be used for nonnumerical computation, the manipulable abstract objects were extended to sets, sequences, tuples, Boolean (truth) values, formulas, and many other types of data. In spite of the fact that many of these data types are nonnumerical, they are *mathematical* in the sense that they are *precisely defined* and *abstract* (i.e., "formal" in the sense that they are nonmaterial).

Second, because abstractions do not exist physically, computation must be accomplished by the manipulation of physical surrogates. Examples of these surrogates are the written digits of manual arithmetic, the beads of an abacus, and physical position on a slide rule or nomograph. Even "mental computation" (e.g., mental arithmetic or algebra) displays this characteristic, though less obviously, since the digits or other signs are represented by physical states in the brain that are manipulated with little regard for their meaning (see following).

Third, computation is a mechanical procedure, which depends on reliably determinable physical properties of the surrogates (i.e. computation is "formal" in the sense that it depends on the form but not the meaning of the surrogate objects). For example, manual arithmetic can be carried out without concern for the meaning of the digits. This, of course, is the property that permits automatic computation, but even in the days before computing machinery, it allowed sophisticated calculations to be carried out by skilled, but mathematically ignorant, human "computers," who were organized in large scale parallel arrays for scientific computations and simulations.

Finally, we observe that throughout most of the history of computation, both discrete ("digital") and continuous ("analog") surrogates have been used. For example, in ancient Greece we find both the abacus (discrete) and the use of geometric devices and constructions (continuous). The recent history of computers, which has been dominated by discrete computation, biases us against continuous computation, but that is a serious mistake if our concern is computation in the brain (and probably even if our interest is limited to computer technology [12, 16, 17]).

I have argued elsewhere [26] that the difference between analog and digital computation has nothing to do with a supposed analogy, possessed by the former but not the latter, between the physical and abstract systems. On the contrary, both kinds of computation depend on a systematic correlation — an analogy — between the states and processes of the physical and abstract systems, as is explained below. For this reason I will use the terms *discrete* and *continuous* computation instead of "digital" and "analog" computation, which might be misleading.

In summary: *Computation is the physical realization of an abstract process for the purpose of manipulating abstract objects*.

2.2 Computation in General

Since computation accomplishes an abstract process through the medium of a surrogate physical process, it's necessary to say precisely what is meant by *process*. To stress the parallels between discrete and continuous computation, the definitions will be stated as far as possible in "topology-neutral terms," that is, in terms inclusive of both discrete and continuous computation.

The future behavior of a process is completely determined by its *input* (the independent variables) and its *state* (the dependent variables).[4] In the case of a discrete process the state comprises one or more variables with values chosen from discrete spaces (e.g., integers, Booleans or floating-point numbers). In the case of a continuous process, the state might be a discrete set of continuous-valued variables or a continuum of continuous values, as in a physical field (the latter actually being a special case of the former). The input to a process may be discrete or continuous in the same way as the state.[5]

The moment-to-moment behavior of a process is determined by a law that relates its

[4] For simplicity I'm restricting attention to *deterministic* processes; though nondeterministic processes are also important, they are not relevant to the present discussion.

[5] In principle, hybrid discrete/continuous systems are possible, but they are subject to restrictions imposed by other properties, such as continuity.

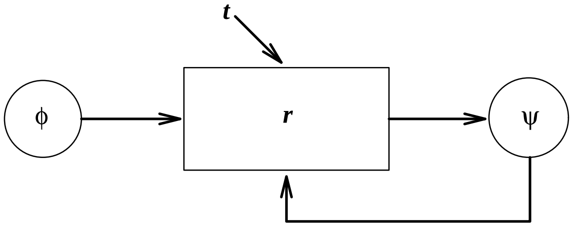


Figure 1: A Process

state changes to the present state and input; this law is itself permitted to change in time. Specifically, let $\psi(t)$ be the state and $\phi(t)$ be the input, both at time t. Then $\psi'(t)$, the new state at time t, is given by $\psi'(t) = r[t, \psi(t), \phi(t)]$, where the state transition function r is the law determining the process (Fig. 1). For a discrete-time process, the new state is given at the next time step $\psi'(t) = \psi(t+1)$; for a continuous-time process, the new state is given at the next instant, $\psi'(t) = \psi(t+\mathrm{d}t)$. A consequence of this is that the law r takes the form of (generalized) difference equations for a discrete process and (generalized) differential equations for a continuous process.[6]

Having said what a process is, we consider what it means for a physical process to realize an abstract process. In brief a realization is a homomorphism from the physical system to the abstract system. Roughly, a realization is a systematic correspondence between the states of the two systems and between the state changes of the two systems, such that the behavior of the abstract system is recoverable from that of the physical system. I'll explain this more carefully (Fig. 2).

Let $\mathcal{A}$ be an abstract (mathematical, formal) system with a transition operator $r : \mathbf{R} \times \mathcal{S} \times \mathcal{I} \to \mathcal{S}$, where $\mathcal{S}$ is the state space, $\mathcal{I}$ is the input space, and $\mathbf{R}$ is the time continuum. Similarly the physical system $\mathcal{P}$ has the transition operator $R : \mathbf{R} \times \tilde{\mathcal{S}} \times \tilde{\mathcal{I}} \to \tilde{\mathcal{S}}$. Then we say that $\mathcal{P}$ is a realization of $\mathcal{A}$ if there is a homomorphism $\mathcal{H} : \mathcal{P} \to \mathcal{A}$; a homomorphism is a mapping that preserves some, though not necessarily all, of the mapped system (see below for specifics). This is because an abstract system is an abstraction of reality, and so the abstract objects will in general have many fewer properties than real (physical) objects. Thus a homomorphism $\mathcal{H} : \mathcal{P} \to \mathcal{A}$ maps some of the physical properties of $\mathcal{P}$ into abstract properties in $\mathcal{A}$, but others, irrelevant to the computation, are ignored. To specify the homomorphism more precisely we must consider the equations of the two systems:

$$\psi'(t) = r[t, \psi(t), \phi(t)],$$
$$\Psi'(t) = R[t, \Psi(t), \Phi(t)].$$

The homomorphism $\mathcal{H} : \mathcal{P} \to \mathcal{A}$ actually comprises four homomorphisms, $\mathcal{H}_\mathrm{s} : \mathcal{S} \to \tilde{\mathcal{S}}$ on the state spaces, $\mathcal{H}_\mathrm{i} : \mathcal{I} \to \mathcal{I}$ on the input spaces, $\mathcal{H}_\mathrm{r}$ on the transition maps, and $\mathcal{H}_\mathrm{t}$ between time in the two systems, though usually the distinction is clear and the subscripts

[6] This applies even for states containing nonnumerical variables, since generalized difference equations can be defined over any discrete space, and in this sense any digital computer program is a set of generalized difference equations [13]. Generalized differential equations can be defined over Banach spaces.

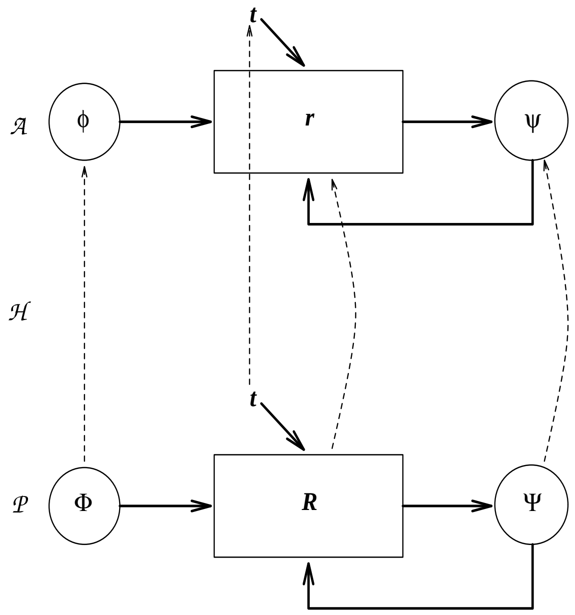


Figure 2: Physical Realization of an Abstract Process

are omitted. Thus we write $\psi = \mathcal{H}\{\Psi\}$, $\phi = \mathcal{H}\{\Phi\}$ and $r = \mathcal{H}\{R\}$. Figures 3 and 4 show two simple examples of realizations.

We can now state the condition that must be satisfied for the physical system $\mathcal{P}$ to realize the abstract system $\mathcal{A}$. Since a homomorphism preserves structure,

$$\psi' = \mathcal{H}\{\Psi'\} = \mathcal{H}\{R(t, \Psi, \Phi)\} = \mathcal{H}\{R\}(\mathcal{H}\{t\}, \mathcal{H}\{\psi\}, \mathcal{H}\{\phi\}) = r(\mathcal{H}\{t\}, \mathcal{H}\{\Psi\}, \mathcal{H}\{\Phi\}).$$

That is, the mapping is a homomorphism if

$$\mathcal{H}\{\Psi'\} = r(\mathcal{H}\{t\}, \mathcal{H}\{\Psi\}, \mathcal{H}\{\Phi\}). \tag{1}$$

Notice that in Eq. 1 time in the abstract system $\mathcal{H}\{t\}$ need not correspond to time in the physical system t; if they do, $t = \mathcal{H}\{t\}$, then we have a *realtime* system. If they are not the same, then we must distinguish the "simulated time" of $\mathcal{A}$ from the time required for its realization in $\mathcal{P}$. (However, we still require $\mathcal{H}_{\mathrm{t}}$ to be monotonic so that time proceeds in the same direction in the abstract system and its realization.)

2.3 Computation in Brains

I've suggested defining computation as the physical realization of an abstract process for the purpose of manipulating abstract objects. The phrase "for the purpose of manipulating abstract objects" is essential, since every physical process is the realization of some abstract process — in fact, of many, since any mathematical model of a physical process is realized by it. Therefore it is critical that the purpose of the physical manipulations be to realize

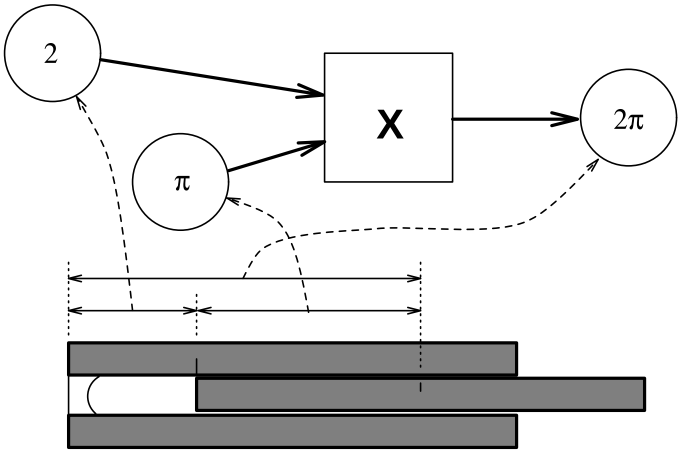


Figure 3: Realization of Continuous System

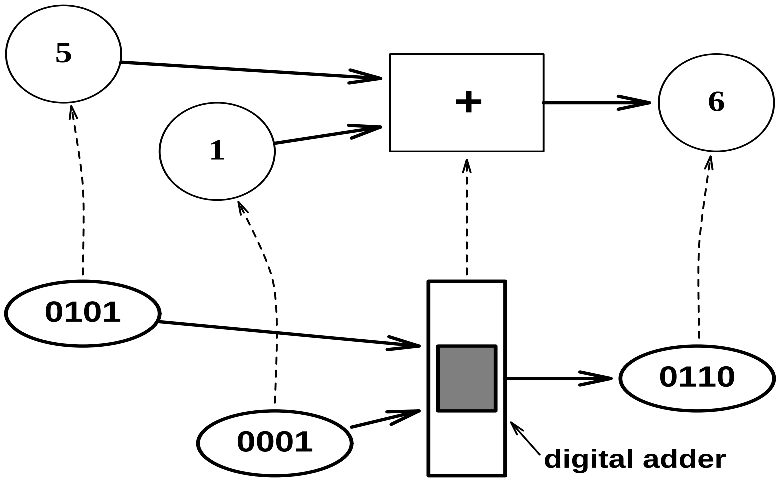


Figure 4: Realization of Discrete System

an abstract process. We may use a desktop computer for a space heater, a slide rule for a straight edge, or an abacus for a rattle, but these physical processes are not computation.[7]

By referring to purpose, the proposed definition of computation has been made inherently teleological. This is perhaps not a problem when we're dealing with manufactured devices, such as slide rules and desktop computers, for their purposes are often stated explicitly. On the other hand, if we are interested in *natural computation*, that is, computation in naturally occurring systems, such as the brain, then the teleological condition becomes problematic; certainly teleological definitions are suspect in the natural sciences.

Since my concern here is the brain, I can evade the general problem of teleological definitions and adopt a pragmatic solution, for in the context of biology the purpose and function of systems is often identifiable. We can objectively attribute purposes to hearts, lungs, stomachs, eyes without running the risk of inviting too many ghosts into our machines. So also for systems in the brain, and if a brain system can be shown to be accomplishing a basically mathematical task (i.e., an abstract process), then we may call it computational (contra Searle [26, 34]). While identifying computational brain systems is important to cognitive neuroscience, it is also central to psychological and philosophical debates about whether cognition is computation; for example, Searle has stressed that his Chinese Room Argument applies only to computational systems, and Harnad often reminds us that a robot cannot be purely computational, since it must interact with the physical world [5].

In fact I doubt there are many purely computational modules in the brain or elsewhere in organisms. Nature is far too opportunistic, it seems, to miss the chance of making a biological system serve multiple purposes.

2.4 Simulacra

A central theoretical construct underlies both the theory of (digital) computation and the use of symbolic (i.e. discrete) knowledge representation in artificial intelligence and cognitive science; it is the idea of a *calculus*, which is an idealized model of *discrete formal systems*. By isolating the essential characteristics of discrete information representation and processing systems, the idea of a calculus unifies investigations in all of these fields. I've argued elsewhere [19, 23, 24, 25] that connectionism lacks an analogous construct, and that this lack hinders our understanding of the range, limits and possibilities of connectionist knowledge representation. In an attempt to remedy this situation, I've proposed the *simulacrum* as a theoretical construct embodying the essential characteristics of *continuous* information representation and processing.

Just as a *calculus* comprises *formulas*, made of *tokens*, that are manipulated by *discrete processes* that can be defined by *rules*, so also a *simulacrum* comprises a *state*, made of *images*, that is transformed by *continuous processes* that can be defined by *graphs*. By "image" I mean, roughly, any element of a continuum, so for example, a single real num-

[7] This may seem a frivolous point, but the question of whether, for example, the jostling molecules in my wall could be said to be computing, arises in the context of philosophical debate, such as Searle's Chinese Room Argument [34, 35]. Putnam has used a similar argument [33].

ber is an image, as is a vector of numbers, or a *field* (a spatially continuous distribution of numbers) [11, 14], such as a visual image. (I do not intend, however, that this term be restricted to perceptual images; it includes motor images and various internal representations, of arbitrary abstractness and complexity, so long as they are drawn from a continuum.)

In mathematics, the *graph* of a function is the set of all its input-output pairs; in the simplest case it defines a curve in two dimensions, but it could also be a more complex surface or shape. It is analogous to a rule in that, in its most basic form, it defines an input-output relation by ostension, by showing the output resulting from each input. However, just as discrete computation systems allow rules to be abbreviated and combined in various ways, so also we may abbreviate and combine graphs for continuous computation.

We have identified a number of postulates satisfied by all simulacra, which we list here for completeness (detailed discussions and justifications can be found elsewhere [19, 24]):

Postulate 1 *All image spaces are path-connected metric spaces.*

Postulate 2 *All image spaces are separable and complete.*

Postulate 3 *Maps between image spaces are continuous.*

Postulate 4 *Formal processes in simulacra are continuous functions of time and process state.*

Postulate 5 *Interpretations of simulacra are continuous.*

For the purposes of this paper, an intuitive understanding of simulacra is sufficient.

3 INVARIANCES

3.1 Behavioral Invariances

As suggested by de Santillana's remark, which heads this paper, the concept of *invariance* is fundamental to science.[8] It is also the basis for our perception of a world of stable objects [31, Ch. 5], since the parts of an object retain an invariant relation to one another under transformations, such as rotation and translation, but their relations to the background are not invariant. Invariance also underlies category or concept formation, since we can say that an organism recognizes a category when it responds invariantly to all individuals belonging to that category.[9] For example, if a monkey, when hungry, always responds to a banana by eating it, then we can say that it recognizes the category *banana* (in the context

[8] As also suggested by his remark, invariance is a *myth*: that is, although it's not literally true, it contains an important truth.

[9] Mathematicians may be confused by my use of the term *category*, which, throughout this paper, will refer to a psychological construct also known as a *concept* or *universal*. Some of the objects discussed here will be mathematical categories, but that's coincidental.

of being hungry). Because the notion of invariance is so important, in this section I will characterize it in general terms, appropriate to continuous knowledge representation, and in the next section I'll consider several simple, biologically plausible mechanisms by which organisms can extract invariances from their environments.

I will be defining "invariance" in terms of the input-output behavior of a system. The goal of these definitions will be questions such as: In what ways can I transform the input to the system, but leave its output invariant? If I transform the input to a system in certain ways, is there a corresponding transformation of the output? The resulting definitions are *behavioral*, but they are not *behavioristic*. First, the system whose input-output behavior is in question need not be an entire organism. For example, we might consider the translation or scale invariance of a particular module in the visual system. Nevertheless, I think it is important to ground cognitive terms, such as *category* or *concept*, in the behavior of the whole organism in its natural environment (i.e. an ecological approach), and many of my examples will relate to an organism's behavior, since they are easier to understand. Thus I may take as an example of invariant behavior: "if a hungry monkey sees a banana (any banana) at location $\mathbf{x}$ in its visual field, then it responds by grabbing the banana at $\mathbf{x}$."

The preceding example illustrates that invariance may depend on the history, or prior state, of the system. (If the monkey isn't hungry, it may not grab the banana.) In the context of our general model of processes (Fig. 1), we must take account of invariance in the face of variation of the internal state as well as variation of the input. In these cases, the transition function of the process (r in Fig. 1) will be treated as a system without memory (state), since this simplifies the analysis, but does not limit its applicability to the more general case. Since in such a case the "input" to the transition function $S(\psi, \phi)$ comprises both the internal state of the system ψ and the input proper ϕ, I will use the term *cause* to refer to the current state/input pair (ψ, ϕ), since this comprises (by hypothesis) everything that can affect the future state of the system. Similarly, I will call the "output" of $S(\psi, \phi)$ its *effect*, since it comprises the new state as well as whatever external behavior (output proper) the system exhibits. Thus we will be considering invariances in maps $S : \mathcal{C} \to \mathcal{E}$ from a space $\mathcal{C}$ of causes to a space $\mathcal{E}$ of effects.

An invariance rarely holds over the entire space of causes to a system. For example, a hungry monkey may in general respond to a banana in a way that is invariant with the size of the banana image, but there must be limits to this scale invariance, since too small an image will not be visible, and too large an image will not be recognizable as a banana (the visual field will be filled with yellow). Thus we must in general specify some *range* $\mathcal{R}$ of causes over which an invariance holds. In actual practice it may be very difficult to describe the range of an invariance, but for theoretical purposes all we need to know is that the invariance holds over some $\mathcal{R} \subseteq \mathcal{C}$.[10]

[10] Furthermore, continuity requires that $\mathcal{R}$ have an indefinite boundary; that is, although the invariance holds absolutely within $\mathcal{R}$, it must also hold approximately for causes arbitrarily near to $\mathcal{R}$ but outside of it. This topic is addressed later.

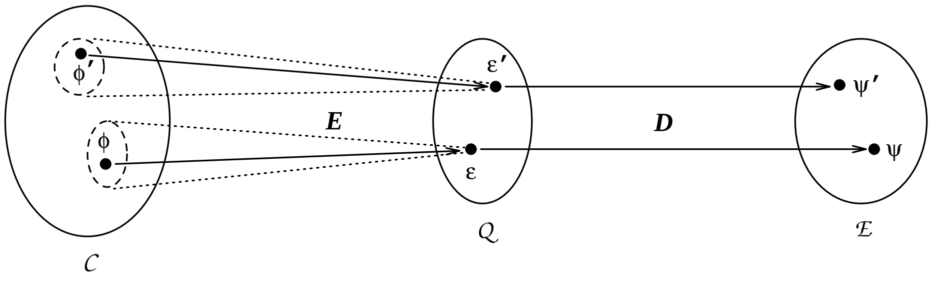


Figure 5: Factoring the Transition Map

3.2 Effect Equivalence

The simplest invariance is when certain changes to the cause leave the effect unchanged. For example, we might suppose that a mouse flees whenever it sees a cat, but that changes in the color or size of the cat do not alter its response. In formal terms, the cause may be changed from φ to φ' but the effect ψ remains the same, $S(\varphi) = \psi = S(\varphi')$. Looked at the other way, we can ask what class of causes results in a given effect. Thus, for each effect ψ there is a class of causes φ such that $S(\varphi) = \psi$ (in mathematical terms, this class in the *inverse image* of the effect, $S^{-1}[\psi]$), and so each effect (in the range under consideration) has an associated *effect-equivalence class* of causes. Thus there is a basic category structure observable in the behavior of every system (though some of the categories might be singleton, i.e., contain only one cause). The effect-equivalence classes represent categories of causes that differ only in behaviorally irrelevant properties (in the range under consideration).

By a well-known theorem in mathematics, any function $S : \mathcal{C} \to \mathcal{E}$ can be "factored" into a composition of two functions, $S = D \circ E$ (Fig. 5). The first function, $E : \mathcal{C} \to \mathcal{Q}$, maps causes into their effect-equivalent classes, or into any other space in a one-to-one relation with these classes (e.g., a space of "symbols" for the classes[11]). This intermediate space is called the *quotient space*, $\mathcal{Q}$. The second map, $D : \mathcal{Q} \to \mathcal{E}$, takes an effect-equivalence class (or its surrogate) into the corresponding effect, so $S(\varphi) = D[E(\varphi)]$. Because the equivalence map E eliminates all effect-irrelevant differences, the effect-map D establishes a one-to-one relation between equivalence classes and their effects (since, by construction, no two effect-equivalence classes have the same effect).

The effect-equivalence classes constitute the most basic invariance in the behavior of a system. It can, in principle, be determined by observing the cause-effect relationships of the system, though in practice we can get only an approximation, since determining the entire invariance would require observing an infinite number of cause-effect pairs. In principle, however, we do not have to look inside the system to discover this invariance. The quotient space would seem to be an ideal locus for understanding the emergence of

[11] Note, however, that this space of "symbols" must be a continuum (if it has more than one point), since it is the continuous image of a continuum (by the simulacra postulates). Thus any such surrogates for the effect-equivalence classes are not symbols in the familiar, discrete sense.

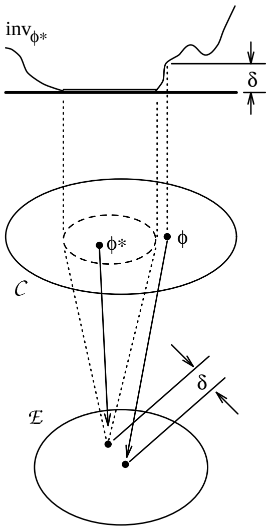


Figure 6: Indefinite Effect-equivalence

invariances through self-organization.

By definition, all causes within a effect-equivalence class lead to the same effect. However, since all maps on simulacra are required to be continuous (by postulates 3 and 4), we know that there are causes outside of the equivalence class whose effects are arbitrarily close to those inside (Fig. 6). Thus it is more accurate to think of effect-equivalence classes as having indefinite boundaries, and to think of invariance as a matter of degree rather than an absolute property. Thus:

Definition The *indefinite equivalence class* of a cause φ^* is:

$$\mathrm{inv}_{\varphi^*}(\varphi) = \delta[S(\varphi^*), S(\varphi)],$$

where δ is the metric (distance measure) on the effect-space.

Thus inv_{φ^*} will be 0 for all members of the (definite) effect-equivalence class of φ^*, for which the invariance is absolute, and will increase gradually as behavior becomes non-invariant outside the effect-equivalence class.[12] Quantifying the degree of invariance is important when addressing the emergence of invariant behavior through adaptation and learning, since the invariances appear gradually.[13] Nevertheless, the following discussion is couched in terms of absolute invariance, since the extension to degrees of invariance is straight-forward.

[12] Note that inv_{φ^*} is analogous to a fuzzy set, though for the usual fuzzy sets the membership function is 1 for full members and decreases to 0 as their membership decreases. There are a several simple ways, which are described elsewhere [25], for converting our indefinite equivalence class into a standard fuzzy set.

[13] Other traditionally absolute mathematical properties, such as being a group or being orthogonal, must also be redefined as matters of degree, so that we may, for example, discuss the gradual emergence of orthogonality or group structure.

3.3 Transform Invariance and Object Perception

An especially important kind of invariance occurs when a system responds the same to a stimulus in spite of it being transformed in some systematic way. For example, a banana may be recognized as a banana in spite of its image being rotated, translated and scaled. Also a melody may be recognized even though it has been changed in absolute pitch. Invariance under various kinds of transforms is the basis for our perception of stable *objects* in the world [31, Ch. 5]. For example, contextual information may allow a monkey to reliably select the larger of two objects regardless of their relative placement, and hence the relative size of their retinal images [38]. Such an instance of *size constancy* shows that the animal is perceiving absolute size, which is a property of the object itself as opposed to its appearance (sensory image). A complex of properties invariant under transformation is what constitutes a collection of phenomena as a perceptual *entity*. Different classes of transformations lead to different kinds of entities. For example, invariance under rigid body transformations, such as translations, rotations and scalings, leads to perception of physical objects. Invariance under absolute pitch transformation leads to perception of melodies, for a melody (as opposed to an absolute pitch sequence) can be defined as that entity that is invariant under change in absolute pitch.

We will consider transformations mapping causes into causes, which depend, in most cases, on a parameter that controls the change effected by the transform. For example, a rotation has a scalar parameter determining the angle of rotation, and a translation has a vector parameter determining the direction and distance of translation. I'll write $T_\alpha(\varphi)$ for a transformation, of type T (e.g. translation) and parameter α, applied to cause φ. In accord with postulate 3 of simulacra, $T_\alpha(\varphi)$ is required to be continuous in both α and φ. Observe that the parameter often varies in time, $\alpha = \alpha(t)$, as the organism moves, thus generating a continuous trajectory $T_{\alpha(t)}(\varphi)$, which is a rich source of information for separating entities from the environment [31, pp. 93–98]. *Variation is necessary to detect invariance.*

My present goal is to characterize transform invariant behavior, without any discussion of mechanisms for achieving it or by which it may emerge. Possible mechanisms by which a system could develop transform invariances, and thus come to perceive a class of entities, will be presented later ("Self-organization").

> **Definition** A system S is *invariant with respect to transformation* T, or T-*invariant*, whenever, for each allowable cause φ, the set of all allowable transforms of it, $T_\alpha(\varphi)$, are effect-equivalent for S (Fig. 7).

Therefore, for all allowable φ and α,

$$S[T_\alpha(\varphi)] = S(\varphi).$$

(As usual, invariances cannot in general be expected to be universal, so we must specify some range over which the invariance holds. Also, although a system may be simultaneously invariant under several kinds of transforms T_α, U_β, $\ldots$, for simplicity I'll consider just one.)

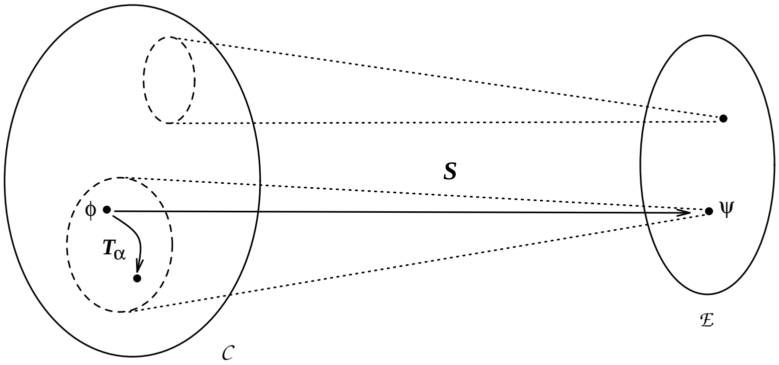


Figure 7: Transform Invariance

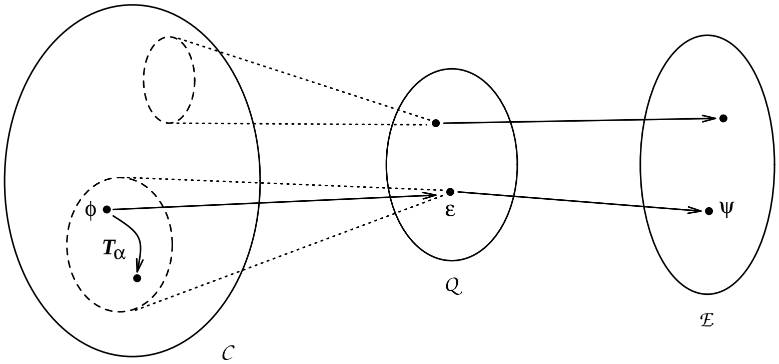


Figure 8: Factored Transform-invariant System

We can get additional insight into transform invariance by factoring S in the same way we did before (Fig. 8). The effect-equivalence classes $\varepsilon = E(\varphi)$ capture all aspects of the causes that are relevant to behavior, and discard all aspects that are not. This is accomplished by the equivalence map E which identifies all behaviorally-indistinguishable causes, thus abstracting only what is relevant to behavior, which is represented by the members of $\mathcal{Q}$. It remains for the second part of the system to map, one for one, these behaviorally relevant abstractions to their effects, $D(\varepsilon) = \psi = S(\varphi)$.

When this analysis is applied to a transform invariant system (Fig. 8), we see that, within the allowable range, E "projects out" all of the effects of the transforms, and passes on the constant entity representing all transformed causes. That is, if $\varepsilon = E(\varphi)$, then for all allowable α, $E[T_\alpha(\varphi)] = \varepsilon$. Thus the elements $\varepsilon \in \mathcal{Q}$ represent the entities recognized by the system: those and just those aspects of the causes that are constant under the transformations and affect the behavior of the system. For example, suppose a moderately hungry monkey responds to ripe fruit by smacking its lips, to unripe fruit by wrinkling its nose, and to other things by ignoring them. Then, for stimuli in the appropriate range, the space $\mathcal{Q}$ will represent only the fruitness and ripeness of the stimuli, since all other characteristics, including the location, size, orientation, and kind of the fruit, are discarded by the equivalence map E. As a result, an entity $\varepsilon \in \mathcal{Q}$ is an abstraction representing a particular degree of fruitness and ripeness.

Since the "entity space" $\mathcal{Q}$ is the nexus of the system's behavior, we might expect that this behavior can be described by rules expressed in terms of the entities $\varepsilon \in \mathcal{Q}$. This can be done, but only to a limited degree. To see the method, pick an abstract entity $\varepsilon \in \mathcal{Q}$. We know that the effect-map D yields a unique effect $\psi = D(\varepsilon)$ for this entity and for just this entity. Further suppose that, over some range, S is T-invariant; thus, if φ is any cause corresponding to ε, then we know that all of its allowable transforms, $T_\alpha(\varphi)$, also correspond to ε. As a consequence, the following rule expresses the behavior of S on all the transforms of φ:

$$T_\alpha(\varphi) \Rightarrow \psi.$$

We can interpret the rule as follows: "If the cause matches $T_\alpha(\varphi)$ for some α (in the appropriate range), then produce the effect ψ."

We cannot assume that the above rule expresses the behavior of the system S over the entire equivalence class leading to ψ, since S may be invariant with respect other properties besides T. However, if it is invariant to only T, or, equivalently, if all the members of the effect-equivalence class are mutually transformable via T, then the above rule completely captures the behavior of S on that equivalence class, and we call it *the rule corresponding to ε*. In this case the cause φ is said to *generate* the entire equivalence class.

Rules are normally expected to be finite in size, and this will be the case provided that the transform T and applicable range of parameters α are understood. Then the rule can be completely specified by exhibiting the particular concrete image φ, a generative cause, and ψ, the corresponding effect. For example, if we understand $\mathsf{rotate}_\theta(\varphi)$ to rotate φ counterclockwise by θ, then the behavior of mapping a square, in any orientation, to a circle, is completely and finitely specified by the rule:

$$\mathsf{rotate}_\theta(\square) \Rightarrow \bigcirc.$$

The question now arises of whether the behavior of a system that is only T-invariant can be completely specified by a finite sequence of rules of the form $T_\alpha(\varphi_k) \Rightarrow \psi_k$. Each such rule completely expresses the behavior of the system over an equivalence class, but it's easy to see that these rules cannot capture all of the system's behavior. This is because the quotient space $\mathcal{Q}$ is a continuum (since it is the continuous image of the image space $\mathcal{C}$, which is a continuum by postulates 1 and 2), and so there is an uncountable number of $\varepsilon \in \mathcal{Q}$. Since all these ε are necessary to the behavior of S, no finite number of rules, nor even a countably infinite number of rules, corresponding to these ε can suffice to express S. The rules accurately describe the behavior of the system over their corresponding equivalence classes, but they say nothing about its behavior on the rest of the space. On the other hand, just as an irrational number can be approximated arbitrarily closely by fractions, so also (by postulate 2) the behavior of S can be approximated arbitrarily closely by longer and longer lists of rules.

This result, which shows the irreducibility of a continuous process to rules, corresponds to observations of expert behavior, which cannot be completely captured by finite rules;

situations falling in the “cracks between the rules” are either unspecified, or must be filled in by some arbitrary — and ultimately inexpert — interpolation rule [3].

3.4 Structural Invariance

Our analysis so far has treated causes and effects as wholes. Therefore, the entities in the quotient space are likewise abstractions of the entire cause, which reflect behavior invariant with respect to any change in the cause. For example, if we think of the system as an animal, then a cause φ comprises all of the stimulus and the animal’s internal state, and the corresponding entity $\varepsilon = E(\varphi)$ is treated as an indivisible abstraction of the stimulus + state, which is the unique abstraction corresponding to its effect.

Although this approach captures the holistic character of an animal’s response to the current situation, it’s often illuminating to analyze behavior in terms of one or more components of the cause. For example, we might take the stimulus to comprise two figures, a banana and a tree, and the ground on which they appear. Then we would consider invariances with respect to changes in the particulars of the banana as an independent component of the cause. Likewise, we might analyze the internal state as a feeling of hunger on a background of other internal perceptions, and consider invariances with respect to just this component of the state. We would expect such behavior to be describable in terms of abstractions, such as *banana* and *hunger* corresponding to the components of the cause, and abstract relations among these components. In other words, the cause may be analyzed into a number of components in some relation to one another, in which case the entities and effects might be similarly analyzed. In this section, I’ll consider the general approach to these *structural invariances*.

Suppose $\beta \in \mathcal{F}$ is concrete image, say, a banana image, and $C : \mathcal{F} \to \mathcal{C}$ is a map that puts a figure on a specific ground, so $C(\beta)$ is a complete cause, which includes β as a component. Factor the system map as before, $S = D \circ E$, where $E : \mathcal{C} \to \mathcal{Q}$ and $D : \mathcal{Q} \to \mathcal{E}$. Now define $E_{\mathcal{F}} : \mathcal{F} \to \mathcal{Q}$, a map from figures to entities, by $E_{\mathcal{F}} = E \circ C$. Each entity $E_{\mathcal{F}}(\beta)$ abstracts all and just the characteristics of the figure β that, in the context provided by C, influence the effect; the corresponding equivalence class of particular figures is $E_{\mathcal{F}}^{-1}[\varepsilon]$ (Fig. 9).

Notice that the same space $\mathcal{Q}$ accommodates the abstractions of the figures in $\mathcal{F}$ and the causes in $\mathcal{C}$. This is because, for a fixed ground or context, the 1-1 relation between abstract figures and effects corresponds to the 1-1 relation between abstract causes and effects. Since the entities in $\mathcal{Q}$ are completely abstract, they may be thought of as abstract figures or abstract causes, as we like, though the figure equivalence map $E_{\mathcal{F}}$ may not be capable of generating all the entities in $\mathcal{Q}$. For example, suppose that, in a given context, a certain monkey will smack its lips at a banana but wrinkle its nose at a mango. Then, if $C(\beta)$ is the operation of putting a particular piece of fruit in this context, then the entities in $\mathcal{Q}$ generated by the figure equivalence map $E_{\mathcal{F}}$ correspond to the abstractions *banana* and *mango* (and all the borderline cases in between), and the corresponding effect-equivalence classes $E_{\mathcal{F}}^{-1}[\varepsilon]$ in $\mathcal{F}$ are classes of particular pieces of fruit. These same entities in $\mathcal{Q}$

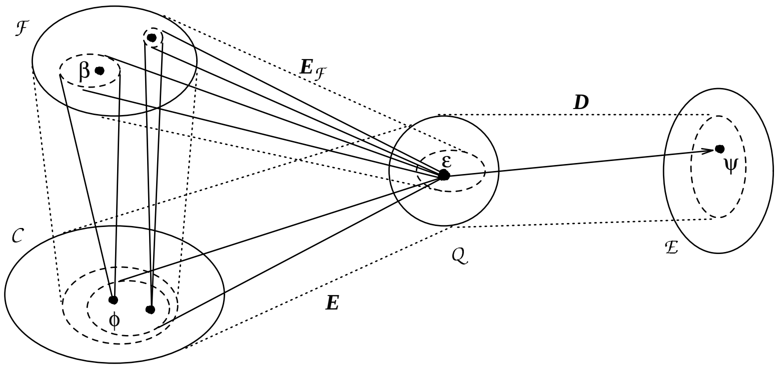


Figure 9: Structural Invariance with One Component

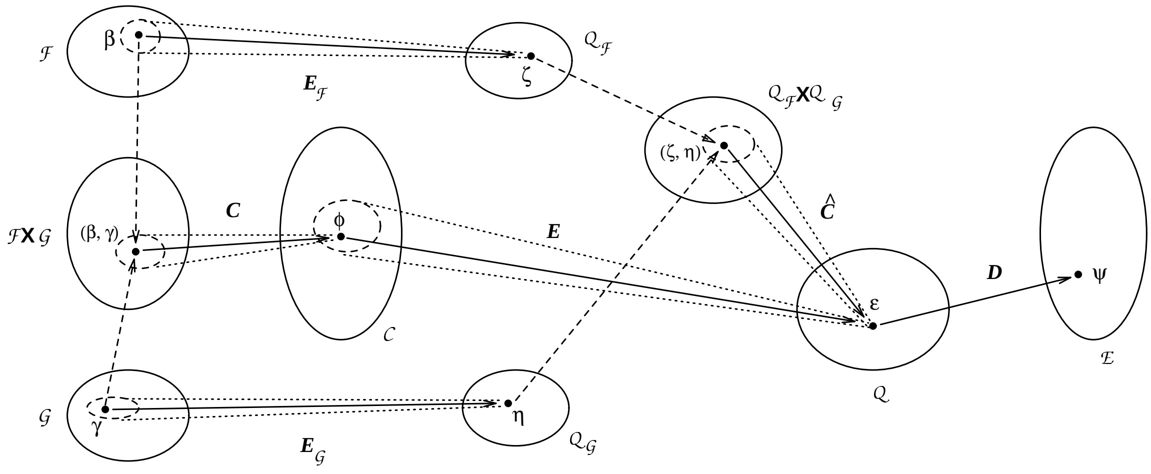
may be viewed as abstractions of complete causes including the fruit as a component, and the corresponding effect-equivalence classes $E^{-1}[\varepsilon]$ in $\mathcal{C}$ are classes of complete particular causes. There may also be entities $\varepsilon \in \mathcal{Q}$ with corresponding equivalence classes $E^{-1}[\varepsilon]$ of causes that are not results of $C(\beta)$, and so do not have corresponding classes of figures.

Commonly, structural invariances involve two or more components of the cause, for example, behavior may be invariant to changes in several figures, or to changes in the background as well as the figures. We'll consider the case of structural invariances involving two components, and the general case will be clear.

Suppose that $\beta \in \mathcal{F}$ and $\gamma \in \mathcal{G}$ are two images, which might correspond to two components of a cause, or to a figure and ground within a cause. Next consider any function $C : \mathcal{F} \times \mathcal{G} \to \mathcal{C}$, which assembles the components β and γ into a cause $C(\beta, \gamma)$. For example, if β is a particular banana image and γ is a particular background image, then $C(\beta, \gamma)$ might place them in a particular context (e.g., relative position) to yield a composite image of that banana on that background.

Our goal then is to characterize the invariant behavior of a system $S : \mathcal{C} \to \mathcal{E}$ in terms of equivalence classes of images in $\mathcal{F}$ and $\mathcal{G}$. To accomplish this, define two quotient classes $\mathcal{Q}_{\mathcal{F}}$ and $\mathcal{Q}_{\mathcal{G}}$ as follows: define two images $\beta, \beta' \in \mathcal{F}$ to be effect-equivalent provided that they always yield the same effect, that is, for all allowable $\gamma \in \mathcal{G}$, $S[C(\beta, \gamma)] = S[C(\beta', \gamma)]$. Then, as before, let $\mathcal{Q}_{\mathcal{F}}$ be the space of all such equivalence classes (or any surrogates in a 1-1 relation to them), and let $E_{\mathcal{F}} : \mathcal{F} \to \mathcal{Q}_{\mathcal{F}}$ be the equivalence map from a particular figure β to the abstraction $E_{\mathcal{F}}(\beta)$ representing all and just those properties that can influence the effect $S[C(\beta, \gamma)]$. Thus, if S is a monkey who responds invariantly to banana images in any context, then the equivalence class corresponding to this response will be an abstraction of all banana images, independent of their context. Exactly the same construction can be applied to $\mathcal{G}$, yielding a corresponding effect-equivalence space $\mathcal{Q}_{\mathcal{G}}$ and equivalence map $E_{\mathcal{G}} : \mathcal{G} \to \mathcal{Q}_{\mathcal{G}}$.

One might suppose that the abstract causes in $\mathcal{Q}$ correspond one-to-one with pairs of abstraction in $\mathcal{Q}_{\mathcal{F}} \times \mathcal{Q}_{\mathcal{G}}$, but this is not to case. To see why, suppose that a monkey grabs a banana if it's either on a tree or a dish, but grabs a mango only if it's on a tree. Since the response to bananas is sometimes different from that for mangos, both abstractions must



$\mathcal{F}$
β
$\mathcal{Q}_\mathcal{F}$
ζ
$E_\mathcal{F}$
$\mathcal{Q}_\mathcal{F}$X$\mathcal{Q}_\mathcal{G}$
(ζ, η)
$\hat{C}$
$\mathcal{F}$X$\mathcal{G}$
(β, γ)
C
ϕ
E
D
ψ
ε
$\mathcal{C}$
$\mathcal{E}$
$\mathcal{G}$
γ
η
$\mathcal{Q}_\mathcal{G}$
$\mathcal{Q}$
$E_\mathcal{G}$

Figure 10: Structural Invariance with Two Components

be in $\mathcal{Q}_\mathcal{F}$; likewise, since the response to things in trees and in dishes differ, they must be distinct members of $\mathcal{Q}_\mathcal{G}$. Therefore, the abstract space $\mathcal{Q}_\mathcal{F} \times \mathcal{Q}_\mathcal{G}$ must represent (at least) the four possibilities: banana/tree, banana/dish, mango/tree and mango/dish. Nevertheless, three of these combinations map to the same effect. Therefore, to describe the behavior of the system in terms of the abstractions in $\mathcal{Q}_\mathcal{F}$ and $\mathcal{Q}_\mathcal{G}$ it's necessary to have a second map $\hat{C} : \mathcal{Q}_\mathcal{F} \times \mathcal{Q}_\mathcal{G} \to \mathcal{Q}$ which maps pairs of abstract components into the corresponding abstract causes (Fig. 10). It turns out that such a map always exists; the effect of $\hat{C}(\varepsilon, \eta)$, for $\zeta \in \mathcal{Q}_\mathcal{F}$ and $\eta \in \mathcal{Q}_\mathcal{G}$, is equivalent to (1) replacing the abstractions ζ and η by any concrete instances β and γ, such that $E_\mathcal{F}(\beta) = \zeta$ and $E_\mathcal{G}(\gamma) = \eta$, (2) composing these into a concrete cause $\varphi = C(\beta, \gamma)$, and (3) finding the corresponding abstract cause, $\varepsilon = E(\varphi)$.[14]

In summary, if some or all of the causes in $\mathcal{C}$ are a result of a mapping on the images of component spaces $\mathcal{F}_1, \ldots, \mathcal{F}_n$, then there are corresponding abstract spaces $\mathcal{Q}_1, \ldots, \mathcal{Q}_n$, which represent all and just the distinctions of the component spaces that can influence the effect. There is also an operation $\hat{C}$ which takes n-tuples of these entities into the corresponding effect-equivalence classes, which determine the system's behavior. Thus the behavior can be described at the abstract level (provided we know the corresponding equivalence classes).

3.5 Transformation-Preserving Systems

Next we consider invariances in which there is a systematic relationship between variations in the causes and effects. For example, a monkey may respond to a banana at a particular location in its visual field by reaching toward a *corresponding* location in its motor field. More precisely, in *transformation-preserving systems* the effect is not invariant under changes to the cause; rather the effect varies systematically with variation of the cause.

[14] Mathematically, $\hat{C} = S \circ C \circ (F_\mathcal{F} \times F_\mathcal{G})$, where $F_\mathcal{F} : \mathcal{Q}_\mathcal{F} \to \mathcal{F}$ is any right-inverse of $E_\mathcal{F}$, and $F_\mathcal{G} : \mathcal{Q}_\mathcal{G} \to \mathcal{G}$ is any right-inverse of $E_\mathcal{G}$. By definition of $\mathcal{Q}$, the choice of right-inverses does not alter $\hat{C}$.

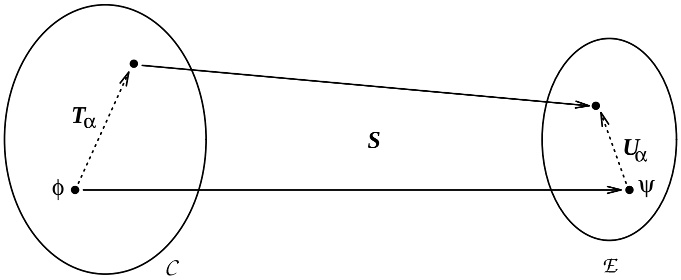


Figure 11: (T_α, U_α)-Transformation-Preserving Behavior

What is invariant is the *relationship* between between transformations of the cause and transformations of the corresponding effect. In this sense it is a higher order of invariance than those hitherto considered. For example, if α is a vector in space, then T_α might be the scaling and translation of a visual image that results from moving the corresponding object. Likewise U_α might be the corresponding change in joint positions to reach an object whose position has been changed by α.

> **Definition** Suppose, as usual, that $S : \mathcal{C} \to \mathcal{E}$ is the system under consideration. Further suppose that for $\alpha \in \mathcal{P}$ a parameter space, $T_\alpha : \mathcal{C} \to \mathcal{C}$ is a transformation on the cause space, and $U_\alpha : \mathcal{E} \to \mathcal{E}$ is a transformation on the effect space. We say S is (T_α, U_α)-*transformation preserving*, or (T_α, U_α)-*preserving*, provided, for all allowable parameters $\alpha \in \mathcal{P}$ and causes $\varphi \in \mathcal{C}$,
>
> $$S[T_\alpha(\varphi)] = U_\alpha[S(\varphi)]. \tag{2}$$

See Fig. 11, which depicts (T_α, U_α)-preserving behavior.

The transformations T_α are an algebra under composition, that is, they are a set having some algebraic structure when their actions are combined. For example, translations form a commutative group, since (1) translations can be done in any order ($T_\alpha \circ T_\beta = T_\beta \circ T_\alpha$), (2) there is an identity translation ($T_0(\varphi) = \varphi$), (3) each translation has an inverse translation ($T_\alpha \circ T_{-\alpha} = T_0$), and (4) translation is associative ($T_\alpha \circ [T_\beta \circ T_\gamma] = [T_\alpha \circ T_\beta] \circ T_\gamma$). (Indeed, any set of transformations closed under composition is a monoid, i.e., satisfies properties 2 and 4.) Similarly, rotation and scaling are commutative groups, the most common algebraic structure. From this perspective we can interpret Eq. 2 in a different way, since it implies a relation between the algebraic structures of the cause and effect transformations. For example, from the commutativity of the cause transformation ($T_\alpha \circ T_\beta = T_\beta \circ T_\alpha$) and Eq. 2 we can conclude

$$U_\alpha \circ U_\beta \circ S = U_\beta \circ U_\alpha \circ S. \tag{3}$$

We *cannot* conclude from this that the effect transformation is commutative ($U_\alpha \circ U_\beta = U_\beta \circ U_\alpha$), but Eq. 3 tells us that the effect transformation is commutative on those effects that can be produced by the system. This is reasonable, since Eq. 2 does not in any way constrain the behavior of U_α on elements of $\mathcal{E}$ not generated by S. This is an example of the *directly induced algebraic structure* of the effect transformations.

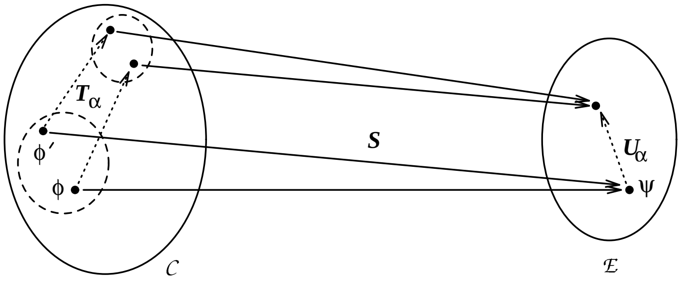


Figure 12: Equivalence Preserving Transformation

Conversely, if the effect transformations commute ($U_\alpha \circ U_\beta = U_\beta \circ U_\alpha$), then we can conclude that

$$S \circ T_\alpha \circ T_\beta = S \circ T_\beta \circ T_\alpha,$$

which does not say that the cause transformations commute, but that the system's behavior is invariant under any differences that result from commuting the cause transformations T_α and T_β. This is an example of the *inversely induced algebraic structure* of the cause transformations. Analogous results hold for other algebraic properties, such as associativity and the existence of identity and inverse transformations.

In the preceding cases we defined, for given cause transformations T_α and effect transformations U_α, what it means for a system to be (T_α, U_α)-transformation preserving, and we saw the conditions under which algebraic properties of the cause transformations are induced in the effect transformations, and vice versa. Now we consider the conditions under which a cause transformation T_α induces an effect transformation U_α such that a system is (T_α, U_α)-preserving, and conversely the conditions under which an effect transformation U_α induces a cause transformation T_α such that a system is (T_α, U_α)-preserving. To this end we will need the following definition:

> **Definition (Equivalence preserving)** For a system $S : \mathcal{C} \to \mathcal{E}$, a transformation $T_\alpha : \mathcal{C} \to \mathcal{C}$ is *S-equivalence preserving* if, for all allowable causes $\varphi, \varphi' \in \mathcal{C}$ and all allowable transformation parameters $\alpha \in \mathcal{P}$, $S(\varphi) = S(\varphi')$ implies $S[T_\alpha(\varphi)] = S[T_\alpha(\varphi')]$.

In other words, if two causes are indistinguishable to the system, then their transformations must also be indistinguishable (Fig. 12). For example, a transformation that magnifies an unnoticeable difference into a noticeable difference would not be preserved by the system, which loses essential information.

It will also be useful to define a generalized notion of an inverse mapping:

> **Definition (Effective Inverse)** Let $S : \mathcal{C} \to \mathcal{E}$ be any mapping, and factor it as usual into a surjection (onto-map) $E : \mathcal{C} \to \mathcal{Q}$ and an injection (1-1 map) $D : \mathcal{Q} \to \mathcal{E}$ such that $S = D \circ E$. Let $C : \mathcal{E} \to \mathcal{Q}$ be any left-inverse of D and let $F : \mathcal{Q} \to \mathcal{C}$ be any right-inverse of E; that is, $C \circ D = \mathrm{I}$ and $E \circ F = \mathrm{I}$); these are guaranteed to exist. Then the *effective inverse of S* (relative to C and F) is $S^\star = F \circ C$.

An effective inverse satisfies the following properties (readers may notice some similarities to the Moore-Penrose pseudoinverse):

$$\begin{aligned} S \circ S^\star \circ S &= S, \\ S^\star \circ S \circ S^\star &= S^\star. \end{aligned}$$

That is, $S^\star \circ S$ is an "effective identity" for domain elements of S, and $S \circ S^\star$ is an "effective identity" for the range elements of S.

> **Proposition (Forward-induced Transformation)** For any system $S : \mathcal{C} \to \mathcal{E}$, if the cause transformation $T_\alpha : \mathcal{C} \to \mathcal{C}$ is S-equivalence preserving, then S is (T_α, U_α)-transformation preserving for $U_\alpha = S \circ T_\alpha \circ S^\star$, where $S^\star$ is any effective inverse of S.

Proof: It is necessary to show that for all allowable φ and α, $S[T_\alpha(\varphi)] = U_\alpha[S(\varphi)]$. By hypothesis,

$$U_\alpha[S(\varphi)] = S\left(T_\alpha\{S^\star[S(\varphi)]\}\right) = S[T_\alpha(\varphi')],$$

where $\varphi' = S^\star[S(\varphi)]$ is some φ' such that $S(\varphi') = S(\varphi)$. Therefore, since T_α is S-equivalence preserving, $S[T_\alpha(\varphi')] = S[T_\alpha(\varphi)]$. Hence, $U_\alpha[S(\varphi)] = S[T_\alpha(\varphi)]$. □

Thus, "well behaved" transformations of the cause space induce corresponding transformations on the effect space.

> **Proposition (Backward-induced Transformation)** For any system $S : \mathcal{C} \to \mathcal{E}$ and effect transformation $U_\alpha : \mathcal{E} \to \mathcal{E}$, the system S is (T_α, U_α)-transformation preserving for $T_\alpha = S^\star \circ U_\alpha \circ S$, where $S^\star$ is any effective inverse of S. Further, T_α is S-equivalence preserving.

Proof: It is is required to show $S[T_\alpha(\varphi)] = U_\alpha[S(\varphi)]$ for all allowable φ and α. By hypothesis,

$$S[T_\alpha(\varphi)] = S\left(S^\star\{U_\alpha[S(\varphi)]\}\right).$$

Let $\psi = U_\alpha[S(\varphi)]$, so $S[T_\alpha(\varphi)] = S[S^\star(\psi)]$. By definition $S^\star(\psi)$ is some φ' such that $S(\varphi') = \psi$. Therefore,

$$S[T_\alpha(\varphi)] = S(\varphi') = \psi = U_\alpha[S(\varphi)].$$

Next it is required to show that T_α is S-equivalence preserving. Therefore suppose $S(\varphi) = S(\varphi')$. Observe:

$$\begin{aligned} S[T_\alpha(\varphi)] &= S\left(S^\star\{U_\alpha[S(\varphi)]\}\right) \\ &= S\left(S^\star\{U_\alpha[S(\varphi')]\}\right) \\ &= S(\varphi''), \end{aligned}$$

where $\varphi'' = S^\star\{U_\alpha[S(\varphi')]\}$ is some φ'' such that $S(\varphi'') = U_\alpha[S(\varphi')]$. Hence,

$$S[T_\alpha(\varphi)] = U_\alpha[S(\varphi')] = S[T_\alpha(\varphi')],$$

since S is (T_α, U_α)-transformation preserving. □

Thus *any* transformation on the effect space will induce a corresponding "well behaved" transformation on the cause space.

4 SELF-ORGANIZATION

4.1 Emergence of Discrete from Continuous

Discrete knowledge representation and processing systems are (superficially, at least) very good at manipulating mathematical and logical formulas, and similar discrete structures, according to precise rules; they have been less successful at subsymbolic processes, such as perception, recognition, association, control and sensorimotor coordination. In contrast, connectionist approaches are well-suited to subsymbolic tasks, but there has been doubt about how well they can operate at the symbolic level. This naturally suggests hybrid architectures, with symbolic tasks accomplished by discrete (digital) computation and subsymbolic tasks by continuous (analog) computation.

This is not the way the brain works, however, for in the brain discrete, symbolic processes emerge from underlying continuous, subsymbolic processes. There is reason to believe that this is not just an accident of biological intelligence, but that this underlying continuity imparts to the emergent symbolic processes the flexibility characteristic of human symbol use [16, 17]. The problem is then to understand how approximately discrete representations and processes can emerge from continuous representations and processes.

It is an oversimplification to treat language as a discrete system. For example, although we accept the space between written words without question, anyone who has done continuous speech recognition knows that we cannot depend on spaces in the sound stream. Furthermore, in ancient Greek, when written language was not considered an autonomous means of expression, but was viewed as a visual representation of the sound stream, we find words run together without intervening spaces, just the way we speak. Seneca claimed that Latin writers sometimes separated words because there was a *difference in speech rhythm* between Greek and Latin speakers, namely, that Latin speakers left a pause after each word [37]. Havelock [6] points out that the alphabet was in use in Greece for 300 years before Greek had a word for ‘word’. Apparently the concept of a word, as a discrete, indivisible unit of the sound stream is not so obvious as we now take it.

In summary, the phenomenological salience of the word is partially a result of our use of an alphabetic writing that separates words, and of the cultural practices that go along with it, such as dictionaries, indices, and word-oriented reading instruction [37]. Nevertheless, the concept of a word is neither illusory nor arbitrary, since as a matter of fact speakers tend to treat certain segments of the sound stream as units, and it is the recurrence and semi-independence of these segments that form the basis of the ‘word’ idea. Therefore, it seems that the emergence of approximately-discrete, symbolic processes from the underlying continuous, subsymbolic processes will be illuminated by considering the self-organization of processes for recognizing recurring parts of images (such as words in the sound stream).

4.2 The Dendritic Net as a Linear System

In the remainder of this paper I will consider several simple examples of self-organization based on so-called Hebbian learning rules, which are more accurately termed correlational learning rules. I have concentrated on correlational mechanisms because (1) there is more biological evidence for correlational learning than for other processes, and (2) correlational learning tends to be fast, and so may better account for rapid learning in people and other animals.

My first example shows how correlational learning in the dendritic net can lead to the self-organization of recursive, matched filters for spatiotemporal patterns. The end result is that dendritic nets can "resonate" when presented with signals that match those that have previously triggered a learning signal. This can be shown through linear systems analysis, which, I have argued elsewhere [22], is a good model of dendritic interactions; here I will take the model for granted and summarize the analysis.

We analyze the electrochemical dynamics of dendritic nets in terms of three spatiotemporal fields of time-varying electrochemical quantities distributed over the synapses and ephapses of the dendritic net. The input field ϕ at times t, $\phi(t)$, represents computationally relevant variables determined externally to the net; it comprises the variables $\phi(z,t)$ at interaction sites z (typically synapses and ephapses).[15] The output field $\omega = \omega(t)$ represents computationally relevant variables $\omega(x,t)$ that affect processes external to the dendritic net, but not internal to it. The state field $\psi = \psi(t)$ represents those computationally relevant variables $\psi(y,t)$ that are determined internally to the net and whose direct effects are confined to the net. Some of these variables will correspond to physical quantities, such as ion currents and neurotransmitter concentrations; others are introduced for mathematical convenience and represent, for example, derivatives of physical quantities. Since, we have argued, dendritic information processing approximates a linear process [22], the dendritic process can be described by the equations:

$$\begin{aligned} \omega(t) &= E\phi(t) + G\psi(t), \\ \dot{\psi}(t) &= D\phi(t) + F\psi(t). \end{aligned}$$

The matrices D, E, F and G are essentially fixed fields representing the (linear) coupling strengths between the computationally relevant quantities. For example, E_{yz} represents the coupling strength from input variable $\phi(z,t)$ to state variable $\psi(y,t)$; that is, if $\phi(z,t)$ varies by δ and everything else is held constant, then $\psi(y,t)$ will vary by $E_{yz}\delta$.

To analyze the system we take the Laplace transform of the state evolution equation:

$$s\Psi(s) - \psi(0) = D\Phi(s) + F\Psi(s),$$

where $\Psi(s) = \mathcal{L}\{\psi(t)\}$ and $\Phi(s) = \mathcal{L}\{\phi(t)\}$ ($\mathcal{L}\{\}$ being the Laplace transform). This

[15] I will be systematically vague about whether the interaction sites form a continuum or a discrete set, and so, for the most part, the analysis will be independent of the denseness of the set of interaction sites. Integration over a set of sites reduces to summation if the sites are a discrete set.

equation can be solved for $\Psi(s)$, the Laplace transform of the state:

$$\Psi(s) = R(s)[D\Phi(s) + \psi(0)],$$

where the transformed transition matrix is given by $R(s) = (s\mathrm{I} - F)^{-1}$. To find out the *forced response* of the system (its response with no energy in the initial state), set $\psi(0) = 0$. Then the Laplace transform of the system's output is then given by:

$$\Omega(s) = E\Phi(s) + GR(s)D\Phi(s).$$

If we define the *transfer function matrix,*

$$H(s) = E + GR(s)D,$$

then we can see the relation between the input and output in the transform domain: $\Omega(s) = H(s)\Phi(s)$. When this matrix product is expanded and transferred back to the space-time domain, it becomes a superposition of convolutions ($\star$), which we may write:

$$\omega(x,t) = \sum_z h_{xz}(t) \star \phi(z,t).$$

That is, the signal $\omega(x,t)$ at output site x is given by the sum of each input signal $\phi(z,t)$ convolved with $h_{xz}(t)$, where the *impulse response* $h_{xz}(t)$ is the signal we see at output site x when an impulse is injected into input site z (all other input sites being clamped to 0).

4.3 Self-organization of Recursive Matched Filters

It's well known that for normalized signals ζ and ϕ, the inner product $\langle \zeta, \phi \rangle$ is maximized when $\zeta = \phi$; this makes it a simple basis for neural network pattern matching.[16] It's also well-known that the inner product is the final value of reverse convolution:

$$\langle \zeta, \phi \rangle = \zeta(T - t) \star \phi(t)\big|_{t=T}\,. \tag{4}$$

Therefore, if a linear system has an impulse response that is the time-reverse of the pattern, $h(t) = \zeta(T - t)$, then it will function as a pattern matcher for ζ, since its output will be maximized by (normalized) signals matching ζ. Such a system is a *matched filter* for ζ.[17]

Since the dendritic tree of a neuron is an approximately linear system, we can conclude that it's a matched filter for the pattern that is the reverse of its impulse response. If the threshold at the axon hillock is a little less than the norm of this pattern, then the neuron

[16] The requirement for normalization might seem to be biologically unrealistic, but Sokolov [36] has shown that in the rhesus monkey color is coded by constant-length vectors; previous work by Sokolov and his colleagues indicates that in humans color is likewise coded by position on the surface of a four-dimensional hypersphere [7, 28].

[17] The time T, which is measured from the time when the input signal begins to add energy to the filter, defines the effective duration of the pattern.

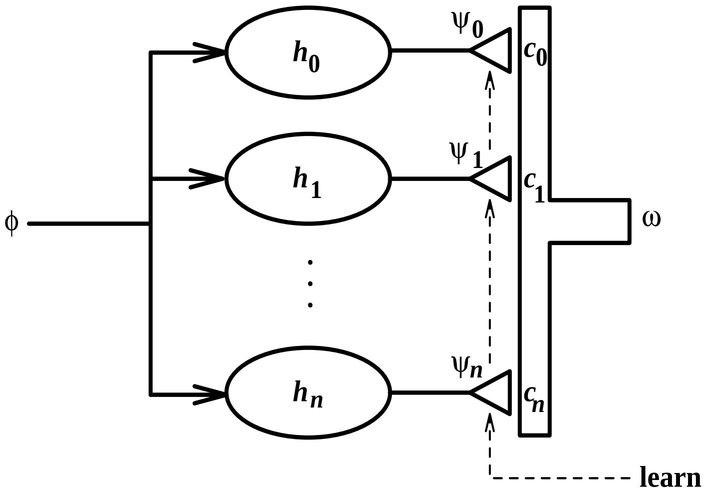


Figure 13: Adaptive Recursive Filter in Dendritic Net

will generate an action potential only when its spatiotemporal input signal is sufficiently similar to the pattern. There are several ways that a dendritic net could self-organize into a matched filter for a given pattern signal [22]; just one will be considered here.

Our goal is a network that self-organizes into a linear filter with impulse response $h(t) = \zeta(T - t)$. To accomplish this, expand $\zeta(T - t)$ in a generalized Fourier series in terms of a complete orthonormal set of basis functions ϱ_k over the interval $(0, T)$:

$$\zeta(T - t) = \sum_{k=0}^{\infty} c_k \varrho_k(t),$$

The coefficients c_k needed to implement the filter are given by the integral,

$$c_k = \int_0^T \zeta(T - t)\varrho_k(t)\mathrm{d}t, \tag{5}$$

which is just the final ($t = T$) value of the convolution:

$$c_k = \zeta(t) \star \varrho_k(t)|_{t=T} \,.$$

That is, coefficient c_k is the output signal, at time T, of a linear filter with impulse response $\varrho_k(t)$.

Figure 13 shows a possible neural implementation of this self-organizing filter. The input signal $\phi(t)$ is distributed to a bank of linear filters h_k, each of which has an impulse response equal to one of the basis functions, $h_k(t) = \varrho_k(t)$. (Since the input signal is band-limited, only a finite number of the basis functions need to be represented.) The graded outputs $\psi_k(t)$ of these filters are passed through synapses onto a common neuron, which sums the transmitted signals to produce the output signal $\omega(t)$. When the input signal is the pattern to be learned, $\phi(t) = \zeta(t)$, then the required coefficients will be available

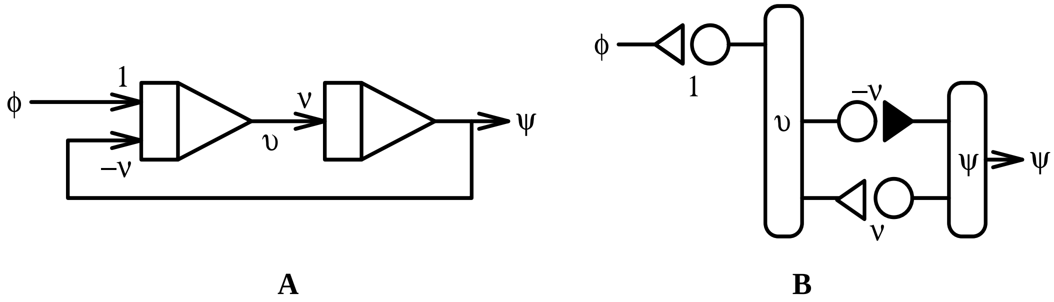


Figure 14: Sine Filter

presynaptically at time T, $c_k = \psi_k(T)$. Therefore, if some "learning signal" (indicated by "learn" in Fig. 13) causes this presynaptic activity to determine the efficacy of its synapse, then the coefficients c_k will become the connection weights (as shown in Fig. 13). The result is a filter matched to the pattern ζ:

$$h(t) = \sum_{k=0}^{n} c_k h_k(t) = \zeta(T - t),$$

There are several possible sources for the learning signal. It could, of course, be some kind of global reinforcement signal indicating the salience of recently received signals. Another interesting possibility is that it results from an *antidromic electrotonic pulse* in the postsynaptic neuron. This occurs when the postsynaptic neuron generates an action potential, for the sudden depolarization causes an electrical pulse to spread backward from the axon hillock out into the dendrites, where it is efficiently transferred into the dendritic spines [22]. It seems plausible that this pulse on the postsynaptic side could trigger changes in the synapse so that its efficacy reflects the graded activity on the presynaptic side. Therefore, if several adaptive filters of this kind converged on a single neuron, then the firing of that neuron could cause the filters to adapt to the input pattern, thus tuning the neuron to the particular set of patterns that caused it to fire.

It remains to say a few words about the filter bank h_k, which must implement the basis functions ϱ_k. This is not so difficult as it might appear. For example, to implement the familiar trigonometric basis:

$$\varrho_0(t) = 1, \quad \varrho_{2k}(t) = \cos(2\pi kt/T), \quad \varrho_{2k-1}(t) = \sin(2\pi kt/T),$$

it's sufficient to implement the differential equations:

$$\begin{aligned} \ddot{\varrho}_{2k} &= \dot{\phi} - \nu_{2k}^2 \varrho_{2k}, \\ \dot{\varrho}_{2k-1} &= \phi - \nu_{2k-1}^2 \int \varrho_{2k-1} \mathrm{d}t, \end{aligned}$$

where $\nu_k = 2\pi k/T$. These differential equations are implemented by the biologically plausible dendritic nets shown in Figs. 14 and 15. It may seem unlikely that the correct

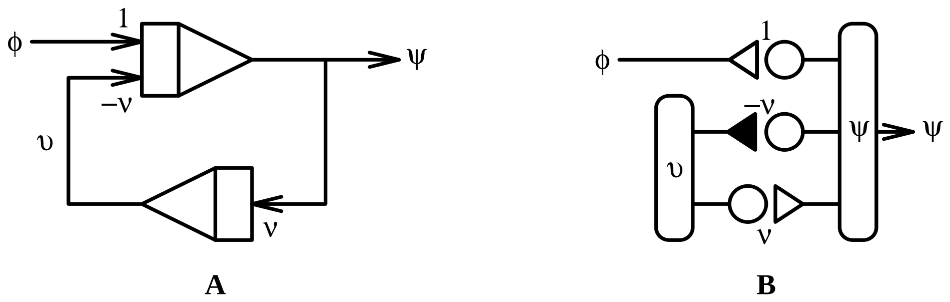


Figure 15: Cosine Filter

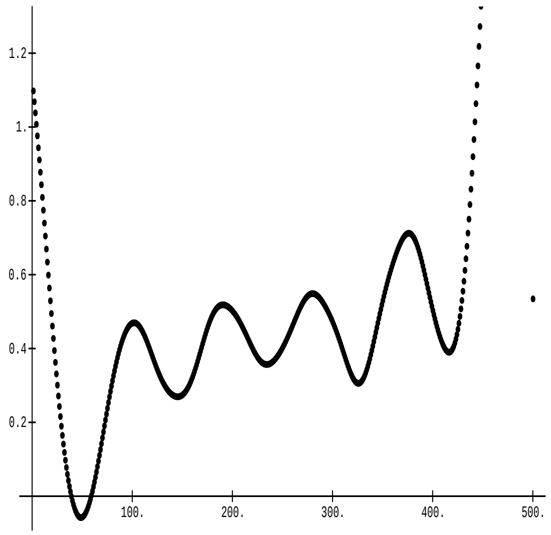


Figure 16: Initial Impulse Response (Random Coefficients)

coefficients, $\nu_k = 2\pi k/T$ $(k = 0, 1, 2, \ldots)$, which determine the resonant frequencies, would occur in a biological system, but the relationship $\nu_k = k\nu_1$ could result from a simple growth process. Furthermore, simulation experiments have shown that it is not necessary to have an accurate orthonormal basis, and filter banks with randomly chosen coefficients ν_k often do quite well.

Figures 16–20 show results from a simulation of this self-organizing filter. Figure 16 shows the initial impulse response $h(t)$ with randomly initialized coefficients c_k. The training pattern $\zeta(t)$, normalized for ease of comparison, is shown in Fig. 17. The training pattern is input to the filter, and the presynaptic activities at time $T = 500$ msec. are transferred to the coefficients, $c_k = \psi_k(T)$. The result is that the filter has adapted to the pattern, as seen from its new impulse response shown in Fig. 18 (normalized and time-reversed for ease of comparison), which is very similar to the training pattern (Fig. 17). We can also see the self-organization of the filter by comparing its response to the pattern before and after adaptation (recall that a matched filter produces maximum positive output at time T —

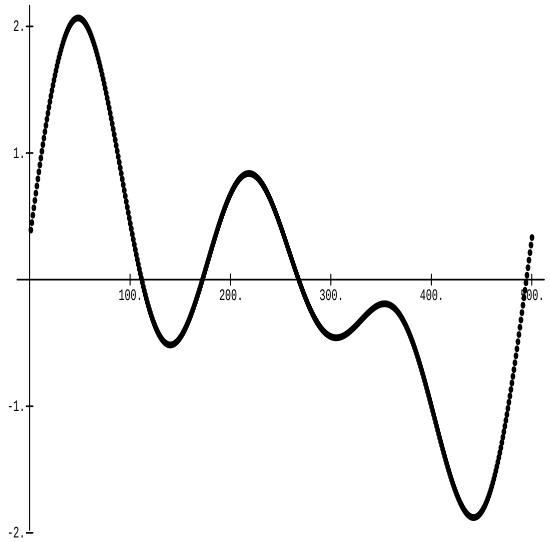


Figure 17: Example Training Signal (Normalized)

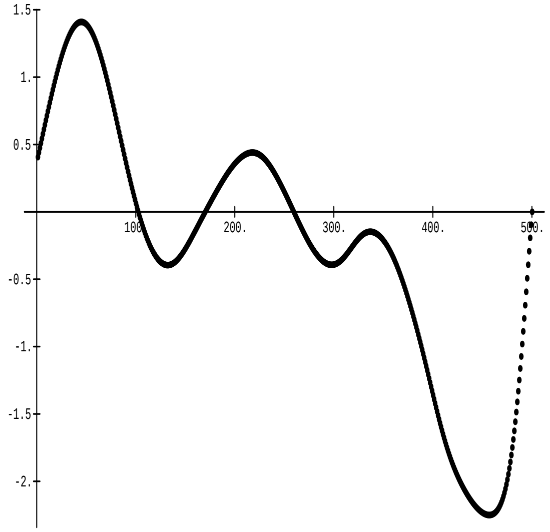


Figure 18: Final Impulse Response (Normalized and Reversed)

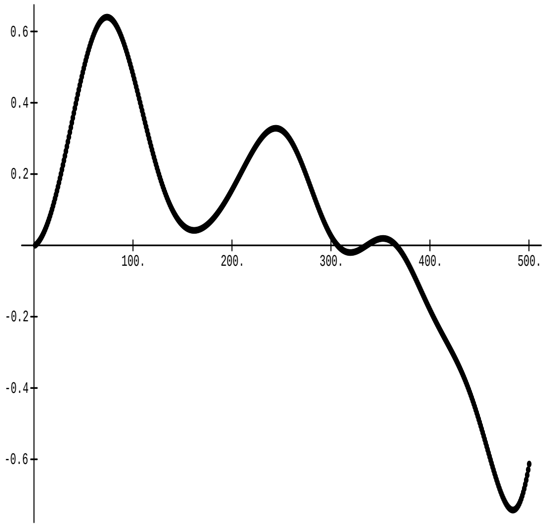


Figure 19: Initial Response to Pattern Signal

ideally an impulse — for the signal to which it is matched). Figure 19 shows the response to the pattern $\zeta(t)$ before adaptation; note that the output at time $T = 500$ msec. is quite negative. On the other hand, Fig. 20 shows that after adaptation the filter's response to the same signal is positive and approximates an impulse at time T.

4.4 Nonlinear Correlational Learning

One way a system can adapt to algebraic invariances, such as symmetry groups, is to learn an arbitrary nonlinear mapping, given samples of the input and corresponding targets (images of correct completion). Therefore, suppose $\mathcal{I}$ is the space of input images and $\mathcal{O}$ is the space of output and target images; our problem is to learn an arbitrary $S : \mathcal{I} \to \mathcal{O}$. This is easily accomplished if we know a *complete* set of functions $\xi_k : \mathcal{I} \to \mathcal{O}$, $k = 1, 2, \ldots$, since then, by definition, every such S can be represented (in at least one way) by a linear combination of the ξ_k. However, the representing functions need not be normalized or even orthogonal, so the representation may not be unique and cannot be calculated by simple inner products. (These characteristics are typical of some wavelet representations, such as Gabor wavelets [15].) Fortunately, complete sets of neurally plausible functions are not hard to find; various sorts of radial basis functions, which coarse-code the input space, are examples. Since we require completeness but not orthogonality or normality, the functions can often be generated randomly.

Therefore, suppose we have a complete set of functions $\xi_k : \mathcal{I} \to \mathcal{O}$, $k = 1, 2, \ldots$; then any $S : \mathcal{I} \to \mathcal{O}$ can be written (in at least one way) as a linear combination $S = \sum_k c_k \xi_k$.

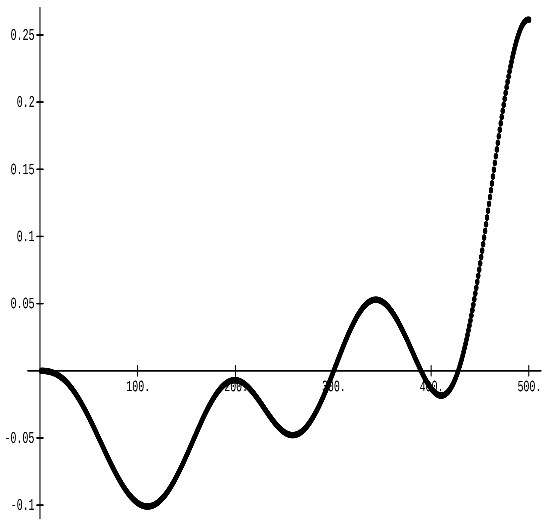


Figure 20: Final Response to Pattern Signal

Our task is to learn the coefficient vector **c** that will give some desired behavior S.[18] Suppose the transformation to be learned is exemplified by the (spatiotemporal) input-output pairs $\phi^j \mapsto \zeta^j$. For the least-squares solution, define as usual the $\mathcal{L}_2$ error as a function of the coefficients: $E(\mathbf{c}) = \sum_j \|\delta^j\|^2$, where $\delta^j = \zeta^j - S(\phi^j)$. To find where the error is minimized, set the derivative to zero:

$$0 = \mathrm{d}E/\mathrm{d}c_k = \sum_j \langle \delta^j, \upsilon_k^j \rangle, \tag{6}$$

where $\upsilon_k^j = \xi_k(\phi^j)$. Since the error is quadratic in the coefficients and the representation is complete we know that the global minimum is $E = 0$ (possibly degenerate) and that there are no local minima. Since $\delta^j = \zeta^j - S(\phi^j)$ we separate the summation:

$$\sum_j \langle \zeta^j, \upsilon_k^j \rangle = \sum_j \langle S(\phi^j), \upsilon_k^j \rangle. \tag{7}$$

Define the pattern vector $z_j = \zeta^j$ and the representation matrix $U_{kj} = \upsilon_k^j$. Then the left-hand side of Eq. 7 is given by:

$$a_k = \sum_j \langle z_j, U_{kj} \rangle,$$

which we abbreviate $\mathbf{a} = U\mathbf{z}$. The right-hand side of Eq. 7 can be expanded to $\sum_{ji} c_k \langle \upsilon_i^j, \upsilon_k^j \rangle$, which can be expressed in terms of the matrix:

$$B_{ij} = \sum_k \langle U_{ik}, U_{jk} \rangle,$$

[18] Corresponding techniques have been used for training radial basis function networks [1, 27].

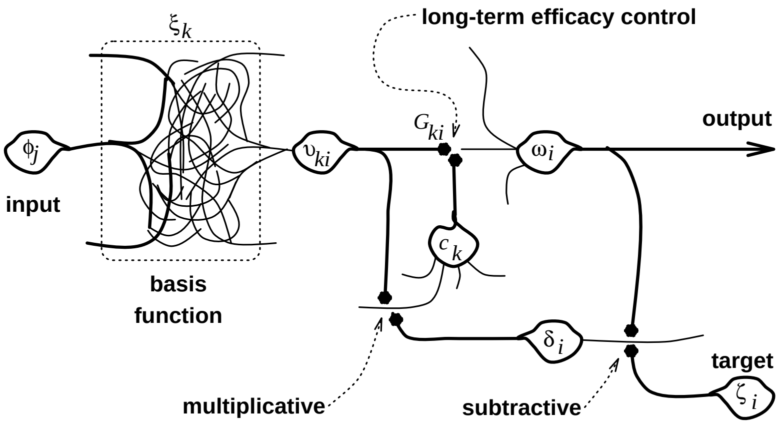


Figure 21: Neural Implementation of Nonlinear Correlational Learning

which we abbreviate $B = UU^{\mathrm{T}}$. Now Eq. 7 can be expressed by the equation $\mathbf{a} = B\mathbf{c}$ and we can solve directly for the optimal coefficients, $\mathbf{c} = B^{-1}\mathbf{a}$. Of course, since the representation is only complete, the coefficients may be underdetermined, because there may be many equally good solutions.

To avoid the necessity of inverting the B matrix, we can use gradient descent; since there are no local minima it cannot get trapped. From Eq. 6 we know

$$\mathrm{d}E/\mathrm{d}c_k = \sum_j \langle \delta^j, \upsilon_k^j \rangle,$$

which we abbreviate $\mathrm{d}E/\mathrm{d}\mathbf{c} = U\boldsymbol{\delta}$. For gradient descent we set $\dot{\mathbf{c}} = -\eta \mathrm{d}E/\mathrm{d}\mathbf{c}$, for learning rate η, so that $\dot{E} = (\mathrm{d}E/\mathrm{d}\mathbf{c})\dot{\mathbf{c}} = -\eta \|\mathrm{d}E/\mathrm{d}\mathbf{c}\|^2 \leq 0$.

The effect of the gradient descent rule can be seen by expanding B^{-1} in a Neumann series:

$$B^{-1} = \mathrm{I} + (\mathrm{I} - B) + (\mathrm{I} - B)^2 + \cdots,$$

which converges provided $\|\mathrm{I} - B\| \leq 1$. Therefore,

$$\mathbf{c} = \mathbf{a} + (\mathrm{I} - B)\mathbf{a} + (\mathrm{I} - B)^2\mathbf{a} + \cdots.$$

To see the relation to gradient descent, observe $\dot{\mathbf{c}} = -\eta U(\mathbf{z} - \mathbf{c}U) = -\eta(\mathbf{a} - U\mathbf{c}U)$. Since $U\mathbf{c}U = B\mathbf{c}$,

$$\dot{\mathbf{c}} = -\eta(\mathbf{a} - B\mathbf{c}). \tag{8}$$

Therefore, gradient descent will compute successive approximations to the Neumann series, $\mathbf{c}_k = \sum_{j=0}^{k}(\mathrm{I} - B)^j\mathbf{a}$, if we choose $\mathbf{c}_0 = \mathbf{a}$ for the initial coefficient vector, and $\mathbf{c}_{k+1} = \mathbf{a} + (\mathrm{I} - B)\mathbf{c}_k$, which is gradient descent (Eq. 8).

Fig. 21 shows a neurally-feasible implementation of the gradient descent process. The representation functions ξ_k are computed by fixed dendritic networks on the left; as noted previously, completeness is a rather weak property, and an appropriate set of functions, such as radial basis functions, can be computed by randomly generated networks. The resulting field ω is compared with an image ζ of correct completion to generate the difference field δ. Multiplicative synapses combine the difference fields with the intermediate fields υ_k to control the long-term efficacy coefficients $G_{ki} = c_k$.

4.5 Orthogonal Representation of Discrete Categories

Next I'll consider a simple mechanism by which a system can self-organize so that discrete stimulus classes are represented orthogonally. A basic hypothesis is that (approximately) discrete categories are a consequence of the need for discrete responses (e.g., fight or flight), which in turn are often dependent on discreteness in the environment, as encountered. For example, suppose green lizards are good to eat but blue lizards are poisonous, and that all lizards in the environment are distinctly green or distinctly blue. Under these circumstances we expect discrete categories. On the other hand, if greenish lizards are pretty good to eat and blue ones are not so good, and if the lizards in the environment come in all shades between green and blue, then we expect graded categories, since whether or not we eat a lizard will depend on its exact shade as well as contextual factors, such as our hunger.

I will consider the consequences of a correlational learning rule operating on a simple kind of three-layer network. The input field ϕ is coarse-coded by an intermediate field $\upsilon(\phi)$. The coarse-coding neuron υ_κ responds most strongly to an input $\phi = \kappa$. For convenience I will assume that all these neurons have a receptive field profile θ with radius of support r; thus $\theta(\lambda) = 0$ for $\|\lambda\| > r$. Therefore, the activity of a coarse-coding neuron is given by $\upsilon_\kappa(\phi) = \theta(\kappa - \phi)$. The output field ω is simply a linear combination of the coarse-coded input, $\omega_\lambda = \int G_{\lambda\kappa}\upsilon_\kappa \mathrm{d}\kappa$, which is abbreviated $\omega = G\upsilon$.

Next consider the consequences of correlational learning in the interconnection field G, which will result from a correlation of the coarse-coding field $\upsilon(\phi)$ and the target response field $R(\phi)$. Thus, $\dot{G}_{\lambda\kappa} = R_\lambda(\phi)\upsilon_\kappa(\phi)$, which is the outer product, $\dot{G} = R(\phi) \wedge \upsilon(\phi)$. If $\phi(t)$ represents the time sequence of stimuli and $R[\phi(t)]$ the corresponding targets, then the interconnection field after time T will be:

$$G(T) = G(0) + \int_0^T R[\phi(t)] \wedge \upsilon(\phi)\mathrm{d}t.$$

If T is sufficiently long, then stimulus ϕ will appear with its stationary probability $p(\phi)\mathrm{d}\phi$, and the limiting interconnection field is

$$G^\infty = c\int_{\mathcal{S}} R(\phi) \wedge \upsilon(\phi)p(\phi)\mathrm{d}\phi,$$

where c is a scale factor determined by the learning duration.

Suppose we have n disjoint classes of stimuli $\mathcal{C}_1, \ldots, \mathcal{C}_n \subset \mathcal{S}$ and that these classes are "sufficiently separated" (in a sense to be made clear shortly). Furthermore, let $\mu_k(\phi)$ be the probability density of ϕ in $\mathcal{C}_k$, that is, μ_k is a fuzzy membership function for $\mathcal{C}_k$. Clearly, then

$$G^\infty = c\sum_{k=1}^{n}\int_{\mathcal{S}} R(\phi) \wedge \upsilon(\phi)\mu_k(\phi)\mathrm{d}\phi.$$

Now, to see the effect of discrete target responses, suppose the target response $R(\phi)$ is a constant ϱ_k for all $\phi \in \mathcal{C}_k$. In this case,

$$G^\infty = c \sum \varrho_k \wedge \int_{\mathcal{S}} \upsilon(\phi) \mu_k(\phi) \mathrm{d}\phi.$$

That is, $G^\infty = c \sum \varrho_k \wedge \varepsilon_k$, where $\varepsilon_k = \int_{\mathcal{S}} \upsilon(\phi)\mu_k(\phi)\mathrm{d}\phi$.

It's easy to see that ε_k is the convolution $\theta \star \mu_k$; that is ε_k is the category membership function μ_k "blurred" by the coarse-coding function θ. This blurring has a maximum radius r. Therefore, if the original membership functions were separated by at least $2r$ then the blurred functions ε_k will still be disjoint (i.e. have nonoverlapping support). Under these conditions it's straightforward to show that the ε_k are orthogonal fields; that is, $\langle \varepsilon_j, \varepsilon_k \rangle = 0$ for $j \neq k$. In summary, coarse-coding results in an orthogonal representation if each of the discrete responses is associated with stimuli that are sufficiently unambiguous.[19]

It's also easy to see that the network has self-organized to behave correctly, that is, to produce response ϱ_k for stimuli in $\mathcal{C}_k$. For, suppose $\phi \in \mathcal{C}_k$, then the output is $G\upsilon(\phi) = c \sum \varrho_k \langle \varepsilon_k, \upsilon(\phi) \rangle$. Now observe that $\langle \varepsilon_k, \upsilon(\phi) \rangle = \langle \theta \star \mu_k, \upsilon(\phi) \rangle$ is nonzero, since θ spreads μ_k. Conversely, for all $j \neq k$ we know $\langle \theta \star \mu_k, \upsilon(\phi) \rangle = 0$, since μ_j is separated from μ_k by at least $2r$. Therefore the output elicited by $\phi \in \mathcal{C}_k$ is the correct response ϱ_k, scaled by its "fuzzy" category membership. Note that $\langle \varepsilon_k, \upsilon(\phi) \rangle$ amounts to a fuzzy intersection of ε_k (the blurred category) and $\upsilon(\phi)$ (the singleton ϕ blurred by coarse coding).

4.6 Metric Correlation and Group Invariances

Some of the most important invariances respected by perceptual and motor systems are *algebraic invariances*, especially *group* invariances. While it is true that some of these mechanisms are innate and have evolved through "blind variation and selective retention" [2], others are learned, and so we need to consider more systematic mechanisms by which such invariances may emerge. One of the attractions of correlational and convolutional methods (such as holography) is that they simultaneously accomplish translation-invariant pattern recognition in parallel with pattern location [31, Ch. 5]. For example, if ζ is a pattern and ϕ is an image, which may contain instances of ζ, the the correlation $\zeta \otimes \phi$ will be an image with "bright spots" wherever the pattern occurs in the input image ϕ, with the location of the spot showing the location of the pattern instance, and the "brightness" of the spot indicating the closeness of the match (by the Euclidean metric). The result of the correlation is reminiscent of the so-called "what" and "where" channels in primate vision (the occipitotemporal and occipitoparietal pathways). There is no mystery about how the correlation (or convolution) accomplishes this: it simply forms, in parallel, inner products between ϕ and every possible translation of ζ.

Although translation invariance is important, there are many other transformations for which we would like a system to respond invariantly, while simultaneously indicating how

[19] If there is some ambiguity, as a result of coarse-coding or noise, then the representation will be correspondingly only approximately orthogonal.

the input has been transformed. Therefore, in this section I will generalize the familiar correlation operation so that it works on arbitrary transformations and metrics.

As discussed above, the speech stream is essentially continuous, and, to a first approximation, words are segments of the sound stream that can be treated as discrete units, that is, relocated as wholes. Conversely, recurrence of the same signal in a variety of contexts is evidence that it is a meaning unit, often a word. Therefore, we suspect that one component of language learning is the detection of such recurrences. However, a word will rarely recur exactly; it will be transformed in duration, pitch, amplitude, etc., or by the context of surrounding sounds. This suggests metric correlation as a mechanism for the self-organization of a word-recognition system [18].

The familiar correlation attempts to match one signal to all possible translations of another signal, and returns a signal showing how well these matched. The correlation of images φ and ψ is defined:

$$[\varphi \otimes \psi]_\alpha = \int_\Omega \varphi(t-\alpha)\psi(t)\mathrm{d}t.$$

The structure of this is more apparent when we realize that $\varphi(t-\alpha)$ is φ translated to the right by an amount α. Therefore we introduce the operator T_α to mean a rightward translation by α, and rewrite the correlation:

$$[\varphi \otimes \psi]_\alpha = \int_\Omega T_\alpha \varphi(t)\psi(t)\mathrm{d}t.$$

It is then apparent that the correlation is the inner product of the translated image $T_\alpha\varphi$ and the image ψ. So we write it that way:

$$[\varphi \otimes \psi]_\alpha = \langle T_\alpha \varphi, \psi \rangle.$$

The significance of the inner product is that it measures the similarity of normalized images. Therefore we write $\sigma(\varphi, \psi) = \langle \varphi, \psi \rangle$ to emphasize that its purpose is to measure similarity:

$$[\varphi \otimes \psi]_\alpha = \sigma(T_\alpha \varphi, \psi).$$

Thus, in general terms, the value of the field $\chi = \varphi \otimes \psi$ at a point α is the similarity to ψ of the α-translate of φ.

The common correlation can be generalized to other classes of transformation (rotation, scaling, perspective distortion, etc.) as well as to other measures of similarity or dissimilarity. If T is any parameterized class of transforms and σ is any similarity metric, then define $\sigma(\varphi \lhd_T \psi)$, the *metric correlation* with ψ of all T-transforms of φ, by:

$$[\sigma(\varphi \lhd_T \psi)]_\alpha = \sigma(T_\alpha \varphi, \psi).$$

Sometimes it is easier to work in terms of difference rather than similarity, in which case we write:

$$[\delta(\varphi \lhd_T \psi)]_\alpha = \delta(T_\alpha \varphi, \psi),$$

where δ is a (difference) metric. When the metric or transform is clear from context it will be omitted; thus in general $\varphi \lhd \psi$ is the metric correlation of all transforms of φ with ψ. Sometimes it is more convenient to consider the metric correlation of φ with all T-transforms of ψ, so we write:

$$[\rho(\varphi \rhd_T \psi)]_\alpha = \rho(\varphi, T_\alpha \psi),$$

where ρ is either a similarity or difference metric. Clearly, $\varphi \lhd \psi = \psi \rhd \varphi$. Finally we define the operator which correlates all T-transforms of φ with all U-transforms of ψ:

$$[\rho(\varphi \diamond_{T,U} \psi)]_{\alpha,\beta} = \rho(T_\alpha \varphi, U_\beta \psi).$$

Notice that the correlation field resulting from this operation is indexed over two parameter spaces. These operations satisfy many simple identities, most of which are obvious, and so omitted.

It is often useful to consider metric correlations under multiple transformations. For example, if X_α for $\alpha \in \mathcal{E}^2$ (2D Euclidean space) is a translation by α, and R_β for $\beta \in S^1$ (a topological circle) is a rotation through an angle β, then $T_{\alpha\beta} = X_\alpha R_\beta$ is a rotation followed by a translation (the order doesn't matter; they commute). Thus $(\varphi \lhd \psi)_{\alpha\beta}$ measures the correlation between ψ and the α-translation, β-rotation of φ.

Often the transformations applied to images have algebraic structure; frequently they form a topological group. In these cases the metric correlations inherit the structure; for example, if the transformations are an abelian group:

$$(T_\alpha \varphi \lhd \psi)_\beta = (\varphi \lhd \psi)_{\alpha+\beta} = (\varphi \lhd \psi)_{\beta+\alpha} = (T_\beta \varphi \lhd \psi)_\alpha.$$

The reader may suppose that metric correlations are computationally too expensive to have much significance to cognitive processing, but this need not be so. First observe that they are not much more expensive than the usual (inner product) correlations. Second, the limited precision of neural computation (one or two digits) will limit the number of transforms to be computed in parallel to a dozen or so, for each real parameter. A combination of two transforms might require several hundred to be computed (in parallel).

5 CONCLUSIONS

I have argued that the foundation of expertise and skillful behavior is knowledge represented as a *structured continuum* as opposed to a discrete structure. The reason is that continuous knowledge representation permits more flexible behavior and avoids failures due to brittleness. As a consequence it is crucial that we understand the principles of continuous representation so that they can be applied in artificial intelligence and neuroscience. To this end I have compared and contrasted continuous (“analog”) and discrete (“digital”) computation, and have suggested a theoretical framework (*simulacra*) for continuous representations.

This theoretical framework was used to characterize several kinds of *invariance* that occur in natural and artificial systems, including *effect equivalence*, *transform invariance*, *structural invariance* and *transform preservation*. Included as special cases are psychologically significant invariances, such as symmetry groups and other algebraic invariances involved in object perception. Several ways of quantifying *degree of invariance* were noted *en passant*, since quantification is a necessary prerequisite to understanding the emergence of algebraic structure in perceptual and cognitive processes.

I reviewed several simple mechanisms, based on correlational learning, that lead to the emergence of invariances. These included (1) the self-organization of matched filters for spatiotemporal signals by means of approximately linear processes in dendritic nets, (2) the approximation of nonlinear functions by bases that are complete but not necessarily orthogonal, and (3) the emergence of orthogonal representations through coarse coding. Finally, I proposed *metric correlation*, a generalization of the familiar correlation and convolution operations, which accomplishes transform invariant pattern recognition. The foregoing is only a hint, however, of what is needed: a comprehensive theory of the means by which (approximately) discrete representations may emerge from continuous processes to form the structured continua that are the foundation of skillful, intelligent behavior in the world.

References

[1] D. S. Broomhead & D. Lowe, “Multivariable Functional Interpolation and Adaptive Networks,” *Complex Systems*. vol. 2, pp. 321–355, 1988.

[2] D. T. Campbell, “Blind Variation and Selective Retention in Creative Thought as in Other Knowledge Processes,” in: Gerard Radnitzk & W. W. Bartley, III (Eds.), *Evolutionary Epistemology, Theory of Rationality, and the Sociology of Knowledge*, pp. 91–114.

[3] H. L. Dreyfus & S. E. Dreyfus, *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*, New York: Free Press, 1986.

[4] M.-L. von Franz, *Alchemy: An Introduction to the Symbolism and the Psychology*, Toronto: Inner City Books, 1980, pp. 166, 238, 264.

[5] S. Harnad, “Grounding Symbols in the Analog World with Neural Nets: A Hybrid Model,” *Think*, vol. 2, pp. 12–20, 74–77, June 1993.

[6] E. A. Havelock, E. A. *The Literate Revolution in Greece and its Cultural Consequences*, Princeton: Princeton University Press, 1982, p. 8.

[7] Ch. A. Izmailov & E. N. Sokolov, “A Semantic Space of Color Names,” *Psychological Science*, vol. 3, pp. 105–110, 1992.

[8] C. G. Jung, *Psychology and Alchemy*, Bollingen Series XX, second ed., tr. R. F. C. Hull, Princeton: Princeton University Press, 1968, pp. 327–330.

[9] C. H. Kahn, *The Art and Thought of Heraclitus: An Edition of the Fragments with Translation and Commentary*, Cambridge: Cambridge University Press, 1979, pp. 43, 117–118.

[10] G. S. Kirk, J. E. Raven & M. Schofield, *The Presocratic Philosophers*, second ed, Cambridge: Cambridge University Press, 1983, pp. 36–37.

[11] B. J. MacLennan, "Technology-Independent Design of Neurocomputers: The Universal Field Computer," in *Proceedings, IEEE First International Conference on Neural Networks* (June 21–24, 1987), Vol. III, pp. 39–49.

[12] B. J. MacLennan, "Logic for the New AI," in J. H. Fetzer (Ed.), *Aspects of Artificial Intelligence* (pp. 163–192). Dordrecht: Kluwer, 1988, p. 189

[13] B. J. MacLennan, *The Calculus of Functional Differences and Integrals* (Technical Report CS-89-80), Knoxville: Computer Science Department, University of Tennessee, 1989.

[14] B. J. MacLennan, *Field Computation: A Theoretical Framework for Massively Parallel Analog Computation; Parts I – IV*, University of Tennessee, Knoxville, Computer Science Department Technical Report CS-90-100, February 1990.

[15] B. J. MacLennan, *Gabor Representations of Spatiotemporal Visual Images*, University of Tennessee, Knoxville, Computer Science Department technical report CS-91-144, September 1991.

[16] B. J. MacLennan, "Flexible Computing in the 21st Century," *Vector Register*, vol. 4, 3, pp. 8–12, 1991.

[17] B. J. MacLennan, *Research Issues in Flexible Computing: Two Presentations in Japan* (Report No. CS-92-172). Knoxville: University of Tennessee, Computer Science Department, September 1992.

[18] B. J. MacLennan, *Image and Symbol: Continuous Computation and the Emergence of the Discrete*, University of Tennessee, Knoxville, Department of Computer Science Technical Report CS-93-199, December 18, 1992 (revised August 5, 1993).

[19] B. J. MacLennan, "Characteristics of Connectionist Knowledge Representation," *Information Sciences*, vol. 70, pp. 119–143, 1993.

[20] B. J. MacLennan, "Visualizing the Possibilities" (review of Johnson-Laird & Byrne's *Deduction*), *Behavioral and Brain Sciences*, vol. 16, 2, pp. 356–357, 1993.

[21] B. J. MacLennan, “Grounding Analog Computers,” *Think*, vol. 2, pp. 48–51, 74–77, June 1993.

[22] B. J. MacLennan, “Information Processing in the Dendritic Net,” in: Karl Pribram (Ed.), *Rethinking Neural Networks: Quantum Fields and Biological Data*, pp. 161–197. Hillsdale, NJ: Lawrence Erlbaum, 1993.

[23] B. J. MacLennan, “Field Computation in the Brain,” in: Karl Pribram (Ed.), *Rethinking Neural Networks: Quantum Fields and Biological Data*, pp. 199–232. Hillsdale, NJ: Lawrence Erlbaum, 1993.

[24] B. J. MacLennan, “Continuous Symbol Systems: The Logic of Connectionism,” in: D. S. Levine and M. Aparicio IV (Eds.), *Neural Networks for Knowledge Representation and Inference*, pp. 83–120. Hillsdale: Lawrence Erlbaum, 1994.

[25] B. J. MacLennan, “Image and Symbol: Continuous Computation and the Emergence of the Discrete,” in: Vasant Honavar and Leonard Uhr (Eds.), *Artificial Intelligence and Neural Networks: Steps Toward Principled Integration, Volume I: Basic Paradigms; Learning Representational Issues; and Integrated Architectures*, New York: Academic Press, in press.

[26] B. J. MacLennan, “Words Lie in Our Way,” *Minds and Machines*, to appear.

[27] J. Moody & C. Darken, “Learning with Localized Receptive Fields,” in: D. Touretzky, G. Hinton & T. Sejnowski (Eds.), *Proceedings of the 1988 Connectionist Models Summer School*, San Mateo: Morgan Kaufmann, 1989, pp. 133–143.

[28] G. V. Paramei, Ch. A. Izmailov & E. N. Sokolov, “Multidimensional Scaling of Large Chromatic Differences by Normal and Color-deficient Subjects,” *Psychological Science*, vol. 2, pp. 244–248, 1991.

[29] S. Parpola, “The Assyrian Tree of Life: Tracing the Origins of Jewish Monotheism and Greek Philosophy,” *Journal of Near Eastern Studies*, vol. 52, 2, pp. 161–208, July 1993.

[30] J. W. Perry, *Lord of the Four Quarters: Myths of the Royal Father*, New York: Braziller, 1966.

[31] K. H. Pribram, *Brain and Perception: Holonomy and Structure in Figural Processing*, Hillsdale: Lawrence Erlbaum, 1991, ch. 5.

[32] R. A. Prier, *Archaic Logic: Symbol and Structure in Heraclitus, Parmenides, and Empedocles*, The Hague: Mouton, 1976, ch. 3.

[33] H. Putnam, *Representation and Reality*, Cambridge: MIT Press, 1988, pp. 94–103, 121–125.

[34] J. R. Searle, “Is the Brain a Digital Computer?” Presidential Address, *Proceedings of the American Philosophical Association*, 1990.

[35] J. R. Searle, *The Rediscovery of the Mind*, Cambridge: MIT Press, 1992, pp. 208–209.

[36] E. N. Sokolov, “Vector Coding in Neuronal Nets: Color Vision,” these proceedings.

[37] J. P. Small, “Historical Development of Writing and Reading,” *Psycholoquy*, `92.3.61.`
`reading.10.small`, 1992.

[38] L. Ungerleider, L. Ganz & K. H. Pribram, “Size Constancy in Rhesus Monkeys: Effects of Pulvinar, Prestriate and Infero-temporal Lesions,” *Experimental Brain Research*, vol. 27, pp. 251–269, 1977.